2021

Transitions in new technology and market structure: applications and new methods for discrete choice model estimation

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

TRANSITIONS IN NEW TECHNOLOGY AND MARKET STRUCTURE: APPLICATIONS AND NEW METHODS FOR DISCRETE CHOICE MODEL ESTIMATION

by

SHUANG WANG

B.A., Zhejiang University, 2013
M.A., Duke University, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2021

Approved by

First Reader

Marc Rysman, PhD
Professor of Economics

Second Reader

Hiroaki Kaido, PhD
Associate Professor of Economics

Third Reader

Jihye Jeon, PhD
Assistant Professor of Economics

Fourth Reader

Jordi Jaumandreu, PhD
Senior Academic Researcher of Economics

Acknowledgments

I am very grateful to my dissertation committee for the guidance they have provided throughout my studies at Boston University. I was blessed to have Marc Rysman as my main advisor. He set a role model of an excellent researcher and respectable person, and led me to all precious opportunities of research assistantships, co-authorships and internships, which molded my skills of doing rigorous research and passion to apply academic works to real-world problems. I thank Hiroaki Kaido for the instructions and encouragement when my research needed econometrics solutions, in particularly for giving me that assignment of replicating Ciliberto and Tamer (2009) for EC711, which put me under a typical pressure but ignited my ambition and pursuit for research on moment inequality estimation. I also own thanks to Jihye Jeon and Jordi Jaumandreu for all the generous support and constructive suggestions that helped grow my ideas into this dissertation.

I have benefitted from advices and research resources kindly offered by my colleagues from the Office of the Chief Economist at Microsoft. Jacob LaRiviere, as the co-author of the third chapter and my manager during the two summers when I worked as a Microsoft research intern, trusted me with the opportunity of working with prestigious economists on high-stake business problems and always has my best interests at heart.

I also thank my PhD colleagues and friends Undral Byambadalai, Lester Chan, Shuowen Chen, Fernando Payró Chew, Eric Hardy, Youming Liu, Siyi Luo, Guang Zhang and many others for their invaluable contributions to my work. The days when I could see you and talk to you around the hallway of the SSW basement will always be a cherishable memory of my life.

My greatest supporters throughout this journey have been my parents, Ping Yuan and Yuemin Wang. They supported all my choices, bear the costs and share the cheerful moments with me no matter what.

Shuang Wang

Doctoral Candidate

Department of Economics

TRANSITIONS IN NEW TECHNOLOGY AND MARKET STRUCTURE: APPLICATIONS AND NEW METHODS FOR DISCRETE CHOICE MODEL ESTIMATION

SHUANG WANG

Boston University, Graduate School of Arts and Sciences, 2021

Major Professor: Marc Rysman, PhD
Professor of Economics, Department Chair

ABSTRACT

My dissertation consists of three chapters that evaluate the social welfare effect of either antitrust policy or industrial transition, all using discrete choice model estimation as the front end for counterfactual analysis. In the first chapter, I investigate the economic impact of the merger that created the world's largest hotel chain, Marriott's acquisition of Starwood, thereby shedding light on the antitrust authorities' performance in protecting competitive markets for the benefit of consumers.

Different from traditional merger analysis that focuses on the tradeoff between the upward pricing pressure and the cost synergy among the merging parties while fixing the market structure, I endogenize firms' entry decisions into an oligopoly price competition model. To tackle the associated multiple equilibria issue, I use moment inequality estimation and propose a novel lower probability bound that reduces the computational burden from being exponential to being linear in the number of players. It also adds to the scant empirical evidence on post-merger cost synergy by showing that every one more affiliated hotel in the local market reduces a hotel's marginal cost by up to 2.3%. Then a comparison between the simulated with-merger and without-

merger equilibria indicates that this merger enhances social welfare. In particular, for those markets that are previously not profitable for any firm to enter, because of the post-merger cost saving, Marriott or Starwood would enter 6% - 24% of them, which provides a new perspective for merger reviews.

The second chapter, joint with Mingli Chen, Marc Rysman and Krzysztof Wozniak, studies the determinants of the US payment system's shift from paper payment instruments, namely cash and check, to digital instruments, such as debit cards and credit cards. With a 5-year transaction-level panel data, for the first time in the literature, we can distinguish the short-term effects of transaction size from the long-term changes in households' preferences. To do so, we incorporate a household-product-quarter fixed effect into a multinomial logit model. We develop a new method based on the Minorization-Maximization (MM) algorithm to address the prohibitive computational challenge of estimating over one million fixed effects in such a nonlinear model. Results show that over a short horizon (within a quarter), the probability of using card increases with transaction sizes in general but exhibits substantial household heterogeneity. While over long horizon (five-year period of the data), with the estimated household-product-quarter fixed effects, we decompose the increase in card usage into different channels and find that only a third of it is due to the changes in household preferences. Another significant driver is the households' entry and exit into the sample.

In the third chapter, my coauthors Jacob LaRiviere, Aadharsh Kannan, and I explore the "death of distance" hypothesis with a novel anonymized customer-level dataset on demand for cloud computing, accounting for both spatial and price competition among public cloud providers. We introduce a mixed logit demand model of spatial competition estimable with detailed data of a single firm but only aggregate sales data of a second. We leverage the Expectation-Maximization (EM) algorithm to tackle the customer-level missing data problem of the second firm. Estimation results and counterfactuals show that standard spatial competition economics hold even when distance for cloud latency is trivial.

