Title
Inferring Structural Models of Travel Behavior: An Inverse Reinforcement Learning Approach

Author
Feygin, Sidney

Publication Date
2018

Peer reviewed|Thesis/dissertation

Inferring Structural Models of Travel Behavior:
An Inverse Reinforcement Learning Approach

by

Sidney Feygin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Civil & Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alexey Pozdnukhov, Chair
Professor John Canny
Professor Joan Walker

Spring 2018

Inferring Structural Models of Travel Behavior:
An Inverse Reinforcement Learning Approach

Abstract

Inferring Structural Models of Travel Behavior:
An Inverse Reinforcement Learning Approach

by

Sidney Feygin

Doctor of Philosophy in Engineering – Civil & Environmental Engineering

University of California, Berkeley

Professor Alexey Pozdnukhov, Chair

Large volumes of digital human trajectories at high spatiotemporal resolution have become increasingly available to researchers and public entities. Derived from anonymized cellular records and social network postings, fine-grained mobility traces present exciting opportunities for longitudinal studies of daily travel and activity planning decisions. Through the application of automated and efficient data-mining techniques, researchers in machine learning, transportation engineering, and related disciplines have been able to use these movement microdata to model and forecast daily traffic conditions at metropolitan scales with unprecedented accuracy (González and Hidalgo, 2008; Lin Z. et al., 2017; Widhalm et al., 2015; Yin M. et al., 2017).

However, state-of-the-art machine-learning and discrete-choice frameworks do not consider the dynamics of daily mobility decisions at the individual level. Existing methods also do not take into account strategic, interdependent interactions between representative agents, complicating the cost-benefit analysis of innovative decentralized policy instruments such as induced peer-to-peer influence. Interpretable structural models that can provide consistent and disaggregate estimates of replanning behavior are needed in order to evaluate the impacts of these novel regulatory measures.

Therefore, in order to take better advantage of future and emerging technologies as tools to forge cooperative and sustainable relationships between citizens, governments, and the built environment, this thesis develops a framework for data-driven city management that bridges established travel demand planning practices with innovations in big data, reinforcement learning, and strategic decision-making. The work described herein is comprised of three major components. First, we develop a two-stage game theoretic model of peer pressure to investigate feedback between social, geographic, and temporal dimensions of agent choices in a hyper-realistic microsimulation of urban travel behavior. Second, in order to learn representations of dynamic agent utility functions, we extend *inverse reinforcement learning* (IRL) algorithms to novel activity and travel planning environments and estimate associated structural parameters. Finally, we investigate the strength of modern high-dimensional

imitation learning techniques to train flexible and accurate models of schedule composition and activity duration. Results from applications of the empirical methods developed herein suggest that our contributions could effectively complement the microsimulation and discrete choice modeling techniques used in disaggregate urban infrastructure planning frameworks such as activity-based transportation demand models.

To Nahum and Anna Tsenter

Contents

List of Figures

List of Tables

Acknowledgments

It was only through the encouragement of many intelligent, caring, and patient people as well as the generous support of forward-thinking educational institutions that I've completed this dissertation. It is with great joy that I take this opportunity to express my appreciation for the advisers, mentors, colleagues, friends, and family members who, in countless ways, enriched my life during my graduate program at Berkeley.

First, I am grateful to Berkeley and the Civil Engineering graduate program for providing me with the opportunity to pursue my intellectual passions. Thanks to Steve Glaser for helping me get started on my journey and to Shelley Okimoto for making sure that I didn't get lost along the way.

Alexey Pozdnukhov, my primary research advisor, has stoked my curiosity in urban data science, machine learning, and artificial intelligence. Throughout my studies, he has encouraged me to follow my intuition, while not letting me go too far afield. I greatly appreciate the many opportunities he's taken to share his considerable knowledge and experience with me. My sincere thanks go to my other committee members, Joan Walker and John Canny for taking time to understand and provide insightful advice on extending the relevance of my research.

My initial research experience at Berkeley was with Raja Sengupta and Shankar Kariv who both very much shaped my thinking about how to measure human behavior and consider how technology, for better or worse, changes behavior. I am grateful to them for planting seeds of inquiry that would inspire many of the topics covered in this dissertation. During this time, I also had the privilege of working with many talented, knowledgeable, and fun fellow students on xMobile and our spin-off emotion and decision-making project: Aluma Dembo, Orianna DeMasi, Alex Mead, Andre Carrel, Nachi Mehta, Andrew Campbell, and Dounan Tang.

Many thanks go to my co-authors in Alexey's group: Madeline Sheehan, Andrew Campbell, Ziheng Lin, Mogeng Yin, Sudatta Mohanty, Max Gardner, Colin Sheppard, and Danqing Zhang. Thanks also to my collaborators and mentors at Lawrence Berkeley National Laboratory: Anand Gopal, Rashid Wariach, Colin Sheppard, and I feel very fortunate to have worked with such an inspiring and gifted group. I feel no less lucky to have had such excellent office mates in 116 McLaughlin. Thanks to Sreeta Gorripaty, Feras El Zarawi, Allan Ogwang, and Timothy Brathwaite for sharing in the occasional pains of learning as well as the more frequent joys of success.

None of my achievements could have been possible without the endlessly patient and unconditionally loving support of my family. The values of intellectual curiosity and stubbornness that my parents have imbued in me have helped me to persevere through the most challenging times of this experience while expanding my inquiry well outside of traditional thinking in my discipline. To my sister and brother: your friendship and encouragement over the past few years have been priceless. This dissertation is dedicated to my grandparents who first dreamed that I would one day be a scholar.

1

Introduction

Don't let us forget that the causes of human actions are usually immeasurably more complex and varied than our subsequent explanations of them.

– Fyodor Dostoevsky, *The Idiot*

1.1 Motivation

Addressing wicked problems in smart cities through information architecture

The *smart city* paradigm envisions the seamless automation and regulation of instrumented infrastructure using data gathered from *cyber-physical systems* (CPS)[1] (Batty et al., 2012; Cassandras, 2016; Kitchin, 2014). Coordinated information flows from heterogeneous data streams will be used to power novel analytic techniques, models, and simulations, enabling policy analysis and participatory urban planning platforms to optimize public services according to their efficiency, equity, and contribution towards improved quality of life for all citizens (Batty et al., 2012) (see Figure 1.1). While more robust privacy safeguards have liberated previously siloed corporate data and efficient data-mining techniques have been developed to collect and collate *digital exhaust*[2], much work remains to be done in defining the intelligence functions that will operationalize smarter cities.

Towards this end, researchers in city science and engineering disciplines have focused on developing scalable methods to organize passively-collected geo-tagged records from social

[1] According to the NSF, cyber-physical systems consist of "physical and software components [that] are deeply intertwined, each operating on different spatial and temporal scales, exhibiting multiple and distinct behavioral modalities, and interacting with each other in a myriad of ways that change with context"(National Science Foundation, 2018).

[2] Coined by Glaeser et. al., digital exhaust is defined as "the trail of data left online through everyone's day-to-day use of the Internet"(Glaeser, Edward L and Kominers, Scott Duke and Luca, Michael and Naik, 2018). Examples include Craigslist postings, Zillow ads, and search engine queries.

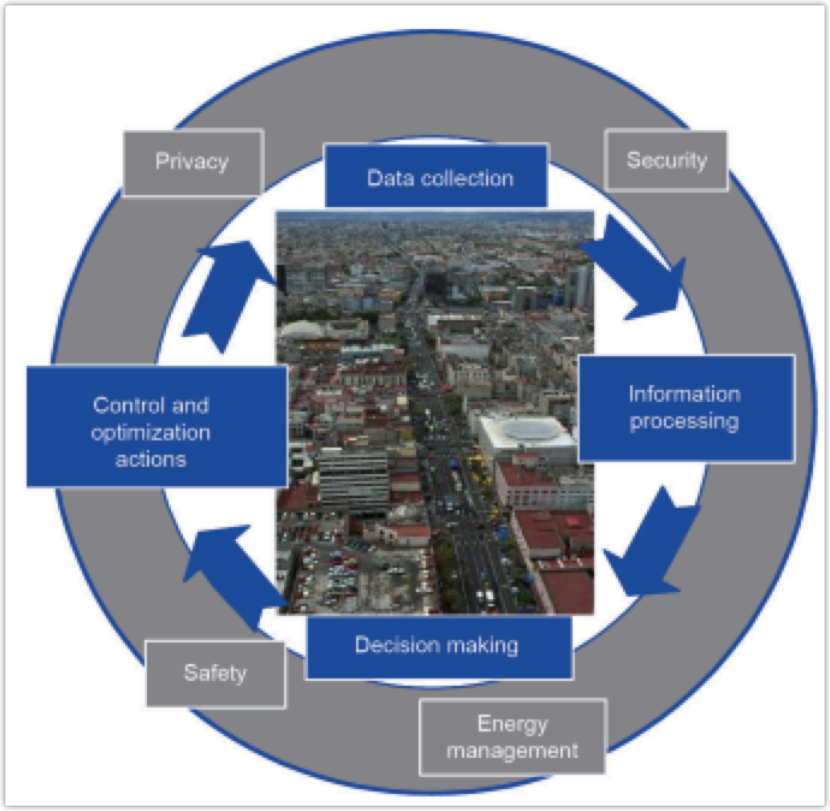


Figure 1.1: Smart cities as cyberphysical social systems (Cassandras, 2016).

and physical networks into actionable knowledge about *spatial behavior* (Song et al., 2016). These vast streams of movement data offer unprecedented insight into the fundamental patterns that underlie individual mobility as urban populations make their way through complex morphologies (González and Hidalgo, 2008). Many of the resulting innovations in urban analytics rely on state-of-the-art machine learning (ML) systems employing deep neural networks (DNNs). An important benefit of DNNs is that they enable automated *representation learning* through *feature discovery*, which reduces the need for domain experts to specify explanatory variables (also known as factors of variation) in statistical models of data-generating processes (Goodfellow et al., 2016). Meanwhile, advances in distributed data warehousing, efficient routing, and parallel processing (collectively termed *big data* technologies, (Laney, 2001)) are facilitating the rapid integration of the rich information streams produced by these analyses into protocols that automate the management of critical infrastructure (Lheureux et al., 2017).

In the transportation sector, digital movement traces and activity sequences derived from anonymized call detail records (CDRs) are serving as inputs into traffic flow models based on generative models of commuter behavior (Çolak et al., 2016; Pozdnoukhov A. et al., 2015; Widhalm et al., 2015). For example, synthetic daily activity patterns generated using DNNs

trained on cellular data have enabled agent-based microsimulation software to replicate traffic volume counts across large metropolitan areas with high fidelity (Lin Z. et al., 2017; Yin M. et al., 2017). As network carriers begin to continuously exchange cellular data with transportation managers, short-term forecasts of traffic flows can be used to update demand models with consistent predictions seconds to hours ahead of time (Vlahogianni et al., 2014; Ye et al., 2012; Yin et al., 2018). Advanced traveler information services (ATIS) such as variable message signs can then use predictions based on perturbed flow scenarios to provide travelers with warnings of expected as well as non-recurring changes in traffic conditions such as adverse weather or significant public events (Klein and Ben-Elia, 2016). Drivers often react to unanticipated traffic conditions by updating their current routes, increasingly relying on in-vehicle navigation systems or GPS-enabled smartphones to automate rerouting decisions (Varga, 2014). The digital traces gathered from the connected devices that facilitate adaptive routing are, in turn, powering crowd-sourced social navigation systems such as Waze and Google Maps, leading to increasingly accurate representations of prevailing traffic conditions as well as predictions of traffic flows (Ben-Elia and Avineri, 2015; Zheng and Van Zuylen, 2013).

Somewhat counterintuitively, better informed travelers do not necessarily imply better off travelers. Theory suggests that only in the unrealistic case of perfect information do drivers make more efficient use of limited-capacity road networks (Arnott et al., 1991). Experimental studies bear this theoretical development out; demonstrating that providing inaccurate, sub-optimal, or counterfactual routing alternatives could result in individuals making decisions that result in lower utility (Abdel-Aty et al., 1997; Ben-Elia and Avineri, 2015; Ben-Elia et al., 2013; Lu et al., 2011). On the other hand, models of strategic interaction on road networks suggest that uncoordinated reductions in information asymmetry propagated by central controllers drive traffic flows away from *system optimum* (SO) and towards a more selfish *user equilibrium* (UE), leading to an increase in the price of anarchy (Klein et al., 2018; Papadimitriou, 2003; Roughgarden and Tardos, 2000; Varga, 2014, 2015). That is, if everyone relies on traffic data from Google Maps more time could be spent by all in congestion. Providing travel information in the interest of improving social welfare demonstrates characteristics of a "wicked problem" Rittel and Webber (1973). The complex behavioral response to information and resistance to 'quick-fix' technological solutions implied by the dilemma of too much data requires a principled approach to policy analysis: one that balances both individual desires and social good.

The prevailing behavior theoretic paradigm in neoclassical welfare economics posits that consumers are rational actors making decisions under perfect information. Historically, most econometric models used in justifying social policy assume that decisions are representative of individual's behaving in a manner consistent with their preferences (Hicks, 1939; Mas-Colell et al., 1995). The decision to drive alone imposes certain *external costs*[3] on society such as congestion, emissions, noise, pollution, and accidents. The externalities arising

[3]An externality is defined as an experienced cost or benefit due to the failure of a rational economic actor to take into account the consequences of their behavior on others (Rothengatter, 1994; Verhoef, 1994). Externalities can be characterized as *positive* or *negative* depending on whether they benefit or harm parties

from following a putatively rational preference to drive alone can be mitigated through, for example, *Pigouvian taxes*[4].

While Pigouvian taxes can be justified by the desire of societies to reduce external costs due to self-regarding preferences, a complementary, individual-centric approach towards the regulation of socially suboptimal behavior presumes that people make choices that may not be in their self-interest. Strongly paternalistic theories of social authority propose that individual preferences may result in choices that impose self-harm or harm others (Dhami, 2016). Under strong paternalism, this presumed misalignment of self-interest and social values results in policies that limit individual autonomy in order to reassert perfectly rational behavior. Thus, the compulsion to wear seatbelts, helmets for motorcyclists, drug laws, limitation on working hours, and other non-consensual policies are often justified on the basis that a government has a better idea of what is in the long-term best interests of its constituents (Camerer et al., 2003).

Growing bodies of evidence from studies in experimental and behavioral economics, psychology, sociology, and other disciplines have demonstrated that consistent errors in decision-making contradict the predictions of the neoclassical model, inspiring choice theories based on *bounded rationality*[5] and *social preferences*[6] (Kahneman and Tversky, 1979; Ng and Tseng, 2017; Rubinstein, 1998; Simon, 1972). Concern that bounded rationality would lend further justification to coercive, strongly paternalistic policies has led to theorists to pursue alternative modalities of social authority such as *asymmetric paternalism*[7] and *libertatian paternalism*[8]. Surging interest and political acceptance of policies based on libertarian paternalism in the form of *nudges*, have created novel opportunities for transportation planners to strategically structure mobility choice architectures that actuate and stabilize socially-cooperative attitudes (Avineri, 2012; Leonard et al., 2008; Thaler and Sunstein, 2008).

---

external to the action in question. Making individuals aware of the effects of their decisions on others in order to reduce externalities is known as *internalization*.

[4]Pigouvian mechanisms are market-based approaches that attempt to internalize externalities by taxing goods resulting in net disbenefits or subsidising goods that result in net benefits (Pigou, 1920). First-best Pigouvian internalization mechanisms for transport involve charging individuals with the negative external costs for which they are directly responsible. See Mas-Colell et al. (1995) for more information

[5]Bounded rationality implies a relaxation of expected discounted utility maximization in choice situations.

[6]In interdependent choice situations, individual preferences may be conditional on the outcomes of others' decisions. These are known as social preferences.

[7]The propensity to make decisions that deviate from from rationality may be imperfectly distributed through society. Thus, corrective policies would unfairly punish those who are already behaving rationally. Asymmetric paternalism aims to reduce costs to rational actors while counteracting the mistakes of irrational ones (Camerer et al., 2003)

[8]Thaler and Sunstein 2003 argue that, while there are no feasible alternatives to paternalistic forms of corrective social authority, paternalism can be limited to non-coercive policies. Paternalistic policies that are maximally libertarian understand that humans are (to varying extents) limited in the cognitive abilities necessary to make optimally rational choices. In order to contend with this empirical reality, libertarian paternalistic policies influence rather than compel rational decision-making by assisting decision-makers in making choices that they themselves would have made had they been better informed of the consequences of the alternatives (Thaler and Sunstein, 2003).

However, empirically-justified behavioral models of welfare are required if regulators are to create *information architectures* such that predictive travel applications can reduce congestion. Unfortunately, the majority of ML-based studies of travel demand offer limited explanatory insight into the underlying dynamics and incentives influencing decision-making (Chen et al., 2016b). This critical limitation stems from the observation that ML and econometrics methods have been specialized to solve different types of problems (Athey, 2017; Glaeser, Edward L and Kominers, Scott Duke and Luca, Michael and Naik, 2018; Mullainathan and Spiess, 2017). Whereas ML algorithms excel at pattern recognition, microeconometrics aims to specify structural models[9] compatible with a theoretical understanding of human behavior (Holmes and Sieg, 2014; Mullainathan and Spiess, 2017). Thus, in order for big data to take part in achieving an inclusive and progressive vision of smart cities, there remains a need to develop *interpretable* machine learning-based models of decision-making.

Scaling up individual models of adaptive transportation decision-making

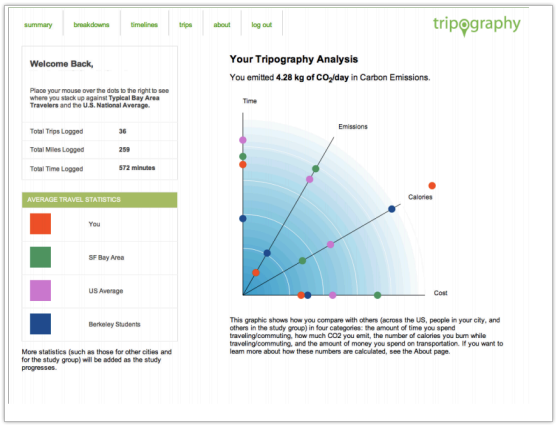
As highlighted above, travelers are becoming increasingly reliant on smartphones for an on-demand, personalized and geo-localized stream of information and suggestions on where to go, who to meet, and what to do. Connected devices help simplify or even solve complex spatiotemporal decision-making problems, presenting a more limited, yet salient set of choices to users. Smartphone applications developed as persuasive technologies[10] have the potential to enact nudges by, for example, inspiring awareness of the potential benefits of alternative transportation modes, or alerting commuters to cognitive biases that may influence noncooperative travel decisions (Gaker et al., 2010; Jariyasunant et al., 2015; Shankari et al., 2014). These applications typically work by collecting movement data from participant smartphones, performing analytics (often on remote servers), and offering online visualizations of the participant's short and long-term mobility behavior as well as the environmental consequences thereof. Figure 1.2 shows two recent examples of web-based dashboards that provide quantitative feedback to individual commuters about their mobility habits in relation to those of their peers.

In order to validate the effect of behavioral policies designed to promote the emergence of cooperative mobility decisions in socioeconomically diverse cities, it has become increasingly necessary to model the preferences of individual travelers. In the past, four-step transport demand models were trained using cross-sectional datasets[11]. The potential for spurious ecological correlations due to the Yule-Simpson effect (Udny, 1903) as well as limited ap-

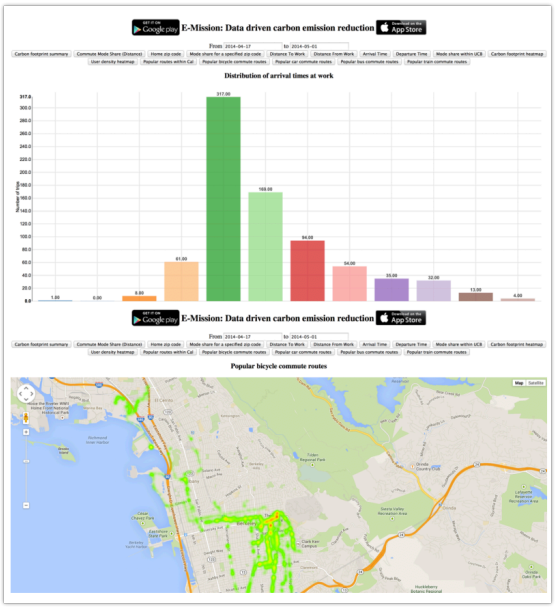
[9] i.e., functional forms that rely on consistent estimates of interpretable parameters to inform policy counterfactuals.

[10] In the sense of Fogg (2002).

[11] Briefly, in the four-step method, trips are generated and distributed between each traffic analysis zone in an urban area to form an origin-destination (OD) matrix. Modal splits are then computed, followed by assignment of aggregate flows to the physical network. See (Ortúzar and Willumsen, 2011) for a comprehensive review of classical four-step as well as more modern transport demand models.



(a) Tripography web interface



(b) eMission web interface

Figure 1.2: Examples of smartphone-based transportation behavior change personalized data visualization dashboards (Jariyasunant et al., 2015; Shankari et al., 2014).

plicability in economic policy appraisal analysis has led the transportation engineering and urban planning communities to adopt increasingly disaggregated transport demand modelling techniques.

For example, state-of-the-art activity-based travel demand models (ABTDMs) (Castiglione et al., 2014; Ortúzar and Willumsen, 2011) use discrete choice analysis (DCA) in order to predict activity sequence, frequency, timing, location, mode choice, as well as joint-trips that involve coordination with other agents' schedules. Discrete choice models operate at the individual level, permitting analysts to estimate utility functions from observed decision-making as well as under counterfactual policy scenarios (Ben-Akiva and Lerman, 1985).

However, ABTDMs often include the simplifying assumption that people plan their days from a fixed set of activity-travel patterns[12]. One limitation of this approach is that it does not account for schedule adaptation due to evolving choice sets and preferences. Consequently, models of individual travelers do not take into account external conditions that impose scheduling restrictions or introduce exploration of new travel modes. These assumptions were merited when computational complexity and, particularly, high resolution, *longitudinal* data, was more of a constraint for individual-level mobility analyses than it is

[12] A daily *activity-travel pattern* consists of several *tours*, which may further be comprised of one or more *trips* or, sometimes, *sojourns*. Discrete trip components (i.e., travel mode changes) are referred to as *legs* or *stages*.

today[13].

Evaluating the impact of innovative policy proposals such as income-dependent road pricing or smartphone-based social norming would be difficult given the static models of decision-making available in modern travel demand models. Improved methodological approaches are therefore necessary in order to rationalize adaptive human mobility patterns at scale. Thus, in accordance with the desire (expressed in the previous section) of estimating interpretable machine-learning-based models of dynamic mobility, this Ph.D project aims to improve the verisimilitude of activity-based travel demand models by inferring the *(re)scheduling* preferences of individual commuters. While recent work on this topic proposes an econometric approach using dynamic discrete choice methods (Jonsson and Karlström, 2005; Väastberg et al., 2016), we investigate the potential for state-of-the-art research in reinforcement learning (Sutton et al., 1999) and inverse reinforcement learning (Ng and Russell, 2000) to learn flexible models of daily activity-travel planning behavior.

The influence of social dynamics on travel mode choice

In contrast to the passive processes governing diffusion of social influence (e.g., via the smartphone-based approaches described above) and adherence to norms, individuals can, at some cost in utility to themselves, actively influence each others' choices through *peer pressure* (Calvó-Armengol and Jackson, 2010; Mani et al., 2013; Pentland and Reid, 2013). In particular, when one person's choices result in visible negative external costs to his community, his peers may persuade him to make decisions that internalize the consequences of his actions (Lazaer and Kandel, 1992). The decision to apply pressure may be viewed as a strategic one, since pressure is often costly, but may result in net social benefits accruing to society. Given the diffuse global costs arising from negative urban transportation system externalities (e.g., congestion, air pollution) as well as strong structural factors (i.e., urban sprawl, advertising, social status) promoting automobile dependency, individuals may not receive immediate or sufficiently strong feedback signals from sanctioning actions, resulting in under-actuated enforcement of socially-cooperative norms (Van Vugt et al., 1996). Such *second-order free-rider effects* are often observed when sanctioning is costly, since individuals may prefer that others enforce cooperative norms (Van Vugt et al., 1996).

The need for improved strategies to internalize transport externalities stems from existing difficulties implementing *centralized* reward and punishment mechanisms such as congestion charging or GHG emissions pricing. While effective in theory, and demonstrable in simulation, public support for congestion or emissions pricing remains lukewarm (Eriksson et al., 2006; Hårsman and Quigley, 2010). Equity may also be a concern, as users with lower in-

[13]Note that while initial studies using CDRs had shown that human mobility is predictable with approximately 93% accuracy (Eagle and Pentland, 2009; Song et al., 2010), more recent work has questioned these findings (Smith et al., 2016), indicating that this upper bound should be revised downwards. Furthermore, as with models built using the survey data in the above discussion, these predictions do not account for changes in prevailing travel conditions.

comes may feel the effects of a toll disproportionately to more affluent users[14] (Viegas, 2001; Walter and Suter, 2003). In addition, it is difficult for a central controller to allocate revenues from GHG emissions or traffic internalization, as the social costs of these externalities may extend well beyond the major metropolitan areas considered for policy changes (Verhoef, 1994).

According to Ostrom (1990), centralized sanctioning schemes often result in perceived inequality, decaying collective trust and fraying the social fabric that counteracts resource over-appropriations (Chen, 2013; Ostrom, 1990). Field and lab evidence from experimental economics and other disciplines have demonstrated that subsidizing *decentralized* sanctioning mechanisms may lead to more widespread adoption of self-sustaining cooperative behavior (Chaudhuri, 2011). Developing targeted choice architectures or asymmetrically paternalistic reward or punishment mechanisms that take into account the sociodemographics of spatially-embedded urban social networks could make better use of financial capital resources available to municipal governments by leveraging and, in effect, preserving the social capital inherent in regulated communities.

Towards principled models of individual utility in urban microsimulations

Cities have often been characterized as complex adaptive systems (Batty, 2007); generating emergent global phenomena due to interactions among many local entities. More specifically, cities are "people systems" (Jacobs, 1961), consisting of social networks of individuals, aggregated into communities that co-evolve with the physical morphologies of urban infrastructure (Heppenstall et al., 2016). Proponents of smart cities have advocated for data-driven agent-based modelling methods to simulate the interaction of policy proposals with the complex social factors and diverse spatial temporal scales governing urban behavioral processes (Batty et al., 2012). The visualization and analysis of simulated scenarios could justify regulatory interventions to stakeholders.

Simulations at micro, meso, and macro-scales have become an important component of transport demand models (Castiglione et al., 2014). Agent-based microsimulation software such as MATSim (Horni et al., 2016a) permit analysis of urban mobility at a highly disaggregate level. The flexibility offered to demand modelers and planners by such software has made it possible to study the interaction of social dynamics with transportation behavior (Dubernet and Axhausen, 2013; Hackney and Axhausen, 2006; Illenberger, 2012) as well as changes in mobility decisions due to hypothetical policy instruments (Agarwal and Kickhöfer, 2015a; Kaddoura et al., 2014). Simulations of cyber-social influence on travel

[14]While one could argue that *efficient* allocations would subsume poor users as well, if a driver simply cannot afford to pay a congestion charge, she may end up having to miss out on an important activity. The personal and social costs of this loss are difficult to quantify using current methods. Although understanding how policy interventions affect individual time valuation is part of the problem that work in this dissertation attempts to address, we also argue that there are more effective and equitable methods to control traffic than congestion taxes.

decision-making (such as the peer pressure mechanism indicaated above) may help transit agencies understand the implications of peer-to-peer influence as a decentralized means of encouraging socially cooperative mobility decisions.

Unfortunately, current functional representations of agent preferences in travel demand microsimulations have little empirical basis. Thus, a "guess and check" approach is often used to model the correspondence of synthetic agents to real-life travelers (Horni et al., 2016a). Choice dimensions such as transport mode, departure time, and route are randomly perturbed in agent plans, which are then simulated on a virtual representation of the physical road network. Upon completion of a single day's schedule, the agents are scored based on the cumulative sum of utility experienced via participation in activities less the cumulative sum of utility experienced during travel episodes[15]. This rather heuristic replanning process bears little in common with psychologically-grounded theories of experiential learning (Erev and Barron, 2005), particularly in the presence of exogenous information processes (Lu et al., 2011). Novel models representing the dynamic heterogeneity of daily travel planning in urban environments are necessary in order to better simulate innovative policy measures intended to actuate pro-social travel decision-making.

Problem Statement and Approach

Motivated by the opportunity to combine computational methods emerging from the world of applied machine learning and artificial intelligence with existing techniques to understand human decision-making, this thesis explores the potential for data-driven decision-support systems to model and predict how social interactions and temporal interdependencies induce adaptive changes in the daily activity-travel scheduling behavior of individual agents. In particular, we evaluate the potential for game-theoretic and reinforcement learning-based frameworks to augment existing urban travel demand modelling and simulation practices.

The first component of this thesis explores the potential for microsimulation frameworks to function as transportation *policy laboratories*. We model the tendency of transit users to pressure solo drivers to switch away from automobile use as a many player game taking place over social, road, and public transit supernetworks. Scaling game theoretic models to travel demand scenarios presents the opportunity to investigate the impact of novel policy levers that encourage pro-social and pro-environmental behavior in among urban commuters (Klein and Ben-Elia, 2016; Klein et al., 2018).

We illustrate our peer pressure model on an agent-based microsimulation of commuters using car and public transit in the San Francisco Bay Area in California. While experiments focus on the effect of public transportation as an alternative mode, the game-theoretic setting and experimental simulation framework developed as part of this work covers a wide class of situations characterized by a high price of anarchy. Based on the outcome of agent decisions in our synthetic social dilemma of daily mode choices, we observe how neglected

[15]Travel episodes are often quantified according to their disutility as a result of the foregone opportunity cost of activity participation as well as the intrinsic unpleasantness of the travel itself, see Chapter 2, Section 2.1 for more information on this topic.

hidden effects on decision making can significantly affect transportation system equilibrium conditions.

At the same time, our scenario demonstrates the conceptual limitations of extending random utility modelling methods to account for the complexity of human travel behaviors. The scoring functions used to evaluate daily schedules in microsimulation software such as MATSim (Horni et al., 2016a) are heuristically parameterized and calibrated. Without a sound theoretical basis, it is not possible to estimate utility functions for virtual agents from data. Thus, in the second component, we address several of the limitations identified in the first component as well as augment the travel demand forecasting tools used therein with a scalable technique to rationalize the preferences motivating sequential choices of activities and travel from cellular data. Towards this end, we adopt the approach of inverse reinforcement learning (IRL).

IRL is a highly active area of current research within the artificial intelligence community, which is specifically concerned with developing methods to estimate the parameters of utility functions from observed sequences of choices made by an expert agent acting in a high-dimensional, stochastic environments. These utility functions may then be used to simulate decision-making policies in the original scenario as well as observe how perturbations in the choice context (e.g., via regulatory interventions) lead to plan adaptations. Recent successes in reinforcement learning using deep neural networks (Mnih et al., 2013; Silver et al., 2016) and subsequent applications to IRL methods (Finn et al., 2016c; Wulfmeier et al., 2015, 2016) have helped to automate feature design from high-dimensional data.

We initiate our investigation into adaptive decision-making algorithms through the use of the maximum entropy inverse reinforcement learning framework (MaxEnt IRL), as it bears strong similarity to models of dynamic discrete choice currently used in econometric practice (Ermon et al., 2015; Ziebart and Maas, 2008; Ziebart et al., 2009).

The experimental evaluations of the frameworks described in this thesis involve inference of the determinants of travel behavior from passively-collected CDRs. In addition to the previously described advantages of cellular data over traditional survey methods, the use of CDRs provides two important advantages with respect to generalizability:

1. As the trajectories derived from cell phone records have broad spatial scope *within* cities, we are able to adapt policy instruments to communities with diverse demographics and values.

2. Our methods are intended to be applicable *across* cities as well due to the expanding market penetration of smartphones worldwide.

1.2 Thesis Organization And Summary of Contributions

Chapter 2 continues this dissertation with a review of prior art as well as preliminaries on relevant technical topics in travel demand modelling, computational social science, and econometric analysis. The following chapters build on this background material while encapsulating the distinct contributions of this research.

Chapter 3 explores the factors governing the emergence of pro-environmental behavior in transportation systems by adapting the peer pressure game of Mani et al. (2013) to a social dilemma of travel mode choice. We implement our game-theoretical model in the microsimulation software, MATSim, using the daily activity plans of a social network of tens of thousands of synthetic agents. Equilibrium between supply and demand is first calibrated according to road network sensors. This establishes a baseline that we can use to observe how peer pressure changes the incentive structure for decision-making. Our representation of individual agent micro-dynamics using this principled, agent-based approach permits us to study heterogeneous interactions between social, geographic, and temporal dimensions in a way that may be more realistic than purely statistical methods. A discussion of the spatiotemporal and social dynamics of the system both with and without pressure is provided. In order to understand the effects of system components that do not appear to have likely real-world analogs, we perform comprehensive sensitivity analysis. We conclude from our results that policies subsidizing peer pressure have the potential to significantly reduce greenhouse gas emissions and congestion. This further suggests that mobility trajectories inferred from CDRs may lend greater ecological validity to decentralized, pro-environmental policy interventions. On the other hand, our analysis indicates that variability in the marginal cost of pressure results in complex spatial and behavioral dynamics, highlighting the importance of explicitly modeling and estimating individual utility functions that incorporate social and travel mode choice preferences.

Towards this end we model within-day activity and travel decision-making as an inverse planning problem, using reinforcement learning as a model of behavior. In Chapter 4, we describe our representation of activity-travel planning as a Markov decision process, estimation approach using the MaxEnt IRL dynamic programming algorithm, and empirical evaluation using agent mobility traces. The work presented in this chapter treats a general absence in the transport demand modelling literature of individualized, forward-looking models of travel pattern selection under an estimation framework consistent with maximization of individual utility. Our results demonstrate that our specifications of utility functions are in agreement with theoretical motivations behind agent preferences. That is, we are able to interpret the trade-offs in utility between being at home vs. being at work vs. participation in some other activity. However, we find that the *decision-rules* induced by the utility representations that we estimate do not generally induce behavioral dynamics that are similar to those of actual commuters. We identify the issue stems from a somewhat unrealistic decision-making environment, in which the agent has to make a choice of staying at her current activity or

departing at high-frequency intervals between decision-making epochs.

In order to address this limitation and extend our work to more realistic models of travel scheduling behavior, in Chapter 5, we expand the decision-space to allow agents to choose activity duration in addition to timing. Doing so introduces computational complexity into the MaxEnt IRL algorithm, making it intractable for use in large populations. We address this deficiency by exploring the use of the probabilistic imitation learning (IL) frameworks based on generative adversarial networks (GANs, (Goodfellow et al., 2014)). Specifically, we use the generative adversarial imitation learning, (GAIL, (Ho and Ermon, 2016)) and information maximizing GAIL (InfoGAIL, (Li et al., 2017)) algorithms to model and simulate the behavior of commuters from cellular traces. We find that our proposed method is able to reproduce activity timing, identity, and duration from a relatively small number of agent demonstrations. We perform a quantitative comparison between IL and IRL methods applied to the simulation of agent behavior under realistic, albeit computationally demanding travel planning environments. Compared with MaxEnt IRL, we find that the GAIL-based algorithm recovers agent behavior more faithfully and efficiently. Generally, we find that results demonstrate that the use of model-free IRL algorithms could permit more flexible specification of decision-making contexts for individual travelers.

Chapter 6 summarizes the work presented herein and proposes directions for future research. We conclude with final thoughts on the strengths and limitations of big data and machine learning enabled techniques applied in smart city regulatory environments, advocating a measured and interdisciplinary approach towards fusing automated reasoning with urban policy analysis.

2

Background and Related Work

> *To deal with these problems - of world population and hunger, of peace, of energy and mineral resources, of environmental pollution, of poverty - we must broaden and deepen our knowledge of nature's laws, and we must broaden and deepen our understanding of the laws of human behavior.*
>
> – Herbert Simon

Models of decision-making in this thesis consider day-to-day preferences over possible activity-travel patterns. Central to the understanding of how individual preferences over different schedules can be quantified, the discrete choice analysis framework represents an important starting point for several theoretical developments and empirical methods described in this thesis. Methods from discrete choice analysis figure into a large part of the typical workflow of travel demand modelling and analysis. Thus, Section 2.1 situates our work in the context of the activity-based travel demand models that comprise the current state of practice in Transportation Engineering and Science. We also include a brief review of relevant concepts in the appraisal of economic policy to mitigate transportation externalities.

Part of the purpose of this thesis is to connect emerging computational trends in the Machine Learning and Computer Science disciplines with well-validated empirical models of rational decision-making originating in the Transportation Engineering and Economics literatures. The mathematical formalism of Markov decision processes (MDPs) underlies two theoretical and algorithmic frameworks commonly used to model sequential decision-making problems; one from Econometrics: dynamic discrete choice models (DDCM), and one from artificial intelligence: inverse reinforcement learning (IRL). In addition to a brief review of the technical aspects of DDCM and IRL, Section 2.2 reviews prior applications of both of these frameworks models to the of the transport demand modeller's workflow are presented.

As indicated in the introduction, the influence of social preferences on travel comprises an important component of this work. In Section 2.3 we summarize some of the ways in which a person's societal relationships may influence transportation decisions. The purpose

of this section is not a comprehensive review of all social influences on travel behavior, but, rather, a focus on topics relevant to this thesis.

2.1 Static Decision-Making Frameworks: Theory And Applications To Travel Demand Modelling.

To study travel behavior at the individual level, planners typically turn to the methods of discrete choice analysis (DCA): a statistical, econometric, and psychological framework of decision-making. As we will see in subsequent sections, many of the assumptions underlying DCA in static choice contexts will be equally applicable in dynamic decision-making environments.

Overview. The central actor in a discrete choice problem is a *representative* individual (or group) who seeks to make *rational* choices from a *finite set of mutually exclusive alternatives*. (McFadden et al., 1973). We say that an individual is representative in the sense that their choices are prototypical of a much larger group of people with similar socioeconomic characteristics. However, in order for a decision-model to itself be representative of the behavior of diverse populations, the data sample should ideally consist of individuals from various socioeconomic strata. In this way, DCA permits *aggregate* predictions of travel behavior to be made from data describing the outcomes of individual choice scenarios (Bowman, 1998; Ortúzar and Willumsen, 2011). Compared to methods that focus on directly deriving population-level mobility statistics, *disaggregate* methods such as DCA can inform the design of counterfactual policy scenarios.

Discrete choice models assume that a *rational* individual consumer n selects a unique option from a mutually exclusive and exhaustive finite *choice set*, I of *feasible* options (such as travel mode, destination, vehicle, etc.) I_n in accordance with her preferences and subject to budgetary, time, and any other applicable constraints (we will further address the idea of rational preferences below). A researcher observes the decisions of all individuals $n \in N$ and collects data on the *characteristics* and *attributes* of all alternatives, $i \in I$. Discrete or continuous measures of input data represent the vector of potential explanatory variables, X_{in}, that may enter into model *specifications*.

Rational choice In what way do discrete choice models assume decision-makers are rational? Neoclassical economic theory is grounded in the assumption that consumers behave *as if* optimizing utility by maximizing the outcome of choice situations. Empirical research in microeconomics quantifies preferences according to utility functions, $U : I \to \mathbb{R}$, i.e., expressing the outcomes of choice situations as real numbers. For example, consumer n is assumed to select the i that has maximum utility U_{in}. Formally, for every $j \in I_n$, $j \neq i$, alternative i is selected if and only if $U_{in} \geq U_{jn}$. For example, a person, i may say that they prefer driving a car to work as opposed to commuting via bus or subway because they

find that the car provides a greater sense of reliability and comfort. Consequently, for n's choice of commute mode, i, we write $U_n(i = \text{bus}) \leq U_n(j = \text{car})$. Models based on rational choice have proven to be robust across a wide variety of travel decision-making contexts (Castiglione et al., 2014).

Probabilistic choice models While intended to be comprehensive, choice data can only reveal so much about what factors drive individual decisions. Experimental analyses of travel decision-making models have demonstrated a degree of randomness inconsistent with the theoretical conditions defining utility maximization. That is, when making repeated decisions in identical choice situations, decision-makers are observed to make different choices. The inability of deterministic decision-rules to capture the empirically observed preferences of consumers in discrete choice situations precipitated the application of probabilistic choice theories to discrete choice analysis.

Rationality further requires that individuals posses full knowledge of the available alternatives and that alternatives are considered equally such that they may be ranked in preference consistent order (as just described). As discussed in the introduction, travelers must often make decisions under imperfect information. The expected utility theory (EUT) of von Neumann and Morgenstern 1953 (further developed by Luce and Raffia in 1957), extended rationality to preferences under uncertainty; that is, when only the probabilities, or, prospects, of choice outcomes are know to the decision-maker. Accordingly, models of rational choice based on EUT posit *expected utility functions*, which are computed as the product of constant utility and outcome probabilities (Von Neumann and Morgenstern, 1953).

The process by which economic agents arrive at the requisite "universally exhaustive and mutually exclusive set of alternatives"; however is not specified according to EUT. In activity-travel planning, the number of possible ways that an individual can plan a day is immense (Bowman, 1998). Experimental and behavioral economic research provides compelling evidence that individuals make decisions under bounded rationality (Fredrickson and Kahneman, 1993; Kahneman and Tversky, 1979; Simon, 1972), suggesting the salience of choice rules inconsistent with EUT.