Contents

List of Tables

List of Figures

List of Abbreviations

ACS	American Community Survey
ADR	Average Daily Rate
AME	Average Marginal Effect
ASPD	Anonymized Self-reported Performance Data
AWS	Amazon Web Services
BFGS	Broyden–Fletcher–Goldfarb–Shanno Algorithm
BLP	Berry, Levinsohn and Pakes (1995)
BR	Business Register
CDF	Cumulative Distribution Function
CI	Confidence Interval
CMT	Ciliberto, Murry and Tamer (2018)
CPU	Central Processing Unit
CT	Ciliberto and Tamer (2009)
DC	Data Center
DGP	Data-generating Process
DOJ	Department of Justice
DS	Dominant Strategy
EM	Expectation-Maximization
FA	Franchise Agreement
FC	Fixed Cost
FTC	Federal Trade Commission
GCP	Google Cloud Platform
HHI	Herfindahl-Hirschman Index
HOT	Hotel Occupancy Tax
IHG	Intercontinental Hotel Group
IPP	Incidental Parameters Problem
IaaS	Infrastructure as a Service
IIA	Independence of Irrelevant Alternatives
IT	Information Technology
KFF	Henry J Kaiser Family Foundation
M&A	Merger and Acquisition
MCMC	Markov Chain Monte Carlo

ME	....	Marginal Effect
MLE	....	Maximum Likelihood Estimation
MM	....	Minorization-Maximization
MSA	....	Metropolitan Statistical Area
MWh	....	Megawatt-Hour
OLS	....	Ordinary Least Squares
OO	....	Outside Option
PaaS	....	Platform as a Service
PC	....	Personal Computer
PIN	....	Personal Identification Number
PJM	....	Pennsylvania, New Jersey, and Maryland
PPHI	....	Pakes et al. (2015)
RAM	....	Random-access Memory
RUM	....	Random Utility Model
SaaS	....	Software as a Service
SEC	....	Securities and Exchange Commission
SKU	....	Stock Keeping Unit
SPNE	....	Subgame Perfect Nash Equilibrium
STR	....	Smith Travel Research
STRDC	....	Census Database from Smith Travel Research
UPC	....	Universal Product Code
UPP	....	Upward Pricing Pressure
VM	....	Virtual Machine

Chapter 1

Price Competition with Endogenous Entry: The Effects of Marriott & Starwood's Merger in Texas

1.1 Introduction

Merger and acquisition (M&A) activities usually bring anti-competitive concerns, and protecting competitive markets for the benefit of consumers is a central challenge for antitrust authorities. For enforcement success, product price and cost are two key metrics. Specifically, a merger is usually considered anticompetitive if the price is expected to increase even when cost savings are taken into account; however, if such a profit margin lift induces entry of rival firms, the upward pricing pressure (UPP) may be eventually offset. Therefore the *2010 Horizontal Merger Guidelines* emphasizes that the agencies consider entry into the relevant markets as part of their full assessment of competitive effects. However, rigorous and quantitative analysis on entry is rarely found in practice (Li et al., 2018).

This work contributes to filling this inadequacy in merger evaluations that incorporate entry by providing empirical evidence and useful tools. I study a canonical "mega-merger", Marriott International's $13 billion acquisition of Starwood Hotels & Resorts Worldwide in 2016. It turned the united Marriott into the worlds' largest hotel chain which owns over 5,800 properties and 1.1 million rooms in more than 110 countries. Since Marriott announced its plan of acquisition, many consumer groups, such as Travelers United, had urged the competition agencies to block the deal; [1] however, neither the Department of Justice (DOJ) nor the Federal Trade Commission (FTC) pursued any challenges against Marriott's filing. One explanation can the agencies' expectation for a significant cost synergy, another one is the counter-UPP effect of with-merger new entry. And the latter one seems consistent with the ex post facto that over 23.5% of the local markets in Texas have new hotel openings after the merger by Dec, 2018. Therefore, this paper explores the role of entry in compensating for the lost competition caused by the Marriott and Starwood merger.

In particular, there are two ways in which entry of new properties could remedy a merger's market power effect. First, entry of rival firms into the markets where the merging parties are currently in, if timely and sufficient, can offset the increase in market concentration (see Bougette, Hüschelrath and Müller, 2014; Hosken, Olson and Smith, 2016). Such entry effect has been seen addressed in DOJ and FTC's previous merger evaluations. However, the second influence channel, entry from the merging parties into the markets that would have been unprofitable but for their with-merger cost synergy, has not been well-studied in either the *2010 Merger Guidelines* or in academic literature. For instance, the merging firms can achieve a reduction in marginal cost from economies of scale. Conventional merger evaluations often take entry as exogenous, therefore overlooking the potential social welfare improvement in the new markets (see Williamson, 1968; Focarelli and Panetta, 2003; Gugler and Siebert, 2007).

To fully capture the effect of this merger, I take the firms' entry decisions as endogenous when estimating the price competition in the industry. I model the hotel chains' competition as a two-stage oligopoly game. In the first stage, firms simultane-

[1] The president of United Travellers, Charles Leocha, said that "there is nothing positive for consumers. Less competition not only means higher prices but also allows companies to do things that are less visible and would hurt buyers". For more details, see `https://www.mlexwatch.com/articles/2185/print?section=ftcwatch`

ously make entry decisions based on future profitability with full information, meaning that all the firms' profit determinants, even if unobservable to the econometricians, are common knowledge among the firms. In the second stage, profit-maximizing entrants play a Nash-Bertrand pricing game. Therefore, the market structure, including the entrants' number and identities, affects the price competition outcomes and vice versa.

The full information assumption allows selection based on the exact price competition outcomes, which involves the unobserved demand and marginal cost shocks. This could cause self-selection bias in the estimates. Resembling the Heckman correction in single-agent problems, I address this bias in a multi-agent setting by including the shocks in the price competition stage into the firm's entry function and allowing the unobservables across the two stages to be correlated.

Estimating full-information multi-agent entry game can be challenging. One complexity is the multiple equilibria problem rooted in its discrete choice nature. Consider a two-firm example, it may be an equilibrium for only one firm to enter, but a Nash equilibrium concept does not determine which one. Therefore, the probability of the observed outcome is unspecified without arbitrary assumptions on the equilibrium selection mechanism. Regular estimation methods such as maximum likelihood estimation (MLE) or method of moments are therefore infeasible. Following Ciliberto and Tamer (CT, 2009) and Ciliberto, Murry and Tamer (CMT, 2018), I use moment inequalities for estimation, which is based on the insight that the empirical probability of an equilibrium should be greater than the probability of it being unique and less than that of it being one of the multiple equilibria.