Models of probabilistic choice differ in how they account for the source of the stochasticity in choice data. In the *constant utility* approach, the randomness is assumed to be due to suboptimal behavior of the decision-maker. That is, a consumer makes choices that maximize her utility with high probability, but has a non-zero probability of choosing lower utility options. According to the alternative *random utility* approach, randomness in decision outcomes is hypothesized to be attributable to components of the choice situation unobserved by the researcher (Marschak, 1959; Thurstone, 1927).

The random utility model is applied more commonly to the travel demand modeling scenarios we will be examining in this work, although it is possible to derive equivalent choice models using either framework. From the random utility perspective, the modeler hypothesizes that the net utility U_{in} of I_i to individual n is decomposed as the sum of a *systematic component*, denoted as $V_{in} = V(X_{in}; \beta)$, and a *random disturbance*, ϵ_{in}, measuring

specification errors made by a researcher analyzing i's behavior:

$$U_{in} = V(X_{in}; \theta) + \epsilon_{in}.$$

The vector of random parameters, θ, is unknown and must be estimated from data. Combining the above structural equation with a measurement equation of choice indicators,

$$y_{in} = \begin{cases} 1 & \text{if } U_{in} = \max_j U_{jn} \\ 0 & \text{otherwise} \end{cases},$$

enforces the maximum likelihood criterion, while lending the desired probabilistic interpretation to decision-making.

The *multinomial logit* (MNL) distribution is obtained by treating ϵ_{ij} as independently and identically distributed type I extreme value. The form of the systematic utility function in the MNL model is generally linear in the unknown parameters (i.e., $V_{in} = \theta^\top X_i n$), resulting in choice probability:

$$P(y_{in} = 1 \mid X_{in}; \theta) = \frac{e^{\theta^\top X_{in}}}{\sum_{j \in I_n} e^{\theta^\top X_{jn}}} \tag{2.1}$$

An important property of MNL (as well as related choice models) is the *independence from irrelevant alternatives* (IIA) condition Luce (1959). The IIA axiom stipulates that for any two alternatives i and j, the odds that n will select i over j is constant relative to the addition of alternatives to $\mathcal{C}_n$. Specifying a model with the IIA property simplifies the data collection task of the modeller in the sense that choice probabilities for multiple alternatives may be estimated from binary choice experiments. However, a common criticism of IIA is that, when many of the attributes of alternatives are identical to each other, IIA biases the choice probabilities, resulting in inaccurate predictions of choice outcomes. This problem is best illustrated by the classic "Red Bus/Blue Bus" (McFadden et al., 1973) example in which an individual, when faced with the choice of commuting by red bus or car, is observed to select both options with equal probability. Supposing that a new blue bus alternative is added, under IIA, the probability of selecting any alternative remains constant, (i.e., $P(C_{\text{blue bus},n}) = P(C_{\text{red bus},n}) = P(C_{\text{car},n}) = 1/3)$). However, it is more likely that the researcher would observe equiprobable selection between car and bus *modes* irrespective of the identity of the alternatives, (i.e., $P(C_{\text{blue bus},n}) = P(C_{\text{red bus},n}) = 1/4$, and $P(C_{\text{car},n}) = 1/2$).

For MNL models, we see that IIA holds:

$$\frac{P(C_{in} \mid \mathcal{C}_n)}{P(C_{jn} \mid \mathcal{C}_n)} = e^{\theta} \tag{2.2}$$

The *nested logit* (NL) model better capturing hierarchical substitution patterns among choices, wherein ϵ is correlated among mutually exclusive groups of alternatives (termed *nests*). Further relaxing the IIA assumption, models such as probit trade nearly unlimited

flexibility in the structure of the variance-covariance matrix of error terms in exchange for increased computational complexity. Unobserved heterogeneity among discrete or continuous segments of the population may be captured using mixture models (Kamakura and Russell, 1989).

Model Specification. In DCA, the attributes of alternatives and *socioeconomic characteristics* of decision-makers enter as explanatory variables in the utility function for each alternative. These variables are weighted by coefficients (also known as *parameters*) that define how important a variable is to the net utility of an option[1]. Part of the modeller's job is to specify salient variables that are *shared* among alternatives and use data to estimate the corresponding parameters. Specifications for alternatives typically include relevant policy terms, socioeconomic variables, as well as an alternative-specific constant term capturing the value of attributes not otherwise listed in the utility function.

Estimation. Parameter *estimation* in discrete choice models typically involves maximizing the *likelihood function*, assuming an independently and identically distributed (iid) data sample. Details of *maximum likelihood estimation* (MLE) as well as alternative, sampling based methods may be found in a standard reference such as (Ortúzar and Willumsen, 2011) or (Train, 2003). Upon convergence, outputs from estimation results often include standard errors for individual parameters and corresponding t-statistics. Model specifications as a whole may be tested for goodness of fit using indicators such as rho-squared. Alternative specifications that trade off parsimony with goodness-of-fit are typically compared with one another.

Software Implementations. Several well-regarded libraries exist for structural estimation and evaluation of discrete choice models. In Python, the `pylogit` library (`https://github.com/timothyb0912/pylogit`) provides a recent, efficient implementation that focuses on usability and comfortable transformation of data using the `Pandas` library (Brathwaite and Walker, 2016). The venerable `BioGeme` (`http://biogeme.epfl.ch/home.html`) in its Python and Bison implementations has been validated through a number of published studies (Bierlaire, 2016). The `mlogit` package (`https://cran.r-project.org/web/packages/mlogit/index.html`) for the `R` programming language has similar capabilities to its Pythonic brethren in addition to excellent documentation (Croissant et al., 2012).

Scheduling decisions in activity based travel demand models.

While convenient in its simplicity, the traditional four-step approach to demand modeling fails to incorporate the well-established basis for travel as derived demand arising from an individual's need to participate in a variety of activities that are not always co-located at a

[1]Note that discrete choice models are members of the class of *generalized linear models*, which permit nonlinear terms or interactions to enter into the utility function.

unique facility (Bowman, 1998; Chapin, 1974; Hägerstrand, 1970; Ortúzar and Willumsen, 2011). Moreover, many schedules include slack or flexibility to respond to uncertain external conditions such as weather events, stochasticity in transportation system performance, and unforeseen changes in the plans of friends, family, and co-workers (Bowman, 1998; Ortúzar and Willumsen, 2011). In order to better serve the planning community as a tool for policy analysis, modern *activity-based travel demand models* (ABTDMs) explicitly consider how the interdependence of activities as well as their time and space constraints influence decision-making at the *individual* level[2]. ABTDMs employ discrete choice models that generally obey time and space constraints when estimated. However, as we shall see, even state-of-the-art logit-based frameworks lack the flexibility to respond to novel policy interventions.

In ABTDMs, daily activity and travel preferences are conditional on more slowly changing lifestyle attributes such as activity priorities, jobs, habits, household roles and concomitant commitments. Lifestyle factors, in turn, may affect mobility decisions such as vehicle purchases, home or work relocation, as well as participation in employer- or government-sponsored commuter programs, scheduling a regular carpool, and the frequency of telecommuting. Given the outcomes of lifestyle and mobility decision processes, on any given day, more or less mandatory (typically known as *primary*) activities with fixed locations and schedules (e.g., work and school) constrain the selection of more flexible *secondary* activities (e.g., lunch and shopping), planned arrival and departure times, in addition to the mode of travel between activities (e.g., car, subway, or bicycle) whenever alternatives beyond a single option are *accessible*.

In contrast to older, trip-based models, ABTDMs integrate a sequence of trips that begin and start at the same location into a single *tour*. Tours may be characterized by a central activity or purpose. For example, a trip from home to work and back again constitutes a single *home based work* tour. The spatial and temporal restrictions imposed by subdivision of the day into tours can be used to enforce realistic vehicle use constraints (e.g., a personal car won't be available for a lunchtime appointment if a traveler commuted to work by tram) as well as coordination between members of the same household. Furthermore, ABTDMs enforce the temporal constraint that plans contain only those tours that can be completed within the timeframe of a single day. As we will see below, temporal constraints have important implications for the modelling of utility associated with daily activity-travel plans as well as the *value of time* (VOT) derived from estimates of utility function parameters.

The *day-activity schedule model system* of Bowman (1998) reflects the interdependence between spatial and temporal decisions inherent in an activity-based model. A daily *activity schedule* specifies a set of tours that are bound together with an *activity pattern*. Individual tour choices are themselves conditioned on the choice of activity pattern. Consequently, the probability of schedule choice is equal to the product of the pattern probability and the conditional probability of tour attributes given the choice of pattern for each tour in the pattern. Patterns define the overall structure of the day according to a primary activity,

[2]A detailed reference for design considerations and implementation details of ABTDMs is available in (Castiglione et al., 2014).

whether this activity occurs at home or away, tour type for primary activity, number and purpose of secondary tours and at-home episodes. Tour attributes include details of timing (i.e., departure time to and from the primary activity of the tour), activity location, and choice of travel mode.

The utility trade-offs inherent in the choice of a daily-activity schedule may be expressed using a nested logit model wherein the probabilities of lower tier nests (e.g., time of day and destination for car driver tours) are conditional on choices made at a higher level (e.g., choice of day activity pattern). Preferences over patterns are dependent on both their inherent relative utility as well as the expected maximum utility of associated tours. The *logsum* term refers to the maximum expected utility of lower tier choices.

Valuation of time An individual is assumed to derive positive utility from time spent engaging in scheduled activities. Unless she works at home, this person must also spend some amount of time in transit between activities. Traveling to an activity; however, leaves less time for activity participation and thus typically results in accrual of negative utility. These simple assumptions guide development of most of the existing full day utility modeling approaches used in ABTDM.

While time-of-day constraints such as departure time and duration have been integrated into the day-pattern model, (Vovsha and Bradley, 2004), the expected value of time later in the day does not influence decision-making occurring earlier in the day[3]. For example, the choice of departure time for work in the morning may vary depending on expected time spent in congestion. The actual arrival time at work will likely influence decisions made later in the day such as when to leave work or whether to go shopping before returning home. Understanding such trade-offs is necessary in order to forecast the effects of interventions designed to spread out peak traffic flows. The frameworks introduced in Section 2.2 may be applied to activity scheduling in order to estimate structural models of decision-making that are *dynamically consistent* with time-of-day constraints.

Microsimulations of travel demand. State-of-the-art *microsimulation* software is able to approximate the level of complexity needed to examine the policy forecasting implications of ABTDMs. By synthesizing the planned activities and transportation choices of sociodemographically heterogeneous populations and then realizing these plans on a virtual representation of physical road networks, microsimulations permit the resolution of feedback loops and spatiotemporal constraints operating between tour purposes, road network congestion, household vehicle availability, and infrastructure levels of service. Comparisons between business-as-usual and policy cases are typically computed using an iterative Monte Carlo process that randomly selects a subset of plans to mutate and then execute together with the other, unchanged plans. Outputs of microsimulations may be used to communicate policy alternatives to stakeholders. Visualizations of congested roadways with millions of

[3]we say that the two measures are *dynamically inconsistent*

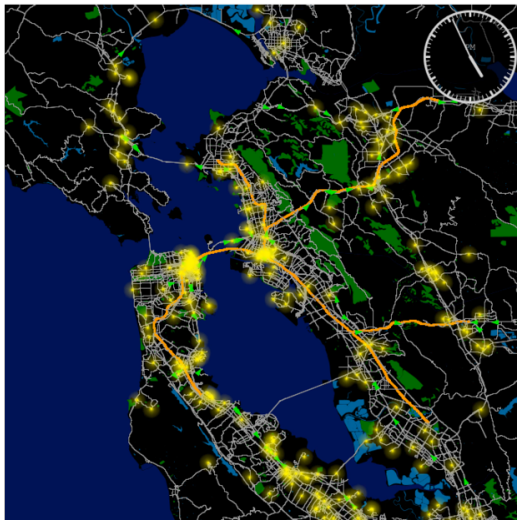


Figure 2.1: Visualization of agent-based microsimulation of travel demand.

agents behaving independently are a particularly compelling method to communicate infrastructure interventions Figure 2.1. More information on the operating routines of the simulation software MATSim (Horni et al., 2016a) will be presented in Chapter 3.

To date, the formulations of the full day utility function used to evaluate preferences between alternative plans and drive the Monte Carlo engine in utility-based simulations like MATSim are unable to account for within-day adaptive decision-making. The dynamic formulations presented herein could be incorporated into microsimulation software, permitting virtual agents to react realistically in response to, for example, variable road toll schemes or weather events.

Data development and aggregation. In the past, ABTDMs have used household-level travel surveys in order to estimate the parameters of discrete choice structural decision-making models (Arentze et al., 2000). Collecting travel diary samples is often time-consuming and expensive. In addition, sample sizes are often insufficient to estimate choice models of appropriate complexity. These shortcomings make it impractical to revise ABTDMs more often than once every five to ten years.

With improved spatial coverage, longer time horizons, and reduced turn-around time compared with traditional survey-based methods, passively-collected cellular records and social media postings are inspiring new data-driven strategies for travel demand modellers and urban planning practitioners to adaptively monitor, predict, and manage rapid changes in urban physical and social ecosystems (Song et al., 2010). Recent research efforts in transportation engineering have benefited from mobile technologies and data mining as accessible tools to monitor demand fluctuations in real time (González and Hidalgo, 2008; Pozdnoukhov and Kaiser, 2011; Widhalm et al., 2015; Zilske and Nagel, 2014).

For planning applications, incorporating CDR-derived mobility patterns into microsimu-

lations as done in (Lin Z. et al., 2017; Yin M. et al., 2017) has a number of advantages over mere extraction of activity locations and frequencies. These include: use of actual network (as opposed to straight-line) distances, population scaling, privacy preservation, inference of mode choice, and policy analysis. Ground truth validation and sampling bias continue to be major limitations to the more extensive use of CDRs in mobility assessments (Chen et al., 2016a).

Aggregate population assignment often involves the use of census data products mapped to socioeconomic characteristics in order to synthesize realistic travellers. Calibration of network flows (representing one of the outputs of microsimulations) is usually performed using count data from sensors embedded in road infrastructure (Castiglione et al., 2014).

Cost benefit analysis in travel demand models.

Discrete choice models are often used by urban planners to aggregate preferences when performing a cost-benefit analysis (CBA) of transportation policy and infrastructure investment alternatives. In fact, the logit assumptions of the MNL model described above make computation of consumer surplus particularly straightforward (Train, 2003).

The typical objective of the social planner is improvement of social welfare according to a criterion such as *Kaldor-Hicks efficiency*[4]. Faithful structural models of behavior permit forecasts of changes due to policy measures at aggregate and disaggregate levels. They involve the computation and analysis of elasticities[5] of travel demand, a necessary quantity for CBA. The value of travel time savings (VTTS) meanwhile captures the variation in value of time (VOT) across alternatives.

The research presented in chapter 3 extends related research using MATSim to perform CBA for policy incentives that internalize the environmental externalities of transportation (Agarwal and Kickhöfer, 2015b; Axhausen, 2007; Kaddoura and Kroeger, 2015; Kickhöfer, 2016; Kickhöfer and Agarwal, 2015). In transportation settings, the scope of externalities varies from local to global. For example, the noise pollution from a loud motorcycle engine may irritate people in the immediate vicinity of the vehicle. In contrast, CO_2 emissions from fossil-fuel powered automobile exhaust contribute to global climate change, which has a more diffuse social cost. External costs need not be directly monetary in nature. The increase in travel delays caused by an additional driver on the road network affects commuters in accordance with their value of time (Small, 2012).

As the scope and uncertainty associated with transportation externalities increases, *Pigouvian mechanisms* become more politically contentious, since attributing external costs to their sources becomes less precise. Pigouvian mechanisms also fail to take advantage

[4]In welfare economics, a policy achieving *Pareto improvement* implies that no one is worse off by a policy and that at least one person's welfare improves. Kaldor-Hicks efficiency, meanwhile, is a more realistic criterion, qualifying a policy as more efficient than the status-quo as long as winners can potentially pay the losers (Hicks, 1939; Kaldor, 1939)

[5]Price elasticity of demand is a measure of the change in demand for a good due to a change in price. Elasticities are unitless values computed relative to the status-quo.

of the social effects that often shape individual behavior and preferences (Ostrom, 1990). Policy mechanisms involving directly charging an agent with the social costs that he was responsible for have been evaluated by (Kaddoura et al., 2014; Kickhöfer and Nagel, 2013). Negative perceptions of the distributive justness of a behavioral policy intervention may reduce its effectiveness (Kamas and Preston, 2012; Konow, 2010). The mechanism through which this operates likely has to do with heterogenous values of time (VOT) for individual agents. Values of time exhibit sociodemographic variation (e.g., according to income), so the assumption of an homogeneous value of time when combining utilities across a population segment may result in aggregation bias when computing value of travel time savings (VTTS) (Castiglione et al., 2014; Small, 2012). However, cost-benefit analysis of policy alternatives using a differentiated VOT may result in overweighting individuals with a high wage rate (Kickhöfer, 2014), potentially activating tendencies towards inequality aversion, as explained in Section 2.3, below (Kamas and Preston, 2012).

2.2 Sequential Decision-Making Frameworks.

Unexpected circumstances such as adverse weather, automobile accidents, vehicle breakdowns, or significant public events (to name a few) may cause deviations from planned activities. More severe disturbances, such as the addition of a new rapid bus service or a road closure due to a snowstorm may induce periods of more concentrated schedule re-evaluation. The resulting schedule modification may be as prosaic as selecting a slightly different route to work, or as extreme as switching to a new mode of travel altogether.

Additionally, to a varying degree, people tend to seek out opportunities that could lead to an increase in life fulfillment. Boredom with local lunch spots may lead one to explore unfamiliar neighborhoods. A friend's invitation to a fitness class may inspire a new hobby with cascading changes to one's repertoire of schedules.

The DCA-based models described in Section 2.1 are limited to static contexts wherein agents do not account for the impact of future opportunity costs when making near-term decisions. That is, at the time of decision, agents do not incorporate known social, structural, and temporal constraints that may affect subsequent choices. This section describes two interrelated frameworks of dynamic decision-making and their use in travel demand modelling.

Markov decision processes.

Markov decision processes, as used in the dynamic choice models described in this thesis, are defined at the level of an individual, forward-looking agent (representing a single decision-maker or a group making joint decisions) observing a signal representing discrete *states* of the *world* (also known as an *environment* or *task environment* in the reinforcement learning literature). At successive time steps, the agent makes decisions (sometimes denoted *actions*

in the reinforcement learning community) that change the state of the world as well as produce a numeric *reward* signal.

In general, the current state is a function of the history of all past decisions. Under the *Markov assumption*, it is sufficient for the state to encode all of the information necessary for the agent to compute the next action and its consequences (i.e., the next state and the utility of the next state)[6]. When the state or action space consists of multiple attributes, it is often useful to use vector representations. While the dimensions of action spaces are often rather small, state spaces may be large or continuous. In terms of computational complexity, the so-called "curse of dimensionality" renders dynamic programming techniques infeasible unless simplifying assumptions or approximations can be made (Bellman, 1954).

The *goal* of the agent is to, at each time step, choose actions according to a sequence of decision rules, collectively known as a *policy*, that maximizes the cumulative sum of the expected value of future rewards, referred to as *returns*, over a (typically finite, possibly infinite) time *horizon*. This objective may be contrasted with that of an agent in a static decision problem, i.e., to maximize the expected utility of a single choice. The *dynamics* of the environment are governed by a set of transition probability functions, which give the probability of transitioning to a new state, given the current state and action. These dynamics may be assumed or estimated from data.

A salient decision-theoretic assumption of the *normative* application of MDPs is that immediate gratification is preferable to the decision-making agent. In addition, the uncertainty surrounding *exogenous* future events introduces an element of risk that, if uncompensated, diminishes the value of delaying payoffs. Together, these assumptions imply that future utility may be *discounted* by a factor, which is typically a constant between zero and one.

Dynamic discrete choice models

First formalized by John Rust in 1987, *dynamic discrete choice models* (DDCMs) address the shortcomings of DCA to model adaptive decision-making behavior by extending random utility theory to a dynamic environment (Rust, 1987a). In DDCMs, the state is partitioned into two terms, one observed by both the researcher and decision-maker and another, random term, representing components of the state known only to the agent. As with static models, the goal is to estimate the structural parameters governing choice preferences. However, one may also estimate transition probabilities. By assuming that the random factors are additively separable, and conditionally independently and identically distributed over time

[6]Note that a full knowledge of the state by the agent is a necessary condition for rationality. If any part of this state cannot be observed by the agent, then we term the MDP a *partially observable Markov decision process* (POMDP) (Kaelbling et al., 1998). In such a situation, it is more convenient to speak in terms of *observations* of the state, which result in the agent maintaining a *belief state* characterizing the extent of an individual's state of knowledge of the world. While a full treatment of the additional complexity introduced by POMDPs is beyond the scope of this documentation, their consideration in DDCMs and (inverse) reinforcement learning models is an active and evolving topic of research, particularly in the reinforcement learning literature.

Extreme Value type I, it becomes possible to estimate agent preferences from data via a nested fixed-point maximum likelihood algorithm (Rust, 1987a).

Prior applications of DDCM to travel-activity scheduling problems

While there have been several empirical studies of DDCMs in the transportation domain (see (Cirillo and Xu, 2011) for a review), herein we describe prior research on dynamic decision-making in activity-travel (re)planning contexts.

In (Jonsson and Karlström, 2005), the authors use dynamic discrete choice within their proposed SCAPES framework, which models an individual daily activity-planning process as an MDP with transitions occurring at evenly spaced decision epochs over the course of the day. State is represented using a vector that incorporates the current time, location (from a finite set), car availability, escort (i.e., child drop-off) responsibility, food stock maintenance level, and work hour maintenance level. Uncertainty in state transitions (delay) during the morning and evening hours is used to represent congestion. The optimal policy is found through backward induction on the value function, (only a single sweep is necessary, as state transitions are from the current time step to a future step only). Example deterministic choice model formulations are presented for finite and infinite planning horizon models, where the infinite horizon problem represents a continuation between days of the optimal policy defined on the product state space of the maintenance variables (food stock and work flex time level). Here, the form of the reward function is treated as given and simulations of realistic agents[7] with different levels of stock variables are performed in order to determine their effect on departure time choice. Simulations were minimally calibrated in order to ensure that infeasible states were unreachable. (Väastberg et al., 2016) extends (Jonsson and Karlström, 2005) by providing a sampling-based utility modeling and estimation framework.

Reinforcement learning and inverse reinforcement learning.

Reinforcement learning (RL)-based frameworks share many similarities with the DDCM formulation presented above. However, as the name suggests, RL is more concerned with how agents learn to make optimal sequential decisions in episodic and non-stationary environments. With an explicit notion of exploration, habit-formation, and goal orientation that has both psychological and neurological antecedents, RL is a suitable general framework for adaptive decision-making under uncertainty.

Reinforcement learning.

As in the DDCM framework, RL algorithms permit inference of sequential decision rules as well as the reward functions motivating observed behavior. Compared to the DDCM framework; however, RL offers a wider array of tools to deal with learning adaptive, optimal policies in complex, uncertain environments.

[7] Characteristics were based on a prior congestion charging policy in Stockholm.

In RL, the objective of an untrained agent navigating a new task environment is to learn a sequence of actions (or distribution over sequences) leading to states that maximize the agent's experience of a reward signal over time. RL algorithms have been developed for a variety of environment models, admitting both continuous and discrete representations of states and actions. Goal-directed behavior may be modeled with the use of a terminal, absorbing state.

The adaptive decision rule governing the agent's behavior is known as a policy. Depending on the nature of the learning task, policies may be deterministic or stochastic and may be flexibly parameterized using a universal function approximator such as a neural network or a Gaussian process.

In model-free methods, agents are not assumed to know the transition kernel. This is in contrast to dynamic discrete choice models, which often assume that the agent has some existing understanding of which actions lead to which outcomes. However, without initial knowledge of the system dynamics, must traverse the environment as if blind: learning about the environment and adjusting its policy based on the actions and resultant next states and reward signals. In off-policy methods such as Q-learning, the agent may accumulate experience in memory to train an optimal action-value (Q) function, which, for each state in an environments state space, represents the action leading to states that accumulate maximum expected rewards.

In addition to learning an optimal policy, the agent may learn a model of the environment in what is known as model-based RL. Model-based RL permits simplified adaptation to alternative goals. A further benefit of learning a model of the environment is that salient changes in the system dynamics may be detected, which could also result in more rapid convergence to a new optimal policy. In a daily activity-planning problem, for example, an agent could memorize multiple contingency plans in order to better adapt alternative policies conditional on new information. However, even with a model of the environment, it is important to understand what rewards motivate agent behavior. The next subsection deals precisely with this topic.

Inverse reinforcement learning.

Inverse reinforcement learning (IRL) is concerned with developing methods to estimate the parameters of a reward function from a dataset of states and actions known as *demonstrations* (alternatively referred to as *trajectories* or *traces*). While initially developed to train robots, IRL research has more recently sought to harmonize the strong behavioral and neuroscientific underpinnings of RL with structural estimation of dynamic human behavior. The following brief overview of IRL is provided in order to inform a subsequent discussion on the merits and limitations of the IRL methodology compared to DDCMs.

Individual demonstrations are assumed to be sequential state and action pairs, tracing the behavior of a *rational* agent navigating a (potentially high-dimensional) state and action space over a (typically finite) time horizon. The assumption of rationality does not; however, preclude the agent from making mistakes on the way to its goal, producing a suboptimal

trajectory as its output. More precisely, instead of using a deterministic decision rule at each timestep, the agent may be presumed to follow a stochastic policy that prefers high net reward trajectories over lower ones and randomly selects between trajectories with identical rewards.

In a complimentary sense, IRL can be seen as a method to build a *generative* model of MDPs in the form of a probabilistic graphical model (PGMs). In the context of goal-oriented behavior, solutions to an MDP modeled as a PGM enable analogies to be made between probabilistic inference and planning. Accordingly, it is possible to give behavioral interpretations to a variety of standard PGM inference tasks. Using this perspective, we can sample random trajectories that approximate the behavior of the agent(s) who contributed their data to the dataset.

As is the case with policies, reward functions learned using IRL admit flexible linear and nonlinear representation. If rewards are intended to be interpretable; however, it is still incumbent on domain experts to specify the relevant features of the reward function for estimation purposes.

Prior work in RL and IRL.

Existing works in the travel-demand literature have focused on applying so-called value-based RL methods (e.g., Q learning, (Watkins, 1989)) to the activity scheduling problem (Davy, 2006; Ma and Gerber, 2016; Yang et al., 2014). Early insight into the MATSim structural equation model was developed through Q-learning based approaches (Charypar and Nagel, 2005b; Vanhulsel et al., 2007). A more comprehensive dynamic public transit path and activity choice model was developed in (Medhat et al., 2008) as the MILITRAS system. All of these formulations use a predetermined reward (utility) function; a significant limitation to the use of "forward" RL methods alone to learn decision rules for activity scheduling.

While our application of IRL to the inference of activity-travel plan scheduling is novel, several studies have successfully adapted IRL as a tool to rationalize sequential decision-making behavior observed in transport settings. Route choice and destination prediction from observations of taxi trajectories was highlighted in the original presentation of maximum entropy IRL by Ziebart and Maas (2008). In (Vogel et al., 2012), a hybrid vehicle control policy is optimized for fuel efficiency while taking into account driver preferred routes. A related work by Ondruska and Posner (2014) determined range maps for electric vehicles. Demonstrating the ability of IRL methods to model boundedly rational agents, Ratliff and Mazumdar (2017) have recovered preference for risk aversion using Uber Movement data.

The surprising success of deep reinforcement learning in replicating complex agent behavior from observations (Mnih et al., 2013; Silver et al., 2016) as well as studies showing many neural correlates of RL algorithms (Niv, 2011) has spurred a great deal of research into novel RL methods using artificial neural networks as general-purpose function approximators. Works published within the last year have investigated the emergence of cooperative behavior using deep RL with multiple agents (Hausknecht et al., 2016; Leibo et al., 2017). The study by Leibo et al. (2017) is notable in that it generalizes matrix games of social

dilemmas to more complex state and action spaces where cooperative or non-cooperative behavior does not involve purely atomic decisions. IRL has also begun to benefit from deep learning models, with extensions of MaxEnt IRL demonstrating efficient recovery of complex reward models from high-dimensional data (Finn et al., 2016b; Wulfmeier et al., 2015, 2016).

The growing interest in IRL methods has inspired collaboration between Computer Science and social science disciplines. In (Ermon et al., 2015), it is shown that under certain conditions, the logit Dynamic Discrete Choice formulation (logit DDC) of Rust (1987b) and Maximum Entropy IRL trajectory distributions are equivalent. The authors demonstrate that utility functions from high-dimensional spatiotemporal data can be recovered in order understand long-term migratory behavior. IRL has also been used in multiagent scenarios. Game-theoretic formulations of IRL and MaxEnt IRL were proposed early on (Syed and Schapire, 2008; Waugh et al., 2008). Inquiry into the use of inverse reinforcement learning to infer social norms in multiagent social dilemmas has also been initiated (Ho and Ermon, 2016; Kleiman-Weiner and Tenenbaum, 2016). We take up the topic of social interactions with transport behavior in the following section.

2.3 Social Dynamics In Transportation Choice Settings

Our motivating application considers the setting of interdependent commute mode choice as a social dilemma. The determinants of behavior in social dilemmas assume the existence of altruistic agents (S. Frey and Benz, 2001). As previously discussed, modern ABTDMs using DCA typically assume rational, expected discounted utility-maximizing travelers.

The transportation choice modelling literature has recognized the importance of relaxing the assumption that agents are purely self-interested (Avineri, 2012; McFadden, 2005). Although several studies have extended discrete choice methods to model a variety of social influences on travel decisions, discrete choice models typically do not account for social influence (Dugundji and Walker, 2005; Páez and Scott, 2006; Scott, 2004). Furthermore, the difficulty and cost of collecting large survey samples with sufficient social connectivity limits the statistical representativeness of DCA-based models in aggregate analyses used for planning applications.

This section discusses the relevant typology of social preferences, game theoretic models of decision-making in groups and communities, as well as related empirical applications and simulation experiments.

Social preferences and social dilemmas.

One may dichotomize social preferences as *reciprocal*, defining an individual's tendency towards altruistic or myopic behavior, or *distributive*, capturing notions of equity or fairness (Kamas and Preston, 2012). Examples of reciprocal social preferences may take the form

of norm activation mechanisms. As conceptualized by Schwartz (1977), personal norms define feelings of moral obligation that operate within a social-psychological theory of altruism (Bamberg and Schmidt, 2003). Norm activation mechanisms attempt to raise awareness of the moral implications of socially uncooperative actions such as driving alone in order to inspire incipient feelings of responsibility for ones actions. As an example of distributive social preferences, individuals with a tendency towards *inequity aversion* sacrifice personal utility to ensure that the distribution of payoffs in an interdependent choice situation is equal (Fehr and Schmidt, 1999).

In social dilemmas, players (economic actors or agents) have a rational incentive to behave selfishly; however, all players stand to gain if cooperation is achieved. Long-standing conflicts between individual and collective well-being in shared resource settings have often been framed in terms social dilemmas over *common pool resources*. By definition, common pool resources are freely-available and ungoverned (Ostrom, 2010). *The tragedy of the commons* may occur when a common pool resource is depleted to the extent that each‘ additional unit of consumption reduces the value of the resource for the entire society (Hardin, 1968; Ostrom, 1990; Ostrom and Walker, 1991).

Social dilemmas are often modeled using game-theoretic variations on Prisoner's Dilemma (PD), public goods games (PGG), (Dawes, 1980; Ledyard et al., 1997) and commons pool resource (CPR) games (Ostrom, 1990). For example, in the N-player PGG, players are allocated tokens and must secretly decide on how much to contribute towards a central pot, which may be characterized as some "public good". Tokens in the pot are multiplied by some factor $1 < r < N$ and re-distributed evenly among the players. A player's winnings are equal to the tokens received from re-allocation plus any remaining private tokens not contributed. Players have a rational incentive to "free-ride", that is, contribute nothing, while still receiving the public allocation. The Nash equilibrium of the PGG predicts zero contributions from each player, yet, in experimental settings, the pure Nash equilibrium is rarely seen (unless the multiplication factor is set extremely low) (Camerer and Fehr, 2004; Leyton-Brown and Shoham, 2008).

Repeated play with many random pairings of individual agent strategies has led to various interpretations of evolutionarily stable strategy (ESS) concepts (Smith and Price, 1973). Popularized in the iterated PD tournaments of Axelrod and Hamilton (1981), N-player matrix games (e.g., PD, PGG, and CPR) have been frequently studied in the economics, control, and distributed artificial intelligence communities to model interdependent appropriations of telecommunication, wildlife, and global climate resources (Diekert, 2012; Ostrom, 1999; Parkes and Ungar, ????; Saha and Sen, 2003; Turner and Turner, 1992). For example, common formulations of commute mode choice as a social dilemmas are as an N-Person Prisoner's Dilemma or as an N-person Chicken Dilemma Game (depending on whether pro-environmental or accessibility-related values are hypothesized to be predominate decision-making in the study population) (Van Vugt, 1996). Modern, interdisciplinary extensions have explored the spatial, social, and learning dynamics of these games to understand mechanisms and behavioral patterns guiding the evolution of cooperation (Capraro, 2013; Nowak and May, 1992; Rand and Nowak, 2013). Viewing social influ-

ence on travel choice through the lens of Behavioral Economics or game theory may serve as a useful complement to the DCA methods used in transportation demand modeling practice insofar as game theoretic techniques can be used to understand the factors motivating the emergence of socially-cooperative travel decisions.

Empirical studies of social preferences on travel decision-making.

Social preferences have been shown in laboratory and field experiments to play an important role in determining whether individuals are more likely to prefer socially-cooperative outcomes in choice situations (S. Frey and Benz, 2001; Van Vugt, 1996). Understanding the extent to which social preferences influence pro-environmental decisions is of particular importance in the study of transportation behavior, as growing concern over the contribution of greenhouse gas (GHG) emissions from fossil fuels to climate change and increasing access to low-cost, alternative-energy transportation modes, has resulted in urban commuters switching to public transit, electric vehicles, and ridesharing services at rising rates (Biel and Thøgersen, 2007; SFCTA, 2010). For example, recent work studying automobile purchase decisions shows to what extent adoption of a new technology (such as electric vehicles) is driven by the spread of attitudes and behaviors (e.g., pro-environmental mode choice) as they become social norms (Gaker et al., 2011; Nordlund and Garvill, 2003).

Individuals can shape their preference for alternatives via social comparison with their peers, basing their own perception of an alternative on the advice or apparent well-being others derive from outcomes in similar situations. From a psychological perspective, greater satisfaction with the attributes of a transportation mode (e.g., reliability or travel time) relative to members of a reference group was a significant determinant of personal utility for commuters (Abou-Zeid and Ben-Akiva, 2011). Social comparisons in transport decision-making can lead to diffusion of influence through social networks via so-called "mass-effects" (Abou-Zeid et al., 2013). Imitation and innovation models in transport have often used the famous "Bass-model" (Bass, 1969) as in (Schmöcker et al., 2014), but dynamic discrete choice models have also been used (El Zarwi et al., 2017). As discussed in (Manski, 1993); however, well-identified structural model specifications must account for homophily in order to avoid endogeneity via the so-called *reflection problem*, a criterion that reduces the value of data from observational studies in making causal claims.

Individuals may also learn their preferences based on repeated interaction with other actors in their social network. Transportation researchers have increasingly explored the impact of social preferences on human mobility behavior by developing novel big data collection and analysis methods for location-based social networks (LBSN) (Axhausen, 2007; Carrasco et al., 2008; Grabowicz et al., 2014; Verplanken et al., 2008). Reciprocal call frequency data extracted from CDRs have been used to study human mobility dynamics in social networks of large urban environments (Hidalgo and Rodriguez-Sickert, 2008; Walsh et al., 2013). When combined with census data, LBSNs may also reveal community and influential individuals that could be more or less susceptible to public policy mechanisms as demonstrated, for example, in (Zhang D. et al., 2017).

Simulating social dilemmas

Agent-based simulations of social dilemmas permit theoretical investigations into the mechanisms underlying such social preferences in complex adaptive systems. Sunitiyoso et al. (2011) investigate various theories of social learning in a mode choice scenario, which bears some similarities to our work. The authors use laboratory experiments to investigate several alternative theories of learning from social preferences after a hypothetical employer-based commuter reward program is implemented. The findings from the lab studies are then used to inform large-scale simulation models under counterfactual scenarios. While the use of a empirical data is important in validating theories of social preferences, as stated previously, the representativeness of small sample SP studies is affected by hypothetical and ecological biases.

Research efforts in agent-based microsimulation have explored the effect of social influence on transportation in the domain of joint decision-making (Hackney and Axhausen, 2006). Studies employing synthetic and survey-based social networks of travelers as well as simulations of joint activity choice, vehicle sharing, and household-level coordination of plans have been implemented (Dubernet and Axhausen, 2013; Hackney and Axhausen, 2006; Illenberger, 2012). While our work is similar in that we extend a microsimulation with a social network of traveling agents, the scenario presented herein does not involve joint decision-making by agents during the course of plan evaluation.

3

Peer Pressure Enables the Actuation of Mobility Lifestyles

> *Effective policies are those that support socially valued outcomes not only by harnessing selfish motives to socially valued ends, but also by evoking, cultivating, and empowering public-spirited motives.*
>
> – Adam Smith, *Theory of Moral Sentiments*

In this chapter[1], we present a model of strategic peer pressure behavior, which we have adapted from the original formulation by Mani et al. (2013) to take place within the larger context of an activity-based model of urban travel. Our focus herein is on the impact that active interpersonal influence has on an agent's travel mode choice in the presence of system-wide externalities arising from the concurrent execution of all agent's plans on the physical network. Agents are made aware of externalities such as traffic and CO_2 emissions via penalties to the utility of their realized plans. In response, agents producing the fewest externalities (e.g., public transit users) may choose to exert pressure on their peers (e.g., agents who drive alone) in the hope that they will follow suit. It is expected that daily travel mode choices will vary for individual agents as they induce and respond to peer pressure. However, if localized changes coalesce into cascading network effects and, consequently, a sufficient reduction in externalities is achieved, we expect that, at system equilibrium, individual shifts towards socially cooperative modes will be stabilized in the form of significant increases in socially cooperative mobility lifestyles.

The software implementation of our theoretical model is built on top of the open-source agent-based microsimulation, MATSim, which was chosen for its compatibility with behavioral choice theories, modularity, and ability to handle large heterogeneous populations. The MATSim co-evolutionary algorithm executes the scheduled activity-travel plans of (potentially millions of) agents in a virtual representation of a physical road network, iteratively improving and replanning plans of randomly selected subpopulations until stochastic user

[1] A version of this work was published as (Feygin, S. and Pozdnukhov, A., 2017).

equilibrium (SUE) is achieved (Ortúzar and Willumsen, 2011). A comprehensive introduction to MATSim is available in (Horni et al., 2016b).

While well-established travel demand modelling software such as MATSim can accurately model physical interactions between agents and transport infrastructure, researchers are unable to specify flexible and data-driven structural models of individual decision-making within these software. Thus, the work presented in this section illustrates the strengths of existing behavioral microsimulation frameworks as well as their inherent limitations.

3.1 Methodology

As has already been well-established in the transportation modeling literature, the demand for travel is derived from an agent's need or desire to participate in activities (e.g., shopping or working) (Bowman, 1998; Hägerstrand, 1970). Scheduling a daily activity-travel plan requires individual agents to make several hierarchically-structured decisions that satisfy spatiotemporal constraints, financial restrictions, professional obligations, and meet a variety of other considerations.

Herein, however, we narrow the scope of this complex decision-making process to focus on the impact that active interpersonal influence has on an agent's travel mode choice in the presence of system-wide externalities arising from the concurrent execution of all agent's plans on the physical network. In the present scenario, agents are made aware of externalities such as traffic and CO_2 emissions via penalties to the utility of their realized plans. In response, agents producing the fewest externalities (e.g., public transit users) may choose to exert pressure on their peers (e.g., agents who drive alone) in the hope that they will follow suit. It is expected that daily travel mode choices will vary for individual agents as they induce and respond to peer pressure. However, if localized changes coalesce into cascading network effects and, consequently, a sufficient reduction in externalities is achieved, we expect that, at equilibrium, individual shifts towards socially cooperative modes will be *sustained* in the form of significant increases in socially cooperative *mobility lifestyles*.

The rest of this section is organized as follows. In subsection 3.2 we define the baseline model as a single agent decision problem; albeit one solved simultaneously by many agents connected via a social network. We then describe how, in the presence of multiple agents–each attempting to make optimal decisions–we reformulate the baseline model as a two stage game. In subsection 3.2, we incorporate the effect of peer pressure and describe its effect on agent behavioral preferences. To aid in comprehension, subsection 3.2 presents a toy numeric example. Finally, in subsection 3.2 we describe the details of a full-scale implementation of our modeling framework in the multiagent travel microsimulation. A summary of notation used in this section and throughout this article is provided as Table 3.1.

N	Set of agents
$Nbr(i)$	Set of neighbors for agent i
$\mathcal{X}$	Set of accumulated plans
X	Plan memory
$\mathcal{H}$	Plan history
x^m	Daily activity-travel plan m
$\boldsymbol{a}^m$	Vector of attributes for plan m
β^m	Coefficient weights
V^m	Systematic utility function
$\boldsymbol{x}$	Action profile (set of plans for all agents)
$\boldsymbol{x}_{-i}$	Set of plans for all agents other than i
$\boldsymbol{x}^*$	Equilibrium action profile
$\boldsymbol{x}^\circ$	Action profile optimizing social welfare
ν	Externality function
mode	Mode choice indicator function
U^m	Utility function of plan m
ΔU_i	Utility gap for agent i
$\mathcal{S}$	Social welfare function
$\boldsymbol{P}$	Peer pressure matrix

Table 3.1: Notation summary

3.2 Baseline Model

Agent decision-making behavior may be modeled as a repeated game played sequentially on consecutive days, $t = (1, 2, \ldots)$, by a set of agents $N = \{1, \ldots, n\}$. Agents are interconnected via a social network G $= (N, E)$, where $E \subseteq N \times N$. Each agent, $i \in N$, has at most K peers in their neighborhood, $Nbr(i) = \{j : (i, j) \in E\}$, such that the graph representing the social network is sparse. An ordered pair of vertices $(i, j) \in E$ denotes a directed social tie emanating from i and incident upon another agent j; conversely, the pair $(j, i) \in E$ denotes a directed edge from j to i. In our formalism, the meaning of edges starting and terminating at the same vertex is undefined, so $E = \big\{(i, j) \in 2^N \mid (i \neq j)\big\}$. The social network structure is assumed to be static: links between agents are fixed and link formation and destruction processes are undefined. For simplicity, ties are assumed to be reciprocal in strength.

At the beginning of each day t each agent $i \in N$ chooses a single activity-travel plan x_{it}^m from a finite individual set of accumulated plans $\mathcal{X}_{it}$. The selected plan x_{it}^m represents a *mental model* of i's schedule on day t, which execute in a *physical model* of the network environment. More specifically, the physical model simulates the spatiotemporal dynamics of daily interactions between agents' vehicles on a capacity-constrained transportation network that permits travel between activity facilities at times specified by the schedule. The full history of an agent's executed plans is denoted $\mathcal{H}_i$.

A vector of plans, which we term an *action profile* $\boldsymbol{x}_t$ represents the outcome of the plan

selection process for all N agents. The action profile indicates which plans to *simultaneously* execute in the physical layer. Once a plan has been executed, an agent may choose to modify the plan based on beliefs of future system performance, which themselves are updated conditional on information gathered from their past experiences. Due to the competition of agents for finite road access, subway car space, and other transportation infrastructure capacity constraints, the quality of an agent's plan depends on the decisions $\boldsymbol{x}_{-it}$ of the $N \setminus \{i\}$ other agents, which we denote $-i$. In order to better capture constraints on human abilities to remember and adapt to the specifics of their daily travel experiences in a dynamic, multi-agent urban environment, agents occasionally forget plans that they rarely execute[2]. An agent i's *memory* X_i, is a fixed size vector, containing tuples of previously experienced plans, $x_{it}^m \in \mathcal{H}_i$ and a corresponding utility score U_{it}^m. A *removal rule* is associated with each X_i, which ensures that agents maintain bounds on the cardinality of X_i, *i.e.*, $|X_i| = M$ at any time t.

The solution concept for the physical system is, in this case, an agent-based stochastic user equilibrium (SUE) (Flötteröd and Kickhöfer, 2016). Let $\boldsymbol{x}^*$ denote the steady-state action profile that is consistently selected and executed at equilibrium. Once SUE is achieved, for all t, each agent is assumed to select plans, x_i^m from a fixed memory, X_i^*, that maximize his enjoyment of activities while minimizing other marginal private costs (MPC) associated with scheduling choice dimensions such as travel mode, route, activity destination, departure time, *etc.*. These attributes are represented as a vector, $\boldsymbol{a}_{it}^m$. Once x_{it}^m is selected and executed, the total utility that i derives from the plan over the course of day t is partially governed by a *systematic utility* function, $V_{it}^m = V(\boldsymbol{a}_{it}^m)\ \forall m$. The systematic utility of a single plan for agent i is a linearly-weighted combination of attributes for the plan:

$$V_{it}^m = \beta_{it}^{m\top} \boldsymbol{a}_{it}^m, \tag{3.1}$$

where β_{it}^m is a vector that parametrizes the marginal utility of plan x_{it}^m's attributes. At SUE, agents are generally fully conscious of the attributes governing their own choice of optimal plan, although they may only be vaguely aware of the attributes governing other agents' choice of plan.

Assuming that the random errors for agents associated with selection of plans at SUE are independently, identically distributed type I extreme value, the plan selection probabilities in the baseline model are assumed to be specified by a multinomial logit discrete choice model,

$$P(x_{it}^m \mid X_{it}) = \frac{e^{\mu_i U_{it}^m}}{\sum_{U_{it}^k \in X_{it}} e^{\mu_i U_{it}^k}}, \tag{3.2}$$

where μ_i is a heterogeneous scale factor measuring the agent's preference for higher scoring plans serving as a rationality parameter, where $\mu \to \infty$ corresponds to a deterministic choice

[2]The agents may thus be said to possess *imperfect recall* due to their limited memory (Rubinstein, 1998). Relaxing the strict informational requirements of perfect recall also has beneficial computational implications, since maintaining the experienced plan histories for all n players would be computationally infeasible (Waugh et al., 2008).

of the best performing plan. This assumption corresponds to the standard random utility model (Ortúzar and Willumsen, 2011; Train, 2003).

When planning his day, an agent typically ignores the marginal external costs (MEC) that execution of their preferred plan in the physical environment imposes on other agents. In order to account for agent preferences in the presence of aggregate external costs, we introduce an externality function, $\nu_{it} : \boldsymbol{x}_{-it} \to \mathbb{R}$ representing the disutility experienced by i due to $\boldsymbol{x}_{-it}$. The total utility of plan selection for agent i on day t is then defined as:

$$U_{it}^{m}\left(x_{it}^{m}, \boldsymbol{x}_{-it}\right) := V_{it}^{m} - \nu_{it}\left(\boldsymbol{x}_{-it}\right). \tag{3.3}$$

In order to lighten notation, we drop further indexing on t and m. The sequential nature of simulation and memory effects are highlighted wherever it is germane.

In the physical model, agents travel between activities by either driving a car or by using some more socially cooperative form of transportation (*e.g.*, public transit or walking). At equilibrium, every agent is assumed to have a preferred choice of transportation mode corresponding to the plan $x_i \in X_i^*$ with maximum U_i. We denote $\text{mode}_i(X_i^*) \in \{\text{car}, \text{sc}\}$ as the preferred transportation mode for a single agent at SUE. A driving agent is an agent for whom $\text{mode}_i(X_i^*) = \text{car}$, and, likewise, an agent that prefers to commute using socially cooperative modes of transportation has $\text{mode}_i(X_i^*) = \text{sc}$. While agents may have forgotten the details of previously selected plans when considering a change in the choice of transportation mode used during daily travel, they do remember the utility associated with their best past experience of the different modes used during plan execution[3].

An agent is also assumed to be generally aware of the primary transportation mode that his neighbors $j \in Nbr(i)$ prefer to use. We define $\Delta U_i = \Delta U_i(X_i)$ as the *utility gap*, which expresses the difference between the utility score of i's equilibrium plan and the utility of the best scoring socially cooperative (*i.e.*, transit or walking) plan in i's memory, which we denote

$$x_{i,\text{sc}}^{\circ} := \arg \max_{(x_i) \in \{\boldsymbol{x} | x_i \in X_i^*;\, \text{mode}(x_i) = \text{sc}\}} U_i(x_i, \boldsymbol{x}_{-i}).$$

The social welfare is defined as the sum of utilities experienced by all agents following plan execution:

$$\mathcal{S}\left(\boldsymbol{x}\right) := \sum_{i \in N} U_i\left(x_i, \boldsymbol{x}_{-i}\right). \tag{3.4}$$

The action profile optimizing social welfare is denoted $\boldsymbol{x}^{\circ}$. In the presence of externalities, we know that the social welfare at the equilibrium action profile, $\boldsymbol{x}^*$ is suboptimal, since marginal social costs (defined as the sum of MECs and MPCs) no longer reflect an agent's *willingness to pay* (Verhoef, 1994). Therefore, at equilibrium $S(\boldsymbol{x^*}) < S(\boldsymbol{x}^{\circ})$.

[3]This specification is made in accordance with empirical research in behavioral economics on peak-end bias (Carrel et al., 2013; Fredrickson and Kahneman, 1993)

Modeling peer pressure

By allowing agents to engage in peer pressure, the social costs implicit in the production of externalities can be internalized, bringing the social welfare of our model of the transportation system economy closer to the optimal value. We introduce peer pressure into our model as follows.

Let the matrix $\boldsymbol{P} \in \mathbb{R}_+^{n \times n}$ denote the peer pressure profile, consisting of elements P_{ij} indicating the pressure that i exerts on peer j. The transpose of this matrix, $\boldsymbol{P}^\top$, consists of elements P_{ji}, representing agent j's pressure on i. If $j \notin Nbr(i)$, then $P_{ij} = 0$.

The utility function with peer pressure is

$$U_i\left(x_i, \boldsymbol{x}_{-i}, \boldsymbol{P}\right) = V_i\left(x_i\right) - \nu_{it}\left(\boldsymbol{x}_{-it}\right) - \sum_{j \in Nbr(i)} P_{ji} - c \sum_{j \in Nbr(i)} P_{ij}, \tag{3.5}$$

where the third term is the cost applied if i is pressured, while the fourth term is applied to the utility function as a sum of the costs accrued for i pressuring other eligible peers in his immediate social network. The *marginal cost of peer pressure* for each agent is c utils per unit of pressure. This parameter is indicative of the ease or difficulty with which one agent may pressure another agent. While in this study c is specified to be constant and homogeneous across the population, an agent-specific marginal cost of peer pressure can be considered in future work.

Agent-specific pressure selection strategies initially specify which agents may pressure each other. These strategies may be composed. We specify two such strategies below, followed by a detailed presentation of the peer pressure profile selection algorithms in Section 3.2.

For the first strategy, we specify that an agent who uses public transit or some other socially cooperative mode can pressure any peer in her neighborhood who drives. For any agent i whose equilibrium mode choice, $\text{mode}_i(X_i^*) = \text{car}$, any peer $j \in \{k \in Nbr(i) \mid \text{mode}_k(x_k^*) = \text{sc}\}$ is permitted to pressure i to consider using an alternative to driving. Thus, in this strategy, peer pressure on i can only take effect if

$$\sum_{j \in Nbr(i)} P_{ji} \geq \Delta U_i = U_i(x_i^*, \boldsymbol{x}_{-i}) - U_i(x_{i,\text{sc}}^\circ, \boldsymbol{x}_{-i}). \tag{3.6}$$

Implicit in this criterion is a measure of accessibility to driving alternatives. Thus, people who do not retain a memory of public transit use at SUE would automatically be excluded from being pressured, as it would be too costly for any peer or group of peers to pressure them.

As a further strategy, we specify that an agent i who drives and has a utility gap greater than any agent $j \in \{k \in Nbr(i) \mid \text{mode}_k(x_k^*) = \text{car}\}$'s utility gap, $\Delta U_i > \Delta U_j$, can pressure j. This predicate measures the extent to which captive drivers would pressure other drivers with access to alternative commute modes to shift off driving in order to potentially benefit from reduced congestion.

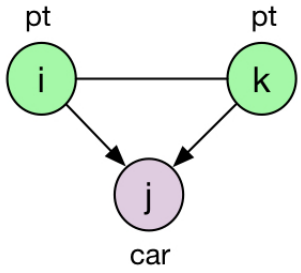


Figure 3.1: Example social network of three agents i, j and k. Agents i and k currently commute via public transit and wish to pressure j, who currently drives to work, to also take public transit.

Example

We now present a numeric example that walks the reader through the computations performed during a round of the peer pressure game set in a fictitious travel environment. For simplicity, we will assume that units of utility are represented as *utils*. Consider a 'society' of three roommates (Figure 3.1), represented by agent i, j, and k, who commute to the same job (i.e., they all have the same home and work facility locations). We assume that each person derives an identical total utility of 100 utils from their daily activity schedule. Agents i and k typically commute via a municipal metro rail line. These two agents suffer from the greenhouse gas contributions of a mutual friend, agent j, who prefers driving an SUV to work over taking the train by $\Delta U_j(\boldsymbol{x}_0) = 20$ utils. Agent j's action results in a 14 utils disutility due to CO_2 emissions, which is felt by i and k equally. As public transit riders, i and k do not produce externality through their respective actions. The net social welfare in this situation is $\mathcal{S}(\boldsymbol{x}) = 100 + 86 + 86 = 272$ utils.

Now, suppose that agents i and k both pressure agent j to leave his SUV in his garage and join them on the train such that $P_{ij} + P_{kj} \geq \Delta U_j = 20$ utils, as indicated in Equation (3.6). In this scenario agent i does not know that agent k will pressure j and vice-versa such that each applies 20 utils of pressure for a total of 40 utils of pressure applied to j. Assuming that the marginal cost of pressure, c is 0.05 utils/unit pressure, the cost to pressure agent j for agents i and k is $c\Delta U_j = 1$ utils each. Agent j then avoids the inconvenience of the 40 utils of peer pressure that he would otherwise lose by indicating that he will join agents i and k on the train in the near future. The initial cost of pressure together with j's disutility for taking the train amounts to a social welfare loss of 22 utils, which is balanced by a social welfare improvement of 28 utils due to reduced externalities, resulting in a net social welfare gain of 6 utils.

Implementation overview

In order to implement the peer pressure game in a setting that allows for its economic evaluation in large, disaggregated urban transportation environments, we extend an existing open source activity-based travel microsimulation platform, MATSim (Horni et al., 2016b). This subsection briefly outlines the simulation runtime cycle, the operation of the congestion and emission externality modules, as well as the implementation of the peer pressure extension developed for this work. Scenario evaluation with peer pressure differs significantly from the state-of-the-art use of MATSim-based tools and is described below in detail.

Simulation platform.

The core MATSim system is an open source development effort that facilitates demand modeling, dynamic traffic assignment, mobility simulation, and analysis. Several extension modules provide additional functionality applicable to modeling a variety of policy analysis scenarios. The extant MATSim API, employed extensions, and software developed as part of the current study are written in Java with associated analysis scripts developed in Python. We have developed our tool as an open-source module extending the main MATSim library (GPL Version 3)[4]

Every scenario simulation execution proceeds in an iterative manner according to the following four steps:

1. **Preparation.** During a single iteration, a simulated population of agents execute plans representing the typical daily home-based work tours. Plan elements consist of activities alternating with travel legs that describe the routes taken between activities. The activities have attributes of type *(e.g.,* home, work, leisure, etc.), location, start time, and duration, while the trip legs have attributes of mode, departure time, distance, as well as elements describing the traversal of routes through the network *(e.g.,* links used, type of links, and total distance traveled). If a population's initial plans are not available with fully detailed routes, one of the several available shortest path algorithms is used to calculate initial idealized daily trajectories.

2. **Mobility Simulation.** Following initial route assignment, the daily plans of the agent population are executed in the physical layer, which is represented by the links and nodes comprising the virtual road network topology. Public transit supply for modes such as subway or train are modeled on links and nodes that route vehicles separately from the road network (Rieser and Nagel, 2010), while buses share the links with general traffic unless a dedicated transit lane is available.

3. **Scoring.** Agents have a configurable memory that permits them to choose between previously executed plans. In order for agents to simulate decision-making that models

[4]Code to reproduce the simulation and analysis presented in this study is available at http://github.com/sfwatergit/peer_pressure_sim.

human behavior, an econometric utility function assigns a numeric score to executed plans (Charypar and Nagel, 2005a; Horni et al., 2016b). Specifically, for agent i a plan, x_i^m, once executed, is assigned a score U_i^m, according to the time spent performing activities and the time spent traveling to and from activities:

$$U_i^m = \sum_{q=0}^{A-1} U_{\text{perf},q} + \sum_{q=0}^{A-1} U_{\text{trav,mode}(q)}, \tag{3.7}$$

where A is the number of activities and trip q follows activity q, as described in (Horni et al., 2016b). In order to represent a typical 24-hour period, the first and last activity are considered in the same iteration such that there are the same number of trips and activities.

The utility earned due to time spent engaging in activities, $U_{\text{act},q}$, is primarily a function of the time spent performing the activity $\tau_{\text{act},q}$:

$$U_{\text{act},q}\left(\tau_{\text{act},q}\right) = \beta_{\text{act}} \tau_{\text{typ},q} \ln\left(\frac{\tau_{\text{act},q}}{\tau_{0,q}}\right), \tag{3.8}$$

where β_{act} denotes the marginal utility of performing an activity for its typical duration, τ_{typ}. At equilibrium, β_{act} is the same for all activities and is equivalent in magnitude to the penalty applied to being late to an activity. The parameter $\tau_{0,q}$ scales the actual time spent performing the activity $\tau_{\text{act},q}$ by the activity's priority and minimum duration, and may be ignored as long as dropping activities is not permitted.

Agents also receive a penalty for arriving late at an activity according to

$$U_{\text{late},q} = \begin{cases} \beta_{\text{late}}(t_{\text{start},q} - t_{\text{latest ar},q}) & \text{if } t_{\text{start},q} > t_{\text{latest ar}}, \\ 0 & \text{otherwise} \end{cases}$$

where $t_{\text{start},q}$ specifies the start time of activity q, $t_{\text{latest ar}}$ specifies the latest possible time that an agent can arrive at activity q.

Travel in the MATSim physical environment is associated with a utility penalty, which varies according to trip cost, the mode-specific perception of trip travel time, and, potentially, several other factors. To simplify model calibration, we include only the mode-specific cost associated with travel time in the structural equations. The drive-alone mode-related parameters are subscripted with car and the alternative public transit mode is subscripted with pt. While walk-to-transit and walking modes are included in the simulation, they are not detailed here to simplify the notation.

Travel-related utility scores are computed according to the following expressions

$$U_{\text{car},q} = \beta_{\tau,\text{car}} \tau_{q,\text{car}} \tag{3.9}$$

$$U_{\text{pt},q} = \beta_{0,\text{pt}} + \beta_{\tau,\text{pt}} \tau_{q,\text{pt}}, \tag{3.10}$$

which are linear in the alternative-specific time parameters, $\beta_{\tau,\text{mode}(q)}$. In accordance with random utility theory, the β_0 terms are alternative-specific constants (ASCs) that characterize other factors that systematically predispose individuals to choose one alternative over another (Ben-Akiva and Lerman, 1985; Train, 2003).

4. **Replanning.** Following scoring, the most recently executed plan is stored with a configurable number M of previously executed plans, X_i in system memory. At the beginning of the subsequent iteration, agents choose a new plan based on a configurable selection module and an optional route modification module. In this study, innovative modification strategies include: changing the departure time, link sequence (route), and choice of transit-related modes including legs performed by walking. While the default configuration specifies that agents select their current best score for modification, we opt to use a probabilistic sampling strategy to achieve a more realistic distribution of agent plans for selection, as given by Equation 3.2. The plan with the worst score is then dropped from the agent's memory, and the modified plans are simulated again.

   Steps 2-4 are repeated until a stochastic user equilibrium is reached. Please note that the iterative cycle of replanning should not be interpreted as representing a day-to-day dynamic model of human learning behavior. It may only be assumed that consecutive iterations bring the system closer to an equilibrium point. For further discussion on this topic, see (Horni et al., 2016b).

Computing and applying externalities.

Once a baseline calibrated scenario has been derived, the travel behavior of the study population is permitted to evolve in the presence of externalities. That is, the simulation steps described in Section 3.2 are repeated except that agents are made aware of the effects of congestion and emissions due to the decision of other agents to drive. Herein, as described in Section 3.2, we assume external costs are globally distributed, and are consequently applied as in Equation (3.3).

The following paragraphs briefly review design choices used in this study to simulate agent air emission and congestion externalities. The adopted methodology is derived from (Agarwal and Kickhöfer, 2015b) that studies the shift from private to public transit due to emissions and congestion pricing.

Emissions. Costs associated with air pollution due to emission of combustion gases during driving activities are computed following the work of (Kickhöfer and Nagel, 2013). Emissions calculations are performed on a link-by-link basis, tying attributes of a traveler's vehicle and road conditions to air pollution parameters. Since road type and quality affect pollutant levels, initial routing computations are modified to anticipate this additional cost, such that agents may choose to travel on roads that avoid creating excess emissions.

Herein, as opposed to earlier work using this module, we consider CO_2 production only. This study investigates the effects of externalities that result in a more diffuse social cost, and, therefore, are more difficult to internalize through regulation. Other automobile exhaust constituents do, indeed, result in transportation externalities, however we restrict the computations to arguably the most representative one for simplicity. Accordingly, we focus on the global warming potential (GWP) of CO_2, and do not simulate the damages due to other emissions.

Congestion. Road network congestion is computed as in (Kaddoura et al., 2014) by taking advantage of the queue model that underlies the traffic flow simulation. In free flow conditions, agents take τ_{free} to traverse a link. A maximum of c_{flow} agents may leave a link in a given time span. Any link traversal by an agent prevents following agents from accessing the next link until $\frac{1}{c_{flow}}$ has passed, resulting in delays, d_{flow}. Spill-back delays ($d_{storage}$) may also arise if the storage capacity $c_{storage}$ of a link, measured in number of vehicles, is exceeded.

Delays are measured in seconds and computed as the difference between the free speed travel time (τ_{free}) and the travel time experienced by an agent (τ_{exp}):

$$d_{tot} = \tau_{free} - \tau_{act} = d_{storage} + d_{flow} \tag{3.11}$$

During the replanning stage, agents take into account delays due to congestion accrued in the previous iteration, potentially motivating less congested routes or mode shift.

Simulating peer pressure

The eligibility of agents to participate in the peer pressure distribution stage is described in the methodology section 3.2. The algorithm is outlined in Algorithm 1 with the conditions defining an agent's eligibility to pressure and be pressured provided, for clarity, as flowcharts in Figure 3.2. In order to simulate the effect of peer pressure, we modify the mode change strategy to potentially reroute the just completed plan for public transit. Once members of a driving agent's social network sufficiently pressure him to consider an alternative mode, their most recently executed plan is then flagged. Flags expire after a number of iterations equal to the size of the agent's memory. Thus, if, through the plans sampling process, an agent does not choose the flagged plan by the expiration iteration, the plan will no longer be eligible for rerouting until the agent is pressured again. The expiration condition models the idea that the memory of social influence is ephemeral, and that not every attempt of

pressure will be successful.

Algorithm 1: Peer Pressure Algorithm

```
for i ∈ G do
  if isEligibleToBePressured(i) then
    P_total[i] ← 0
    for j ∈ Nbr(i) do
      if isEligibleToPressure(j) then
        P_total[i] ← P_total[i] + ΔU_i
      end if
    end for
    if P_total[i] ≥ ΔU_i then
      U_i ← U_i − ΔU_i
      flag(x_i^*)
    end if
    for j ∈ Nbr(i) do
      if isEligibleToPressure(j) then
        U_j ← U_j − cΔU_i
      end if
    end for
  end if
end for
```

3.3 Case Study

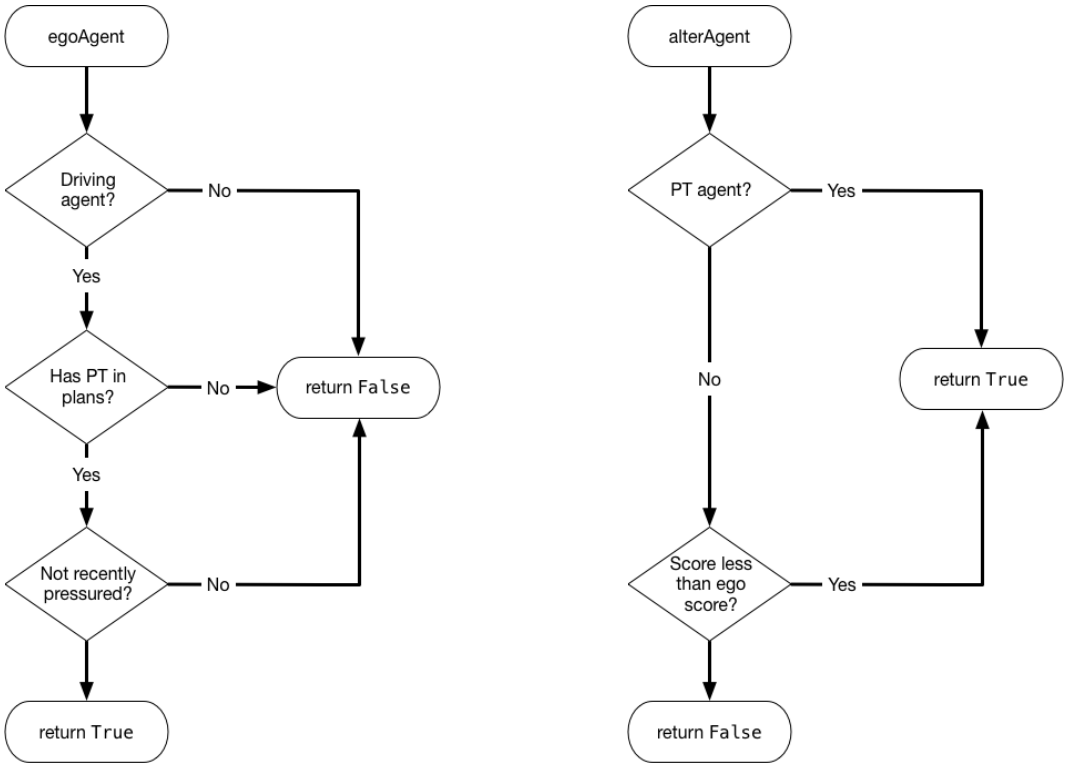
In order to verify the functionality of the peer pressure algorithm on a large scale travel demand scenario, we have applied the framework described in Section 3.1 to a simulation of San Francisco Bay Area daily commute traffic.

Simulation data sources

Network

The road network, consisting of 96,000 links, and representing freeways, state routes, all major arterials, and countryside roads, was generated from Open Street Map data.

We use a fully integrated public transit routing module (Rieser and Nagel, 2010), permitting a highly detailed simulation of Bay Area public transit throughout the course of the day. Physical track and scheduling data for the public transit system are derived from General Transit Feed Service (GTFS) data and include 9 major transit agencies operating light rail, metro and bus routes. The initial modal split has been calibrated to passenger counts obtained from the regional transportation planning authorities, the Metropolitan Transportation Commission. See Figure 3.6 for a map of the transit lines used in this study.



(a) Agent eligibility to be pressured (b) Agent eligibility to apply pressure

Figure 3.2: Pressure decision-making flowcharts: agents eligibility to participate in peer pressure distribution.

Initial Plans.

The population of the study area in 2015 was approximately 7.5 million people. Of the estimated 3.4 million commuters, 75% drive alone, while 11.5% take public transit and 3.5% walk to work. A base population of synthetic commuters, comprising 50% of the Bay Area was adopted from the Bay Area Travel Model One of the Metropolitan Transportation Commission (MTC), and adjusted with anonymized cell phone data logs. Information on the MTC model may be found in the 2012 report on data development (MTC, 2012). Cell phone data records (CDRs) are collected and managed by a major national carrier were recorded at the spatial resolution of cell phone towers. Home and work locations generated from these CDRs are upscaled to match the marginals of the population census, and then sampled to produce a desired number of agents in the synthetic population. A complete methodology of generating activity-based travel demand models from cellular data is described in (Yin M. et al., 2017).

Due to the computational requirements of composing the emission, congestion, and the detailed simulation of public transit, as well as the social network and peer pressure simulation developed as part of this project, a 1% sample of the full synthetic population was

used. The spatial distribution of the synthetic agents home locations split by the commute mode of the initial plan set is illustrated in Figure 3.5. It is instructive to compare modal split to the layout of the transit network in Figure 3.6.

Recognizing the limitations of using a small sample, in practice one has to face heavy computational loads of simulating detailed behaviors of 50,000 interconnected agents. As a result, network flow capacities for network links are scaled down to 1%. Following recommendations found in (Kickhöfer and Agarwal, 2015), storage capacities are scaled to 3% in order to achieve realistic congestion patterns. Rescaling did not substantially affect the attainment of key validation metrics. Particularly challenging for this implementation was the attainment of modal splits, since road network flow capacity does not scale linearly with transit network flows such as those used by Bay Area Rapid Transit (BART). It was necessary to tune these by hand; however a comparison to available data sources indicates that we were successful in this endeavor. To wit, see Figure 3.3 and Figure 3.4 for, respectively, simulated vs. MTC travel model modal split and example BART simulated vs. actual hourly ridership visualizations.

Social network generation

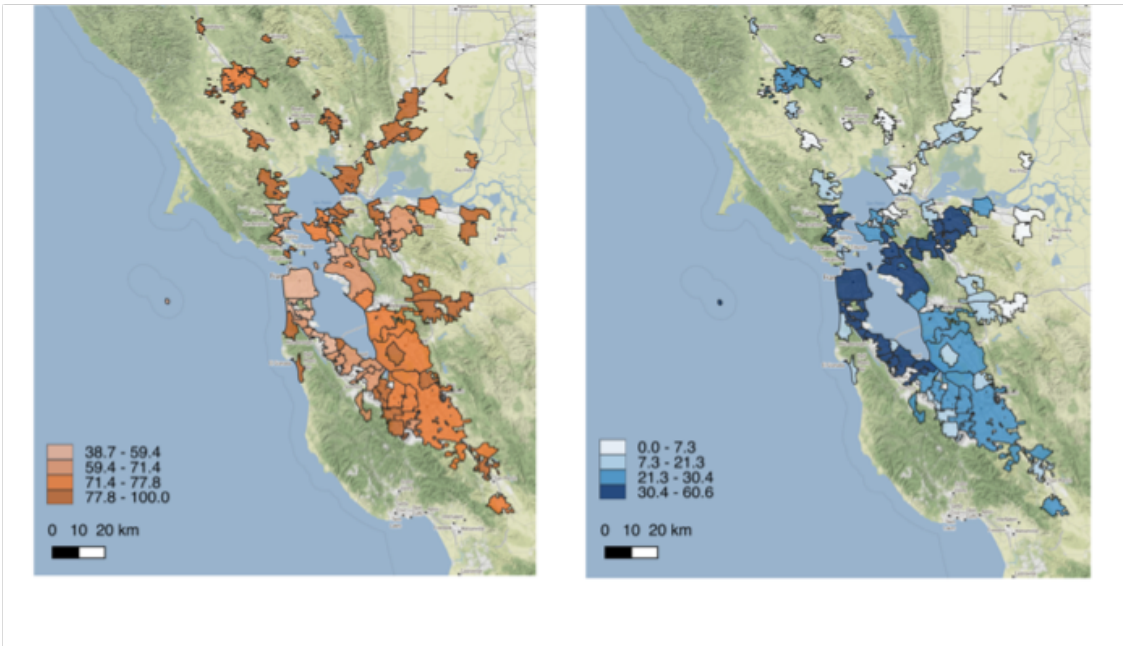
No individual level personally identifiable information was used in the study. A synthetic social network for the population of 50,000 agents was generated. An algorithm used to generate ties in the network respects core statistics of the network graph (node degree distribution, clustering coefficient), as well as the household composition, the marginals of the socio-demographic attributes (age, gender, income) of nodes, and the macro-level spatial patterns of the network community structure. It involves using probabilistic Bayesian networks (Sun and Erath, 2015) to match the conditional distributions of the socio-economic parameters describing the households composition, and the exponential random graph models (ERGM) (Schweinberger and Handcock, 2015) to fit the identified network statistics and community structure parameters. The complete methodology of simulating a required social network was adopted from the algorithms of Zhang D. et al. (2017).

Behavioral parameters

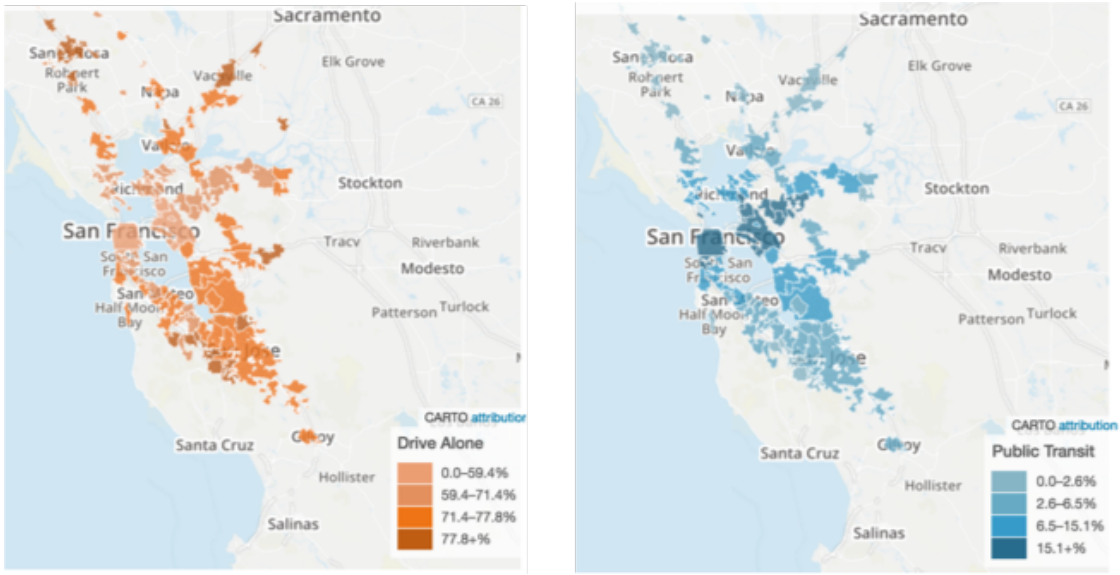
Behavioral parameters used in the simulation and provided in Table 3.2 generally match the accepted specifications described in (Horni et al., 2016b), Chapter 3, except for the values specific to the region in question. Particularly, the alternative specific constants for public transit were adapted to match the observed volumetric passenger counts. Additionally, the marginal utility of money, β_m was derived from survey data used for the San Francisco Mobility, Access, and Pricing Study (SFCTA, 2010).

Emissions module parameters

The existing emission extension developed in (Kickhöfer and Agarwal, 2015) adheres to European standards, practices, and driving conditions. In order to better align with the

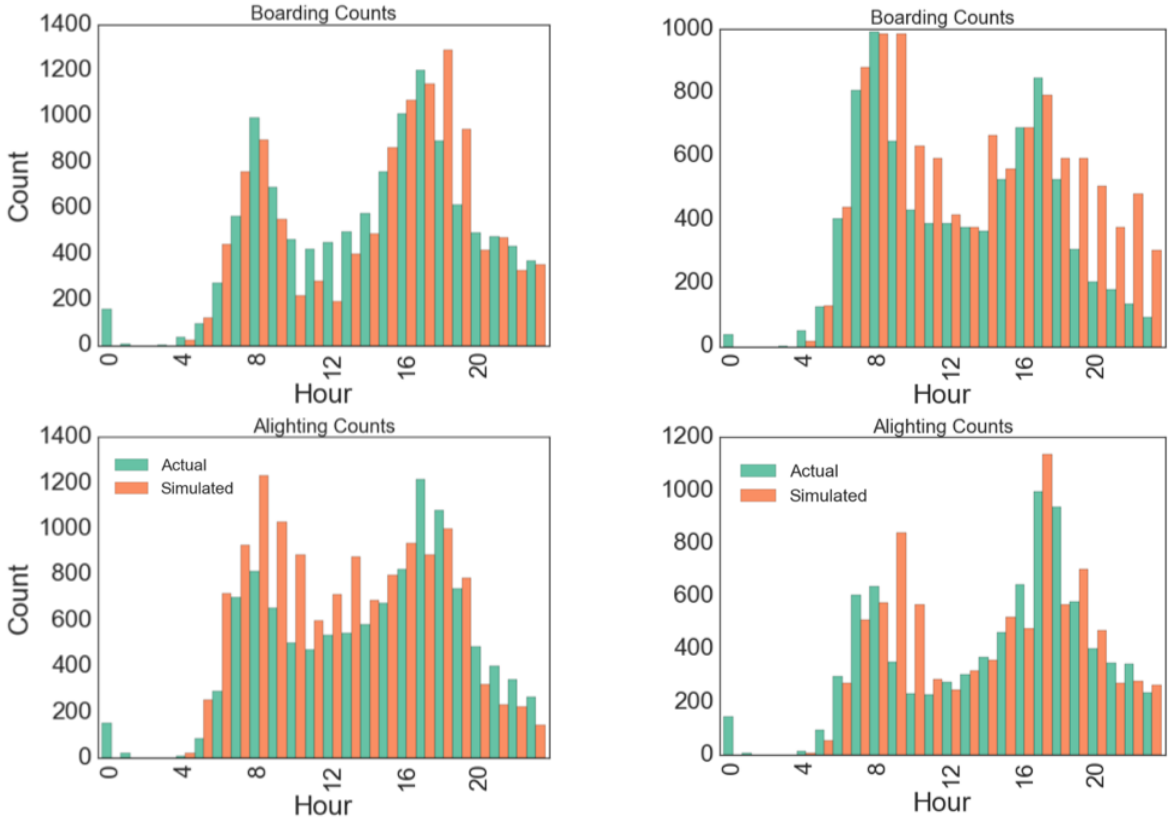


(a) Drive alone, simulated (left) and Transit+Walk, simulated (right)



(b) Drive alone, MTC (left) and Transit+Walk, MTC (right)

Figure 3.3: Comparison of calibrated model output and MTC Travel Model I modal splits between driving alone and socially-cooperative (i.e., transit and walking) modes. Socially-cooperative modes also include walking to transit. MTC figures from MTC vital signs website (Metropolitan Transportation Commision, 2018).



(a) Actual vs. simulated boarding (top) and alighting (bottom) counts at MacArthur BART Station, Oakland, California.

(b) Actual vs. simulated boarding (top) and alighting (bottom) counts at 16th and Mission BART Station, San Francisco, California.

Figure 3.4: Comparison of boarding and alighting counts for calibrated model output and measurements from a single day at two example stations on Bay Area Rapid Transit System.

Parameter	Description	Value	Unit
β_{perf}	Marginal utility of performing activity	1.205	$\text{util} \cdot \text{hr}^{-1}$
β_{late}	Utility of late arrival	-18	$\text{util} \cdot \text{hr}^{-1}$
$\beta_{\tau,\text{car}}$	Marginal utility of time (car)	-0.134	$\text{util} \cdot \text{hr}^{-1}$
$\beta_{\tau,\text{pt}}$	Marginal utility of time (public transit)	-0.16	$\text{util} \cdot \text{hr}^{-1}$
$\beta_{\tau,\text{walk}}$	Marginal utility of time (walking)	-0.29	$\text{util} \cdot \text{hr}^{-1}$
$\beta_{\text{wait,pt}}$	Marginal utility of waiting for public transit	-0.044	$\text{util} \cdot \text{hr}^{-1}$
β_{ls}	Marginal utility of line switch	-0.045	util
$\beta_{0,\text{pt}}$	Alternative specific constant (public transit)	3	util
$\beta_{0,\text{walk}}$	Alternative specific constant (walking)	-1	util
β_m	Marginal utility of money	0.083	$\text{util} \cdot \$^{-1}$

Table 3.2: Behavioral parameters of the utility functions specification.

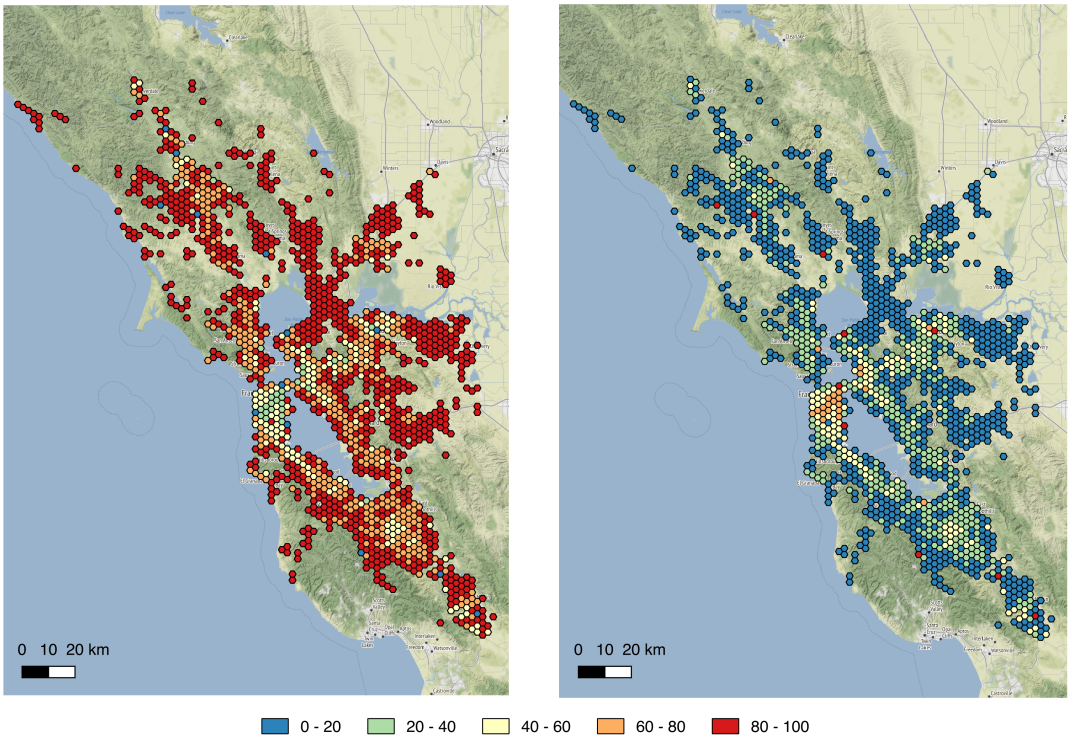


Figure 3.5: Home locations, by commute mode, of the baseline agent population: percent counts of driving agents (left) and socially cooperative modes (right).

local physical and regulatory transportation environment, the module's source code was altered to be compliant with USEPA and (when available) California Air Resource Board (CARB) emission models. Emission factors used in simulation calculations are derived from the CARB's EMFAC2014-LDA passenger vehicle model aggregates for the San Francisco Bay Area Air Quality Basin (California Air Resources Board, 2014). Emission monetary costs are computed using the United States Environmental Protection Agency's Social Cost of CO_2 statistics (IAWG, 2015). These are provided at variable discount rates (1, 3, and 5%). We use the moderate \$36/tonne CO_2 derived using the 3% discount rate as a reasonably conservative measure of the social cost of carbon, noting that a value of as high as \$120/tonne may be used in particularly risk averse scenarios. As in (Kickhöfer and Nagel, 2013), we assume that public transit use has negligible emissions in comparison to automobile travel.