Under the CMT framework, the computational burden of simulating the lower probability bound increases exponentially with the number of players. Correspondingly, in CMT's airline industry application, the number of potential entrants is in each market at most six. To make this framework applicable to the Texas lodging industry, where there are as many as 11 important players, I propose a computationally less costly lower bound. Specifically, rather than using the probability that a Nash equilibrium is unique, I use the probability of an outcome being a strictly dominant strategy (DS) equilibrium. The conditions for being a DS equilibrium are sufficient for uniqueness therefore it is a legitimate lower bound; however, a DS equilibrium, by definition, is the best strategy regardless of what the opponents may play, therefore circumventing the complex interaction among agents. As a result, the computational time is reduced from being exponential to being linear in the number of players. I believe my solution is applicable to a wide variety of settings where a researcher wishes to estimate a multi-agent discrete choice model but is impeded by the computational burden due to the multiple equilibria issue.

In theory, such a less binding bound should produce a less sharp identified set; however, a simulation exercise shows that the DS bounds can give a confidence set with a width comparable to CT bounds, while consuming only about 1% of the computational time needed by CT. This comparability in confidence set is likely because we use the same upper bounds and the DS bounds rely less heavily on simulation therefore have less simulation error. More details can be found in Appendix A.4.

I use data on hotel revenues from the Texas *Hotel Occupancy Tax (HOT) Receipts.* Texas is the only U.S. State that publishes each hotel's monthly revenue to reference the HOT. Due to confidentiality issues, historical price data with hotel identities are usually unavailable to researchers. From *Smith Travel Research* (STR), I have hotel-level historical price ranges and sub-city level monthly average prices. I leverage both to interpolate hotel-specific monthly prices based on a seasonality assumption. I focus on the period from January 2015 until March 2016 when the merger was finalized to mimic the data structure in real merger analysis and avoid any systematic differences between the without- and with-merger scenarios. Besides data availability, Texas is an ideal starting point to understand the U.S. lodging industry for its market size. In 2019, the revenue from the lodging industry in Texas reached $12.8 billion, which is a substantial portion of the total $183 billion across the country.

The endogenous entry model estimation implies a fairly competitive market: a 1% price increase will, on average, lead to a 9.20% - 10.35% drop in market share. Additionally, I find that hotel chains having more affiliated properties within a local market not only win consumers presumably by offering an extensive loyalty program, but also enjoy a lower marginal cost from economies of scale. By comparing these to the results from an exogenous entry model, we see that overlooking the firms' entry decisions causes a downward bias in the price coefficient, and even overturns the efficiency gain from having an extensive affiliation network. Mathematically, the directions of the biases indicate that conditional on entry, price is negatively correlated with the demand shock, and the scale of a firm's affiliation network is positively correlated with the marginal cost shock; intuitively, that means an entrant's positive net profit relies on the co-occurrence of a low price and a high demand shock, and to facilitate a low price, firms with a high marginal cost shock lean more on the cost synergy among their affiliated hotels.

I then evaluate the merger's economic impact with simulation. I construct hypothetical markets with the values of the exogenous variables widespread over their empirical distributions, and then solve for the equilibrium prices, market shares and market structures in both the without- and with-merger scenarios, using the endogenous entry model estimates. In particular, I let the merging parties share their affiliation network; thus the increased number of affiliated hotels affects the equilibrium through both the *quality improvement* (an expanded loyalty program) and the *cost synergy* (economies of scale) channels, in addition to the market power gains from the ownership structure change.

Simulation results suggest a welfare-enhancing merger. In markets that are currently being served, I see that Marriott's and Starwood's prices would generally decrease, which indicates that the upward pricing pressure from market power gains and quality improvement is trumped by cost synergy. That could further impel other firms to lower their price via competition; however, the occasions that the rivals are consequently driven out of the market are lower than 2%. Overall, consumer surplus would increase by 17.14% - 24.03%. Additionally, for markets that have not been served yet, Marriott or Starwood would enter 7.2% - 23.7% of them after the merger. The associated consumer surplus gains in these new markets are comparable to those in the served markets in value but are usually overlooked in traditional merger evaluation. This finding suggests that the antitrust agencies should incorporate the potential entry of the merging parties into new markets into their assessment of competitive effects.

Finally, I conduct the same simulation exercise with the exogenous entry model. Specifically, I fix the market structure to be the same as the one simulated with the endogenous entry model in the without-merger scenario. Then I solve for the equilibrium prices and market shares using the exogenous entry model estimates. Because of the estimation biases and firms' sub-optimal entry decisions, results imply the opposite conclusion: the average price would increase, and consumers would suffer from a 23.96% surplus loss after the merger. This divergence emphasizes the importance of endogenizing firms' entry decisions into any merger evaluation framework.

Overall, this paper makes three contributions. First, I propose a computationally attractive way to estimate a multi-agent discrete choice model with many players with moment inequalities. Second, I am able to identify the with-merger cost synergy in a particular form of network expansion, therefore adding to the scant empirical evidence for this phenomenon. Third, I find that this merger enhances social welfare. Particularly, because of the with-merger cost synergy, the merging parties are likely to enter the new markets that are currently not being served yet, which provides a new perspective for merger reviews, since previously without an applicable endogenous entry model the agencies could only focus on the markets where the merging parties are already in.