Simulation experiments

The workflow for the simulation experiments performed in this study are presented in Figure 3.7. Individual steps are discussed below.

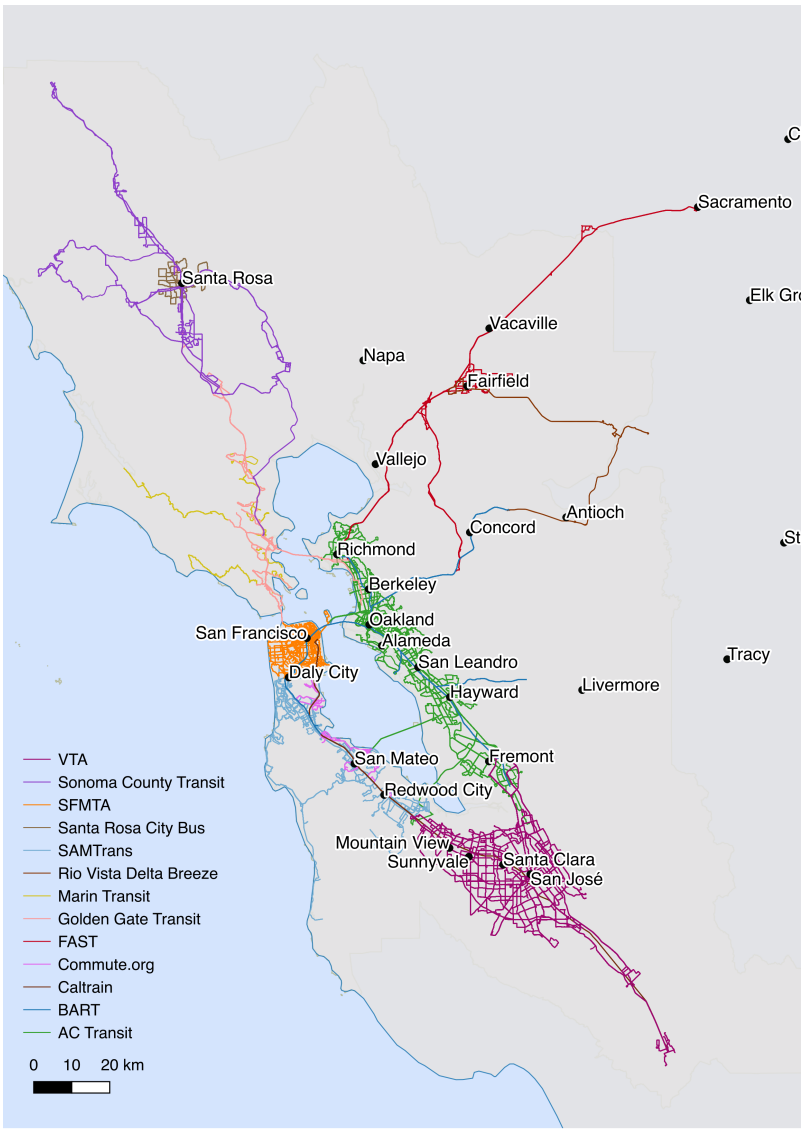


Figure 3.6: Transit lines used in simulation

Baseline scenario

A calibrated base case is first established for policy comparison purposes. To derive a baseline scenario, agents are permitted to adaptively optimize their plans using the MATSim co-evolutionary algorithm described in Section 3.2. Prior to each iteration 20% of agents selected at random will have either their selected plan rerouted, trip departure/arrival times modified, or the travel mode for their daily commute will be shifted from private to public transportation. The simulation continues until the population ensemble average scores reach a stable point, which we found was approximately 200 iterations. Simulated volumetric flows on road network links are compared to data from the California DOT freeway Performance

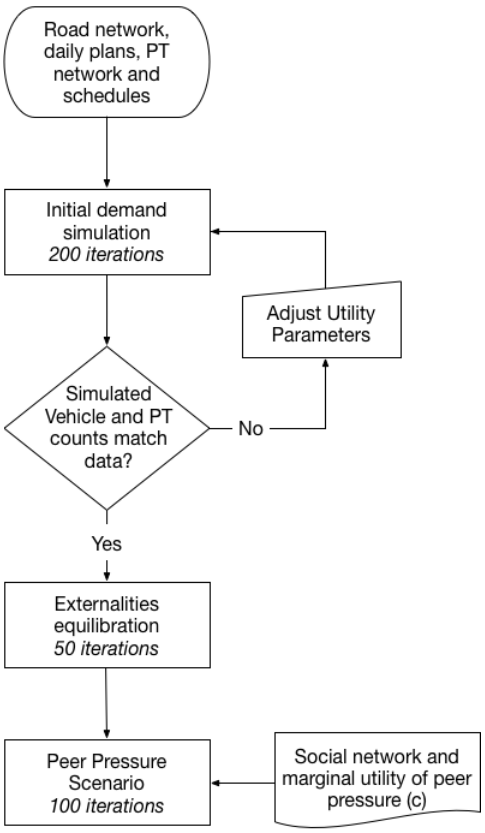


Figure 3.7: Peer pressure simulation stages

Management System (PeMS) as described in (Yin M. et al., 2017). Simulated transit stop entry and exit data from simulated BART agents is also compared to ground truth hourly counts aggregated during October 2013.

Externalities equilibration

After a stable, calibrated baseline has been reached, the set of plans in agent memory are carried forward to the next part of the simulation. Agents are now allowed to modify and reroute their plans in the presence of system-wide externalities, as described in Section 3.2. We find that the simulation reaches a fixed point after an additional 100 iterations.

We calibrate the congestion and emission externalities using linear scaling factors of 10^{-5} and 10^{-4} respectively so that they contribute in the order of 10% of the typical agent score. While quantifying the influence of globally-distributed negative externalities as well as environmental awareness on individual decision making is largely an open research problem, in the present we follow (Agarwal and Kickhöfer, 2015b) to set this order of magnitude.

The output plan data from the final iteration of the base model with externalities are used in welfare comparisons.

Peer pressure scenario

In the peer pressure experiments, utility is assigned according to Equation (3.5). For each run of the policy case, the value of c (marginal cost of pressure) is set beforehand and innovative strategies are maintained as before.

Peer pressure is specified to begin after 5 iterations to verify that the start point of the run is equivalent to the base case end point. Innovative strategies are retained in order to permit agents to modify and optimize their plans in response to peer pressure. We run the simulation with pressure and innovative strategies until iteration 80, at which point plan innovation is turned off. This was done to view how the system relaxes when plans are fixed.

The algorithm used to implement peer pressure in the microsimulation context is provided in Figure 3.2. Recall from Equation (3.5) that the parameter, c, is the marginal cost of pressure. For all agents in a simulation run, we assume a homogeneous value of c. In the present study, all peers, $j \in Nbr(i)$ will pressure i as long as they are eligible to do so. For example, let π_i be the number of peers eligible to pressure an agent i under the first pressure strategy. Then, U_i will be penalized by ΔU_i utils and each $j \in \pi_i$ will be penalized $c\Delta U_i$ utils. Then, the change in social welfare for the system due to peer pressure under action profile $\boldsymbol{x}$ is given by $-\sum_{i \in N}(1 + c \cdot \pi_i)\Delta U_i$ utils.

Since empirical data on the social cost of peer pressure in this context is unavailable, we performed a sensitivity analysis by running the simulation for a range of magnitudes of c, ranging from 0.001 to 10, while holding all other parameters constant.

3.4 Results

Peer pressure: effect on mode shift and system dynamics

In this section, we examine the effect of peer pressure on mode shift and score evolution as well as how system dynamics and target metrics vary with the marginal cost of peer pressure, c.

In Figure 3.8, the number of agents that switch mode between iterations is plotted over time at different values of c. Clearly, for all of the values of c explored, *ceteris paribus*, peer pressure is effective in achieving attenuation in the net number of drivers. Once innovative strategies are turned off, at $t = 80$, the number of shifted agents eventually drops to 0, as expected, since, at this point, agents only select between existing plans in their plansets. Although the simulation with peer pressure was not run to convergence, we observe in Figure 3.8 that the maximum number of agents shifted per iteration does appear to be reaching a fixed point.

When exploring the dynamics of the system as a function of c and $50 < t \leq 80$; however, we observe that a phase transition in the stability of system evolution may occur between $c = 0.01$ and $c = 1$. Specifically, in Figure 3.8, we note that for $c = 0.001$ and for $c = 0.01$ at $t > 50$, the number of agents shifted begins to oscillate around an upward trending baseline.

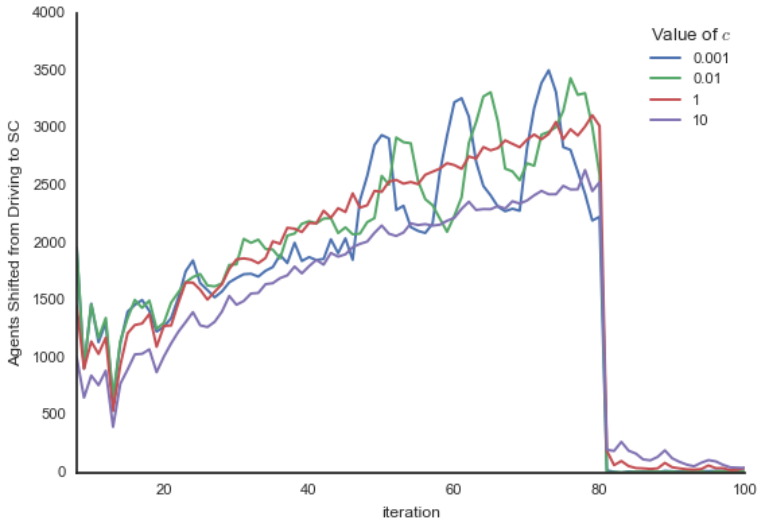
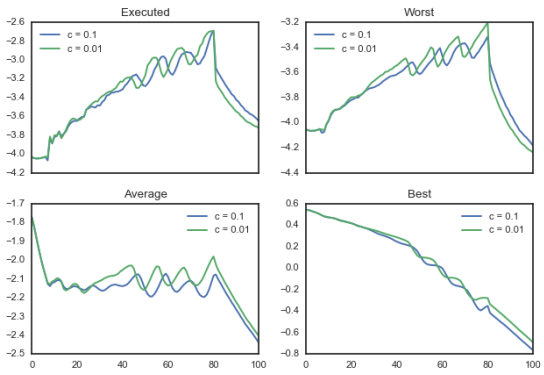


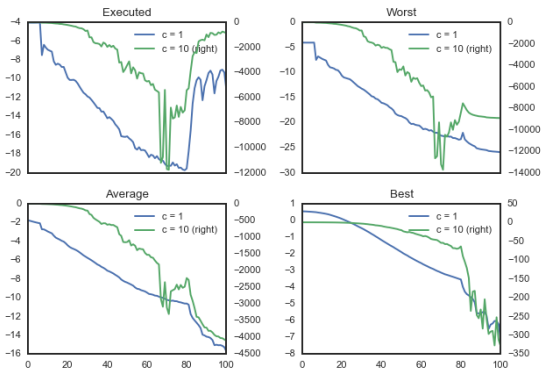
Figure 3.8: Net number of agents shifting to socially cooperative modes for different values of the marginal utility of peer pressure.

These oscillations are on the order of 1×10^3 agents and have a period of 10 iterations. This second order phenomenon is almost entirely absent in the simulation for values of $c \geq 1$.

For $c = 0.01$, Figure 3.10 demonstrates that the number of agents pressured and mode share are roughly covariant. This observation suggests that oscillations in the system evolution occur due to synchronization of pressure-induced mode shift forcing and ephemeral memory effects in a segment of the pressured population. That is, while peer pressure-induced mode shift generally improves utility for many agents (as demonstrated by the



(a) Average scores of agents when $c < 1$.



(b) Average scores of agents when $c \geq 1$.

Figure 3.9: Ensemble average score sensitivity of agents to value of c. Scores are in utils.

overall increased uptake in travel mode), some pressured agents would have been better off driving. Isolating the optimal population that would benefit from pressure will be treated in future work.

The ensemble average score evolution plots (Figure 3.9) present the values of the executed, worst, average, and best plans in an agent's plan set, X_{it} averaged over all agents as a function of the iteration, t. We have separated the plots of ensemble average scores into two subfigures in order to better illustrate how the dynamics of the system evolution vary with c. For values of $c < 1$, mode shift apparently covaries with score evolution, as suggested by the oscillations in scores depicted in Figure 3.9a. For $c \geq 1$; however, Figure 3.9b, indicates a catastrophic collapse in the executed and worst agent scores, with particularly unstable scores observed for $c = 10$.

The precipitous decrease in executed plan scores at high values of c clearly leads to unsustainable dynamics wherein the disutility imposed by peer pressure exceeds the utility of plan execution. The instability is most likely due to the inflexible requirement that agents who have a driving neighbor to pressure are required to pressure that neighbor no matter what the cost to themselves. Clearly this is an unrealistic scenario. We therefore present the rest of our results and analysis for simulation outputs where $c = 0.01$, which we take as a moderate value consistent with the more realistic utility scores observed for lower values of c.

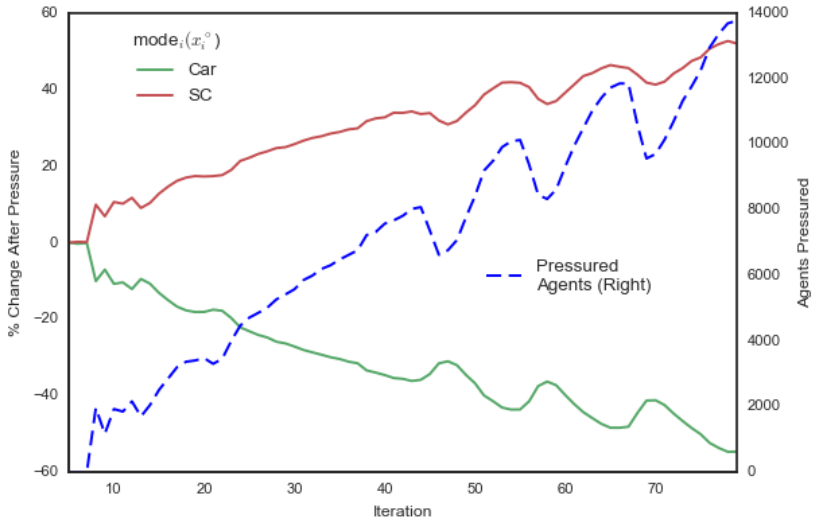


Figure 3.10: Evolution of percent changes in transportation mode share with number of agents pressured following initiation of pressure at iteration $t = 5$ for $c = 0.01$.

	Congestion Delays		CO_2 Emissions	
	(hrs)	($)	(tonnes)	($)
Before Pressure	53,275	770,360	3,085	111,072
After Pressure	15,946	230,584	2,205	79,373
Net Change	-37,329	+539,776	-881	+31,699

Table 3.3: Externality internalization due to peer pressure

NOTE: Values taken at iterations 0 and 80. Value of travel time savings of car mode, $VTTS_{\text{car}}$, taken as 14.52 $\$ \cdot \text{hr}^{-1}$. Social cost of carbon assumed to be $36.00/tonne under a 3% discount rate.

Quantifying changes in externalities

Table 3.3 demonstrates that pressure leads to a reduction in travel delays of 37,329 hours. When multiplied by the value of travel time savings (VTTS) of driving alone[5], the reduction in delays aggregated over all vehicles for the full 24-hour simulation day is equivalent to a net social gain of $539,776. The $31,699 gain from CO_2 abatement is a slightly less significant improvement. As suggested by the results presented in Section 3.4, congestion improvements are due to agents switching from driving alone to transit-oriented modes in response to the active influence of peers in their social group.

A spatial analysis of the redistribution of monetized delays is indicative of the winners and losers of peer pressure as well as where changes in travel time occur. Figure 3.11 illustrates the differences in delay between the business as usual ($t = 0$) and peer pressure ($t = 80$) case experienced by agents visualized as an average over all agents with home locations in a traffic analysis zone (TAZ). Evidently, the greatest improvements in congestion due to peer pressure are experienced by people living in less populated areas. It is instructive to compare total delays experienced by agents on their individual trip routes to delays of all agents on a link-by-link basis[6] (Figure 3.12). The greatest improvements happen on freeways and arterial routes, which is somewhat expected, since the greatest proportion of agents travel over these links. In some rural areas, congestion does appear to increase. Long distance commuters from the rural areas are taking more direct routes to the urban core due to the congestion relief therein, but end up queueing on the approaches. The agents traveling along these routes are unable to pressure their peers to stop driving in order to reduce congestion due to unavailability of alternative modes. This observation suggests that these users may benefit from increased access to public transit, or park and ride facilities.

We observe differences in the distributions of pressured (Figure 3.13) and pressuring agents (Figure 3.14). Initially, pressured agents are, as expected, clustered around public transit. However, over the course of the simulation, as feedback between pressured and

[5]VTTS is computed as in (Agarwal and Kickhöfer, 2015a) by taking the ratio of the marginal utility of travel time ($mUTTS$) and the marginal utility of money, β_m. The marginal utility of travel time is given by $mUTTS = \beta_{\tau,\text{mode}(q)} - \beta_{\text{act}}$.

[6]These delays are measured according to the difference between free-flow travel time and estimated travel time averaged over the 24 hour simulation period

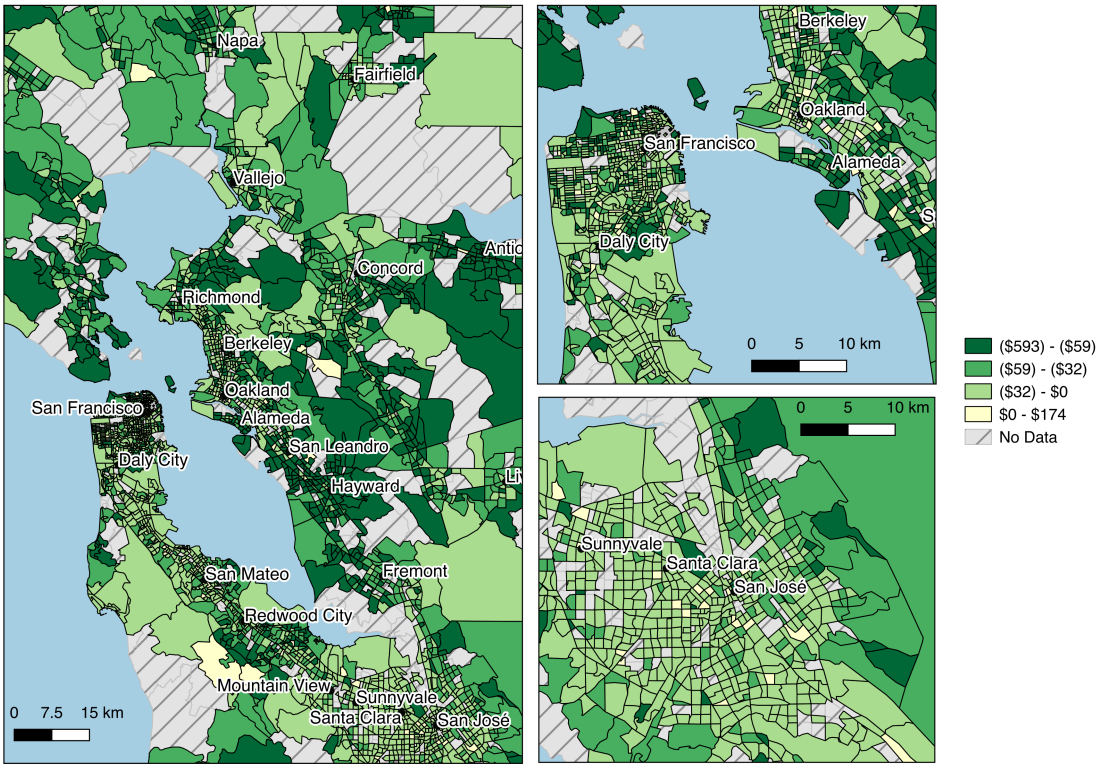


Figure 3.11: Mean monetary delay costs (gains) due to difference between business as usual (iteration 0) and peer pressure (iteration 80) experienced by agents with homes in TAZs as symbolized.

pressuring agents grows, we see that presurees become more evenly dispersed throughout the Bay Area. We also observe multiple clusters of pressuring agents distant from sources of public transportation with concomitant absences of pressured agents in the same locations. Qualitatively, it appears that locations where agents are closer to transit seem to correspond to some of the most improved travel times (Figure 3.12). In Figure 3.13, we can see that, over the course of the simulation, the distribution of pressured agents becomes relatively more concentrated in these areas. More quantitative conclusions about the nature of the spatial variability of pressure and its relationship to transit accessibility in alternative transportation geographies is left as a topic for future research.

3.5 Conclusions

This chapter presents an agent-based simulation framework developed to model the effects of peer pressure on inducing socially-cooperative travel mode choice. By applying the aggregate effects of externalities on agents explicitly and providing a mechanism for agents to,

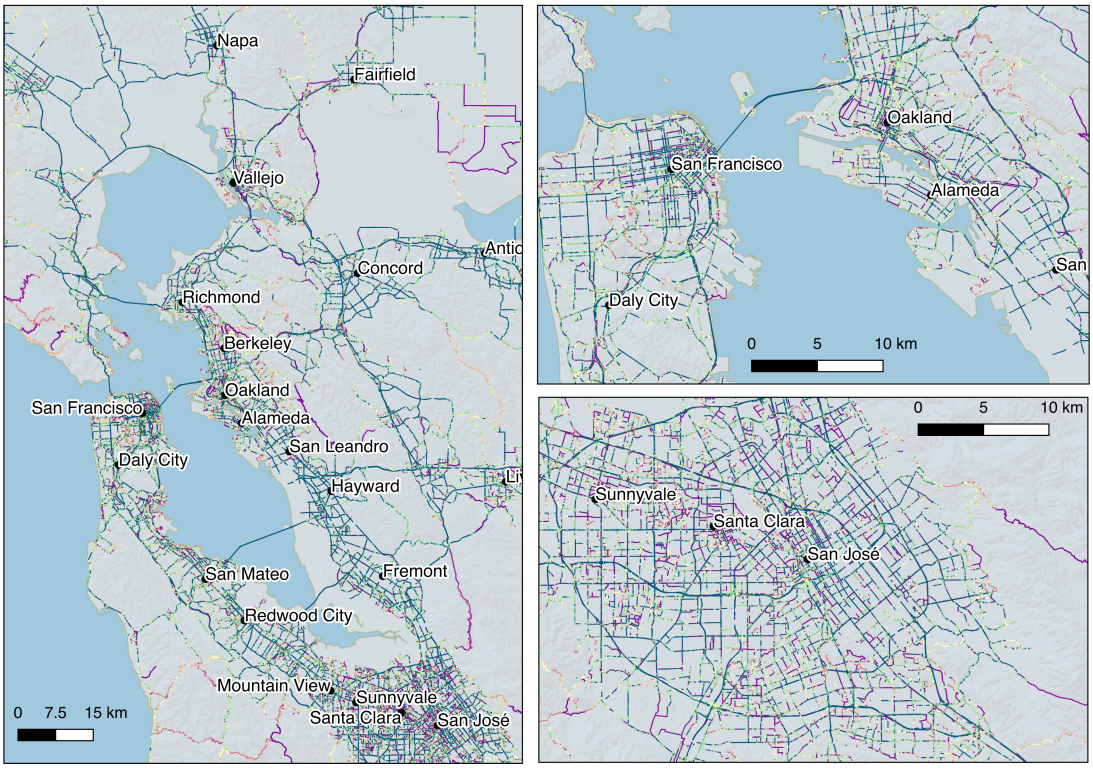


Figure 3.12: Road links with improvements (shown in blue) and delays (shown in purple) based on differences between experienced and free speed travel time between iterations $t = 0$ (business as usual) and $t = 80$ (peer pressure).

effectively, negotiate time valuation, an efficient redistribution of social welfare is achieved.

Due to complex dependence of the peer pressure decisions on social network structure as well as physical infrastructure specifics such as the accessibility of socially cooperative mode choice options, the system as modeled does not admit a closed form or even approximate solution without significant simplifying assumptions. Consequently, it is not possible to develop a closed-form optimal pressure strategy for agents to pursue. The modeling framework that we wish to emphasize in this paper is more akin to that of the emerging game-based modelling (GBM) concept, as described and empirically evaluated in several recent studies (Klein and Ben-Elia, 2016; Klein et al., 2018). Like those authors, we see the use of GBMs as a methodology that could help stakeholders and researchers better understand the conditions under which emergence of cooperation in a complex transportation system might be expected. The rules for our game-based model of peer pressure are not without antecedents that ensure theoretical plausibility: an analytic model of incentivized peer influence (Mani et al., 2013) and an empirically-validated agent-based simulation model of stochastic user equilibrium in transportation networks (Horni et al., 2016b).

Given this modeling framework, many heuristic strategies and solution search algorithms

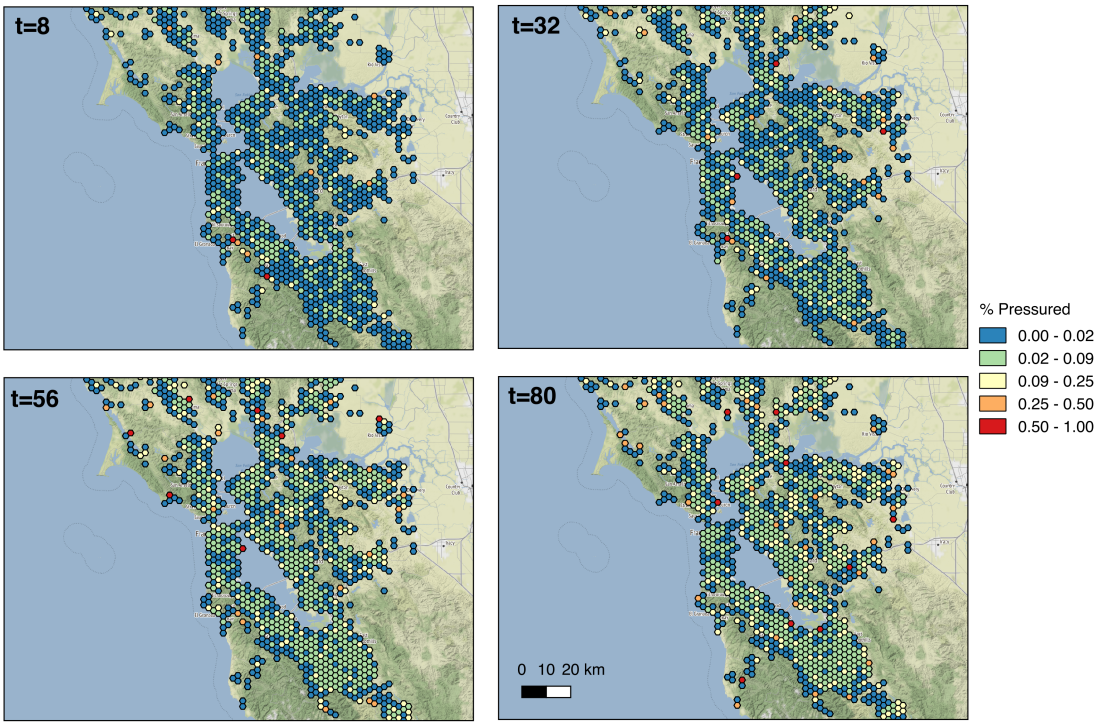


Figure 3.13: Evolution of *pressured* agent spatial distribution through several iterations.

may be explored in order to develop system-optimal pressure behavior. For example, the decision for an agent to apply pressure may be contingent on how many other neighbors the agent can pressure (as well as their pressure costs), whether other neighbors will participate, and, considering that pressure can be applied to the same person in repeated iterations, how successful past attempts were.

Despite demonstrating that peer pressure leads to widespread improvements in congestion and reductions in emissions, the spatial analysis of post-pressure changes shows that some areas are worse off. This finding highlights the need to ensure that policy proposals be sensitive to social justice issues, particularly if travel time and emission improvements are unequally biased towards one demographic or another. Running the simulation with demographic data and heterogeneous preferences accordingly may improve the representativeness of results as well as help communities understand the potential impacts of cyber-social influence.

When reduced externality costs are insufficient to encourage modality shifts away from driving, policy instruments can be used to incentivize agents to pressure their peers. However, the role of governments in achieving cooperative outcomes in social dilemmas need not be a coercive (Ostrom, 1990; Ostrom et al., 1992). In light of the analysis on incentivizing peer pressure described in (Mani et al., 2013), extensions to our framework can be used to design

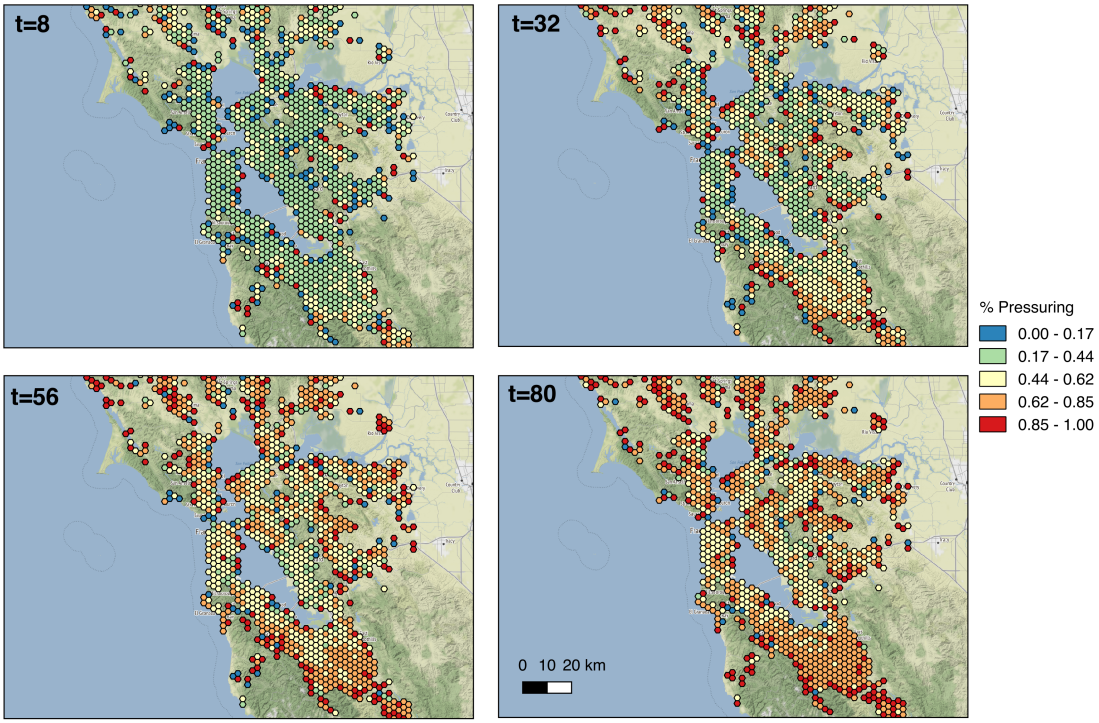


Figure 3.14: Evolution of *pressuring* agent spatial distribution through several iterations.

public transportation policy that subsidizes the social costs of peer pressure with the goal of improving net social welfare. For example, a municipality can encourage positive peer pressure by providing a bonus to drivers who encourage their friends to carpool to work with them.

While designing rewards to subsidize peer pressure is a topic left for future research, the work presented herein is not without its practical merits. Simulating the positive effects of peer pressure on social welfare may motivate citizens to make decisions that equitably address commons problems by demonstrating how social networks spread and stabilize behavior change arising from local interactions. That is, by propagating simulated information from the virtual world to the real world, people can learn under what circumstances the personal cost that they incur in pressuring their peers would result in net personal and social benefits. Alternatively, our framework can be used to inform individuals if peer pressure is not worth the loss in social capital due to excessive free-riding; encouraging policy makers to fund the gap. Providing unbiased and clear information will ensure that policy nudges promote democracy rather than co-opt autonomy.

To translate simulated policy alternatives into information that can be used to guide policy decisions, it is vital that the generalized cost functions motivating agent decision-making to match real-world behavior. The computational framework presented in the next section

represents an opportunity to fuse activity-sequence recognition with the rationalization of decision-making observations using identical data sets. By doing so, we the choice outcomes of simulated agents will more closely reflect the preferences of the study population. In this way, embedded machine learning-based rulemaking can be used to make highly differentiated mobility recommendations and/or choice architectures to individual citizens.

4

Estimating Activity-Travel Plan Utility Functions via Inverse Reinforcement Learning

By three methods we may learn wisdom: First, by reflection, which is noblest; Second, by imitation, which is easiest; and third by experience, which is the bitterest.
– Confucius

While there is an enormous variety of available activity opportunities and travel options within any major city, our daily schedules tend to be fairly routine. Modern activity-based transport demand models thus often make a simplifying assumption that individuals plan their days from a fixed set of activity-travel patterns. These patterns assume a myopic decision-making process: the impact of future opportunities is ignored when making near-term choices. By introducing utility functions with parameters that can be set at the level of individuals, software agents traveling on virtual road networks could represent the inherently heterogeneous dynamics of urban commuting.

In this chapter, we evaluate the use of IRL methods to learn stochastic distributions over paths through the state space that match the expected features observed in expert demonstrations. The resulting reward function can be interpreted as a structural model of agent decision-making behavior Ziebart and Bagnell (2010). When structural parameters are interpretable, it is often possible to reliably predict the real-world implications of *ex post* or *ex ante socioeconomic* policy interventions.

In order to improve understanding and prediction of dynamic individual mobility decisions, this chapter presents a novel inverse reinforcement learning-based approach to infer structural models of activity and travel planning behavior. By formulating daily activity participation dynamics as a MDP, we are able to leverage flexible and efficient IRL methodologies in the recovery of interpretable parameters governing activity participation preferences.

In addition, to better address the challenge of implementing IRL methods at scale, we investigate a reward-sharing and policy-transfer approach with the intent of accelerating the

training of many independent agents sharing an identical task environment.

We evaluate the effectiveness of our proposed methodology by modeling daily activity and travel planning decision-making occurring within the context of a real-world urban travel environment. Individual agent reward functions are estimated using anonymized spatiotemporal microdata collected by a cellular network provider serving millions of customers in the San Francisco Bay Area.

4.1 Background

Reinforcement learning

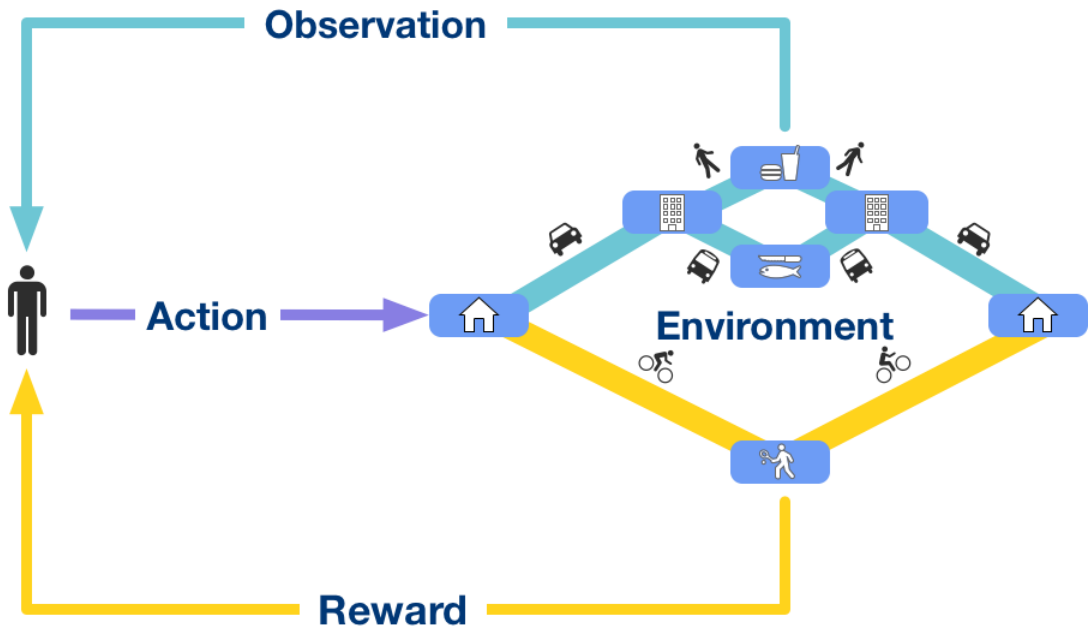


Figure 4.1: A schematic of the general reinforcement learning problem.

In our problem representation, we assume that we can observe individual agents n from a homogeneous population, $\mathcal{N}$, taking actions that stochastically influence a commuting environment. Let $\mathcal{T} = \{0, 1, \ldots, N-1\}$ be the *finite* set of possible times at which an agent can make a decision. At each time $t \in \mathcal{T}$, an agent observes the state of the environment $s_t \in \mathcal{S}$ and makes a decision $a_t \in \mathcal{A}$ based on his observation. As a consequence of the decision, the environment provides the agent with an instantaneous reward $U(s_t, a_t)$ according to a utility function $U : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the agent observes a new state, s_{t+1}. We note that the reward is a signal provided by the environment, but does not necessarily encode the agent's *internal* evaluation of the reward.

In general, we assume that for this MDP, the dynamics of the environment are defined by a transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ such that $T(s, a, s') = P(s' \mid s, a)$ is the probability of transitioning to state s' when in state s and taking action a. In the current context, we consider all actions to be deterministic such that $P(s' \mid s, a) \in \{0, 1\}$.

We assume that agents make choices according to a decision rule known as a stochastic Markov policy $\pi : \mathcal{S} \times A \to [0, 1]$ that at each state $s \in \mathcal{S}$ satisfies $\sum_{a \in \mathcal{A}} \pi(a \mid s) = 1$. In a finite horizon problem, the goal of an agent starting from $s_0 = s$ is to find a policy that maximizes the agent's expected discounted utility over N time steps by optimizing the value function

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{N-1} \gamma^t U(s_t, a_t) \mid s_0 = s\right]. \tag{4.1}$$

Imitation learning

Imitation learning (IL), which is a subset of learning from demonstrations (LfD) problems, refers to the problem of training a policy to match expert behavior associated with expert trajectories[1]. In IL (and, generally, LfD) problems, we take as training data a set of trajectories, $\mathcal{D} = \{\tau^{(q)}\}_{q=1:n}$, where each trajectory,

$$\tau^{(q)} = \left(\left(s_0^{(q)}, a_0^{(q)}\right), \left(s_1^{(q)}, a_1^{(q)}\right), \ldots, \left(s_{N-1}^{(q)}, a_{N-1}^{(q)}\right)\right), \tag{4.2}$$

represents a sequence of states and actions produced by an expert agent. The presumed decision rule used by the expert to generate these policies is π_E.

There are several variants of IL paradigms, which each reflect different approaches to the idea of using auxiliary data to inform policy optimization routines for RL agents:

Behavioral Cloning (BC) In BC, the policy is learned directly from data as a supervised learning problem. In contrast to other IL problems, there is no RL component. Behavioral cloning tends to fail to learn policies that match expert behavior due to covariate shift on test data, which is the problem of small errors metastasizing into catastrophic failures. This outcome makes intuitive sense, as the learned policy is not robust to underrepresented states in the input dataset, which, when encountered, lead to further errors, often making recovery impossible (Bagnell and Ross, 2010). However, BC is still a viable choice when large amounts of training data are available and the policy is parameterized using neural networks or other flexible nonlinear approximators. See, for example, the groundbreaking Autonomous Land Vehicle in Neural Network (ALVINN) of Pomerleau (1991)–a study that continues to inform modern research in autonomous vehicle navigation.

[1] LfD problems simply incorporate behavioral observations into learning algorithms. Thus, LfD represents a much broader class of problems, which may include forward RL problems that are augmented with expert trajectories. These trajectories may be used to, say, warm-start an RL algorithm to accelerate convergence to an optimal policy as was demonstrated in deep Q-learning from demonstrations(Hester et al., 2017).