Related work This paper contributes to two strands of literature. First, the estimation framework builds on papers that endogenize entry decisions or product types into oligopoly demand estimation. When such a discrete choice is endogenized into a multi-agent game, the potential multiple equilibria issue puts point identification in jeopardy unless researchers assume one of the following: (1) an equilibrium selection mechanism (e.g. Bjorn and Vuong, 1984; Kooreman, 1994; Bajari, Hong and Ryan, 2010); (2) that firms are homogeneous or heterogeneous up to types (e.g. Bresnahan and Reiss, 1990, 1991); (3) that multiple outcomes are considered as one event with a well-defined probability (e.g. Berry, 1992); (4) a limited information structure (e.g. Seim, 2006; Sweeting, 2009; Aradillas-Lopez, 2010); (5) a predetermined entry order (e.g. Cohen and Mazzeo, 2007). When researchers are unwilling to impose those assumptions that in some cases lack economic foundations, the model is *incomplete*. It describes the situation where the model prediction is a correspondence rather than a function from the exogenous variables (Tamer, 2003). In that case, one can use inequality restrictions derived from the model for estimation, which is more robust but usually associated with partial identification and computational challenges.

My approach is closest in spirit to Ciliberto, Murry and Tamer (2018). They use inequalities to restrict the empirical probabilities in between the probabilities of being the unique equilibrium and the probabilities of being one of the multiple equilibria. CMT, as well as my model, inherits the full information, heterogeneous firms and simultaneous entry decision settings in Ciliberto and Tamer (2009) and advances it by letting a structural price competition model determine profit.

The full information assumption implies that the exact price competition outcome affects firms' entry decisions. Therefore, researchers have to solve for the equilibrium price and market share for each possible market structure to find the equilibrium ones. That can be computationally burdensome, as the number of possible market structures increases exponentially with the number of players, and in the course of parameter searching this might have to be repeated thousands of times. In CMT's airline industry application, the number of potential entrants per market, defined as airlines that are serving at least one market out of both of the endpoint airports, is at most six. Without further restrictions, the Texas lodging industry has 11 major chains that could potentially enter any local market. My DS lower probability bound not only makes the CMT framework feasible in applications with more players, but also reduces the computational burden of any potential application in real-world antitrust practice.

It is also possible to derive inequalities directly from revealed preference theory, as in Pakes et al. (2015) (PPHI). [2] It is established on the behavioral assumption that each agent's choice must be weakly most profitable given other agents' choices.

A critical difference between my model and the applications of PPHI (e.g. Ishii, 2005; Ho, 2009; Ho and Pakes, 2014a; Eizenberg, 2014) is that they assume limited information. When firms make entry or product decisions, demand and marginal cost shocks are unknown and thus mean independent to the instruments in firms' information sets, therefore the parameters in the price competition stage can be estimated in advance as in Berry, Levinsohn and Pakes (1995) and point identification is guaranteed. For parameters in the entry stage, they either directly assume the fixed

[2] See Pakes (2010) for a comparison between PPHI and the probability-based inequality method, i.e. generalized discrete choice model, such as Ciliberto and Tamer (2009)

cost (FC) shocks are mean-independence, or restrict their level of variation based on industrial facts so that differencing or a uniform bound can be applied. Accordingly, one does not have to recover their distribution and repetitively solve for the equilibrium when the values of these shocks change in estimation. This makes PPHI computationally more appealing, but limiting the selection on unobservables.

Papers that take on the computational burden of estimating full information endogenous entry models are scant. Li et al. (2018) use importance sampling techniques to avoid solving for equilibrium outcome repetitively. However, they leverage a known, sequential order of flight service choice to get around the multiple equilibria issue and assume that demand, marginal cost and fixed cost unobservables are independent. Fan and Yang (2020) also ease the computational burden of estimating a multi-product choice model using a DS equilibrium inspired method, therefore it can be considered as a cocurrent work of mine. They allow selection on fixed cost shocks but not on demand and marginal cost shocks by assuming limited information. Additionally, their inequalities are constructed based on unilateral choice probabilities, so the information carried by the variation in the rivals' choices is not fully utilized. To my knowledge, this paper is the first to estimate an endogenous entry model assuming full information on fully-correlated unobservables, while computationally handling as many as 11 potential entrants.

In addition, this paper adds to the empirical studies on the lodging market in the U.S. and elsewhere. Several papers investigate how market structure affects equilibrium prices and profits in this industry. Nevertheless, almost all of them employ a reduced-form profit function for tractability. Kalnins, Froeb and Tschantz (2017) uses a difference-in-difference design to check the price and occupancy effects of 898 lodging mergers in the United States, with the goal of testing the hypotheses from different models of competition. Mazzeo (2002a) and Mazzeo (2002b) check whether competition-induced price drops are lower when motels are more differentiated. Endogenous product choices are addressed by instruments and estimated separately from price competition. In contrast, I have a detailed model where entry decisions are directly determined by the price competition outcome, and perform a simultaneous estimation which is led by the full information assumption.

Finally, I borrow industrial setting ideas such as the outside option and market size definitions from papers that cover other topics in the lodging industry. Examples includes but are not limited to Fernandez and Marin (1998), Kalnins and Chung (2004), Kalnins (2006), Suzuki (2013), Lewis and Zervas (2016), Leisten (2020).

The remainder of this paper proceeds as follows. In section 2, I introduce the Texas lodging market and describe the data that I use. I introduce my model in section 3. In section 4, I illustrate my estimation strategy and then explain how using the probability of being a DS equilibrium as the lower bound can reduce the computational burden. I show parameter estimates from my endogenous entry model and compare them to the ones from the exogenous entry model in section 5. In section 6, I evaluate the price and social welfare effects of Marriott's acquisition of Starwood with simulation and show that ignoring firms' entry decisions could lead to severe bias. I then conclude in section 7.