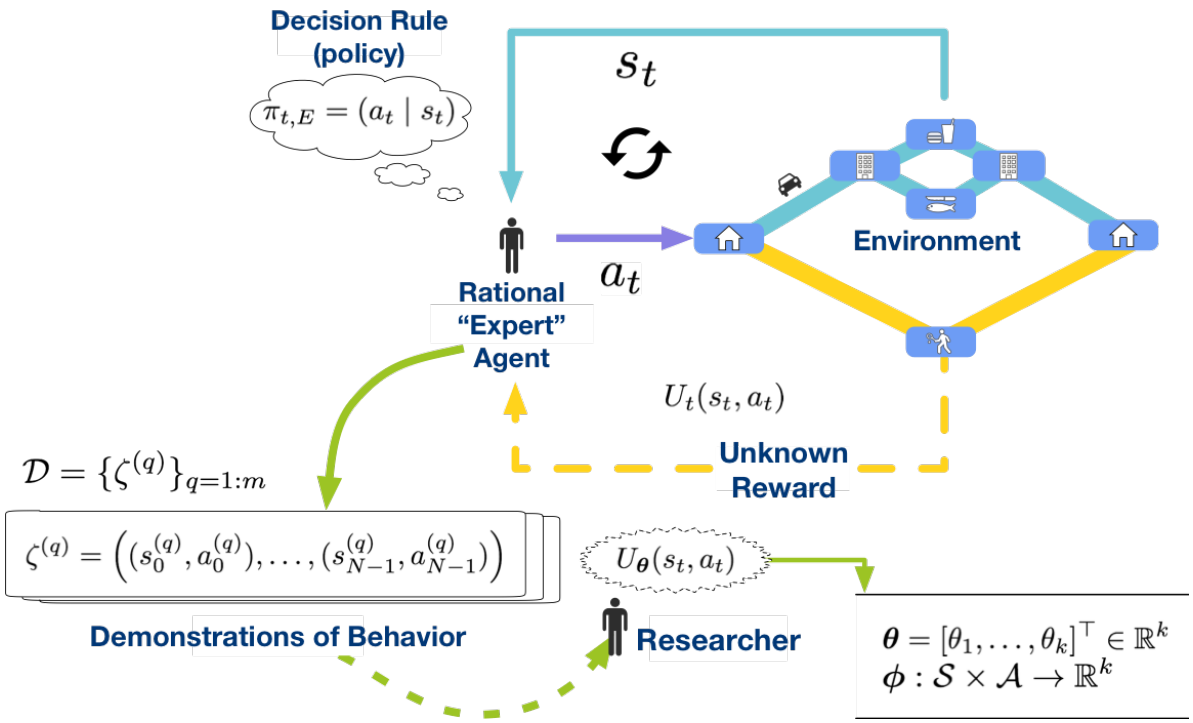


Figure 4.2: A schematic of the general *inverse* reinforcement learning problem.

Inverse Reinforcement Learning (IRL) IRL problems (sometimes referred to as inverse optimal control (IOC) problems in the robotics community) are formulated like reinforcement learning problems except that they are absent reward function. The objective of the IRL problem is primarily to find the reward function that was used to generate the input data.

Apprenticeship Learning (AL) Like IRL, AL is concerned with recovering an optimal reward. However, AL uses the reward function to estimate the policy. In AL, recovery of the reward is often a secondary aspect of training. Sometimes AL and imitation learning are used interchangeably (as in the GAIL algorithm presented in the next chapter). Policies learned using AL/IRL are usually thought to be more robust to covariate shift than those learned using BC, an assumption we will verify in Chapter 5. It is intuitive that learning policies that map from the entire state space to the entire action space will be more robust than those learned via BC, which only learns optimal responses to previously encountered states.

Inverse reinforcement learning: preliminaries

In contrast to the historical review provided in Chapter 2, we now provide a more technical overview of IRL, introducing notation and terminology used in the present study. The reader should refer to Figure 4.2, which illustrates a general IRL framework.

IRL assumes a parametric approximation of the reward function $U_{\boldsymbol{\theta}} : \mathbb{R}^k \rightarrow \mathbb{R}$ with parameter values $\boldsymbol{\theta} \in \mathbb{R}^k$. We associate the set of states and action space with domain-specific feature vectors $\boldsymbol{\phi} = \left\{\boldsymbol{\phi}_{s,a} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}\right\}$, where $k = \left|\boldsymbol{\phi}_{s,a}\right|$ is the number of features in the new embedding. For convenience, we define a feature matrix, Φ, over our state-action space, which has dimensions $|\mathcal{S}|\,|\mathcal{A}| \times |k|$.

Learning agents in IRL solve an optimization problem to find parameters for $U_{\boldsymbol{\theta}}$ that induce behavior in the environment that matches expected empirical feature counts, $\mu_{\mathcal{D}}\left(\tau\right)$, computed as an average over m trajectories,

$$\mu_{\mathcal{D}}\left(\tau^{(q)}\right) = \frac{1}{m} \sum_{\tau^{(q)} \in \mathcal{D}} \sum_{(s_t, a_t) \in \tau^{(q)}} \boldsymbol{\Phi}\left(s_t^{(q)}, a_t^{(q)}\right).$$

The presumptive goal of the expert agent is to maximize the net utility earned over time; however it may be the case that the data collected represents sub-optimal behavior. While (Abbeel and Ng, 2004) show that matching feature expectations is both necessary and sufficient in order to derive policies that emulate expert behavior in the MDP, the IRL problem of recovering reward functions under this constraint is underdetermined. That is, optimal policies where the utility function is all zeros may be recovered. Maximum entropy IRL (reviewed next) was developed by (Ziebart and Maas, 2008) to address this (and other shortcomings) of early solutions to IRL problems.

Maximum entropy inverse reinforcement learning

Maximum entropy IRL addresses the many of the shortcomings of early solutions to IRL problems (Ziebart and Maas, 2008). In this IRL formulation trajectories in the demonstration set, $\mathcal{D}$, are assumed to be sampled from a probability distribution $P_{\boldsymbol{\theta}}^{ME}(\boldsymbol{\tau})$ arising from a potentially large family of such distributions. The principle of maximum entropy (Jaynes, 1955) implies that when suboptimal trajectories are observed the distribution maximizing the entropy of suboptimal expert demonstrations should exhibit no preference over paths beyond matching feature expectations. By applying this principle, an exponential distribution over plan preferences

$$P_{\boldsymbol{\theta}}^{ME}\left(\boldsymbol{\tau}\right) = \frac{1}{Z\left(\boldsymbol{\theta}\right)} \exp\left(U_{\boldsymbol{\theta}}\left(\boldsymbol{\tau}\right)\right)$$

may be defined where Z is the partition function that normalizes this distribution. Thus, paths with higher returns are exponentially preferred to those yielding lower net rewards. Following (Ziebart and Maas, 2008), the maximum entropy objective function maximizes the likelihood of demonstrations according to

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{\tau^{(q)} \in \mathcal{D}} \log P_{\boldsymbol{\theta}}^{ME}(\tau^{(q)}). \tag{4.3}$$

The gradient of the likelihood function for a formulation specifying the reward function as linear in the features, (i.e., $U_{\boldsymbol{\theta}}(s,a) = \theta^\top \boldsymbol{\Phi}(s,a)$), is shown in (Ziebart and Bagnell, 2010)

to be

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \mu_{\mathcal{D}} - \sum_{s \in \mathcal{S}} \sum_{s \in \mathcal{A}} E[\mu(s)] \, \pi_{\boldsymbol{\theta}}(a \mid s) \, \boldsymbol{\Phi}(s, a) \tag{4.4}$$

where $\mu_{\mathcal{D}}$ denotes the empirical state visitation counts and $E[\mu(s)]$ denotes the expected state visitation counts, which represent the estimated frequencies of an agent occupying a state. The first term is computed from the data and the second term is computed over the entire state and action space according to the current setting of the parameters of the reward function.

The state visitation counts may be efficiently computed using dynamic programming. An abbreviated, generic overview of the learning procedure is presented as Algorithm 2, though the reader should refer to (Ziebart and Maas, 2008) or (Wulfmeier et al., 2015) for a more detailed discussion. The parameters defining the reward function U_θ are first randomly initialized. The main loop of the algorithm updates the reward function with the parameter estimate. A corresponding stochastic policy π_θ is then estimated using a form of value iteration, which computes softmax estimates of the cost of reaching a goal from any state in $s \in \mathcal{S}$. This policy is then used to propagate all actions from the initial state distribution, $s_0 \sim \sigma$, through time, yielding the expected probability of a state s at time t as $E_t[\mu(s)]$. The expected state visitation counts are then computed as $E[\mu(s)] = \sum_t E_t[\mu(s)]$. The updated policy and state visitation counts together with the gradients (as computed using Equations 4.4 or 4.5) are used in gradient ascent algorithms to perform the weight updates. For deterministic MDPs, the convexity of the objective function ensures a unique optimal solution will be found.

Algorithm 2: MaxEnt IRL Dynamic Programming Overview

Data: Expert trajectories, $\boldsymbol{\tau}$
Input : $\mu^a_{\mathcal{D}}, \mathcal{S}, \mathcal{A}, T, f, \gamma$
Output: optimal weights, $\boldsymbol{\theta}^*$
$\boldsymbol{\theta}^0$ =`RandomInit()`
for $i = 0, 1, \ldots$ **do**
  $U_{\boldsymbol{\theta}^i}$ = `UpdateReward(`f,$\boldsymbol{\theta}^i$`)`
  $\pi_{\boldsymbol{\theta}^i}$ = `SoftValueIteration(`$U_{\boldsymbol{\theta}^i}, \mathcal{S}, \mathcal{A}, T, \gamma$`)`
  $E[\mu^i]$ = `PolicyPropagation(`$\pi_{\boldsymbol{\theta}^i}, \mathcal{S}, \mathcal{A}, T$`)`
  $\boldsymbol{\theta}^{i+1}$ = `UpdateParameters(`$E[\mu^i], \pi_{\boldsymbol{\theta}^i}$`)` `// see Equations 4.4 (linear)`
  `// or 4.5 (neural network).`
end

Reward functions with flexible function approximators may also be learned using slightly modified MaxEnt IRL objectives (Levine et al., 2011; Wulfmeier et al., 2015). For example, in (Wulfmeier et al., 2015), the gradient of the data likelihood in maximum entropy deep (MaxEnt Deep) IRL with respect to the parameters is shown to be derived as

$$\nabla_{\boldsymbol{\theta}} L_{\mathcal{D}}(\boldsymbol{\theta}) \;\; = \;\; (\mu_{\mathcal{D}} - E[\mu]) \cdot \nabla_{\boldsymbol{\theta}} U(g(\boldsymbol{\Phi}, \boldsymbol{\theta})). \tag{4.5}$$

where the reward function is represented as the output of a neural network:

$$U \approx g(\mathbf{\Phi}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n) \tag{4.6}$$

$$= g_1(g_2(\ldots(g_n(\mathbf{\Phi}, \boldsymbol{\theta}_n), \ldots), \boldsymbol{\theta}_2), \boldsymbol{\theta}_1) \tag{4.7}$$

The MaxEnt Deep IRL objective is fully differentiable with respect to network weights, permitting use of backpropagation to update parameter estimates.

The implementation developed for this research is flexible in that the user may either specify one layer to apply the linear-in-parameters formulation, or more than one layer to recover neural network-based reward functions. While not incorporated into reward learning in the present study, the deep variant can be useful to model high-dimensional features derived from visual data such as maps or LIDAR. Furthermore, for more descriptive economic policy analysis, data-driven end-to-end simulation frameworks may be envisioned wherein IRL is used to infer rewards for RL agents behaving in alternative policy scenarios.

4.2 Accelerating IRL Via Reward Sharing And Policy Transfer

Inspired by the actor-mimic framework of Parisotto et al. (2015), we propose a methodology for policy transfer learning that involves training a deep neural network on the recovered policies of multiple agents. The output of this network can be viewed as a teacher. Additional agents may derive their initial behavior from the policy network of the teacher as a form of semi-supervised pre-training.

In reality, merely mimicking a teacher's behavior does not seem to be as efficient as inferring the motivation for that behavior. By observing the behavior of their more experienced peers, individuals may infer their preferences and thereby shape their own motivations to match those of their instructors. Thus, we also endow our teacher agents with a reward function derived from the aggregate utility function parameters of the agents used to train the teacher's policy network. Our results demonstrate that this form of transfer learning also leads to significant reductions in training time.

Model formulation

We assume access to a set of policies $\{\pi_1 \ldots \pi_N\}$ and reward function parameters, $\{\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_N\}$, trained on the (potentially suboptimal) demonstrations of N expert agents, $E_0, \ldots E_N$, using MaxEnt IRL (as described above) where E_i is the expert associated with policies π_i and reward parameters $\boldsymbol{\theta}_i$. The demonstrations are assumed to occur in identical task environments with identical goal states, although expert agents may have different preferences over trajectory distributions.

In a similar manner to (Parisotto et al., 2015), we define a *teacher* policy network, π_T to be used for transfer learning[2]. Given a state $s \in S$, the actor-mimic's loss function is defined as the cross-entropy between the policy of individual experts and the teacher policy network

$$\mathcal{L}^i_{policy}(\boldsymbol{\theta}_T) = \sum_{a \in \mathcal{A}_{E_i}} \pi_{E_i}(a \mid s) \log \pi_{TP}(a \mid s, \boldsymbol{\theta}_{TP}), \tag{4.8}$$

where $\mathcal{A}_{E_i} \subseteq \mathcal{A}$. The representation of π_{TP} used in this study is as a deep neural network parameterized by $\boldsymbol{\theta}_{TP}$ with 32 hidden units trained using the Adam algorithm (Kingma and Ba, 2015). We train the algorithm with early stopping for regularization purposes.

As a first step towards transfer learning, we simply initialize the policy used in the inner loop of MaxEnt IRL optimization for *student* agents with the policy trained using N experts. The soft value iteration step of Algorithm 2 is simply skipped for several iterations and π_T is used instead. This can be thought of as a warm-start to the policy optimization component of the MaxEnt IRL algorithm.

In addition to training generalized policies, we also considered initializing the reward function of student agents with the average parameters of the expert agents, $\boldsymbol{\theta}_{TR} = \frac{1}{N}\sum_{i=0}^{N} \theta_i$. We hypothesize that this will form an informative prior on the reward parameters of subsequently trained agents, thereby reducing the training time of individual agents.

4.3 Activity-Travel Inverse Planning Problem Formulation

We assume that agents choose an activity α from a finite set of activities $\mathcal{J}$. Travel between activities is performed according to a transport mode m selected from a finite set of available modes of travel $\mathcal{M}$.

State and action space representations. The state, $s_t \in \mathcal{S}$, indexed by the decision epoch, t, represents either participating in an activity, $\alpha \in \mathcal{J}$, or traveling $m \in \mathcal{M}$. The day is subdivided into N intervals of time τ, and $t \in \mathcal{T}$ indicates the time at the point of decision at the end of the interval. It is assumed that the duration of the interval is sufficiently granular to capture behavior, but not so fine that the size of the state space becomes exceedingly large. The augmented state space is $\mathcal{S} = \{\mathcal{J}, \mathcal{M}\} \times \mathcal{T}$. We assume that agents' first and last activity are the same ($s_0 = s_{N-1}$), i.e., agents start and end their days at home. In order to model stationary policies, we ensure that the only available action from the final home state is to itself. Doing so permits use of a small discount factor, $\gamma \in (0, 1)$,

[2]Note that, although (Parisotto et al., 2015) use Q-learning and transform the optimal Q-functions, we have already trained expert agents with softmax policies, thereby obviating the need to perform this transformation. Since the gradient of the loss function will be taken with respect to the teacher network, it is not necessary for the expert policies to be differentiable.

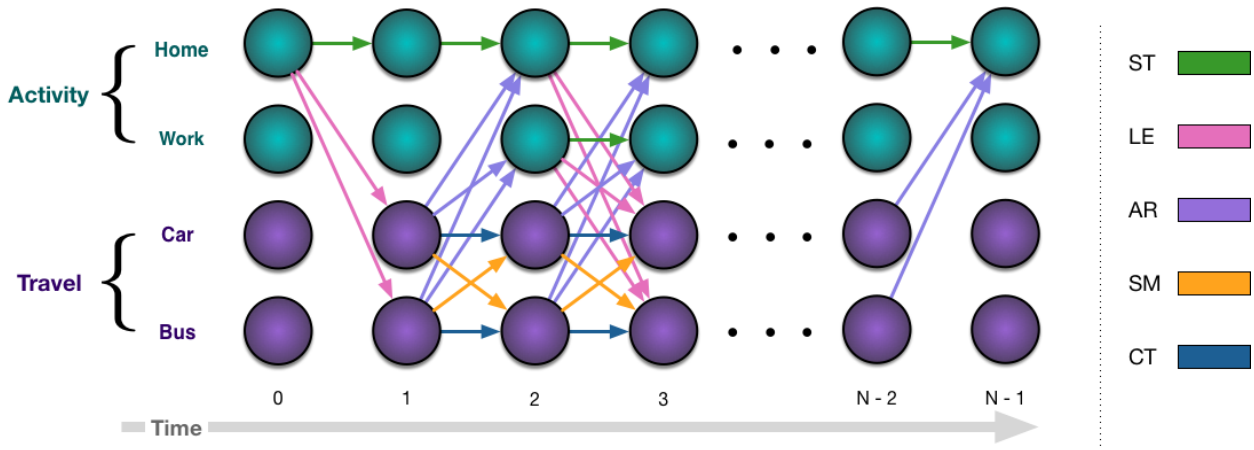


Figure 4.3: Example dynamics describing activity-travel plan MDP for two types of activity and two travel modes. Arrows between activities represent possible choices of next activity or travel given the current state. Note that certain states are not reachable (i.e., car at $t = 0$, work at $t = 1$).

which we find, empirically, rapidly accelerates convergence of dynamic programming (though in reality, people are unlikely to consider within-day temporal discounting).

The set of state-dependent decisions, $a_t \in \mathcal{A}(s_t)$ corresponds to:

ST: staying at the current activity ($s_t = \alpha \implies s_{t+1} = \alpha$),

LE: embarking on a trip to a different activity ($s_t = \alpha \implies s_{t+1} = m$),

AR: arriving at a destination ($s = m \implies s_{t+1} = \alpha$)

CT: continuing a trip in progress using the same mode ($s_t = m \implies s_{t+1} = m$), and

SM: switching to a different travel mode ($s_t = m \implies s_{t+1} = m', m \neq m'$).

See Figure 4.3 for a schematic of possible states and transitions.

Utility function specification. The specification used in this work draws from the Charypar-Nagel generalized utility function specification used in MATSim agent-based microsimulation (Charypar and Nagel, 2005b). The original linear-in-parameters formulation needs to be modified somewhat in order to accommodate the Markovian nature of our problem definition, but the definitions largely follow those described in Chapter 3 of (Horni et al., 2016a).

We assume that representative values of the following attributes are available or can be imputed for each activity, $\alpha \in \mathcal{J}$:

- Earliest end time, t_{end},
- Latest start time, t_{start}, and
- Typical duration, t_{typ}.

The overall utility function specification is

$$U_{\text{plan}} = \sum_{\alpha \in \mathcal{J}} \left(U_{\text{early dep},\alpha} + U_{\text{late arr},\alpha} + U_{\text{perf},\alpha} \right) + \sum_{m \in \mathcal{M}} U_{\text{travel},m}, \tag{4.9}$$

which is comprised of the following components:

- An early departure penalty,
$$U_{\text{early dep},\alpha} = \theta_{\text{early dep}}^{\top} \phi_{\text{early dep}},$$
for not spending enough time at an activity (i.e., leaving before the earliest end time);
- A late arrival penalty,
$$U_{\text{late arr},\alpha} = \theta_{\text{late arr}}^{\top} \boldsymbol{\phi}_{\text{late arr}},$$
for arriving late at an activity (i.e., past the latest possible start time); and
- A participation benefit,
$$U_{\text{perf},\alpha} = \theta_{\text{perf}}^{\top} \boldsymbol{\phi}_{\text{perf}},$$
for time spent performing an activity, α. The activity participation benefit is defined for all $t \in \mathcal{T}$, permitting resolution of marginal preferences at any decision epoch.
- For travel modes, we simply specify a travel time cost feature. That is,
$$U_{\text{travel},m} = \theta_{\text{travel}}^{\top} \boldsymbol{\phi}_{\text{travel}},$$
where
$$\boldsymbol{\phi}_{\text{travel}}\left(s_t, a_t\right) = \begin{cases} \tau & \text{if } s_t \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases}.$$

An additional feature function is added to enforce the constraint that agents end their day at home.

4.4 Experiments

In the following section, we describe evaluation of our framework in a real-world case-study, demonstrating its applicability to estimate demand for transport in large metropolitan centers.

We implement our activity-travel domain using the OpenAI gym API, which has gained rapid traction as a standard benchmark for evaluation of RL and IRL algorithms (OpenAI, 2016). A parallelized implementation of our framework using TensorFlow (Abadi et al., 2016) is available at `https://github.com/sfwatergit/da-irl`.

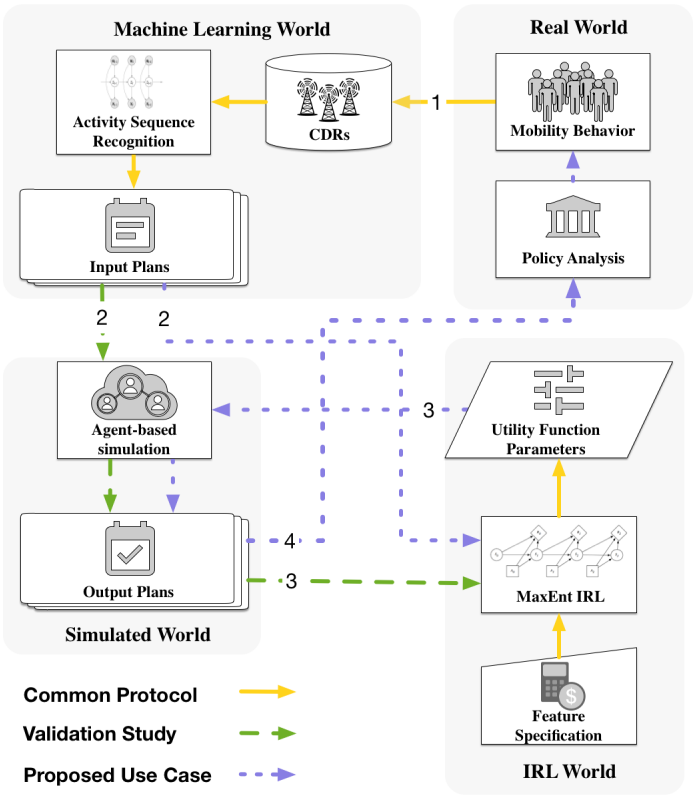


Figure 4.4: Process component interaction for validation study (see Section 4.4). Numbers indicate the order of execution on the respective process flow paths.

Data description and preparation

Input data consists of a large sample of raw, anonymized call detail record (CDR) data for cellular customers in the San Francisco Bay Area. This data is cleaned and aggregated into stay location sequences (tuples of latitude, longitude, start time, and end time) for individual users using an oscillation removal technique similar to that of (Yin M. et al., 2017). We perform spatial clustering of stay point coordinates using DBSCAN (Ester et al., 1996) to identify important locations and use heuristic rules to assign labels to home and work locations. We assign the label of "other" to clusters that cannot be identified as belonging to the primary home or work clusters. Stay sequences are then resampled according to a user-specified discretization measure defining the frequency of decision-making epochs. All training and data preparation took place on a secure cloud-based server with 36×2.20 GHz Intel® Xeon CPUs with 36 GB of RAM.

Initial algorithm validation

In order to validate our approach, we ran the procedure summarized in Figure 4.4 on the travel activity plans derived from cellular data for a population of 463,000 individuals after a MATSim run until convergence (see Chapter 3 for more information on the MATSim operational cycle). We compared our results with the utility parameters configured for the simulation that generated the output plans (i.e., inputs to the MaxEnt IRL procedure). Example results are summarized in Table 4.1.

Parameter	MATSim (util/hr)	MAXEntIRL (util/hr)	Pct. Diff
$\theta_{\text{late arr}}$	-18	-34	-89
θ_{perf}	6	6.54	-9
$\theta_{\text{travel,walk}}$	-0.29	-1.05	-262
$\theta_{\text{travel,car}}$	-0.34	-0.69	-102
$\theta_{\text{travel,pt}}$	-0.16	-1.5	-837.5

Table 4.1: Performance of IRL-based activity scheduling framework in recovering MATSim marginal utility parameters (ground truth). Note that while late departure was part of the original utility specification, the proportion of plans that included agents who arrived late was negligible. Since this feature was therefore relatively uninformative, we do not report it here.

In Figure 4.5, we plot the daily cumulative utility as a function of time graphs for two representative agents in order to illustrate how our function specification maps onto daily decision-making. The agent whose cumulative utility is plotted in the top graph take public transit to work and arrives 15 minutes past the latest arrival time, incurring a *disutility* of $\sim$ 34 utils (per our recovered parameters, Table 4.1). The bottom graph shows the cumulative utility of an agent who drives to work, arriving and leaving on time. The agent's evening commute is clearly longer than his morning commute. Note that, for ease of interpretation, we linearize the logarithmic "performing activity" component of the utility function.

Our validation of the MATSim utility function gives us some confidence in the applicability of MaxEnt IRL to recovery of the parameters guiding daily-activity scheduling. Note that at the time of this writing, these results have not been subjected to more rigorous cross-validation or other statistical robustness checks. The primary purpose of this initial case study was to validate the algorithm in order to proceed with design of more complex features as well as begin work on the extensions and improvements to our approach (described next).

Recovery of agent utility function parameters

In our second experiment, we evaluated the ability of our system dynamics and algorithm implementation to recover interpretable utility functions. We augmented the MaxEnt IRL

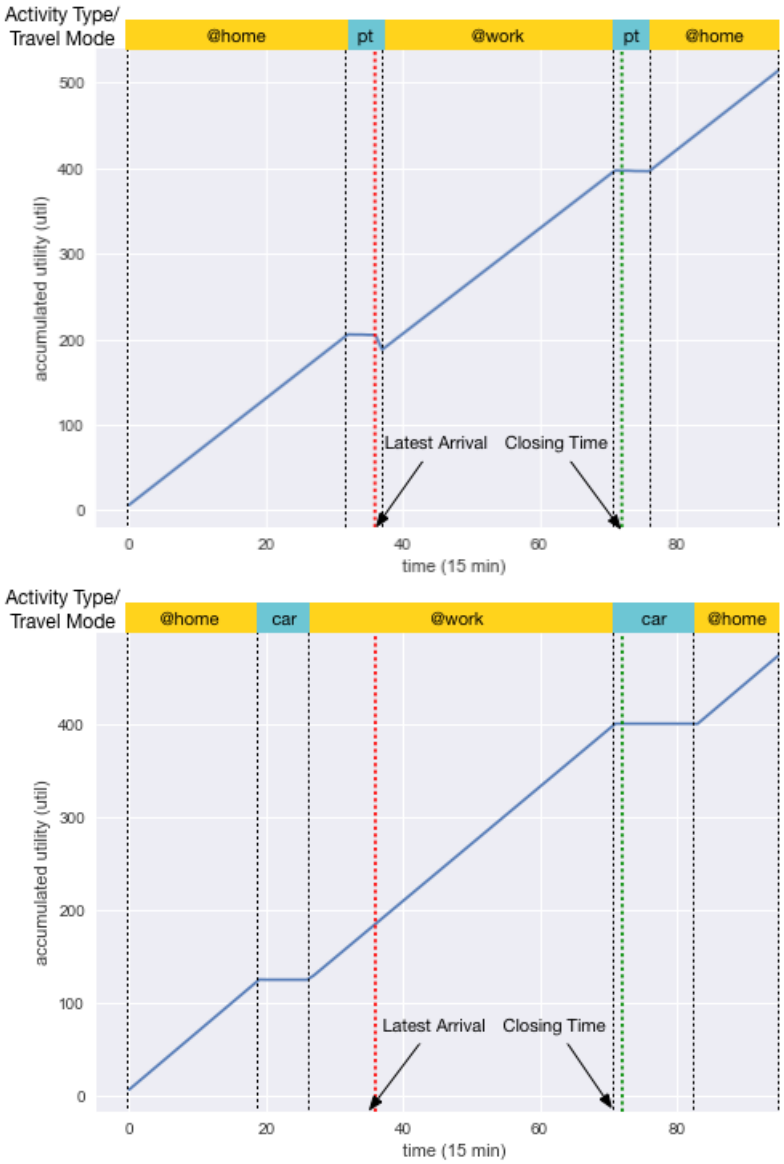


Figure 4.5: Sample utility vs. time plots for two representative agent daily activity-travel schedules. The first agent (top) uses public transit and arrives to work late, incurring a penalty. The second agent (bottom) drives to and from work with a slightly longer evening than morning commute.

objective with L2 regularization, which we found to not only achieve smoother, more interpretable utility curves, but also somewhat faster convergence. Since the scale of the reward function parameters is not identifiable (that is, we cannot make interpretations of the relative scale of the marginal utility between individuals), we normalize their values to 1.0. Parameter optimization is carried out using the Adam algorithm of (Kingma and Ba, 2015).

In Figures 4.6 to 4.9, we provide a representative demonstration of the result of training our algorithm on 50 randomly chosen individuals. The learning curve (Figure 4.6) is nearly identical for all agents. While we observe that the feature matching objective rapidly approaches 10^{-2}, in practice, we find that parameter values typically do not stabilize until $\mu_{\mathcal{D}} - \mathbb{E}[\mu] = 10^{-3}$, which is occurs around the 150th to 175th iteration of gradient descent.

Figures 4.8 and 4.9 are indicative of the shape of marginal utility over the course of the day. As expected, we see in Figure 4.8, a generally strong preference for agents to be located at home in the evening and early morning hours, whereas preference for work is steady through the middle portion of the day and participation in other activities occurs primarily in the evening.

More interestingly, Figure 4.8 demonstrates diminishing marginal utility of work activity participation towards the end of the day. We further note that similarities and differences between agents in marginal utility trade-offs suggests reactive policy interventions could be designed to influence the behavior of individuals based on cross-elasticities between home and work parameter values derived in this manner. A topic for further research will be to investigate socioeconomic characteristics associated with the shape of these curves.

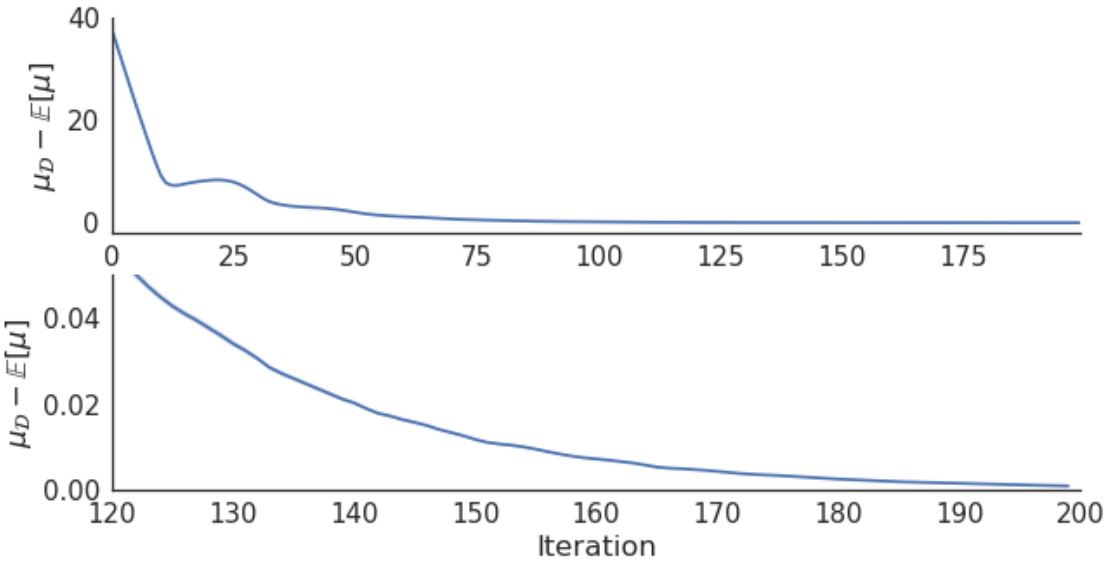


Figure 4.6: Learning curve for 50 expert agents. Top: over all training iterations. Bottom: detail of final 80 iterations.

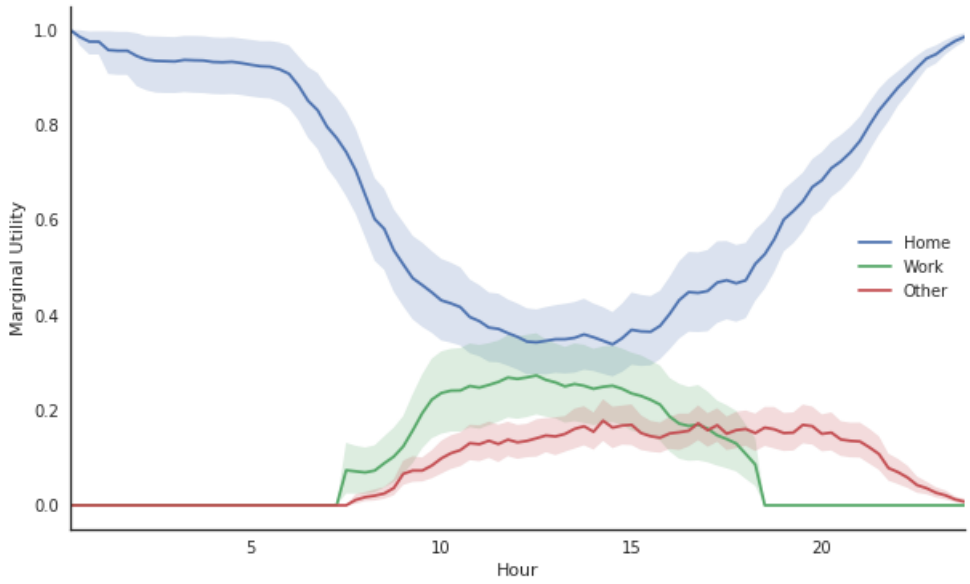


Figure 4.7: Combined utility values for 50 expert agents (mean value with 95% confidence regions).

Performance of policy and reward transfer IRL

We evaluate our transfer-learning approach using 50 randomly selected expert agents, which we then trained using MaxEnt IRL. We then randomly sample the demonstrations of 10 *student* experts and run 100 iteration of the MaxEnt IRL algorithm by itself and augmented with policy transfer, reward transfer, as well as combined policy and reward transfer-learning. In the experimental conditions involving policy transfer, we skip the first 5 iterations of soft value iteration and then resume soft-value iteration thereafter.

As illustrated in Figure 4.10, our results demonstrate significantly improved convergence for all experts when MaxEnt IRL is augmented with aggregate reward pre-training. In practice, we have observed that reward pre-training can reduce the number of iterations required to reach convergence for individual agents by approximately one half.

Unfortunately, we don't observe a significant reduction in convergence over vanilla Max-Ent IRL when introducing policy transfer, and, in fact, the experiment presented herein shows that more, not fewer iterations are required. Future work will investigate refining this approach before abandoning it altogether.

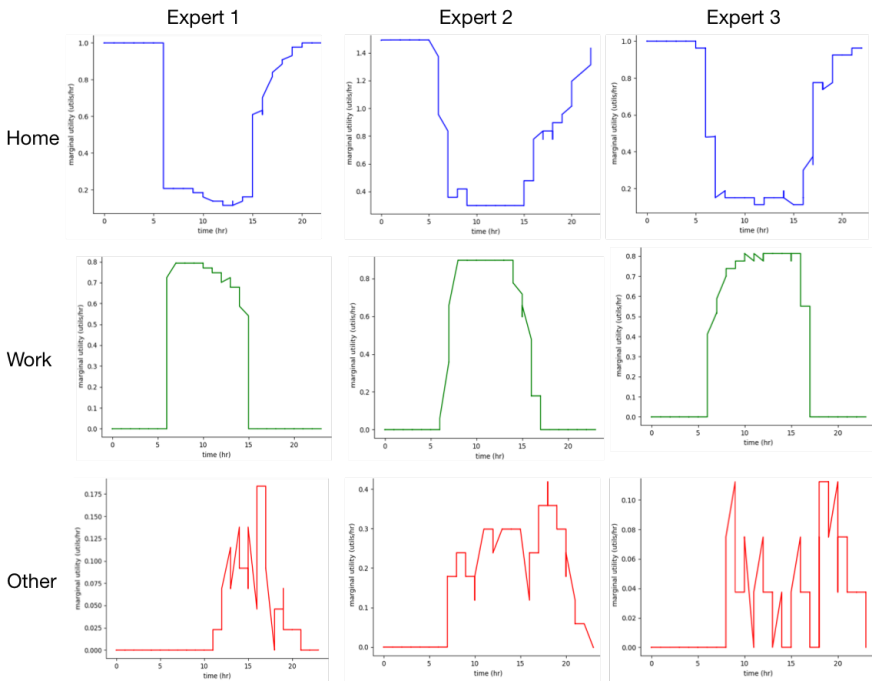


Figure 4.8: Representative utility parameters for 3 expert agents with typical utility profiles

4.5 Future Work

In order to demonstrate the applicability of MaxEnt IRL to inform economic policy appraisal for heterogeneous users of municipal transport systems, it is necessary to rationalize the behavior of diverse demographics. Towards this end, we propose to cluster trajectories according to behavioral classes. Features used for clustering will be based on the travel parameters, and may include inferred activity identities, number of hours worked, opening/closing times of activities, and home-work locations by origin and destination (OD) TAZ. Learning model parameters for different population segments would demonstrate the ability of our framework to capture the revealed preferences of commuters from cellular data at various levels of disaggregation.

An additional improvement would be to formally encode the hierarchical nature of the activity-travel schedule into the reward learning algorithm. For example, travel behavioral preferences are conditional on mode choice. Perhaps more importantly, the diminishing marginal utility of activity performance cannot be effectively encoded in the discretized temporal structure that defines the current environment dynamics. Representation of the environment dynamics as a semi-Markov Decision (SMDP) is one way to proceed, although no IRL frameworks developed thus far easily generalize to SMDPs. Perhaps a simpler approach is to learn temporally-extended actions, otherwise known as options (Sutton et al., 1999). Recently, several works have developed unsupervised algorithms to infer higher-level behavioral structure in trajectories using deep neural networks (Fox et al., 2017; Machado

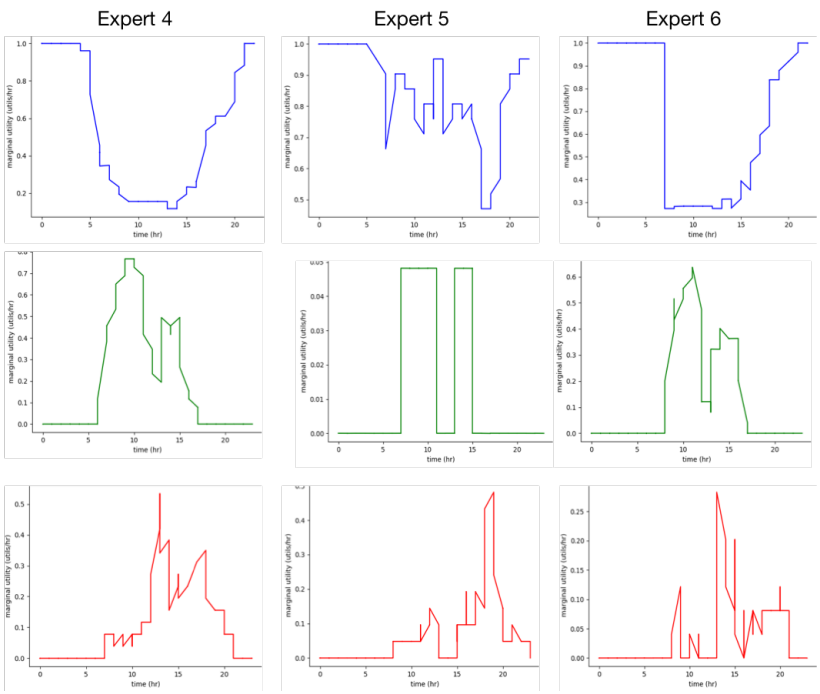


Figure 4.9: Representative utility parameters for 3 expert agents with atypical utility profiles

et al., 2017). Similarly, Krishnan et al. (2016) introduce Sequential-Windowed Inverse Reinforcement Learning (SWIRL) an algorithm to discover local intervals with heterogeneous reward functions. Once the sequence of instantaneous transition points between reward regions are identified (using an unsupervised variant of Gaussian mixture model clustering learned over the temporal activity-travel sequences), the SWIRL algorithm trains MaxEnt IRL separately on intervals between transition points.

4.6 Conclusions

In this chapter we presented a novel formulation of daily activity-travel scheduling as a problem of inverse optimal control. Our basic approach was evaluated in the context of of an end-to-end framework for rationalization of commuter preferences from digital human trajectories derived from large quantities of passively collected cellular data. Our case study demonstrated that MaxEnt IRL is a promising computational methodology for the inference of utility functions motivating daily activity-travel plans. The individual utility function parameters recovered by our methodology provide empirical validation of theoretical all-day utility functions while illustrating interesting heterogeneity across agents.

We also demonstrated how reward and transfer-learning for agents demonstrating behavior in similar environments could be used to accelerate convergence of MaxEnt IRL algorithms. In particular, we find that initialization of reward functions used in the inner

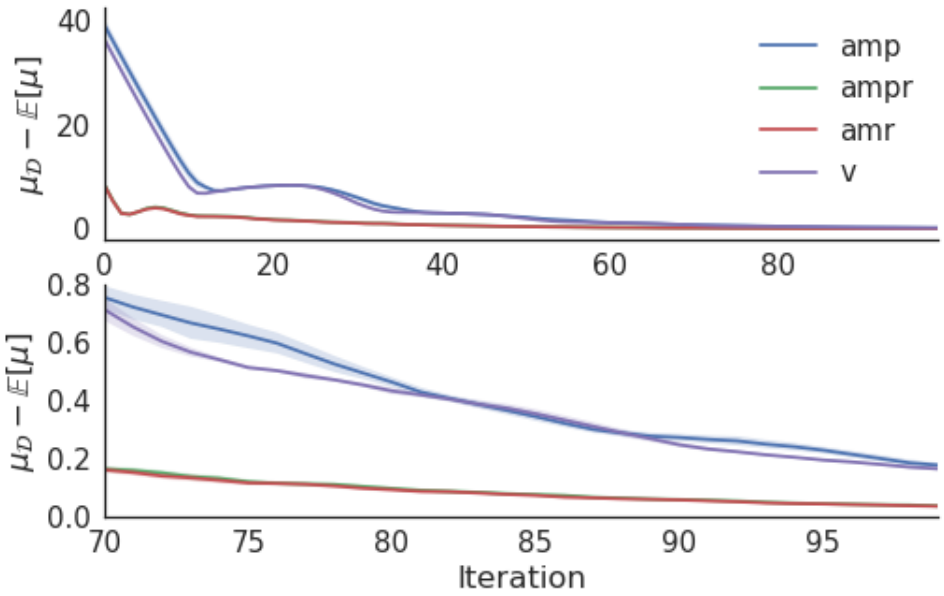


Figure 4.10: Learning curves for vanilla MaxEntIRL (v), policy transfer (amp), reward transfer (amr), and combined policy and reward transfer learning (ampr) for 10 *students*. Top: over all training iterations. Bottom: detail of final 80 iterations.

loop of MaxEnt IRL optimization leads to significant reductions in training time. Future work will seek to improve our policy transfer method as well as benchmark the performance of our method against other approaches to meta-learning reward function parameters and policies described in the literature.

A significant shortcoming of this work is that learned policies did not appear to match demonstrated behavior upon algorithm convergence. We suspect that this was due to the somewhat unrealistic assumption that decision-epochs are evenly spaced at regular intervals throughout the day (every 15 minutes, in this case). Clearly, real people do not reconsider their choice of activity with such high frequency and such exact regularity. This limitation also makes it difficult to specify variables and estimate parameters related to distance between activities and the duration of activities. This observation motivates the refactored environment presented in the next chapter, in which agents may choose activity duration in addition to identity. We will also explore efficient algorithms that are more conducive to larger action spaces than MaxEnt IRL, a necessary consequence of the new environment definition.

5

Generative Models of Activity Sequences and Duration via Adversarial Imitation Learning

Nature is commonplace. Imitation is more interesting.

– Gertrude Stein

The previous chapter presented a somewhat simplified model of daily activity-travel planning choice dynamics in order to investigate the utility of MaxEnt IRL framework for the estimation of structural models of individual preferences from partially-labeled stay-point sequences derived from cellular data. While the parameters learned for the reward function features matched intuitions, limitations imposed by the deterministic transition dynamics prevented several desirable variables (e.g., disutility of late arrival, travel distance cost) from entering into the specification.

In this chapter, we introduce a more realistic version of the model of choice dynamics presented in previous chapters. Whereas the former environment assumed a constant 15 minute interval between decision epochs, the proposed update allows for an arbitrary amount of time to pass from one decision to the next. That is, agents should be able to choose, at the end of each activity, what their next activity should be, how long to spend there, and what mode to take. The crux of this change is a transition kernel that more accurately reflects real-world activity and travel planning contexts and therefore allows for more natural specification of both spatial as well as temporal variables influencing travel behavior. The resulting utility functions could more readily be implemented in microsimulation frameworks, permitting their use in policy analysis under congested, multi-agent contexts. However, this expressive power comes at the cost of a significantly expanded action space, rendering the model-based soft value iteration and policy propagation algorithms of the causal maximum entropy and related dynamic discrete choice framework less viable due to computational costs that scale exponentially with the number of available actions (Ziebart et al., 2008).

The primary goal of this work is to train a reward function using expert demonstrations

of daily-activity travel patterns such that the policy induced by the learned reward function is able to mimic expert activity-travel patterns (i.e., those sampled from expert's empirical trajectory distribution). Towards this end, we explore the ability of the generative adversarial imitation learning (GAIL) algorithm described in (Ho and Ermon, 2016) to train a distribution over activity-travel patterns. We find that our implementation is able to simulate individual behavior sequences that match the distributions of observations in identify, time, and duration. To the best of our knowledge, this is the first work to synthesize daily activity-travel plans using adversarial models.

This chapter is organized as follows: Section 5.1 presents the theoretical background necessary to understand GAIL, including a brief introduction to the powerful, model-free policy optimization methods that make the GAIL algorithm possible. Next, in Section 5.2, we formally describe our planning environment. Results are presented and discussed in Section 5.3. Lastly, selected topics for future studies and conclusions are presented in Section 5.4.

5.1 Background And Preliminaries

Model-free IL and IRL

In order to reproduce the behavior of complex, adaptive agents we turn to recently-developed IL and IRL methodologies that expand the representational power of earlier techniques (Finn et al., 2016a,c; Fu et al., 2018; Ho and Ermon, 2016). In these frameworks, *generative models* of stochastic policies are trained to reproduce expert behavior when dynamics (i.e., the transition model, $P(s' \mid s, a)$ are complex or unknown. Termed *model-free IRL* methods (in analogy to model-free RL methods), the absence of a transition kernel typically requires either some form of *policy optimization* (described in more detail below) or *policy search* (see (Levine and Abbeel, 2014) for details on policy search methods, which will not be treated further here) in order to evaluate improvements in reward functions during training. The training algorithms proposed in these works are similar in that they each draw inspiration from *generative adversarial networks* (GANs, Goodfellow et al. (2014)): a framework that trains generative models to confuse a *discriminative* classifier (GANs and their connection to IRL are described in further detail, below).

Efficient policy gradient algorithms

Before describing the algorithms used in this chapter, we will provide a brief review of the modern gradient-based policy optimization methods, which play a critical role in making model-free IRL possible for use in high-dimensional problems (e.g., the current environment setting). It is important to appreciate that policy gradient algorithms have only become scalable to high-dimensional environments due to very recent (last three years) advances in:

1. natural policy gradient-based optimization methods (see (Kakade, 2001) and (Achiam et al., 2017) for more details on these), and

2. the science and implementation of deep neural network methods.

Like all RL algorithms, the objective of policy gradient methods is to maximize the expected discounted reward over trajectory realizations (often termed rollouts, $\boldsymbol{\tau}$). Gradient-based methods optimize a differentiable, parameterized policy, $\pi_{\boldsymbol{\theta}}$, by taking stochastic gradient steps in the space of policy parameters. In practice, the class of policy gradient methods (i.e., those based on the original REINFORCE algorithm, (Williams, 1992)) was considered to be sample inefficient and liable to suffer from frequent catastrophic performance collapse during stochastic optimization algorithms due to underspecified step-size constraints (Achiam et al., 2017; Schulman et al., 2015). Modern policy optimization routines address these two problems by taking steps in policy space[1], rather than parameter space, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In these algorithms monotonic improvements in step size are guaranteed. The objective enforces a trust region constraint defined using the Kullback-Liebler divergence between policies recovered at successive optimization steps, i.e.,

$$D_{KL}(\pi_{k+1} \,||\, \pi_k)[s] = \sum_{a \in \mathcal{A}} \pi_{k+1}(a \mid s) \log \frac{\pi_{k+1}(a \mid s)}{\pi_k(a \mid s)}.$$

The resulting objective and constraint can be estimated from rollouts under the current policy by computing the so-called *natural gradient*, which requires computation of the Fisher information matrix (inverse Hessian) of the policy (Kakade, 2001). For neural networks with large numbers of parameters, the complexity of computing the Hessian for each step is prohibitive for practical use. Recently, Schulman, et. al., (2015) found that using a trust-region optimization method[2] resulted in significant performance gains. Within only a few years, the resulting algorithm, *Trust Region Policy Optimization* (TRPO), has become a critical component of many recent innovations in reinforcement learning research.

Connections between IL and GANs

Generative adversarial networks: preliminaries

The success of deep neural networks has inspired a separate strand of research in generative model estimation frameworks, inquiry into the properties and efficient training of *variational autoencoders* (VAE, (Kingma and Welling, 2013)) and generative adversarial networks (GANs) being the most prominent. The adversarial network modeling techniques used in

[1]The policy space of all possible policies defined over the state and action space is

$$\Pi = \left\{ \pi : \pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}, \sum_{a} \pi_{s,a} = 1, \pi_{s,a \geq 0} \right\}.$$

[2]Trust region methods are a relatively new class of algorithms used in nonlinear optimization problems. The most common trust region approach uses a conjugate gradient approximation to the Hessian and enforces a pre-defined step-size using backtracking line search. They are often robust and can be applied to ill-conditioned problems.

this thesis involve learning a *generator*, G's, distribution, $p_{\boldsymbol{x}}$ over input data, $\boldsymbol{x}$ by training a differentiable multilayer perceptron (MLP) $G(\boldsymbol{z};\boldsymbol{\theta}_g)$, to produced synthetic data, $\boldsymbol{z}$, from a distribution over noise variables $p_{\boldsymbol{z}}(\boldsymbol{z})$, which are then classified as real or fake using a second MLP known as a *discriminator* $D(\boldsymbol{x};\boldsymbol{\theta}_d)$. The discriminator is trained to maximize the probability of assigning the correct label to samples from both the training data as well as from G, while G is trained to minimize $\log(1 - D(G(\boldsymbol{z})))$. Training continues until D is unable to distinguish true data from fake data generated by G, that is, $p_g = p_{\text{data}}$. The relationship between G and D may be represented as a two-player minimax game,

$$\min_G \max_D V\left(D,G\right) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}\left[\log D\left(\boldsymbol{x}\right)\right] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right],$$

where the goal is to achieve Nash equilibrium between the discriminator and generator.

It is important to note that the initial theoretical results have not necessarily corresponded to high-quality learning, in practice. Achieving Nash equilibrium during GAN training is difficult, and early implementations of GANs suffered from vanishing discriminator gradients and mode collapse, among other problems[3] (Arjovsky and Bottou, 2017; Salimans et al., 2016). However, much better results have been achieved with modest objective reformulations and training improvements (Arjovsky and Bottou, 2017; Salimans et al., 2016). For example, Arjovsky et al. (2017) introduces the Wasserstein GAN (WGAN), which uses an Earth-Mover's (Wasserstein I) distance objective as a replacement for the Jensen-Shannon Divergence used in Goodfellow et al. (2014)'s original formulation. In addition, the WGAN paper introduced weight-clipping to control for vanishing gradients. The WGAN objective can also be augmented with a gradient penalty, as reported in (Gulrajani et al., 2017). Currently, a combination of the Wasserstein objective, weight-clipping, and gradient penalty leads to state-of-the-art image generation with deep neural networks (Gulrajani et al., 2017). Thus, we employ these improvements in our own implementation.

Generative adversarial imitation learning

We now describe the GAIL algorithm, which uses a GAN-inspired approach to perform IRL. The One can think of the algorithm as rewarding a parametric policy, π_θ for producing trajectories $\boldsymbol{\tau}$ that are similar to trajectories $\boldsymbol{\tau}_E$, sampled from an expert policy, π_E, such that a discriminator function, $D_{\boldsymbol{\omega}}$ (represented by a neural network), fails to distinguish between the two.

The GAIL objective is the sigmoid cross-entropy:

$$\max_{\boldsymbol{\theta}} \min_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{\tau}\sim\pi_{\boldsymbol{\theta}}}\left[\log\left(D_{\boldsymbol{\omega}}\left(s,a\right)\right)\right] + \mathbb{E}_{\boldsymbol{\tau}_E\sim\pi_E}\left[\log\left(1 - D_{\boldsymbol{\omega}}\left(s,a\right)\right)\right] + \lambda_H H\left(\pi_{\boldsymbol{\theta}}\right)$$

where $H\left(\pi_{\boldsymbol{\theta}}\right)$ is the entropy of the learned expert policy, π_E and λ_H is a weighting hyperparameter for this entropy regularization term.

[3]Mode collapse refers to the discriminator predicting a certain low dimensional component of the data distribution without generalizing to the surrounding manifold. Mode collapse is occasionally referred to, cheekily, as "the Helvetica scenario" (Goodfellow et al., 2014).

Training proceeds by sampling expert trajectories from π_E and simulating trajectories by performing rollouts of the generative policy $\pi_{\boldsymbol{\theta}}$ on the environment. The reward function used to evaluate rollouts is:

$$U(s_t, a_t, \boldsymbol{\omega}) = -\log(1 - D_{\boldsymbol{\omega}}(s_t, a_t)).$$

Algorithm 3: Generative Adversarial Imitation Learning (GAIL)

Input : $\boldsymbol{\tau}_E \sim \pi_E$, randomly initialized parameters $\boldsymbol{\theta_0}$,$\boldsymbol{\omega}_0$

for $i = 0, 1, 2, \ldots$ **do**

1. Sample trajectories $\boldsymbol{\tau}_i \sim \pi_{\theta_i}$
2. Update $\boldsymbol{\omega}_i$ to $\boldsymbol{\omega}_{i+1}$ following gradient

$$\mathbb{E}_{\boldsymbol{\tau}_i \sim \pi_\theta}\left[\nabla_{\boldsymbol{\omega}} \log\left(D_{\boldsymbol{\omega}}\left(s, a\right)\right)\right] + \mathbb{E}_{\boldsymbol{\tau}_E \sim \pi_E}\left[\log\left(1 - D_{\boldsymbol{\omega}}\left(s, a\right)\right)\right]$$

3. Take a policy step from $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_{i+1}$ using the TRPO update rule with cost function $\log\left(D_{\boldsymbol{\omega}_{i+1}}\left(s, a\right)\right)$ with the following objective:

$$\mathbb{E}_{\boldsymbol{\tau}_i}\left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\left(a \mid s\right) Q\left(s, a\right) - \lambda \nabla_{\boldsymbol{\theta}} H\left(\pi_{\boldsymbol{\theta}}\right)\right],$$

where $Q\left(\bar{s}, \bar{a}\right) = \mathbb{E}_{\boldsymbol{\tau}_i}\left[\log\left(D_{\boldsymbol{\omega}_{i+1}}\left(s, a\right)\right) \mid s_0 = \bar{s}, a_0 = \bar{a}\right]$

end

Towards interpretable imitation learning with InfoGAIL

An efficient approach to interpretable reinforcement learning is to formulate a multimodal training problem (Hausman et al., 2017; Li et al., 2017). In the InfoGAIL algorithm, a latent variable distribution over expert data is defined in order to train policies with outputs conditional on the latent value (Li et al., 2017). Doing so "... allows us to disentangle trajectories that may arise from a mixture of experts, such as different individuals performing the same task" (Li et al., 2017). This statement describes a scenario for our planing environment wherein we receive a set of trajectories from multiple agents and desire to model their behavior as a group and individually. The ability to successfully distinguish and synthesize trajectories from separate agents would greatly speed up learning. For this purpose, there is no need to employ an unsupervised approach as in (Li et al., 2017), since expert trajectories are easily labeled according to the agent that generated them.

InfoGAIL assumes that a single, *multimodal* expert policy may be characterized as a mixture distribution arising from the behavior of many experts $\pi_E = \{\pi_E^0, \pi_E^1, \ldots\}$. Assume that a discrete latent code, c, can be used to distinguish the policies π_E^n from one another according to $p(\pi \mid c)$. Then realizations of expert trajectories $\boldsymbol{\tau}_E \sim \pi_E$ can be described by the following generative process:

$$\begin{aligned} s_0 &\sim \rho_0 \\ c &\sim p(c) \\ \pi &\sim p(\pi \mid c) \\ a_t &\sim \pi(a_t \mid s_t) \\ s_{t+1} &\sim P(s_{t+1} \mid a_t, s_t) \end{aligned}$$

where $p(c)$ is a known prior over latent code c.

The objective of InfoGAIL is to recover the multimodal expert policy $\pi(a \mid s, c)$. A variational lower bound, $L_I(\pi, Q)$ is introduced in order to enforce high mutual information,$I(c; \boldsymbol{\tau})$, between c and state-action pairs in generated trajectories:

$$\begin{aligned} L_I(\pi, Q) &= \mathbb{E}_{c\sim p(c), a\sim\pi(\cdot|s,c)}[\log Q(c \mid \boldsymbol{\tau})] + H(c) \\ &\leq I(c; \boldsymbol{\tau}) \end{aligned}$$

where $Q(c \mid \boldsymbol{\tau})$ approximates the true posterior $P(c \mid \boldsymbol{\tau})$. The objective of InfoGAIL is then:

$$\min_{\pi,Q} \max_{D} \mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] - \lambda_1 L_I(\pi, Q) - \lambda_2 H(\pi)$$

with $\lambda_1 > 0, \lambda_2 > 0$ are hyperparameters controlling the information maximization regularization term and causal entropy terms, respectively.

Rather than optimizing the posterior approximation over trajectories, InfoGAIL instead uses a simplified approximation over states and actions, $Q(c \mid s, a)$. The optimization algorithm then learns the weights, $\boldsymbol{\theta}, \boldsymbol{\omega}$, and $\boldsymbol{\psi}$ for parameterized representations of π, D, and Q, respectively, by maximizing $L_I(\pi_\theta, Q_\psi)$ and updating π_θ and $Q_{\boldsymbol{\psi}}$ using Adam, TRPO, and Adam, respectively (Kingma and Ba, 2015; Schulman et al., 2015). We reproduce the

training procedure from (Li et al., 2017) as Algorithm 4.

Algorithm 4: InfoGAIL

Input : Initial parameters of policy, discriminator, and posterior approximation $\boldsymbol{\theta}_0$,$\boldsymbol{\omega}_0, \boldsymbol{\pi}_0$, and expert demonstrations, $\boldsymbol{\tau}_E \sim \pi_E$

Output: Learned policy $\pi_{\boldsymbol{\theta}}$

for $i = 0, 1, 2, \ldots$ **do**

Sample a batch of latent codes: $c_i \sim p(c)$.

Sample trajectories $\boldsymbol{\tau}_i \sim \pi_{\boldsymbol{\theta}_i}(c_i)$, with the latent code fixed during each rollout.

Sample state-action pairs $\chi_i \sim \boldsymbol{\tau}_i$ and $X_E \sim \boldsymbol{\tau}_E$ with same batch size.

Update ω_i to ω_{i+1} by ascending with gradients

$$\Delta_{\boldsymbol{\omega}_i} = \hat{\mathbb{E}}_{\chi_i} \left[\nabla_{\boldsymbol{\omega}_i} \log D_{\boldsymbol{\omega}_i} (s, a) \right] + \hat{\mathbb{E}}_{\chi_E} \left[\nabla_{\boldsymbol{\omega}_i} \log \left(1 - D_{\boldsymbol{\omega}_i} (s, a) \right) \right].$$

Update $\boldsymbol{\psi}_i$ to $\boldsymbol{\psi}_{i+1}$ by descending with gradients

$$\Delta_{\boldsymbol{\psi}_i} = -\lambda_1 \hat{\mathbb{E}}_{\chi_i} \left[\nabla_{\boldsymbol{\psi}_i} \log Q_{\boldsymbol{\psi}_i} \log Q_{\boldsymbol{\psi}_i} (c \mid s, a) \right].$$

Take a policy step from $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_{i+1}$ using the TRPO update rule with the following objective:

$$\hat{\mathbb{E}}_{\chi_i} \left[\log D_{\boldsymbol{\omega}_{i+1}} (s, a) \right] - \lambda_1 L_I \left(\pi_{\boldsymbol{\theta}}, Q_{\boldsymbol{\psi}_{i+1}} \right) - \lambda_2 H \left(\pi_{\boldsymbol{\theta}_i} \right).$$

end

5.2 Methodology

For this initial implementation, we would like to reproduce timing, duration as well as the activity identity components of the activity-travel pattern. Additionally, we wish to understand how the reward function changes in response to new patterns of behavior demonstrated by the agent due to, say, changes in the external environment.

Environment specification

A graphical representation of the new environment is presented in Figure 5.1. The following paragraphs describe its state and action space in more detail.

State space Our representation assumes that for every possible duration (e.g., 96 for 15 minute intervals, each agent had a choice of (maximally) one trip or activity type (e.g., six for *work*, *home*, *other*, *trip to home*, *trip to work*, *trip to other*, and one terminal activity, if applicable (the final transition to home is only accessible from a home state).

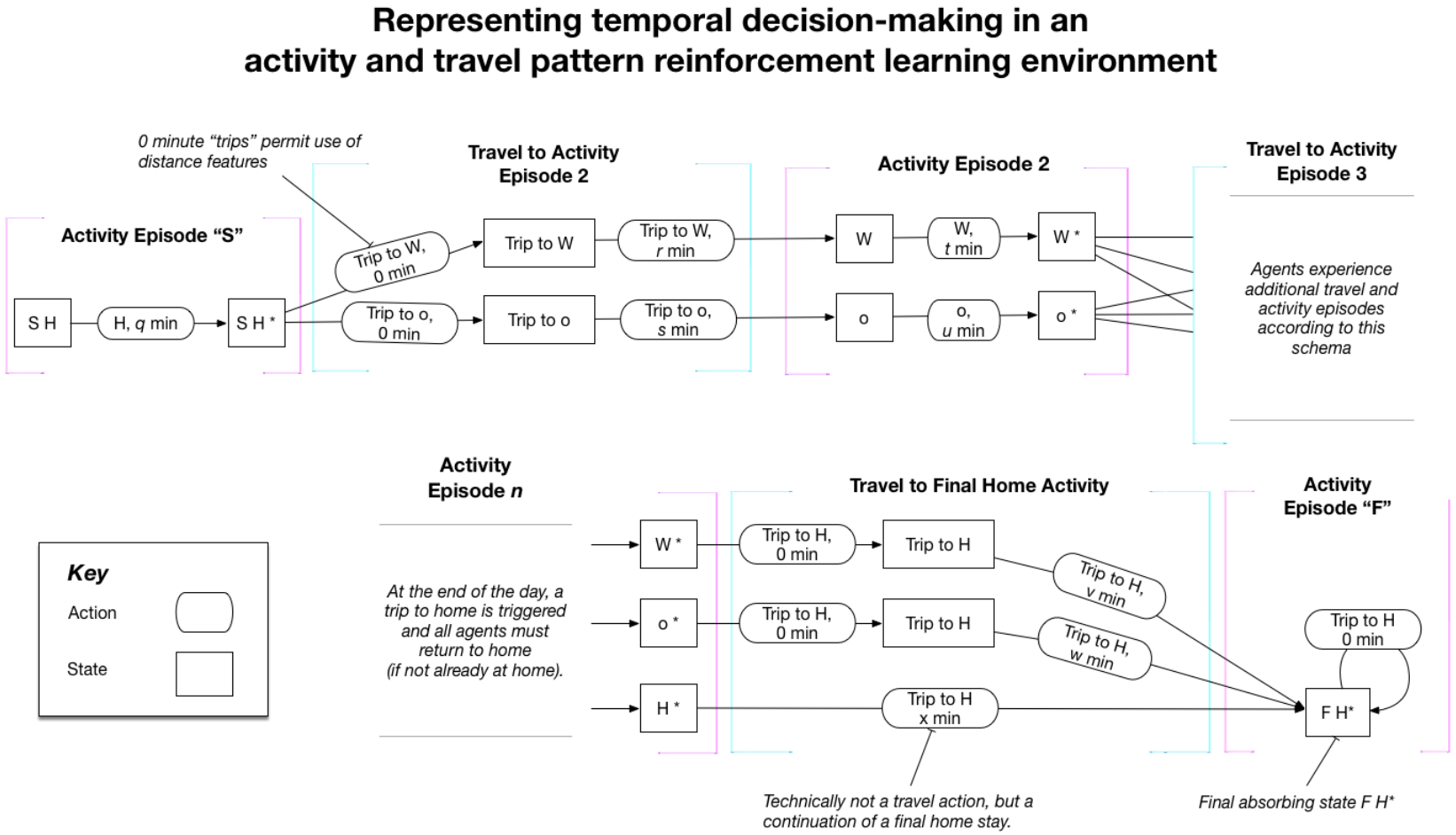


Figure 5.1: Timed activity environment graphic specification

Action space In the new environment, we use a large discrete action space: ($|A| = 2N_{\text{Activities}} \times 96 = 576$). This poses the most challenging component of the optimization problem, as most RL algorithms optimized to work with smaller, continuous action spaces[4].

Implementation details

All algorithms were implemented using `rllab` (Duan et al., 2016) and TensorFlow-based (Abadi et al., 2016) reference code from (Fu et al., 2018)[5]. Significant modifications from these source codes were implemented in order to enforce temporal state-action pair alignment for variable-length trajectories as well as implement regularization techniques to improve generalization in the presence of small individual-level datasets.

Rather than encoding a relatively large state space of ($(2N_{\text{Activities}} + 1) \times N_{\text{Time steps}} \times \{0, 1\} \approx 2 \times 10^3$) states as a one-hot vector, we were able to instead take advantage of the `MultiDiscrete` state space implementation in `rllab`, which concatenates three component vectors corresponding to each dimension of the state space into a single one-hot vector. This reduced the size of the state space to $\approx 10^2$ possible distinct components, which dramatically reduced training times. By leveraging the `rllab` base, we are able to integrate policy optimization algorithms implementations that have been verified and benchmarked by other researchers as well as parallelized to run on modern multicore architectures in cloud computing environments.

As in the environment described in Chapter 4, we implemented this environment according to the OpenAI API.

Network architecture and hyperparameter search. We used deep feedforward and recurrent neural network to parameterize the policy distribution[6]. The output activation of the network is a softmax function. Specifically, given the state s_t, actions are selected from the policy by sampling $a_t \sim \pi_{\boldsymbol{\theta}}(a_t \mid s_t)$. We used a three layer architecture of 64, 128, and 512 units for the policy network with tanh activations, a 2×128 unit discriminator network with relu activations and a batch size times 1 unit output as well as a linear baseline[7]. One

[4] A recent paper has, in fact, addressed the problem of learning in large discrete action spaces using a so-called Wolpertinger architecture with fast approximate nearest neighbor algorithm, but initial evaluations in our setting did not yield particularly compelling results (Dulac-Arnold et al., 2015). This may have been due to an implementation issue that had not been resolved at the time of this writing, so we leave investigating this promising approach to future work.

[5] See `https://github.com/justinjfu/inverse_rl`

[6] We evaluated recurrent neural network (RNN)-based policies (both LSTMs and GRUs); however, found that mode collapse actually increased in these circumstances, unlike as in (Kuefler et al., 2017; Zou et al., 2018), which found positive benefits from RNNs for sequential tasks. One possible improvement may be to implement dropout as a more robust regularizer (Srivastava et al., 2014).

[7] A baseline $B(s)$ is a function of states subtracted from the policy gradient. It is used to reduce variance without increasing bias in policy gradient algorithms. Often, as is done in this work, the state value function, $V^{\pi}(s)$ is selected as $B(s)$.

important aspect of training potentially variable trajectory lengths was to mask out logits (network outputs) prior to computing the loss function.

It is useful to note that a fair amount of parameter tuning was required in order to achieve reasonable results on this task. A hyperparameter search over the space of discount rates, TRPO step sizes, and entropy weights was performed. We found that using a fairly large number of sampled trajectories (3,000 in our case) and using a small ratio of discriminator to generator training iterations (200:1) per epoch achieved the best results. We used the Adam algorithm with $\beta_1 = 0.5, \beta_2 = 0.9$, and a learning rate of 1×10^{-5}. An entropy penalty hyperparameter of $\lambda = 1 \times 10^{-4}$ and $L2$ regularization constant of 0.05 were also used, which we found prevented mode collapse early in training.

5.3 Results

Single agent training The GAIL algorithm was trained on a small dataset of demonstrations of daily activity-travel patterns (see Figure 5.2 for an illustration of the target data). At the endpoint of optimization, the policy was able to reproduce the expert's pattern and corresponding duration distributions with a high level of accuracy. Figure 5.3 illustrates this finding by comparing the empirical distribution over the observed patterns (i.e., those shown in Figure 5.2) to 100 random samples from the optimal policy. Figure 5.4 and Figure 5.5 demonstrate that our algorithm is able to faithfully reproduce activity duration and timing for this agent for each of the 10 possible activity positions illustrated in Figure 5.3 (e.g., the distribution of simulated durations at the 5th possible position in an activity pattern are close to what was actually observed). The most difficulty seemed to occur at the 7th timestep (generally between 14:00 and 16:00 as illustrated in Figure 5.2). Perhaps this is due to the fact that position 7 has the greatest variability in both activity timing and identity, as reflected in Figures 5.3 and 5.5 (i.e., the model must select between 'H', W', and 'o' at different times of day for each possible pattern and activity duration).

It took between 100-200 iterations to train an optimal policy, with further training actually decreasing accuracy. Currently, it takes about 10 minutes to train approximately 120 full GAIL iterations on a 32-CPU machine. Running on GPUs would likely speed up computation somewhat, although the largest bottleneck is the large amount of sample trajectories necessary to achieve appreciable results.

We assume that training difficulty arises from the observation that discrete temporal durations necessarily involve an all-or-nothing selection of the correct action. Thus, in a vanilla encoding, there is no notion of semantic distance between neighboring states and actions. In the future; however, we plan to explore continuous temporal embeddings in order to facilitate more efficient learning.

Behavioral cloning pre-training. In order to accelerate training, we included behavioral cloning as a warm start mechanism. We used maximum likelihood as the BC objective due to

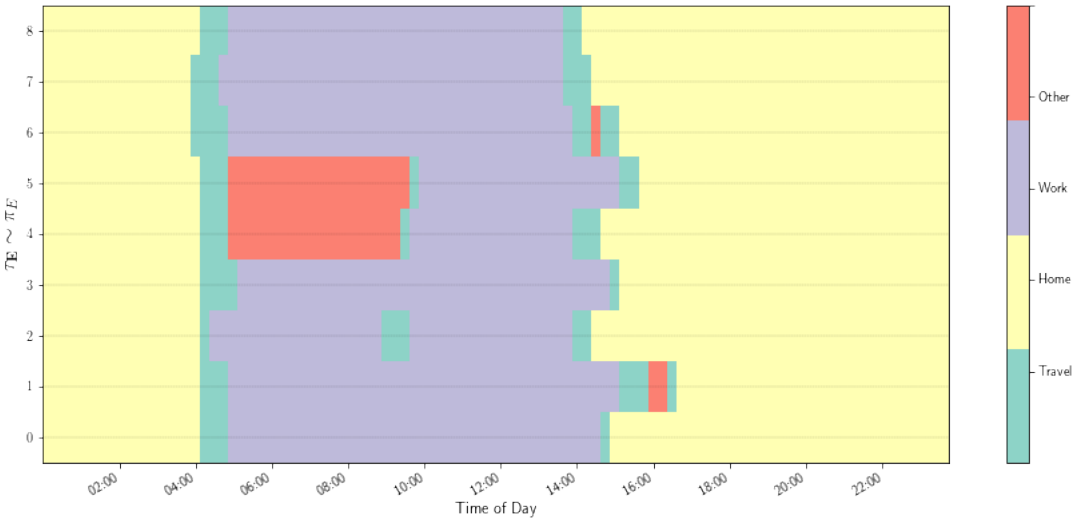


Figure 5.2: Sample expert daily activity-travel patterns (9 trajectories)

the categorical policy distribution[8]. We trained individual trajectories as minibatches using stochastic gradient descent with the Adam optimizer set at a learning rate of 0.01. This seemed to achieved the best results as a pre-training strategy for GAIL training using the same parameters as above (a separate hyperparameter search was conducted at this point; however, the same settings with BC pretraining were optimal as without BC pretraining). Somewhat improved results were achieved through early stopping. We can see from the learning curve of behavioral cloning that convergence to an optimum is achieved fairly quickly Figure 5.6. Based on this curve, we selected a value of 200 as the number of iterations to use for BC pretraining.

Unfortunately, we found that BC pretraining actually encouraged mode collapse in GAIL. As can be seen in Figure 5.9, the simulated state times appear to be more concentrated than with vanilla GAIL, particularly at later timesteps. In addition,duration distributions did not match as well when using pretraining Figure 5.8. We suspect that the small dataset encourages specialization to one type of pattern, making it difficult to retrain parameters that occur later in the day.

Comparison with MaxEnt IRL In a third experiment, we built the MDP over states and actions using efficient, sparse matrix representations. Optimization was performed using the dynamic programming-based algorithm with similar settings to those documented in Chapter 4. Convergence was achieved at around 100 iterations. Optimization took approximately 40 minutes. See Figure 5.10 for a learning curve.

[8]Had the action distribution been continuous, we might have used the more typical mean squared error as the objective.

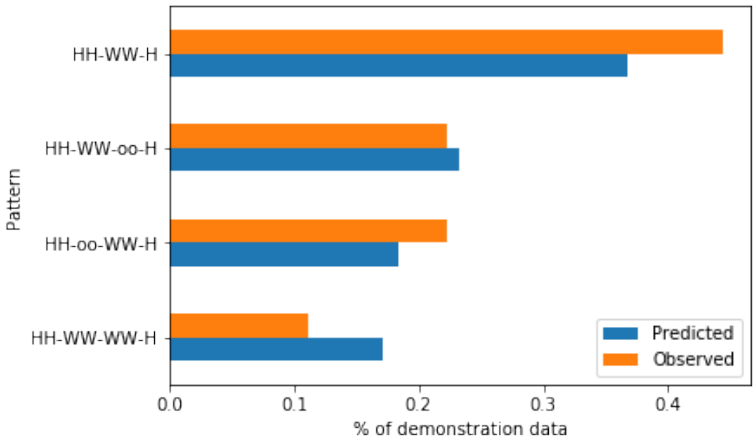


Figure 5.3: Simulated vs. observed activity patterns for a single agent trained using GAIL

Rollouts of the learned policy resulted in, essentially, a uniform distribution over 10-position (maximum possible activity chain) trajectories (see Figure 5.11). The diversity of "correct trajectories" (i.e., those not resulting in termination of the environment prior to reaching the goal state) demonstrates that the chosen feature functions did not adequately constrain the policy. A selection of parameters for the estimated structural equation are provided in Table 5.1. We did not find these values to be readily interpretable. For the current problem, MaxEnt IRL does not provide any benefit in terms of performance or therefore see no additional benefit over GAIL of MaxEnt IRL due to poorer performance.

Multimodal training using trajectory mixtures from many agents In order to evaluate the effectiveness of InfoGAIL in the daily activity pattern setting, we sampled the trajectories of four agents. Training data is illustrated using a state-time diagram as Figure 5.12. We see some diversity of agent behavior, with π_3 spending a relatively long time in the other state on some occasions and π_1 demonstrating short early evening trips between work and home to a distinct other location.

We implemented InfoGAIL with a three hidden layer MLP with a hidden layer architecture of (64|216|512) and leaky-Rectified Linear Unit (ReLU) activations[9] for the policy network. Both the discriminator, D *and* the latent class function, Q were represented as MLPs with two 256 unit hidden layers and leaky ReLU activations. Adam optimization hyperparameters for both latent and discriminator training were identical to those indicated for vanilla GAIL; however, while discriminator training proceeded with an initial learning rate of 0.0001, latent code training was initialized with a learning rate of 0.01. TRPO step size was set at 0.005, or, about half of what it had been for vanilla GAIL. Doing so required

[9]Nonlinear functions that map between units in successive layers.

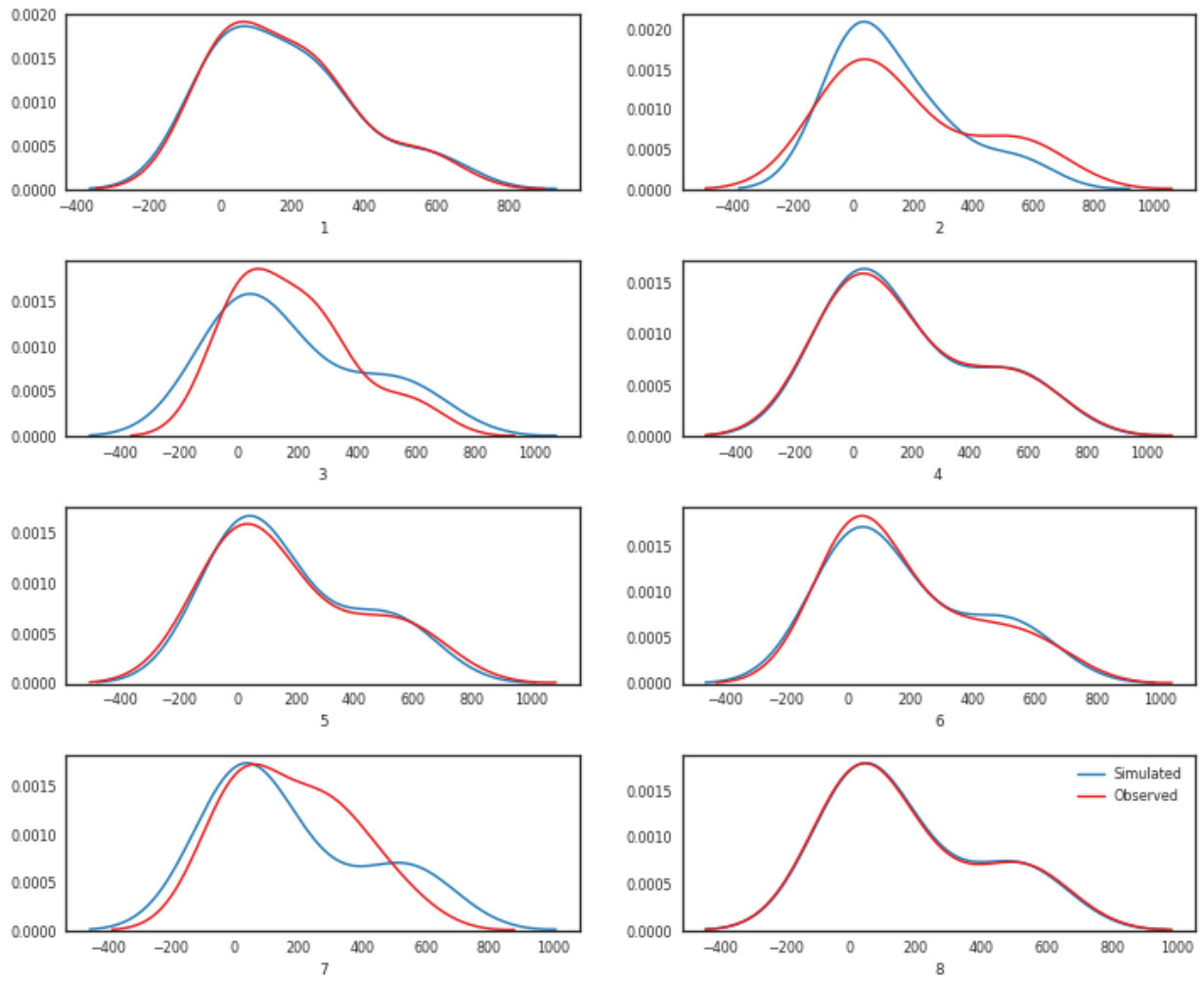


Figure 5.4: Action distribution comparison for GAIL training in the activity planning environment.

that we run training out to 500 iterations, which was about 65% longer than for GAIL. All other parameters were the same as in GAIL. After extensive testing, this combination of hyperparameters seemed to give the best results.

A visualization of the analysis of 200 simulated trajectories is provided as Figures 5.13 to 5.15, showing a comparison of predicted and simulated patterns, state occupancy distribution and action occupancy distributions. InfoGAIL readily recovered the diversity of agent behaviors, in particular, demonstrating a great deal of flexibility in recovering complex duration distributions Figure 5.14.

While vanilla GAIL was also able to recover similar distributions over patterns, actions, and states, InfoGAIL was able to disentangle the components of the mixture over experts. Example state and action components for two different experts (real vs. simulated) are

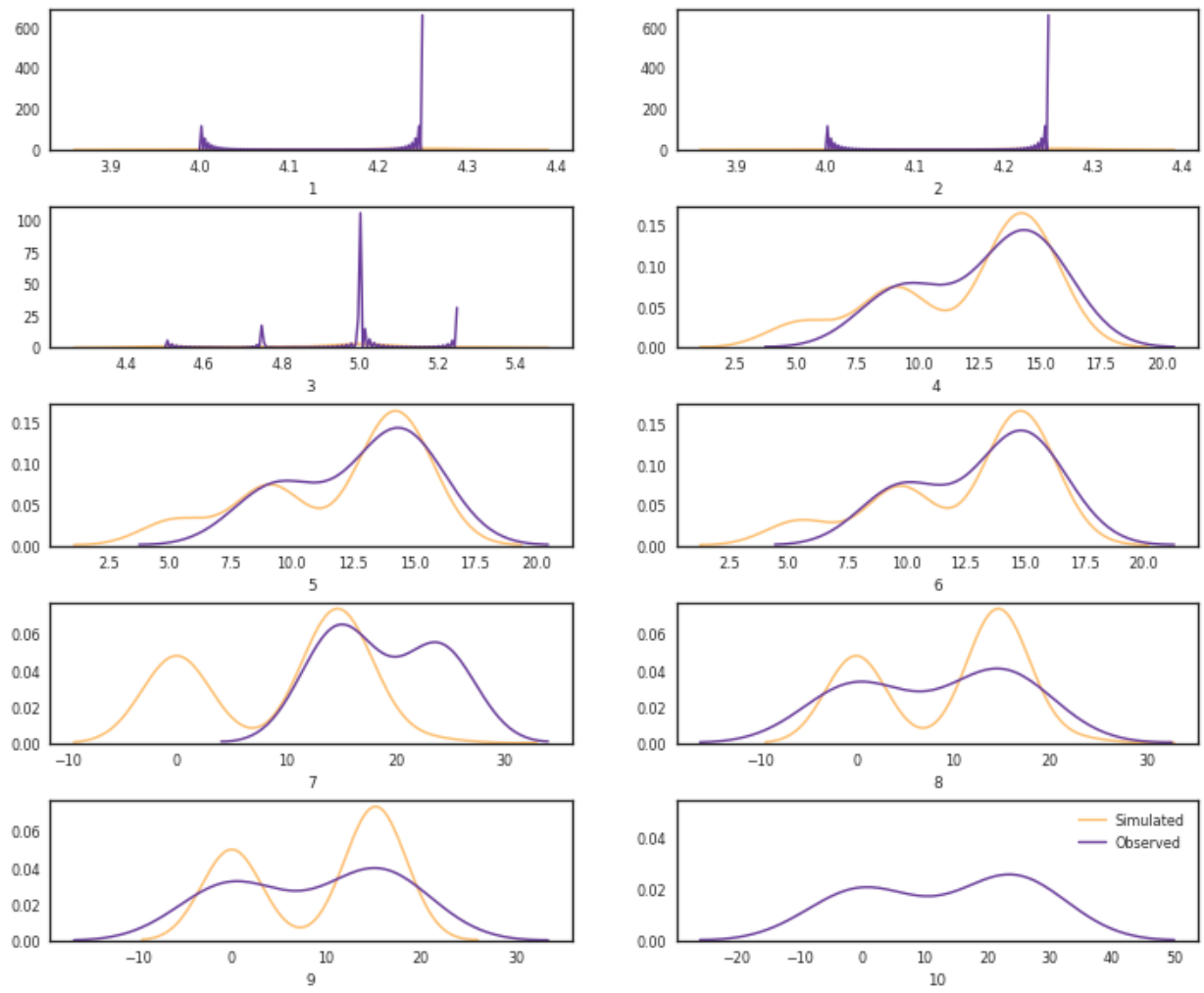


Figure 5.5: State distribution comparison for GAIL training in the activity planning environment.

depicted in Figures 5.16 and 5.17, respectively. The simulated data is the same as that depicted in Figures 5.13 to 5.15. Clearly, it is more straightforward to recover the correct actions as opposed to the states for the different experts. However, the different modes are identified with reasonable accuracy, suggesting that the use of the InfoGAIL could be used to train multiple agents simultaneously. Since the accuracy of multimodal training of agent mixtures is lower than single agent training, with significant failure modes for certain states, in particular, it may be a better option to use InfoGAIL as part of a pre-training routine for many agents with fine-tuning using vanilla GAIL.