1.2 The Texas Lodging Industry and Data

I use two main data sources for this study. One is the hotel census database from *Smith Travel Research* (STRCD),[3] an independent consulting firm specializing in the lodging industry. It offers key information on every hotel in the Texas market, including the hotel's name, location, number of rooms, class, affiliation history, and

[3]See `https://str.com/training/academic-resources/share-center-details`

price range (the historically lowest and highest prices).[4]

The other main data source is the Texas *Hotel Occupancy Tax (HOT) Receipts*, provided by the Texas Comptroller of Public Accounts. HOT is a nationwide tax collected by hotel owners, operators, or managers from their guests who rent a room or space. Among all the U.S. states, only Texas publishes HOT receipts with each hotel' monthly revenue as a reference for determining the tax.[5,6] The data is available for decades, but this paper focuses on the period from January 2015 to March 2016 when the merger was granted clearance, so that the complications rooted in the systematic differences between the without- and with-merger scenarios are avoided.[7]

I merge the two data sets by hotel address, so I have a panel data on hotel revenue at monthly level. In the lodging industry, property owners sometimes change their affiliations. For instance, the DoubleTree by Hilton Hotel at Dallas Richardson opened on Sep 1, 2009, but the same property used to be a Radisson from March 1, 2003 and before that a Clarion. To be consistent at the level of observation, I recover each hotel's historical affiliation also to the monthly level based on the dates of ex-affiliations in STRCD. In the section below, I explain the definitions of market and other variables in the industrial context, along with their construction procedure when they are not directly available in my data sources. Other auxiliary data sources are also introduced when necessary.

[4]In the data, I observe a distinct price range for three types of rooms, single, double and suite. For all-suite hotels, I use the price range of suite; for hotels having multiple types of rooms, I use the price range of double room. A better way of doing this is probably calculating the quantity weighted average across different types. However, since that quantity data, i.e. the number of rooms sold for each type, is not available, I use the price range of double room as an approximate average, given that it usually lies between those of single room and suite.

[5]In the data, the majority of hotels (97.2%) file their HOT monthly, the rest 2.8% file quarterly at least for a period of time. In the latter case, I evenly distribute their quarterly revenues into the three months in that quarter.

[6]The HOT tax rate in Texas is 6% of the cost of a room.

[7]Other papers that also focus on the without-merger period in estimation includes Nevo (2000), Peters (2006) and Björnerstedt and Verboven (2016).

1.2.1 The set of players, market definition and market size

The set of players I consider each parent hotel company as a player and investigate their decisions of entering any local market and setting prices if enter. In the lodging industry, a parent company may hold multiple brands, for example, Marriott International, as a parent company, holds Ritz-Carlton, Marriott, Sheraton, etc. We also see that these brands have hotels at multiple locations. One feature in the U.S. lodging industry is that parent companies are seldom the hotel owners. Instead, they proliferate by franchising.

Here I define franchisee as an owner company operating under a parent company's brand name, and despite of a little ambiguity refer the hotel following such an operation model as a franchisee or a chain-affiliated hotel interchangeably. In reality, these franchisees can be further categorized by their management companies. Specifically, the hotel owners can choose among managing the hotel by themselves, hiring a third-party management company, or letting the parent company manage it directly. In a narrow sense, only the first two cases are considered as franchisees and the last one is distinguished as chain-management hotels. Given that chain-management hotels account for less than 10% of the data and the differences from the perspective of modeling is nuanced, I broadly consider all of them as franchisees.

Also, this paper abstracts from the role that the franchisees play in setting the daily price. It can be justified by the industrial fact that the franchisees usually adjust the price closely following the parent companies' guidance through a centralized revenue management system. Although there are other studies that discuss the pricing behavior in the lodging industry from the franchisee's perspective (e.g. Leisten, 2020), my paper focuses on the competition among parent companies and thus takes a different angle.

The role of the franchisees in determining entry is discussed around the definition of entry later in this section.

Market definition In this paper, a market is vertically restricted within the upscale segment and spatially outlined by neighborhoods, which is roughly equivalent to a census tract in terms of area.

First, industry sources vertically categorize hotels into classes by their *Average Daily Rate* (ADR), broadly labeled by economy, midscale, upscale and luxury. Rigorous analysis should explicitly distinguish among these price-based segments, given that budget-constrained consumers would substitute much more within a scale instead of across them, Technically, for a multi-brand chain hotel, class is associated with the brand, and for each brand, the class is determined by its worldwide rather than local year-end ADR. Therefore, even though hotels in the same class have comparable prices, there is still significant variation within and across local markets to identify the price effect in the estimation.

Table 1.1: A summary of lodging market in Texas by scale

Upper Midscale	Upscale	Upper Upscale	Luxury
	Marriott		
Fairfield Inn (70) TownePlace Suites (38)	Courtyard (91) Residence Inn (67) Springhill Suites (48)	Autograph Collection (2) Gaylord (1) Marriott (25) Marriott Conference Center (1) Renaissance (5)	JW Marriott (4)
	Starwood		
	Four Points by Sheraton (8) aloft Hotel (8) component (2)	Sheraton Hotel (15) Westin (12) Le Meridien (2)	W Hotel (2) Luxury Collection (1) St Regis (1)
	ADR($)		
120	151	215	298

Remark: ADR is calculated by averaging the price ranges of single room, double room and suite, with the data from STRCD, and then report the middle point within the range.

Here I focus on the upscale segment, out of the consideration that the products in this class are more heterogeneous than those in cheaper classes, thereby market concentration increase is more of a threat to the economy. Also, Marriott & Starwood has a substantial overlap in market presence in this class, which is the precondition for exercising any market power. Table 1.1 gives the numbers of their properties in Texas by class in the parentheses behind each brand. Marriott and Starwood altogether own 296 hotels of upscale class, more than the total of all the other three classes.

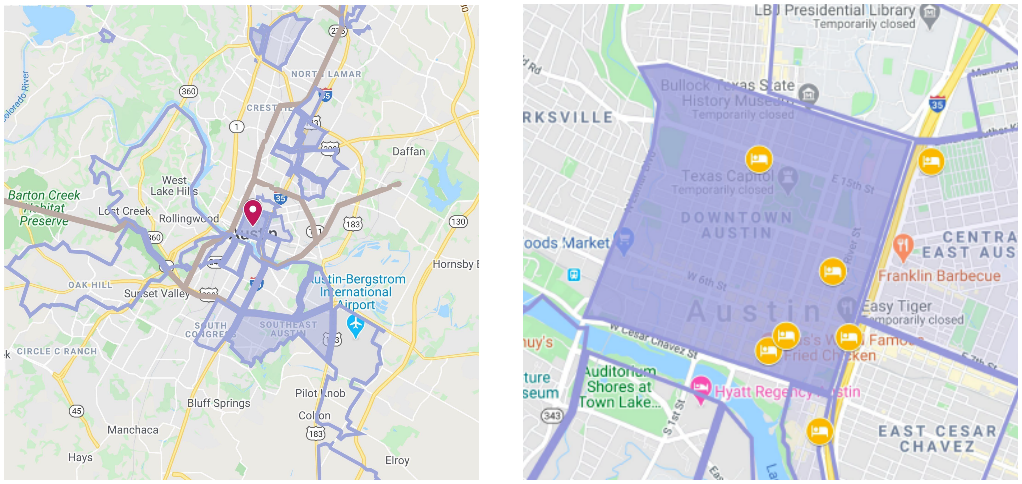
The bottom panel of Table 1.1 demonstrates the sizeable differences in ADR across scales. The numbers are scale-wise averages in Texas, covering the other 9 hotel chains that compete with Marriott and Starwood in the upscale market, i.e. Hilton, Intercontinental Hotel Group (IHG), Hyatt, Best Western Hotels & Resorts, Wyndham Worldwide, Choice Hotels International, Radisson Hotel Group, Great Wolf Lodge and Sonesta International Hotels Corporation. Table 1.2 summarizes the numbers of neighborhoods served and rooms owned by the top hotel chains. The numbers in the brackets indicate percentages. As shown, before the merger, Marriott had already been the largest player in this market. Hilton is its largest competitor, followed by IHG and Hyatt.

Table 1.2: Top Hotel Chains in Texas (Pre-merger)

	Total	Marriott+Starwood	Hilton	IHG	Hyatt	...	0-entry
No. of Neighborhoods	499	114+16 = 121	97	49	29	...	297
		(24.2%)	(19.4%)	(9.8%)	(5.8%)	...	(59.5%)
No. of Rooms	59217	22722+2223 = 24945	17383	7860	4373	...	
		(42.2%)	(29.4%)	(13.3%)	(7.4%)	...	

Second, the choice of neighborhood as local market border follows the *Hotel Investment Handbook* (Rushmore, Ciraldo and Tarras, 2000), which refers to neighborhood as the primary geographic territory that a hotel investor should consider when judging whether another lodging facility represents competition for the subject property. The reason is that a neighborhood consisting of a grouping of complementary land uses *"usually has an observable uniformity and exhibit a greater degree of commonality than the larger market area"*. I retrieve neighborhood information using the Google

Figure 1·1: Austin, TX Neighborhood Map



(a) Neighborhoods in Austin, TX **(b)** Upscale Hotels in Downtown Austin

Map Geocoding API based on the hotels' coordinates. [8,9] There are 202 neighborhoods in Texas that are currently being served by at least one upscale chain hotel. I further define potential market for future entry as neighborhoods that have at least one chain hotel of a class adjacent to upscale, i.e., upper midscale and upper upscale. That gives extra 297 neighborhoods and makes the total 499. The rationale is from their similarity in economic trends.

As an example, Figure 1·1(a) gives the neighborhood layout of Austin, TX. Each neighborhood is shaded by the number of upscale hotels within it. Figure 1·1(b) zooms in the neighborhood *Downtown, Austin*, labeled by a red balloon in Figure 1·1(a). There are five upscale hotels in this neighborhood, labeled by the yellow bal-

[8]The reader is referred to Kahle and Wickham (2013) on how this can be done with an R-package. For those hotels that fall out of the neighborhood boundaries defined by Google Map, I first try to relate them to those hotels with neighborhood information by 5-digit zip code; when such relating is not possible, I group hotels in the same zip code area as a separate local market.

[9]I recognize that compared to city/town/village neighborhood is a relative subjective geographic definition which may fuzzy or even overlapping boundaries; however, according to a Google maps product expert, Google commonly uses the neighborhoods defined by the local government which normally have no overlap. And in practice, I do not see any hotel properties get associated to multiple neighborhoods by Google Maps. More details regarding Google's rule can be found at `https://support.google.com/maps/thread/31414182?hl=en`

loons. On one hand, given their physical proximity, it is reasonable to believe that they are each other's direct competitor; on the other hand, I realize that neighborhood is a relatively narrow definition for local market in the sense that substitution across neighborhoods is not impossible in real world. However, it is certainly less substantial than the within-neighborhood substitution, because neighborhoods are often naturally separated by main streets. In the case of Downtown, Austin, there are two other upscale hotels in the nearby neighborhoods, but they are split by W. Cesar Chavez St and Interstate 35.

Market size I construct the measure of market size as the number of rooms available in each neighborhood-month, using the insight in Lewis and Zervas (2016). For those neighborhoods that are temporarily not being served, I first establish a linear relationship between market size and demographics using the data from the served markets, and then predict the market sizes of those potential markets by plugging their observed demographics into the fitted linear relationship. The demographic indicators are from the American Community Survey (ACS) conducted by Census Bureau, including population/m^2, number of employees in hospitality industry, and median income levels. Since ACS data is at census-tract level, I relate each neighborhood to the census tract with the least distance between their inner points. More details about the fitted linear relationship are in Appendix A.1.

Table 1.3 shows that a served market on average has 1.627 upscale chain hotels, and the market size of a served market is on average higher than that of a market that has not been served yet, which implies selection on the observables.