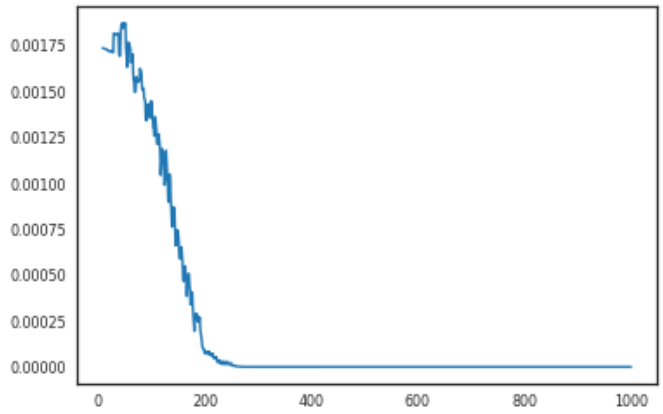


Figure 5.6: Behavioral cloning learning curve smoothed over a window of 10 training epochs.

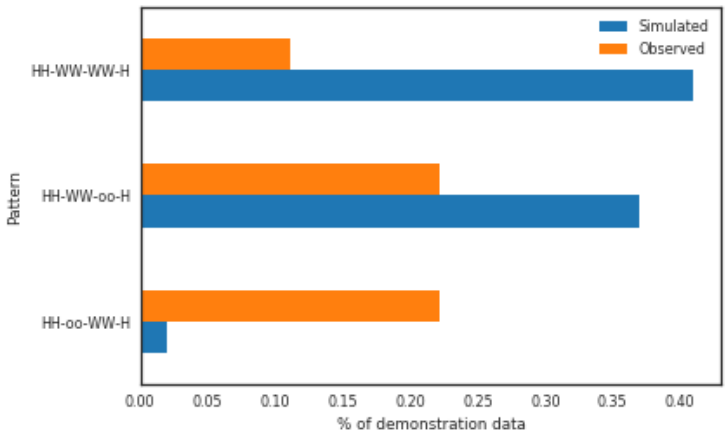


Figure 5.7: Simulated vs. observed activity patterns for a single agent trained using GAIL (with behavioral cloning pretraining).

5.4 Conclusions

In this chapter, we demonstrated the effectiveness of GAIL to imitate potentially long sequences of individual activity patterns in action and state dimensions. While weights learned for the deep neural network models of the reward function of vanilla GAIL were not interpretable, planned work will embed the states and actions in a continuous latent space. Ideally, doing so could permit structural relationships to emerge upon visualizing parameters using dimensionality reduction techniques such as t-SNE (Van Der Maaten and Hinton, 2008).

Our implementation of InfoGAIL in a supervised learning setting permitted disentanglement of a mixture of expert trajectories. While this initial application is promising it would benefit from further experimentation. This initial application of InfoGAIL motivates future work involving unsupervised inference of latent codes. Unsupervised clustering of trajecto-

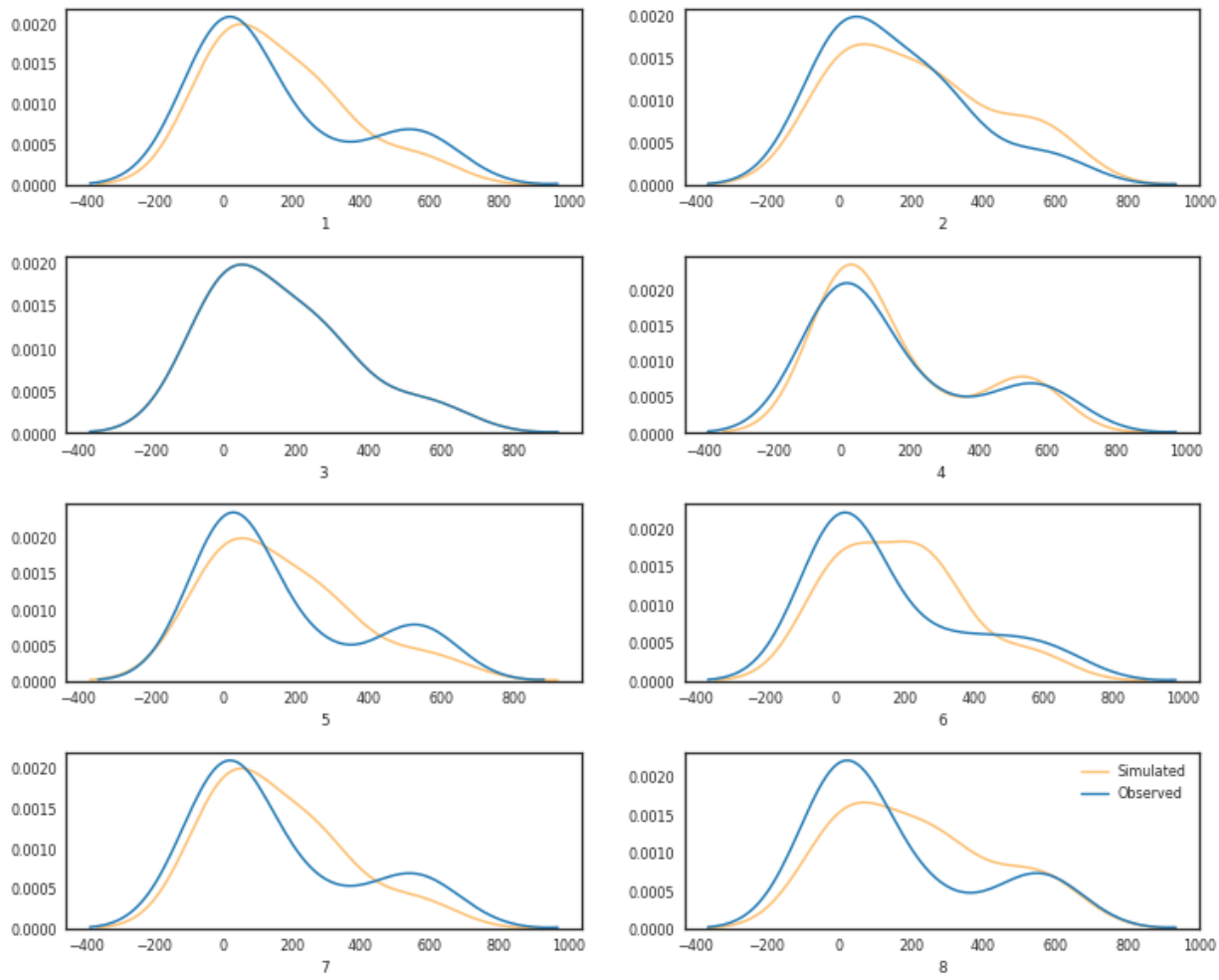


Figure 5.8: Action distribution comparison for GAIL training in the activity planning environment (with behavioral cloning pretraining).

ries could help to segment large populations into similar behavioral classes, permitting more nuanced analysis of duration vis à vis, say, agent sociodemographic attributes. This would represent an important first step towards the use of GAIL and InfoGAIL in policy analysis.

Performance characteristics appear to be a significant limitation of the implementation presented herein. Our attempt to pretrain a model using behavioral cloning was only partially successful: reducing training time, while simultaneously reducing precision. It seems as though faster training may be more likely to induce mode collapse. One possible solution would be to anneal certain regularizing hyperparameters such as the entropy weighting, L2 norm coefficent, as well as the learning rate. In addition, as mentioned earlier, investigating weight dropout may prove helpful. Finally, as was suggested above, simply using InfoGAIL as a warm-start approach or ensemble average reward/policy weights in a similar manner to that described in Chapter 4 may prove to be effective.

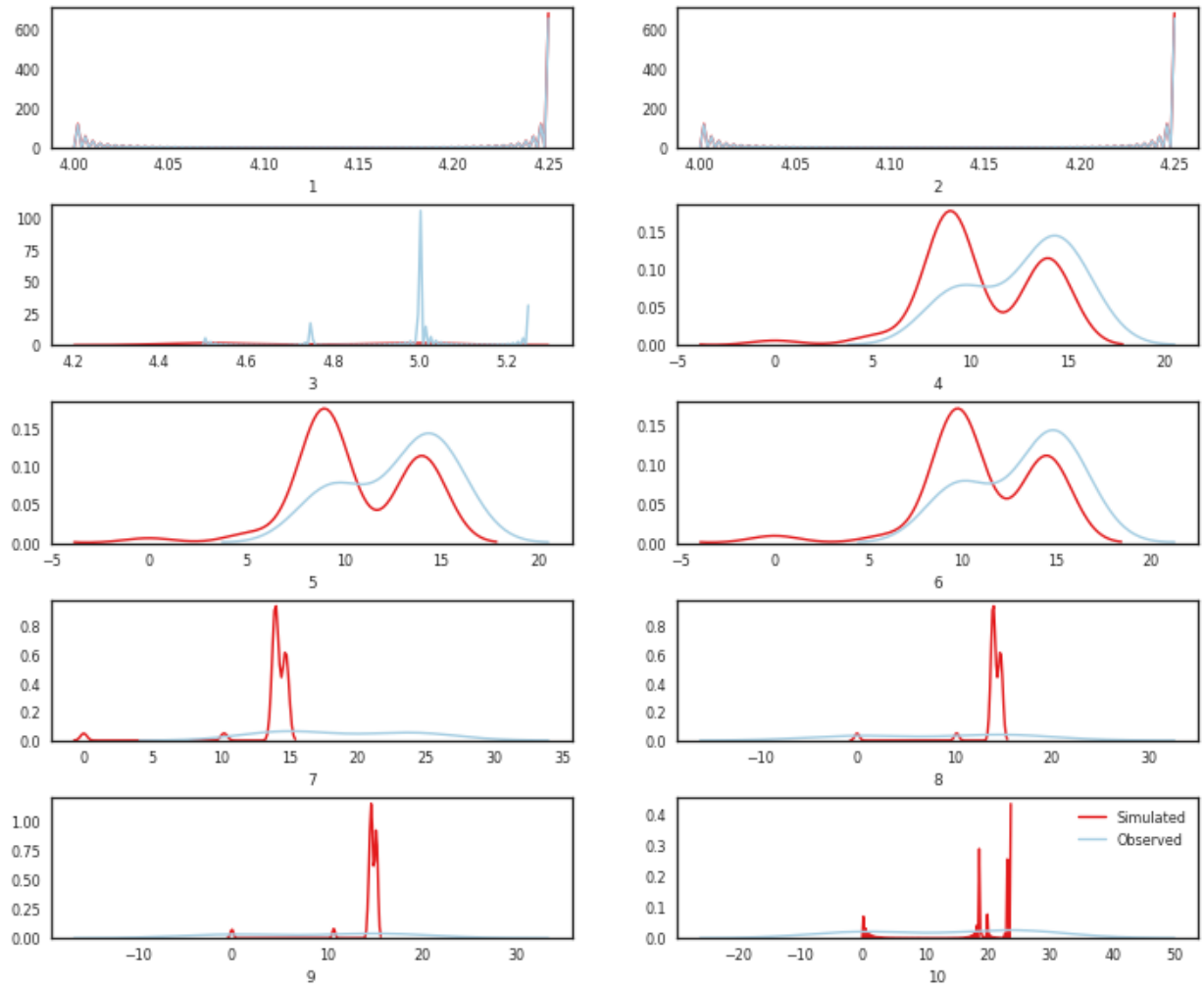


Figure 5.9: State distribution comparison for GAIL training in the activity planning environment (with behavioral cloning pretraining).

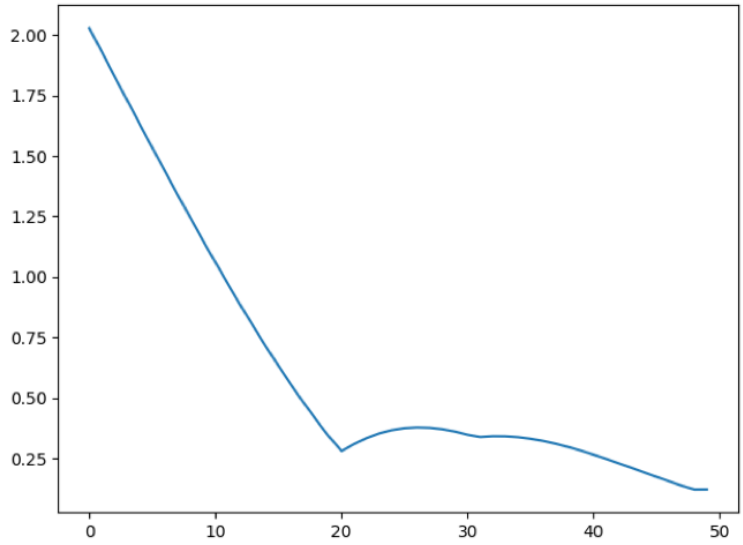


Figure 5.10: Dynamic-programming based MaxEnt IRL learning curve for estimation of durative action structural equation model.

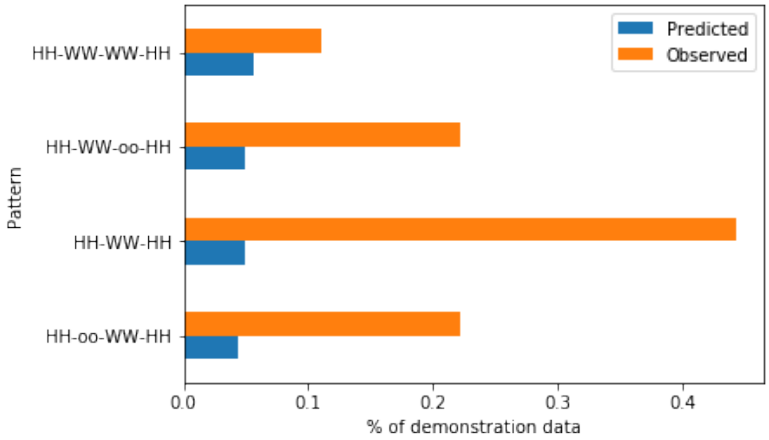


Figure 5.11: Simulated vs. observed activity patterns for a single agent trained using MaxEnt IRL.

Feature Name	Value
- travel time disutility	0.25
S H: Duration feature	-0.73
2 W: Late Arrival feature	0.56
2 W: Early Arrival feature	0.15
2 W: Duration feature	0.24
2 H: Duration feature	0.021
2 o: Duration feature	-0.14
3 W: Late Arrival feature	-0.035
3 W: Duration feature	0.30
3 H: Duration feature	0.11
3 o: Duration feature	-0.0051
4 W: Duration feature	-0.025
4 W: Late Arrival feature	0.15
4 W: Early Arrival feature	-0.19
4 H: Duration feature	-0.064
4 o: Duration feature	-0.16

Table 5.1: Selection of estimated parameter values for structural equation of durative action estimated using MaxEnt IRL

Figure 5.12: Sample expert daily activity-travel patterns for four different agents (62 trajectories).

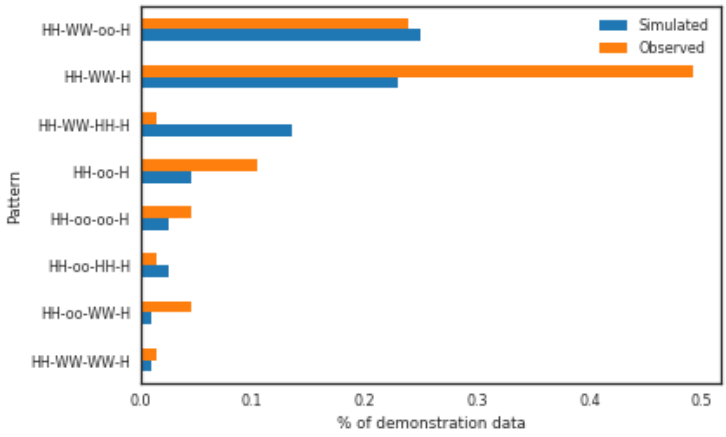


Figure 5.13: Simulated vs. observed activity patterns for multiple (four) agents trained using InfoGAIL

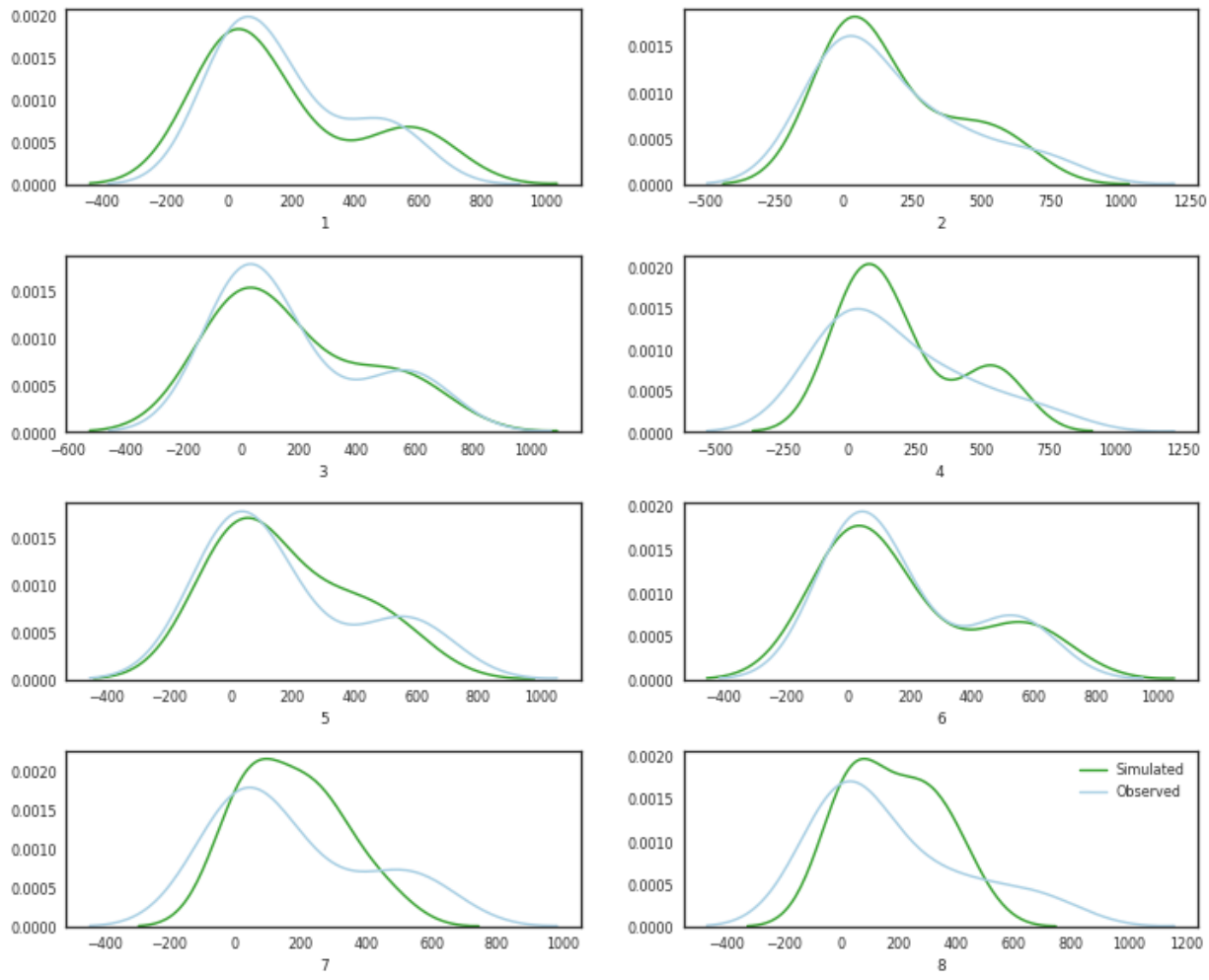


Figure 5.14: Action distribution comparison for InfoGAIL training in the activity planning environment.

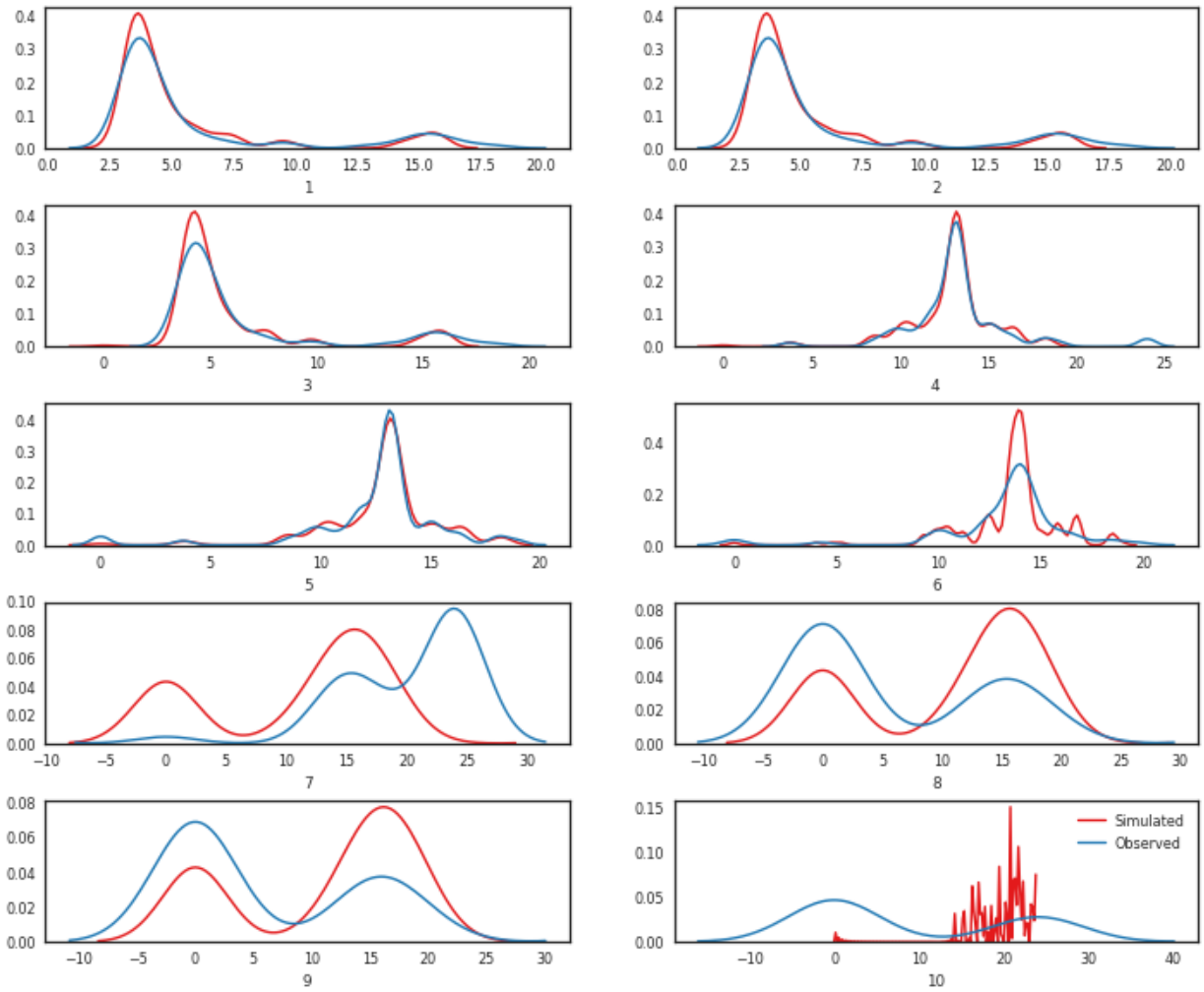
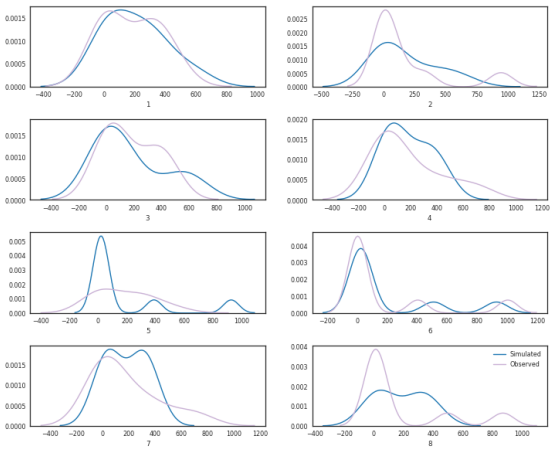
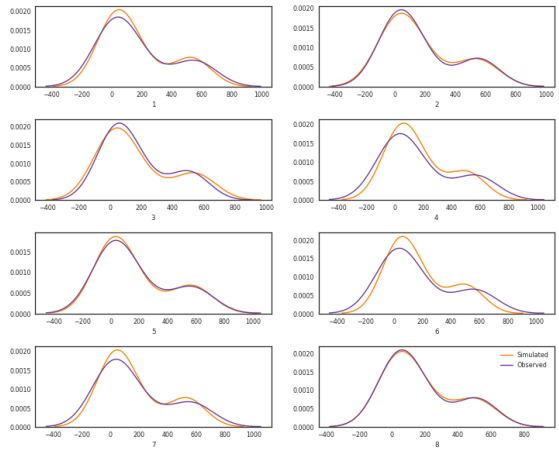


Figure 5.15: State distribution comparison for InfoGAIL training in the activity planning environment.

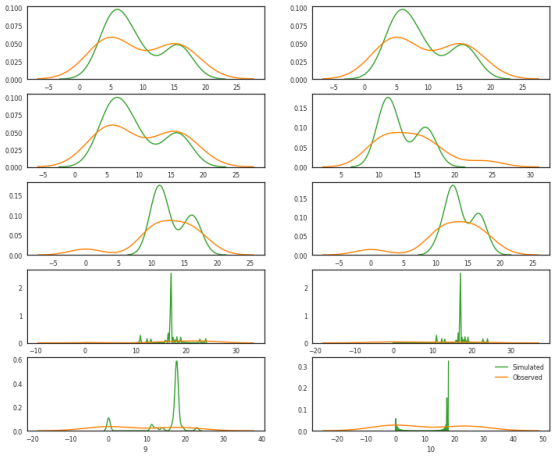


(a) InfoGAIL *action* distribution comparison for expert *2*.

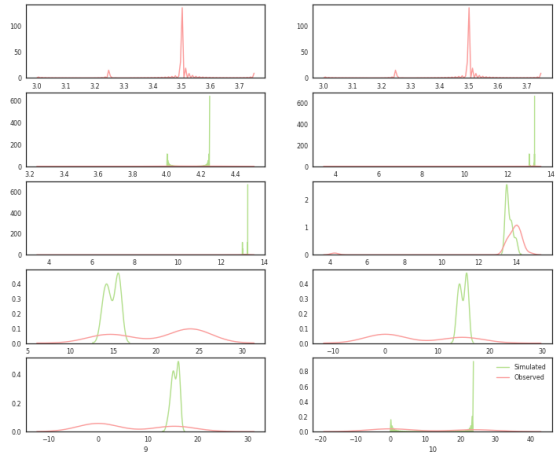


(b) InfoGAIL *action* distribution comparison for expert *3*.

Figure 5.16: Comparison of *action* distributions (simulated vs. observed) for different experts as identified by latent code.



(a) InfoGAIL *state* distribution comparison for expert 2.



(b) InfoGAIL *state* distribution comparison for expert 3.

Figure 5.17: Comparison of *state* distributions (simulated vs. observed) for different experts as identified by latent code.

6

Conclusion

And joy suddenly stirred in his soul, and he even stopped for a moment to catch his breath. The past, he thought, is connected with the present in an unbroken chain of events flowing one out of the other. And it seemed to him that he had just seen both end of that chain: he touched one end, and the other moved.

– Anton Chekhov, *The Student*

6.1 Summary of Contributions

The design of policy instruments intended to achieve socially-optimal utilization of limited-capacity publicly-funded transportation infrastructure and services requires an understanding of the dynamics governing adoption of more sustainable alternatives to driving alone as a primary commute mode.

The effect of peer pressure on socially-cooperative travel decision-making

In Chapter 3, we proposed a novel computational paradigm for the investigation of decentralized sanctioning mechanisms to resolve social dilemmas in the sphere of public utility management and governance. The approach advocated herein makes use of algorithmic tools developed in artificial intelligence and applies them in real-world contexts using very large cellular datasets. Our results show that peer pressure helps in achieving desirable equilibrium properties while reducing congestion and emissions due to sustained mode shift.

Estimating structural models of activity scheduling via IRL

Developing a flexible, time-consistent, and data-driven model of daily human mobility decisions represents a critical next step towards achieving emerging, incentive compatible travel

demand management strategies such as dynamic road pricing. Given demonstrated travel between sequential activities and activity attributes, methodology was developed to infer the weights governing utility functions encoding preferences over alternative activity schedules. Our MaxEnt IRL-based framework was presented in Chapter 4. We described the results of structural estimation of models of dynamic replanning behavior and demonstrated the interpretability of the parameters that resulted from this effort.

Finally, in order to improve the realism of the planning environment implementation demonstrated in Chapter 4, we add activity duration as a choice dimension. We hypothesized that an alternative framework to MaxEnt IRL would help to mitigate performance losses during training due to the dimensional curse imposed by the resultant expansion of the action space. Turning to model-free methods, we used the GAIL algorithm to imitate agent behavior from demonstrations, and confirmed that it performed better than MaxEnt IRL. On the other hand, the deep neural network representation of the surrogate reward function in GAIL made interpretation of estimated weights impossible. We leave the incorporation of more interpretable methods based on the imitation-learning algorithms explored in Chapter 5 for future work.

To summarize, the major contributions of this thesis were:

1. A game-based model of interdependent commuter decision-making to demonstrate the emergence of socially-cooperative travel mode choice behavior in the context of a hyper-realistic agent-based microsimulation;

2. A computationally-efficient and scalable framework to learn sequential daily travel decisions from digital human mobility traces; and

3. An adversarial imitation-learning approach to estimate a model of daily travel planning behavior that is able to replicate both the duration and identity of demonstrated activity sequences with high fidelity.

6.2 Future Research

Human decision-making in settings where cooperation or competition are possible outcomes do not easily reduce to single step or two step games. While our work supports findings that evolutionary learning techniques may lead to the emergence of cooperation, developing methodology to simulate more realistic decision-making to support policy analysis is the ultimate goal of this work. Towards this end, we propose exploring methods that learn strategic cooperative behavior in societies via IRL as an end-to-end process.

Our approach combines the Bayesian theory-of-mind model proposed by Kleiman-Weiner and Tenenbaum (2016) and norm learning via IRL Ho and Ermon (2016). The fusion of these two computational frameworks in iterative game-based travel demand model settings would endow agents with the ability to learn social norms through repeated interactions as well as develop high-level strategies on when and with whom to cooperate.

In applying this methodology, we would need to reformulate our game-based model of peer pressure as a *stochastic game*. Also known as Markov games, stochastic games extend MDPs to strategic, interdependent choice situations. The concept corresponds to a hybrid of dynamic, repeated games and MDPs. We would formulate the multi-agent peer pressure game as a constant-sum stochastic game, wherein agents $i \in \mathcal{N}$ take *joint actions*, $j \in \times_i A^{(i)}$, where $\times_i A^{(i)}$ is the action profile for all agents (Littman, 1994; Shapley, 1953).

Following (Kleiman-Weiner and Tenenbaum, 2016), agents' high-level strategies would take the form of distinct "modes", while low-level parameters guiding structural models of decision-making would be learned through IRL. Agent strategic modes correspond to interpretable behavioral patterns similar to those used in Axelrod and Hamilton (1981) and subsequent IPD tournaments (e.g., "Tit-for-Tat", "Promise", "Grudger", "Win-Stay-Lose-Shift", etc.) (Nowak and Sigmund, 1993; Stewart and Plotkin, 2012).

Social preferences encoding group norms that underlie agents' type assignments would be learned using the approach provided in (Ho and Ermon, 2016). In the feature-based reward function used in (Ho and Ermon, 2016), a norm-based reward is shared among a set of interacting agents. An i^{th} agent's (of a set of $\mathcal{N}$ agents) total parametric reward $U^{(i)}_{\boldsymbol{\theta}}(s)$ is then a combination of the agent's individual utility function and a group norm:

$$U^{(i)}_{\boldsymbol{\theta}}(s) = U^{(i)}_{\boldsymbol{\theta},\text{individual}}(s) + U_{\boldsymbol{\theta},\text{norm}}(s).$$

Using *group IRL*, a *norm-learning agent* estimates the *norm reward function*, $\hat{U}_{\boldsymbol{\theta},\text{norm}}$, based on a history of group interaction, $\mathcal{H} = ((s_0, j_0, s_1), \ldots, (s_{T-1}, j_{T-1}, s_T))$:

$$\hat{U}_{\boldsymbol{\theta},\text{norm}} = \underset{U_{\boldsymbol{\theta},\text{norm}}}{\text{argmax}}\ P(U_{\boldsymbol{\theta},\text{norm}} \mid \mathcal{H}).$$

Features for learning norms represent shared abstract descriptions about the state and/or actions of other agents. An example in the peer pressure game could include spatial distances between home, work, and other activity locations between agents in a group. The size of groups will play a key role in the scalability of our approach. We will explore how different levels of hierarchy and reference groups (e.g., community, TAZ, proximity) affect parameter estimation convergence rates.

6.3 Reflection And Perspective

Clearly, maintaining data-driven intelligence as a sufficient lens through which to gaze upon objective reality risks ceding political accountability to algorithmic command and control (Kitchin, 2014; Söderström et al., 2014). There are constraints on the social contexts that even privacy-aware knowledge-discovery systems are able to infer. The notion that passively-sensed data or heterogeneously-implemented participatory ICT efforts will be sufficient to answer questions of individual and community values seems an inherently flawed objective. Even in the limit of perfect information gleaned from all available digital exhaust, the danger

exists of underrepresented classes that either cannot or choose not to contribute their data towards optimizing intelligent infrastructure. Yet, social welfare is not only maximized over those citizens whose coordinates are captured by smartphones or social networks. Novel governance structures will be necessary in order to maintain adequate firewalls and safety mechanisms to encourage algorithmic fairness and accountability.

This thesis explores the potential for simulations of social dilemmas to play a role in shaping policy to inform cooperative use of limited-capacity transportation infrastructure. Over-reliance on implications from simulation-based policy analysis is certainly not without its risks, and, in order to address emerging issues such as value alignment and safety, we strongly believe that simulated hypothetical choice situations must also be validated through experimental studies of human decision-making. As we continue to learn more about the possibilities and limits of fairness, accountability, and transparency in algorithmic administration of public services, it will be vital for cities that rely on ML/AI systems to keep "humans in the loop": promoting resource-efficient economic progress while aligning government and citizen values. We anticipate that further interdisciplinary inquiry into the design of computational frameworks to plan and manage smart city infrastructure will help societies engage in scalable, equitable, and ultimately democratic conversations about the challenging technical and moral questions raised by regulatory automation.

Bibliography

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. *OSDI*, vol. 16. 265–283.

Abbeel, Pieter, Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. *Proceedings of the twenty-first international conference on Machine learning*. ACM, 1.

Abdel-Aty, Mohamed A, Ryuichi Kitamura, Paul P Jovanis. 1997. Using stated preference data for studying the effect of advanced traffic information on drivers' route choice. *Transportation Research Part C: Emerging Technologies* **5**(1) 39–50.

Abou-Zeid, M, J-D Schmöcker, P F Belgiawan, S Fujii. 2013. Mass effects and mobility decisions. *Transportation Letters* **5**(3) 115–130. doi:10.1179/1942786713Z.00000000011. URL `http://www.maneyonline.com/doi/abs/10.1179/1942786713Z.00000000011`.

Abou-Zeid, Maya, Moshe Ben-Akiva. 2011. The effect of social comparisons on commute well-being. *Transportation Research Part A: Policy and Practice* **45**(4) 345–361.

Achiam, Joshua, David Held, Aviv Tamar, Pieter Abbeel. 2017. Constrained Policy Optimization URL `http://arxiv.org/abs/1705.10528`.

Agarwal, Amit, Benjamin Kickhöfer. 2015a. Agent-Based Simultaneous Optimization of Congestion and Air Pollution : A Real-World Case Study **00**.

Agarwal, Amit, Benjamin Kickhöfer. 2015b. The Internalization of Congestion and Air Pollution Externalities : Evaluating Behavioral Impacts .

Arentze, Theo, Harry Timmermans, Frank Hofman, Nelly Kalfs. 2000. Data needs, data collection, and data quality requirements of activity-based transport demand models. *Transportation research circular* (E-C008) 30–p.

Arjovsky, Martin, Léon Bottou. 2017. Towards Principled Methods for Training Generative Adversarial Networks 1–17doi:10.2507/daaam.scibook.2010.27. URL `http://arxiv.org/abs/1701.04862`.

Arjovsky, Martin, Soumith Chintala, Léon Bottou. 2017. Wasserstein GAN URL http://arxiv.org/abs/1701.07875.

Arnott, Richard, Andre de Palma, Robin Lindsey. 1991. Does providing information to drivers reduce traffic congestion? *Transportation Research Part A: General* **25**(5) 309–318. doi:10.1016/0191-2607(91)90146-H. URL https://www.sciencedirect.com/science/article/pii/019126079190146H?via{%}3Dihub.

Athey, Susan. 2017. Beyond prediction: Using big data for policy problems. *Science* **355**(6324) 483–485.

Avineri, E. 2012. On the use and potential of behavioural economics from the perspective of transport and climate change **24** 512–521. URL http://eprints.uwe.ac.uk/16602/.

Axelrod, Robert, D Hamilton. 1981. The evolution of cooperation. *SCIENCE* **21**(1) 1390.

Axhausen, Kay W. 2007. Activity Spaces, Biographies, Social Networks and their Welfare Gains and Externalities: Some Hypotheses and Empirical Results. *Mobilities* **2**(1) 15–36. doi:10.1080/17450100601106203.

Bagnell, J Andrew, Stéphane Ross. 2010. Efficient Reductions for Imitation Learning. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010* **9** 661–668.

Bamberg, Sebastian, Peter Schmidt. 2003. Incentives, morality, or habit? predicting students car use for university routes with the models of ajzen, schwartz, and triandis. *Environment and behavior* **35**(2) 264–285.

Bass, Frank M. 1969. A new product growth for model consumer durables. *Management science* **15**(5) 215–227.

Batty, Michael. 2007. *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. The MIT Press.

Batty, Michael, Kay W Axhausen, Fosca Giannotti, Alexei Pozdnoukhov, Armando Bazzani, Monica Wachowicz, Georgios Ouzounis, Yuval Portugali. 2012. Smart cities of the future. *The European Physical Journal Special Topics* **214**(1) 481–518.

Bellman, Richard. 1954. The theory of dynamic programming. *Bulletin of the American Mathematical Society* **60**(6) 503–515.

Ben-Akiva, Moshe E, Steven R Lerman. 1985. *Discrete choice analysis: theory and application to travel demand*, vol. 9. MIT press.

Ben-Elia, Eran, Erel Avineri. 2015. Response to Travel Information: A Behavioural Review. *Transport Reviews* **35**(3) 352–377. doi:10.1080/01441647.2015.1015471. URL http://dx.doi.org/10.1080/01441647.2015.1015471.

Ben-Elia, Eran, Roberta Di Pace, Gennaro N Bifulco, Yoram Shiftan. 2013. The impact of travel information ' s accuracy on route-choice. *Transportation Research Part C* **26** 146–159. doi:10.1016/j.trc.2012.07.001. URL http://dx.doi.org/10.1016/j.trc.2012.07.001.

Biel, Anders, John Thøgersen. 2007. Activation of social norms in social dilemmas: A review of the evidence and reflections on the implications for environmental behaviour. *J. Econ. Psychol.* **28**(1) 93–112. doi:10.1016/j.joep.2006.03.003. URL http://linkinghub.elsevier.com/retrieve/pii/S0167487006000250.

Bierlaire, Michel. 2016. PythonBiogeme : a short introduction .

Bowman, John L. 1998. The day activity schedule approach to travel demand analysis. *Metro* 185URL http://www.jbowman.net/theses/1998.Bowman.Thesis.pdf.

Brathwaite, Timothy, Joan Walker. 2016. Asymmetric, Closed-Form, Finite-Parameter Models of Multinomial Choice URL http://arxiv.org/abs/1606.05900.

California Air Resources Board. 2014. California Emission Inventory Data.

Calvó-Armengol, Antoni, Matthew O. Jackson. 2010. Peer pressure. *Journal of the European Economic Association* **8**(1) 62–89. doi:10.1111/j.1542-4774.2010.tb00495.x. URL http://dx.doi.org/10.1111/j.1542-4774.2010.tb00495.x.

Camerer, Colin, Samuel Issacharoff, George Lowenstein, Ted O'Donoghue, Matthew Rabin. 2003. Regulation for Conservatives: Behavioral Economics and the Case for "Assymetric Paternalism". *University of Pennsylvania law review* **151**(3) 1211–1254.

Camerer, Colin F, Ernst Fehr. 2004. Measuring social norms and preferences using experimental games: A guide for social scientists. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies* **97** 55–95.

Capraro, Valerio. 2013. A model of human cooperation in social dilemmas. *PLoS One* **8**(8) e72427.

Carrasco, Juan Antonio, Bernie Hogan, Barry Wellman, Eric J. Miller. 2008. Collecting social network data to study social activity-travel behavior: An egocentric approach. *Environ. Plan. B Plan. Des.* **35**(6) 961–980. doi:10.1068/b3317t.

Carrel, Andre, Anne Halvorsen, Joan Walker. 2013. Passengers' perception of and behavioral adaptation to unreliability in public transportation. *Transportation Research Record: Journal of the Transportation Research Board* (2351) 153–162.

Cassandras, Christos G. 2016. Smart Cities as Cyber-Physical Social Systems. *Engineering* **2**(2) 156–158. doi:http://dx.doi.org/10.1016/J.ENG.2016.02.012. URL http://www.sciencedirect.com/science/article/pii/S2095809916309420.

Castiglione, Joe, Mark Bradley, John Gliebe. 2014. *Activity-Based Travel Demand Models*.

Chapin, F.S. Jr. 1974. *Human Activity Patterns in the City: Things People Do in Time and Space.*. John Wiley and Sons, London.

Charypar, D, K Nagel. 2005a. Q -Learning for Flexible Learning of Daily Activity Plans. *Transportation Research Record: Journal of the Transportation Research Board* 163–169doi:10.3141/1935-19.

Charypar, David, Kai Nagel. 2005b. Q-learning for flexible learning of daily activity plans. *Transportation Research Record: Journal of the Transportation Research Board* (1935) 163–169.

Chaudhuri, Ananish. 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* **14**(1) 47–83.

Chen, Cynthia, Jingtao Ma, Yusak Susilo, Yu Liu, Menglin Wang. 2016a. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies* **68** 285–299.

Chen, Cynthia, Jingtao Ma, Yusak Susilo, Yu Liu, Menglin Wang. 2016b. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies* **68** 285–299. doi:10.1016/j.trc.2016.04.005.

Chen, Jingmin. 2013. Modeling Route Choice Behavior Using Smartphone Data **5649**.

Cirillo, Cinzia, Renting Xu. 2011. Dynamic Discrete Choice Models for Transportation. *Transport Reviews* **31**(4) 473–494. doi:10.1080/01441647.2010.533393. URL `10.1080/01441647.2010.533393{%}5Cnhttp://ezproxy.lib.utexas.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true{&}db=a9h{&}AN=62667369{&}site=ehost-live`.

Çolak, Serdar, Antonio Lima, Marta C González. 2016. Understanding congested travel in urban areas. *Nature communications* **7** 10793. doi:10.1038/ncomms10793. URL `http://www.nature.com/ncomms/2016/160315/ncomms10793/full/ncomms10793.html`.

Croissant, Yves, et al. 2012. Estimation of multinomial logit models in r: The mlogit packages. *R package version 0.2-2. URL: http://cran. r-project. org/web/packages/mlogit/vignettes/mlogit. pdf* .

Davy, Janssens. 2006. Allocating Time and Location Information to Activity-Travel Patterns through Reinforcement Learning (August) 16–20.

Dawes, Robyn M. 1980. Social Dilemmas. *Annual review of psychology* **31**(2) 169–193. doi:10.1146/annurev.ps.31.020180.001125.

Dhami, Sanjit. 2016. The Foundations of Behavioral Economic Analysis.

Diekert, Florian K. 2012. The tragedy of the commons from a game-theoretic perspective. *Sustainability* **4**(8) 1776–1786. doi:10.3390/su4081776.

Duan, Yan, Xi Chen, Rein Houthooft, John Schulman, Pieter Abbeel. 2016. Benchmarking Deep Reinforcement Learning for Continuous Control **48**. doi:10.1109/CVPR.2014.180. URL `http://arxiv.org/abs/1604.06778`.

Dubernet, Thibaut, Kay W Axhausen. 2013. A Framework to Represent Joint Decisions in a Multi-Agent Transport Simulation (April).

Dugundji, Elenna, Joan Walker. 2005. Discrete Choice with Social and Spatial Network Interdependencies: An Empirical Example Using Mixed Generalized Extreme Value Models with Field and Panel Effects. *Transportation Research Record* **1921**(1) 70–78. doi:10.3141/1921-09.

Dulac-Arnold, Gabriel, Richard Evans, Hado van Hasselt, Peter Sunehag, Timothy Lillicrap, Jonathan Hunt, Timothy Mann, Theophane Weber, Thomas Degris, Ben Coppin. 2015. Deep Reinforcement Learning in Large Discrete Action Spaces URL `http://arxiv.org/abs/1512.07679`.

Eagle, Nathan, Alex Sandy Pentland. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology* **63**(7) 1057–1066.

El Zarwi, Feras, Akshay Vij, Joan L Walker. 2017. A discrete choice framework for modeling and forecasting the adoption and diffusion of new transportation services. *Transportation Research Part C: Emerging Technologies* **79** 207–223.

Erev, Ido, Greg Barron. 2005. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review* **112**(4) 912.

Eriksson, Louise, Jörgen Garvill, Annika M. Nordlund. 2006. Acceptability of travel demand management measures: The importance of problem awareness, personal norm, freedom, and fairness. *Journal of Environmental Psychology* **26**(1) 15 – 26. doi:https://doi.org/10.1016/j.jenvp.2006.05.003. URL `http://www.sciencedirect.com/science/article/pii/S0272494406000260`.

Ermon, Stefano, Yexiang Xue, Russell Toth, Bistra Dilkina, Richard Bernstein, Theodoros Damoulas, Patrick Clark, Steve DeGloria, Andrew Mude, Christopher Barrett, Carla P. Gomes. 2015. Learning Large Scale Dynamic Discrete Choice Models of Spatio-Temporal Preferences with Application to Migratory Pastoralism in East Africa. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence Pattern* 644–650.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, vol. 96. 226–231.

Fehr, Ernst, KM Schmidt. 1999. A Theory of Fairness,Competition and Cooperation. *Quarterly journal of Economics* **114**(August) 817–868. doi:10.1162/003355399556151. URL `http://www.jstor.org/stable/2586885`.

Feygin, S., Pozdnukhov, A. 2017. Peer pressure enables actuation of mobility lifestyles **(In review)**.

Finn, Chelsea, Paul Christiano, Pieter Abbeel, Sergey Levine. 2016a. A Connection between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models URL `http://arxiv.org/abs/1611.03852`.

Finn, Chelsea, Paul Christiano, Pieter Abbeel, Sergey Levine. 2016b. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852* .

Finn, Chelsea, Sergey Levine, Pieter Abbeel. 2016c. Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization. *arXiv* **48**(2000). URL `http://arxiv.org/abs/1603.00448`.

Flötteröd, Gunnar, Benjamin Kickhöfer. 2016. Choice Models in MATSim. K Horni, A, Nagel, Axhausen, K W., eds., *The Multi-Agent Transport Simulation MATSim*, 1st ed., chap. 49. Ubiquity Press, London, 337–346. doi:http://dx.doi.org/10.5334/baw.49.

Fogg, B. J. 2002. Persuasive technology: Using computers to change what we think and do. *Ubiquity* **2002**(December). doi:10.1145/764008.763957. URL `http://doi.acm.org/10.1145/764008.763957`.

Fox, Roy, Sanjay Krishnan, Ion Stoica, Ken Goldberg. 2017. Multi-level discovery of deep options. *arXiv preprint arXiv:1703.08294* .

Fredrickson, B L, D Kahneman. 1993. Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology* **65**(1) 45–55. doi:10.1037/0022-3514.65.1.45.

Fu, Justin, Luo Katie, Levine Sergey. 2018. Learning Robust Rewards with Adversarial Inverse Reinforcement Learning 1–14URL `https://openreview.net/pdf?id=rkHywl-A-`.

Gaker, David, David Vautin, Akshay Vij, Joan L Walker. 2011. The power and value of green in promoting sustainable transport behavior. *Environ. Res. Lett.* **6**(3) 034010. doi:10.1088/1748-9326/6/3/034010.

Gaker, David, Yanding Zheng, Joan Walker. 2010. Experimental economics in transportation: focus on social influences and provision of information. *Transportation Research Record: Journal of the Transportation Research Board* (2156) 47–55.

Glaeser, Edward L and Kominers, Scott Duke and Luca, Michael and Naik, Nikhil. 2018. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry* **56**(1) 114–137.

González, M C, C a Hidalgo. 2008. Understanding individual human mobility patterns. *Nature* **453**(June) 1–12. doi:10.1038/nature06958. URL `http://www.nature.com/nature/journal/v453/n7196/abs/nature06958.html{%}5Cnpapers2://publication/uuid/B606D8A9-1F3D-4471-8DD6-3EEDAB194751`.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, Ij, J Pouget-Abadie, Mehdi Mirza. 2014. Generative Adversarial Networks. *arXiv preprint arXiv: ...* 1–9URL `http://arxiv.org/abs/1406.2661`.

Grabowicz, Przemyslaw A, José J Ramasco, Bruno Gonçalves, Víctor M Eguíluz. 2014. Entangling mobility and interactions in social media. *PLoS One* **9**(3) e92196. doi:10.1371/journal.pone.0092196. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3961345&tool=pmcentrez&rendertype=abstract`.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, Aaron Courville. 2017. Improved Training of Wasserstein GANs doi:10.1016/j.aqpro.2013.07.003. URL `http://arxiv.org/abs/1704.00028`.

Hackney, Jeremy, Kay W Axhausen. 2006. *An agent model of social network and travel behavior interdependence interdependence*. August.

Hägerstrand, Torsten. 1970. What about people in Regional Science? *Papers of the Regional Science Association* **24**(1) 6–21. doi:10.1007/BF01936872. URL `http://dx.doi.org/10.1007/BF01936872`.

Hardin, Garrett. 1968. The tragedy of the commons. *science* **162**(3859) 1243–1248.

Hårsman, Björn, John M Quigley. 2010. Political and public acceptability of congestion pricing: Ideology and self-interest. *Journal of Policy Analysis and Management* **29**(4) 854–874.

Hausknecht, Matthew John, Peter Stone, Dana Ballard. 2016. Cooperation and Communication in Multiagent Deep Reinforcement Learning .

Hausman, Karol, Yevgen Chebotar, Stefan Schaal, Gaurav Sukhatme, Joseph Lim. 2017. Multi-Modal Imitation Learning from Unstructured Demonstrations using Generative Adversarial Nets URL `http://arxiv.org/abs/1705.10479`.

Heppenstall, Alison, Nick Malleson, Andrew Crooks. 2016. "Space, the Final Frontier": How Good are Agent-Based Models at Simulating Individuals and Space in Cities? *Systems* **4**(1) 9. doi:10.3390/systems4010009. URL `http://www.mdpi.com/2079-8954/4/1/9`.

Hester, Todd, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Gabriel Dulac-Arnold, Ian Osband, John Agapiou, Joel Z. Leibo, Audrunas Gruslys. 2017. Deep Q-learning from Demonstrations URL `http://arxiv.org/abs/1704.03732`.

Hicks, John R. 1939. The foundations of welfare economics. *The Economic Journal* **49**(196) 696–712.

Hidalgo, Cesar a., C. Rodriguez-Sickert. 2008. The dynamics of a mobile phone network. *Phys. A Stat. Mech. its Appl.* **387**(12) 3017–3024. doi:10.1016/j.physa.2008.01.073.

Ho, Jonathan, Stefano Ermon. 2016. Generative Adversarial Imitation Learning. *Nips* (Nips) 4565–4573. doi:10.1016/j.compeleceng.2013.11.024. URL `https://papers.nips.cc/paper/6391-generative-adversarial-imitation-learning`.

Holmes, Thomas J, Holger Sieg. 2014. Structural estimation in urban economics. *Handbook of Regional and Urban Economics* 1–68URL `http://real.wharton.upenn.edu/{~}duranton/Duranton{_}Papers/Handbook/Structural{_}estimation{_}in{_}urban{_}economics.pdf`.

Horni, Andreas, Kai Nagel, Kay W Axhausen. 2016a. The multi-agent transport simulation matsim. *Ubiquity, London* **9**.

Horni, Andreas, Kai Nagel, Kay W Axhausen. 2016b. The Multi-Agent Transport Simulation MATSim .

IAWG. 2015. Technical Support Document: Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis Under Executive Order 12866 - July 2015 Revision (May 2013) 1–21.

Illenberger, Johannes. 2012. Social Networks and Cooperative Travel Behaviour 190.

Jacobs, Jane. 1961. *The death and life of American cities*.

Jariyasunant, Jerald, Maya Abou-Zeid, Andre Carrel, Venkatesan Ekambaram, David Gaker, Raja Sengupta, Joan L Walker. 2015. Quantified traveler: Travel feedback meets the cloud to change behavior. *Journal of Intelligent Transportation Systems* **19**(2) 109–124.

Jaynes, Edwin T. 1955. Information Theory and Statistical Mechanics. doi:10.1103/PhysRev.108.171.

Jonsson, R D, Anders Karlström. 2005. SCAPES a dynamic microeconomic model of activity scheduling. *European Transport Conference* (March 2015). URL http://www.etcproceedings.org/paper/scapes-a-dynamic-microeconomic-model-of-activity-scheduling.

Kaddoura, Ihab, Lars Kroeger. 2015. An activity-based and dynamic approach to calculate road traffic noise damages 1–24.

Kaddoura, Ihab, Benjamin Universit, Kickhöfer, Technische Universit. 2014. Optimal Road Pricing : Towards an Agent-based Marginal Social Cost Approach Optimal Road Pricing : Towards an Agent-based Marginal Social Cost Approach (JANUARY).

Kaelbling, Leslie Pack, Michael L. Littman, Anthony R. Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence* **101**(1-2) 99–134. doi:10.1016/S0004-3702(98)00023-X. URL http://linkinghub.elsevier.com/retrieve/pii/S000437029800023X.

Kahneman, Daniel, Amos Tversky. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* **47**(March) 263–291.

Kakade, Sham Machandranath. 2001. A Natural Policy Gradient. *Advances in neural information processing systems* 1531–1538doi:10.1.1.19.8165. URL http://www-2.cs.cmu.edu/Groups/NIPS/NIPS2001/papers/psgz/CN11.ps.gz.

Kaldor, Nicholas. 1939. Welfare propositions of economics and interpersonal comparisons of utility. *The Economic Journal* **49**(195) 549–552. URL http://www.jstor.org/stable/2224835.

Kamakura, Wagner A, Gary J Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *Journal of marketing research* 379–390.

Kamas, Linda, Anne Preston. 2012. Distributive and reciprocal fairness: What can we learn from the heterogeneity of social preferences? *Journal of Economic Psychology* **33**(3) 538–553. doi:10.1016/j.joep.2011.12.003. URL http://dx.doi.org/10.1016/j.joep.2011.12.003.

Kickhöfer, B., K. Nagel. 2013. Towards High-Resolution First-Best Air Pollution Tolls - An Evaluation of Regulatory Policies and a Discussion on Long-Term User Reactions. *Networks Spat. Econ.* 1–24doi:10.1007/s11067-013-9204-8.

Kickhöfer, Benjamin. 2014. Economic Policy Appraisal and Heterogeneous Users URL http://opus4.kobv.de/opus4-tuberlin/frontdoor/deliver/index/docId/5348/file/Kickhoefer{_}Benjamin.pdf.

Kickhöfer, Benjamin. 2016. Creating an open MATSim scenario from open data : The case of Santiago de Chile 1–22.

Kickhöfer, Benjamin, Amit Agarwal. 2015. Is marginal emission cost pricing enough to comply with the EU CO 2 reduction targets ? 1–19.

Kingma, Diederik, Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kingma, Diederik P, Max Welling. 2013. Auto-Encoding Variational Bayes (Ml) 1–14. doi: 10.1051/0004-6361/201527329. URL `http://arxiv.org/abs/1312.6114`.

Kitchin, Rob. 2014. The real-time city? Big data and smart urbanism. *GeoJournal* **79**(1) 1–14. doi:10.1007/s10708-013-9516-8.

Kleiman-Weiner, Max, Joshua B Tenenbaum. 2016. Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction **1**.

Klein, Ido, Eran Ben-Elia. 2016. Emergence of cooperation in congested road networks using ICT and future and emerging technologies: A game-based review. *Transportation Research Part C: Emerging Technologies* **72** 10–28. doi:10.1016/j.trc.2016.09.005. URL `http://dx.doi.org/10.1016/j.trc.2016.09.005`.

Klein, Ido, Nadav Levy, Eran Ben-Elia. 2018. An agent-based model of the emergence of cooperation and a fair and stable system optimum using ATIS on a simple road network. *Transportation Research Part C: Emerging Technologies* **86**(November 2017) 183–201. doi:10.1016/j.trc.2017.11.007. URL `http://linkinghub.elsevier.com/retrieve/pii/S0968090X17303182`.

Konow, James. 2010. Mixed feelings: Theories of and evidence on giving. *Journal of Public Economics* **94**(3) 279–297.

Krishnan, Sanjay, Animesh Garg, Richard Liaw, Brijen Thananjeyan, Lauren Miller, Florian T Pokorny, Ken Goldberg. 2016. Swirl: A sequential windowed inverse reinforcement learning algorithm for robot tasks with delayed rewards. *under review at Workshop on Algorithmic Foundations of Robotics (WAFR)*.

Kuefler, Alex, Jeremy Morton, Tim Wheeler, Mykel Kochenderfer. 2017. Imitating Driver Behavior with Generative Adversarial Networks doi:10.1109/IVS.2017.7995721. URL `http://arxiv.org/abs/1701.06699`.

Laney, Doug. 2001. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note* **6**(70).

Lazaer, Edward, Eugene Kandel. 1992. Peer Pressure and Partnerships. *J. Polit. Econ.* **100**(4) 801–817.

Ledyard, John, et al. 1997. Public goods: A survey of experimental research. Tech. rep., David K. Levine.

Leibo, Joel Z, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, Thore Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas .

Leonard, Thomas C, Richard H. Thaler, Cass R. Sunstein. 2008. Nudge: Improving decisions about health, wealth, and happiness. *Constitutional Political Economy* **19**(4) 356–360.

Levine, Sergey, Pieter Abbeel. 2014. Learning Dynamic Manipulation Skills under Unknown Dynamics with Guided Policy Search. *Advances in Neural Information Processing Systems 27* 1–3doi:10.1109/ICRA.2015.7138994. URL `http://papers.nips.cc/paper/5444-learning-neural-network-policies-with-guided-policy-search-under-unknown-dynamic pdf`.

Levine, Sergey, Zoran Popovic, Vladlen Koltun. 2011. Nonlinear Inverse Reinforcement Learning with Gaussian Processes. *Advances in Neural Information Processing Systems* 19–27URL `http://papers.nips.cc/paper/4420-nonlinear-inverse-reinforcement-learning-with-gaussian-processes`.

Leyton-Brown, Kevin, Yoav Shoham. 2008. *Essentials of game theory*, vol. 2. doi:10.2200/S00108ED1V01Y200802AIM003. URL `http://www.gtessentials.org/toc.pdf`.

Li, Yunzhu, Jiaming Song, Stefano Ermon. 2017. Infogail: Interpretable imitation learning from visual demonstrations. *Advances in Neural Information Processing Systems*. 3815–3825.

Lin Z., Yin M., S. Feygin, Sheehan M., Paiement J.-F., Pozdnoukhov, A. 2017. Deep generative models of urban mobility **In review**.

Littman, Michael L. 1994. Markov games as a framework for multi-agent reinforcement learning. *Proceedings of the International Conference on Machine Learning* **157**(1) 157–163. doi:10.1.1.48.8623. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.8293{&}rep=rep1{&}type=pdf`.

Lu, Xuan, Song Gao, Eran Ben-Elia. 2011. Information Impacts on Route Choice and Learning Behavior in a Congested Network. *Transportation Research Record: Journal of the Transportation Research Board* **2243** 89–98. doi:10.3141/2243-11. URL `http://trrjournalonline.trb.org/doi/10.3141/2243-11`.

Luce, R Duncan. 1959. *Individual Choice Behavior*. WIley, New York.

Luce, Robert Duncan, Howard Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey*. Courier Corporation.

Lheureux, Alexandra, Katarina Grolinger, Hany F Elyamany, Miriam AM Capretz. 2017. Machine learning with big data: Challenges and approaches. *IEEE Access* **5** 7776–7797.

Ma, Tai-yu, Philippe Gerber. 2016. A hybrid learning algorithm for generating multi-agent daily activity plans 1–11.

Machado, Marlos C, Marc G Bellemare, Michael Bowling. 2017. A laplacian framework for option discovery in reinforcement learning. *arXiv preprint arXiv:1703.00956* .

Mani, Ankur, Iyad Rahwan, Alex Pentland. 2013. Inducing peer pressure to promote cooperation. *Sci. Rep.* **3** 1735. doi:10.1038/srep01735. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3636514&tool=pmcentrez&rendertype=abstract`.

Manski, Charles F. 1993. Identification of Social Endogenous Effects: The Reflection Problem. *Rev. Econ. Stud.* **60**(3) 531–542. doi:10.2307/2298123.

Marschak, Jacob. 1959. Binary Choice Constraints on Random Utility Indicators. Cowles Foundation Discussion Papers 74, Cowles Foundation for Research in Economics, Yale University. URL `https://ideas.repec.org/p/cwl/cwldpp/74.html`.

Mas-Colell, Andreu, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*, vol. 1. Oxford university press New York.

McFadden, Daniel. 2005. The New Science of Pleasure. *Aging* URL `http://eml.berkeley.edu/wp/mcfadden0105/ScienceofPleasure.pdf`.

McFadden, Daniel, et al. 1973. Conditional logit analysis of qualitative choice behavior .

Medhat, Mohamed, Amin Abdel, Latif Wahba. 2008. MILATRAS MI crosimulation L earning-based A pproach to TR ansit AS signment .

Metropolitan Transportation Commision. 2018. Vital Signs. URL `http://www.vitalsigns.mtc.ca.gov/`.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* .

MTC. 2012. Travel Model Development: Calibration and Validation. URL `http://mtcgis.mtc.ca.gov/foswiki/pub/Main/Documents/2012_05_18_RELEASE_DRAFT_Calibration_and_Validation.pdf`.

Mullainathan, Sendhil, Jann Spiess. 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic PerspectivesVolume* **31**(2Spring) 87–106. doi:10.1257/jep.31.2.87. URL `http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.31.2.87`.

National Science Foundation. 2018. Cyber-Physical Systems | NSF - National Science Foundation. URL `https://www.nsf.gov/funding/pgm{_}summ.jsp?pims{_}id=503286`.

Ng, Andrew, Stuart Russell. 2000. Algorithms for inverse reinforcement learning. *Proceedings of the Seventeenth International Conference on Machine Learning* **0** 663–670. doi:10.2460/ajvr.67.2.323. URL `http://www-cs.stanford.edu/people/ang/papers/icml00-irl.pdf`.

Ng, Irene C L, Lu-ming Tseng. 2017. Learning to Be Sociable The Evolution of Homo Economicus Linked references are available on JSTOR for this article : Learning to be Sociable The Evolution of Homo Economicus **67**(2) 265–286.

Niv, Yael. 2011. Reinforcement learning in the brain. *Learning* 1–38doi:10.1016/j.jmp.2008.12.005.

Nordlund, Annika M., J??rgen Garvill. 2003. Effects of values, problem awareness, and personal norm on willingness to reduce personal car use. *Journal of Environmental Psychology* **23**(4) 339–347. doi:10.1016/S0272-4944(03)00037-9.

Nowak, Martin, Karl Sigmund. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* **364**(6432) 56.

Nowak, Martin A, Robert M May. 1992. Evolutionary games and spatial chaos. *Nature* **359**(6398) 826–829.

Ondruska, Pieter, Ingmar Posner. 2014. The Route Not Taken: Driver-Centric Estimation of Electric Vehicle Range. *Proceedings of the Twenty-Fourth International Conference on Automated Planning and Scheduling* 413–420URL `http://www.aaai.org/ocs/index.php/ICAPS/ICAPS14/paper/viewPaper/7899`.

OpenAI. 2016. Universe. URL `https://blog.openai.com/universe/`.

Ortúzar, Juan De Dios, Luis G. Willumsen. 2011. *Modelling Transport*. 4th ed. Wiley.

Ostrom, Elinor. 1990. Governing the commons: the evolution of institutions for collective action.

Ostrom, Elinor. 1999. Coping With Tragedies of the Commons. *Annu. Rev. Polit. Sci.* **2**(1) 493–535. doi:10.1146/annurev.polisci.2.1.493.

Ostrom, Elinor. 2010. Background Paper to the 2010 World Development Report Climate Change and Individual Behavior Considerations for Policy **15**(October 2009).

Ostrom, Elinor, James Walker, Roy Gardner. 1992. Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review* **86**(2) 404–417. doi:10.2307/1964229.

Ostrom, Elinor, James M. Walker. 1991. Communication in a commons: Cooperation with external enforcement. *Lab. Res. Polit. Econ.* 287–322.

Páez, a., D. M. Scott. 2006. A Discrete Choice Approach to Modeling Social Influence on Individual Decision Making (January).

Papadimitriou, Christos H. 2003. *Computational complexity*. John Wiley and Sons Ltd.

Parisotto, Emilio, Jimmy Lei Ba, Ruslan Salakhutdinov. 2015. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342* .

Parkes, David C, Lyle H Ungar. ???? The Tragedy of the Commons : Pricing Social Welfare in Multiagent Systems Strategic Action in a Multiagent System. *Control* .

Pentland, A, Tg Reid. 2013. Big data and Health. *Kit.Mit.Edu* URL `http://kit.mit.edu/sites/default/files/documents/WISH_BigData_Report.pdf`.

Pigou, Arthur Cecil. 1920. The economic of Welfare 1–323URL `http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Enquiries+Concering+Human+Uderstanding{#}1`.

Pomerleau, Dean A. 1991. Efficient Training of Artificial Neural Networks for Autonomous Navigation. *Neural Computation* **3**(1) 88–97. doi:10.1162/neco.1991.3.1.88. URL `http://www.mitpressjournals.org/doi/10.1162/neco.1991.3.1.88`.

Pozdnoukhov, Alexei, Christian Kaiser. 2011. Space-time dynamics of topics in streaming text. *Proc. 3rd ACM SIGSPATIAL Int. Work. Locat. Based Soc. Networks* 1–8doi:10.1145/2063212.2063223. URL `http://dl.acm.org/citation.cfm?id=2063223`.

Pozdnoukhov A., Campbell A., Feygin S., Yin M., S Mohanty. 2015. *The SmartBay Project: Connected Mobility in the San Francisco Bay Area*, vol. The Multi-. In Press.

Rand, David G, Martin A Nowak. 2013. Human cooperation. *Trends in cognitive sciences* **17**(8) 413–425.

Ratliff, Lillian J., Eric Mazumdar. 2017. Risk-Sensitive Inverse Reinforcement Learning via Gradient Methods URL `http://arxiv.org/abs/1703.09842`.

Rieser, M, K Nagel. 2010. Adding Transit to an Agent-Based Transportation Simulation: Concepts and Implementation. *Vsp* **PhD**.

Rittel, Horst WJ, Melvin M Webber. 1973. Dilemmas in a general theory of planning. *Policy sciences* **4**(2) 155–169.

Rothengatter, Werner. 1994. Do external benefits compensate for external costs of transport? *Transportation Research Part A: Policy and Practice* **28**(4) 321–328.

Roughgarden, T., E. Tardos. 2000. How bad is selfish routing? *Proc. 41st Annu. Symp. Found. Comput. Sci.* 1–26doi:10.1109/SFCS.2000.892069.

Rubinstein, Ariel. 1998. *Modeling Bounded Rationality*, vol. 65. doi:10.2307/1060679. URL http://www.jstor.org/stable/1060679?origin=crossref{%}5Cnhttp://arielrubinstein.tau.ac.il/br/br.pdf{%}5Cnhttp://mitpress.mit.edu/books/modeling-bounded-rationality.

Rust, John. 1987a. Optimal Replacement of GMC Bus Engines : An Empirical Model of Harold Zurcher. *Econometric Society* **55**(5) 999–1033.

Rust, John. 1987b. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society* 999–1033.

S. Frey, Bruno, Matthias Benz. 2001. Motivation transfer effect.

Saha, S, S Sen. 2003. Local decision procedures for avoiding the tragedy of commons. *Distributed Computing: Iwdc 2003* **2918** 311–320. URL {\T1\textless}GotoISI{\T1\textgreater}://000188593300030.

Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen. 2016. Improved Techniques for Training GANs. *Nips* 1–10doi:arXiv:1504.01391.

Schmöcker, Jan-Dirk, Tsuyoshi Hatori, David Watling. 2014. Dynamic process model of mass effects on travel demand. *Transportation* **41**(2) 279–304. doi:10.1007/s11116-013-9460-y. URL http://dx.doi.org/10.1007/s11116-013-9460-y.

Schulman, John, Sergey Levine, Philipp Moritz, Michael I. Jordan, Pieter Abbeel. 2015. Trust Region Policy Optimization doi:10.1063/1.4927398. URL http://arxiv.org/abs/1502.05477.

Schwartz, Shalom H. 1977. Normative Influences on Altruism. *Advances in Experimental Social Psychology* **10**(C) 221–279. doi:10.1016/S0065-2601(08)60358-5.

Schweinberger, Michael, Mark S Handcock. 2015. Local dependence in random graph models: Characterization, properties, and statistical inference. *Journal of the Royal Statistical Society, Series B* .

Scott, Darren M. 2004. Social Influence on Travel Behavior : A Simulation Example of the Decision to Telecommute Social Influence on Travel Behavior : A Simulation Example of the Decision to Telecommute .

SFCTA. 2010. San Francisco Mobility, Access, and Pricing Study (December).

Shankari, Kalyanaraman, David E Culler, Randy H Katz. 2014. E-Mission : Automated transportation emission calculation using smart phones .

Shapley, Lloyd S. 1953. Stochastic games. *Proceedings of the national academy of sciences* **39**(10) 1095–1100.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, Demis Hassabis. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**(7587) 484–489. doi:10.1038/nature16961. URL `http://www.nature.com/doifinder/10.1038/nature16961`.

Simon, Herbert A. 1972. Theories of bounded rationality. *Decision and organization* **1**(1) 161–176.

Small, Kenneth A. 2012. Valuation of travel time. *Economics of Transportation* **1**(1-2) 2–14. doi:10.1016/j.ecotra.2012.09.002.

Smith, J Maynard, GR Price. 1973. Lhe logic of animal conflict. *Nature* **246** 15.

Smith, Megan, Dj Patil, Cecilia Muñoz. 2016. Big Risks, Big Opportunities: the Intersection of Big Data and Civil Rights | whitehouse.gov. URL `https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights`.

Söderström, Ola, Till Paasche, Francisco Klauser. 2014. Smart cities as corporate storytelling. *City* **18**(3) 307–320. doi:10.1080/13604813.2014.906716. URL `http://dx.doi.org/10.1080/13604813.2014.906716`.

Song, Chaoming, Zehui Qu, Nicholas Blumm, Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* **327**(5968) 1018–1021.

Song, Xuan, Hiroshi Kanasugi, Ryosuke Shibasaki. 2016. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level. *IJCAI*. 2618–2624.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15** 1929–1958. doi:10.1214/12-AOS1000.

Stewart, Alexander J, Joshua B Plotkin. 2012. Extortion and cooperation in the prisoners dilemma. *Proceedings of the National Academy of Sciences* **109**(26) 10134–10135.

Sun, Lijun, Alexander Erath. 2015. A bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies* **61** 49–62.

Sunitiyoso, Y., E. Avineri, K. Chatterjee. 2011. On the potential for recognising of social interaction and social learning in modelling travellers' change of behaviour under uncertainty. *Transportmetrica* **7**(1) 5 – 30. doi:10.1080/18128600903244776. URL `http://eprints.uwe.ac.uk/10561/1/`

sunitiyoso-avineri-chatterjee{_}On{_}the{_}potential{_}for{_}recognising.....doc.

Sutton, Richard S., Doina Precup, Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* **112**(1) 181–211. doi:10.1016/S0004-3702(99)00052-1.

Syed, Umar, Robert E Schapire. 2008. A Game-Theoretic Approach to Apprenticeship Learning. *Advances in Neural Information Processing Systems 20* **20** 1–8. URL http://papers.nips.cc/paper/3293-a-game-theoretic-approach-to-apprenticeship-learning.pdf.

Thaler, Richard H, Cass R Sunstein. 2003. Libertarian paternalism. *American economic review* **93**(2) 175–179.

Thaler, Richard H., Cass R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth and happiness*. London, UK: Penguin Group.

Thurstone, Louis L. 1927. A law of comparative judgment. *Psychological review* **34**(4) 273.

Train, Kenneth. 2003. *Discrete Choice Methods with Simulation*. URL http://books.google.com/books?hl=en&lr=&ie=UTF-8&id=F-_gYALlfR4C&oi=fnd&pg=PA1&dq=A+Method+of+Simulated+Moments+for+Estimation+of+Discrete+Response+Models+without+Numerical+Integration&ots=9DWej7aXID&sig=KarLb58YdwvJ_DH83Vg2RjimI3k.

Turner, Roy M, Roy M Turner. 1992. The Tragedy of the Commons and Distributed AI Systems. *Proc. 12th Int. Work. Distrib. Artif. Intell.* 379–390.

Udny, Yule G. 1903. Notes on the Theory of Association in Statistics. *Biometrika* **2**(2) 121–134. doi:10.1093/biomet/2.2.121. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/2.2.121.

Väastberg, Oskar Blom, Anders Karlström, Daniel Jonsson, Marcus Sundberg. 2016. Including time in a travel demand model using dynamic discrete choice (75336).

Van Der Maaten, L J P, G E Hinton. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* **9** 2579–2605. doi:10.1007/s10479-011-0841-3. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed{&}cmd=Retrieve{&}dopt=AbstractPlus{&}list{_}uids=791143147914873454 8related:VOiAgwMNy20J.

Van Vugt, Mark. 1996. *Social dilemmas and transportation decisions*.

Van Vugt, Mark, P. a M Van Lange, Ree M. Meertens. 1996. Commuting by car or public transportation? A social dilemma analysis of travel mode judgements. *Eur. J. Soc. Psychol.* **26**(3) 373–395. doi:10.1002/(SICI)1099-0992(199605)26:3<373::AID-EJSP760>3.0.CO;2-1.

Vanhulsel, Marlies, Davy Janssens, Geert Wets. 2007. Calibrating a new reinforcement learning mechanism for modeling dynamic activity-travel behavior and key events.

Varga, László Z. 2014. Online Routing Games and the Benefit of Online Data. *Eighth International Workshop on Agents in Traffic and Transportation at 13th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-6, 2014* . Paris, France, 88–95.

Varga, László Z. 2015. On intention-propagation-based prediction in autonomously self-adapting navigation. *Scalable Computing* **16**(3) 221–232. doi:10.12694/scpe.v16i3.1098.

Verhoef, Erik. 1994. External effects and social costs of road transport. *Transportation Research Part A: Policy and Practice* **28**(4) 273 – 287. doi:http://dx.doi.org/10.1016/0965-8564(94)90003-5. URL `http://www.sciencedirect.com/science/article/pii/0965856494900035`. Special Issue Transport Externalities.

Verplanken, Bas, Ian Walker, Adrian Davis, Michaela Jurasek. 2008. Context change and travel mode choice: Combining the habit discontinuity and self-activation hypotheses. *J. Environ. Psychol.* **28**(2) 121–127. doi:http://dx.doi.org/10.1016/j.jenvp.2007.10.005. URL `http://www.sciencedirect.com/science/article/pii/S0272494407000898`.

Viegas, Jose M. 2001. Making urban road pricing acceptable and effective: searching for quality and equity in urban mobility. *Transport Policy* **8**(4) 289–294.

Vlahogianni, Eleni I, Matthew G Karlaftis, John C Golias. 2014. Short-term traffic forecasting: Where we are and where were going. *Transportation Research Part C: Emerging Technologies* **43** 3–19.

Vogel, Adam, Deepak Ramachandran, Rakesh Gupta, Antoine Raux. 2012. Improving Hybrid Vehicle Fuel Efficiency Using Inverse Reinforcement Learning. *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* 384–390URL `http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5143/5167`.

Von Neumann, J, Oskar Morgenstern. 1953. Theory of games and economic behavior .

Vovsha, Peter, Mark Bradley. 2004. Hybrid discrete choice departure-time and duration model for scheduling travel tours. *Transportation Research Record: Journal of the Transportation Research Board* (1894) 46–56.

Walsh, Fergal, B Sc, Supervisor Alexei Pozdnoukhov. 2013. The spatial structure of mobile communication networks .

Walter, Felix, Stefan Suter. 2003. Sustainable Transport Pricing : From Theory Towards Application .

Watkins, Christopher John Cornish Hellaby. 1989. Learning from delayed rewards. Ph.D. thesis, King's College, Cambridge.

Waugh, Kevin, Martin Zinkevich, Michael Johanson, Morgan Kan, David Schnizlein, Michael Bowling. 2008. A Practical Use of Imperfect Recall 175–182.

Widhalm, Peter, Yingxiang Yang, Michael Ulm, Shounak Athavale, Marta C. González. 2015. Discovering urban activity patterns in cell phone data. *Transportation* **42**(4) 597–623. doi:10.1007/s11116-015-9598-x.

Williams, Ronald J. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning* **8** 229–256.

Wulfmeier, Markus, Peter Ondruska, Ingmar Posner. 2015. Maximum Entropy Deep Inverse Reinforcement Learning URL `http://arxiv.org/abs/1507.04888`.

Wulfmeier, Markus, Dominic Zeng Wang, Ingmar Posner. 2016. Watch This: Scalable Cost-Function Learning for Path Planning in Urban Environments .

Yang, Min, Yingxiang Yang, Wei Wang, Haoyang Ding, Jian Chen. 2014. Multiagent-based simulation of temporal-spatial characteristics of activity-travel patterns using interactive reinforcement learning. *Mathematical Problems in Engineering* **2014**. doi:10.1155/2014/951367.

Ye, Qing, Wai Yuen Szeto, Sze Chun Wong. 2012. Short-term traffic speed forecasting based on data recorded at irregular intervals. *IEEE Transactions on Intelligent Transportation Systems* **13**(4) 1727–1737.

Yin, Mogeng, Ziheng Lin, Sid Feygin, Madeline Sheehan, Jean-Francois Paiment. 2018. AM/PM: Travel Demand Nowcasting.

Yin M., Sheehan M., S. Feygin, Paiement J.-F., Pozdnoukhov, A. 2017. A generative model of urban activities from cellular data **IEEE Transactions in ITS (to appear)**.

Zhang D., Cao J., **Feygin, S.**, Tang D., Pozdnoukhov A. 2017. Connected population synthesis for urban simulation **In review**.

Zheng, Fangfang, Henk Van Zuylen. 2013. Urban link travel time estimation based on sparse probe vehicle data. *Transportation Research Part C: Emerging Technologies* **31** 145–157.

Ziebart, Brian, Andrew Maas. 2008. Maximum entropy inverse reinforcement learning. *Twenty-Second Conf. Artif. Intell.* 1433–1438URL `http://www.aaai.org/Papers/AAAI/2008/AAAI08-227.pdf`.

Ziebart, Brian, Andrew Maas, Andrew Bagnell, Anind Dey. 2009. Human behavior modeling with maximum entropy inverse optimal control. *AAAI Spring Symposium: Human Behavior Modeling* 92–97URL http://www.aaai.org/Papers/Symposia/Spring/2009/SS-09-04/SS09-04-016.pdf.

Ziebart, Brian D, J Andrew Bagnell. 2010. Modeling Interaction via the Principle of Maximum Causal Entropy. *In Proceedings of the 27th International Conference on Machine Learning* 1255–1262URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.6641{&}rep=rep1{&}type=pdf.

Ziebart, Brian D, Andrew L Maas, J Andrew Bagnell, Anind K Dey. 2008. Maximum Entropy Inverse Reinforcement Learning. *Aaai* 1433–1438.

Zilske, Michael, Kai Nagel. 2014. Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Computer Science* **32** 802–807. doi:10.1016/j.procs.2014.05.494.

Zou, Haosheng, Hang Su, Shihong Song, Jun Zhu. 2018. Understanding Human Behaviors in Crowds by Imitating the Decision-Making Process **1**. URL http://arxiv.org/abs/1801.08391.