1.2.2 Endogenous Variables: Entry, Price and Market Share

Entry I consider entry as a parent company granting the right to use its brand name to a property owner, thereby the property starts to be operated under the brand, and the parent company will do so only if the future profit from operating

Table 1.3: Summary statistics

	Markets Served		All	
	Mean	Std.Dev	Mean	Std.Dev
Market-level Variables				
No. of entrants	1.627	0.980		
Market size	10,707	11,460	10,114	7,584
Fixed cost shifter				
Median monthly gross rent ($)	988.3	322.0	983.9	331.9
	N = 2,723		N =6,986	
	Entrants		All	
	Mean	Std.Dev	Mean	Std.Dev
Market-Chain-level Variables				
ADR ($/room-night)	116.69	22.73		
Marriott	114.62	19.06		
Hilton	121.10	24.54		
IHG	114.70	18.51		
Starwood	107.79	17.46		
Market share (by room-night)	0.339	0.214		
Marriott	0.373	0.186		
Hilton	0.389	0.231		
IHG	0.259	0.169		
Starwood	0.312	0.256		
Demand shifter				
Affiliated upscale hotels in sub-city (abbrev. N^{uc})	3.071	1.854	0.797	1.495
Marriott	4.553	1.850	3.843	1.988
Hilton	3.160	1.238	2.641	1.486
IHG	1.888	0.835	0.972	0.960
Starwood	1.444	0.498	0.282	0.549
Marignal cost shifter				
Affiliated all hotels in sub-city (abbrev. N^{all})	6.594	3.186	4.315	5.107
Marriott	7.699	2.864	6.735	3.127
Hilton	7.106	2.235	6.799	2.333
IHG	7.316	2.851	7.156	2.844
Starwood	2.750	1.464	0.798	1.282
	N = 4,430		N = 76,846	

the hotel is positive. Two complications in real world are avoided in the definition. First, strictly speaking, entry, a new hotel opening, should be a joint decision of the parent company and the owner company, i.e., franchisee and the franchiser. However, since my counterfactual analysis focuses on a merger between franchisers, I suppress the franchisees' role by assuming an excessive supply of them. In other words, I assume that there is always a property owner willing to build a franchiser-franchisee relationship with the chain as long as operating the hotel is profitable.[10] This is in line with the observed much intenser competition among the owner companies compared to that among the parent companies. In my data, for those hotels that have owner company information (51.9% of all the 555 upscale hotels), I see 83 different owner companies, and the average market share by property is only 1.2%, therefore it is reasonable to assume a parent company's market.

Furthermore, I take the parent company as a profit-maximizer and abstract from the complex revenue sharing scheme between the franchiser and the franchisee. This simplification is motivated by tractability. According to a form of Hyatt Place's franchise agreement published by the U.S. securities and exchange commission, franchisees need to pay various fees to their franchisers, including a one-time application fee, a monthly royalty fee and a monthly contribution to the marketing, central reservations and technology fund that are proportional to the hotel's gross room revenue, etc.[11] After all these trivialities, franchisers eventually benefit from franchisees' profit, therefore their guidance on price setting and capacity management should aim at profit maximization, otherwise the franchisees would also be less incentivized to follow.

Price (ADR) Price is probably the most important demand shifter in this market. Given the room-night price may vary by room type and consumer, I use the *average daily rate* (ADR) as the price, following industry standards. There is no good resource for historical proprietary-level price data, and previous empirical works that do not need to identify different parent companies often use the STR *anonymized self-reported performance data* (ASPD), e.g. Suzuki (2013), Lewis and Zervas (2016) and Leisten (2020). However, for my purpose, the hotel-specific ADR is a crucial outcome of the competition among heterogeneous firms and a criterion to distinguish the

[10] Suzuki (2013) relies on the same assumption to investigate the impact of land use regulation on the hotels' entry decisions.

[11] See `https://www.sec.gov/Archives/edgar/data/1468174/000119312509165558/dex1046.htm`

merger's impact on the merging parties vs. their rivals. Leveraging on an assumption on local seasonality, I fill the gap between the anonymized hotel-level ADR and the hotel identities using (1) the hotel-level price range information in STRCD and (2) the sub-city-level monthly ADR provided also by STR. The detailed interpolation procedure is as follows.

First, motivated by the well-known seasonality in hotel prices, I assume that for any hotels j and k in market m, their monthly ADR's, p_{jmt} and p_{kmt}, are *linearly* correlated. Therefore, they are each linearly correlated with the *sub-city* average, $Avgp_{mt}$. Sub-city is equivalent to the concept sub-market constructed by STR. [12] It refers to a subset of a geographic area that is typically made up of a *Metropolitan Statistical Area* (MA), a group of MA's or a group of postal codes, e.g. Austin CBD, TX. In my data, there are 51 sub-cities in total, and each sub-city contains 9.78 neighborhoods on average. As an informal test of the seasonality, I take an anonymized sub-city in ASPD, and then plot the monthly ADR of every upscale hotel in this sub-city as a separate line in Figure 1·2. We see that the common time trend among these prices is prominent.

[12] Given that markets are defined at neighborhood level in this paper, I rename sub-market as sub-city to emphasize that it is actually a geography that usually covers multiple neighborhoods instead of being a subset of it, as the prefix "sub-" indicates.

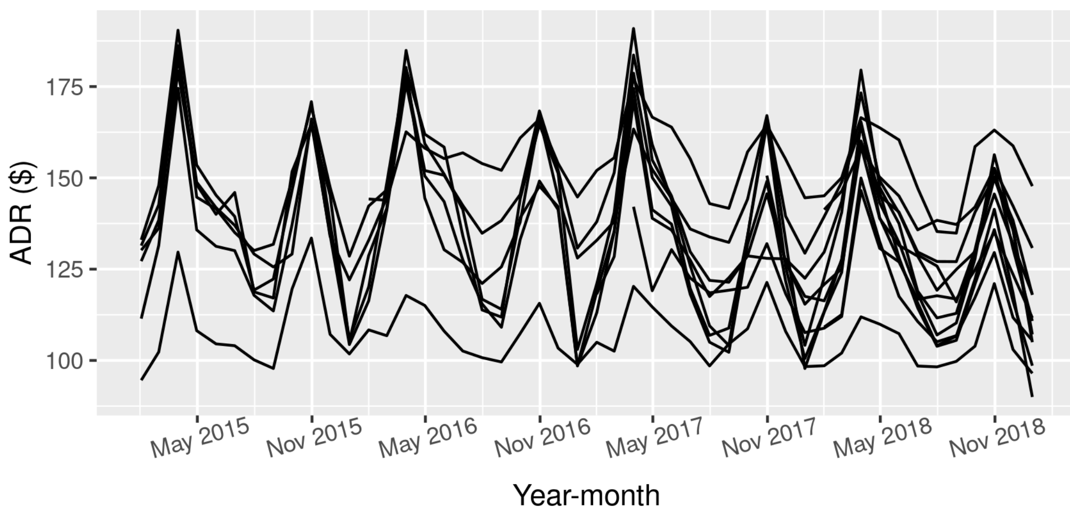
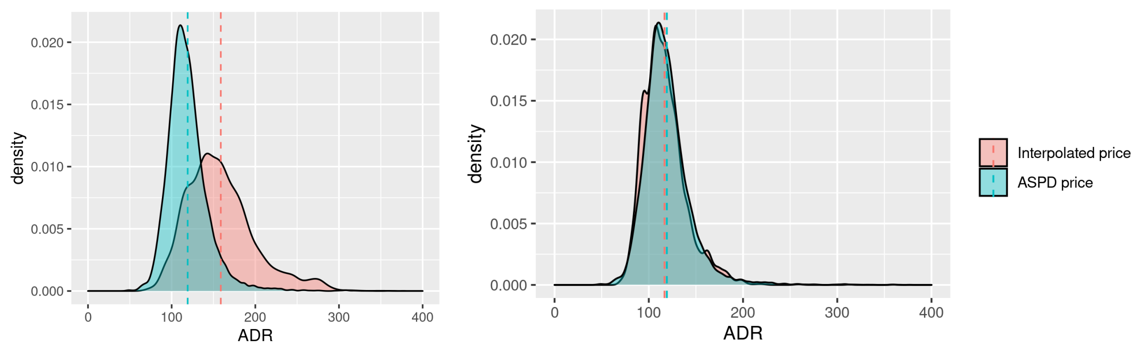


Figure 1·2: Anonymized hotel prices in the same sub-city

Implied by the linearity assumption, the lowest/highest historical prices of firm j, $\bar{p}_{jm}$ /$\underline{p}_{jm}$, during its life span $\mathfrak{T}_{jm}$, should coincide with the lowest/highest sub-city average , $Avgp_{m\bar{t}_j}/Avgp_{m\underline{t}_j}$. Here, $\underline{t}_j = \underset{t\in\mathfrak{T}_{jm}}{argmin}\, p_{jmt}$ and $\bar{t}_j = \underset{t\in\mathfrak{T}_{jm}}{argmax}\, p_{jmt}$ are the months in which $\bar{p}_{jm}$ and $\underline{p}_{jm}$ occur. These two number of pairs $(\bar{p}_{jm}, Avgp_{m\bar{t}_j})$ and $(\underline{p}_{jm}, Avgp_{m\underline{t}_j})$, when considered as two points, pin down the slope of a linear relationship, $\frac{\bar{p}_{jm} - \underline{p}_{jm}}{Avgp_{m\bar{t}_j} - Avgp_{m\underline{t}_j}}$. Then, the hotel-specific ADR's in other months can be interpolated based on its deviation from the lowest point, i.e.,

$$p_{jmt} = \underline{p}_{jm} + \frac{\bar{p}_{jm} - \underline{p}_{jm}}{Avgp_{m\bar{t}_j} - Avgp_{m\underline{t}_j}} \times (Avgp_{mt} - Avgp_{m\underline{t}_j}) \tag{1.1}$$



(a) Before calibration

(b) After calibration

Figure 1·3: Interpolated price vs. ASPD price

I prove the concept plausible by showing that the interpolated prices roughly follow a normal distribution, same as the anonymized prices in ASPD, see Figure 1·3(a). As the last step, I calibrate the interpolated prices towards the distribution of prices in ASPD by mean and standard deviation, and then the two distributions overlap almost perfectly, as shown in Figure 1·3(b).

Market share and consolidation Following the market size definition, I define market share in terms of the number of rooms sold per month, calculated as the hotel's monthly revenue from HOT receipts divided by its calibrated monthly ADR. Then, I can use the *Herfindahl-Hirschman Index* (HHI) to summarize the concentration level of the Texas lodging industry. Conventional ex ante merger evaluation focuses on the incumbents and consider the HHI increment caused by two merging firms operating as one. In that way, the merger led to an HHI increase of 358.34 in Texas from its without-merger level, 2715.64. According to the *2010 Horizontal Merger Guidelines*, markets with an HHI above 2,500 are considered as highly concentrated, and *"Mergers resulting in highly concentrated markets that involve an increase in the HHI of more than 200 points will be presumed to be likely to enhance market power. The presumption may be rebutted by persuasive evidence showing that the merger is* *unlikely to enhance market power"*.[13]

However, as mentioned earlier, we observe a substantial amount of with-merger new entry. Table 1.4 breaks down the probability that a neighborhood has a new hotel by firm. We see that Marriott has the highest probability of entry. Notably, the rivals are also likely to enter. And Starwood, which used to be the 5^{th} in Texas, now has the second highest probability of entry.

Table 1.4: Post-merger new entry

Firms	Marriott	Starwood	Hilton	IHG	Hyatt	Others
% of markets	6.19%	3.98%	3.98%	3.10%	1.77%	2.21%

Consider the potential significant impact of new entry on equilibrium prices, it is necessary to endogenize firms' entry decisions into the demand estimation so that with-merger optimal entry or exit is allowed in merger simulating.

1.2.3 Exogenous variables: demand, marginal cost and fixed cost shifters

Endogenous entry analysis requires the explanatory variables to be observable for both entrants and non-entrants and to vary across markets to account for firms' differentiated entry decisions. Counterexamples can be room amenities or conference room space, since they are unobservable if a parent company has not served a certain neighborhood yet. To address such needs, I connect each neighborhood to the higher level geography, sub-city, and then use sub-city-chain specific characteristics as demand and marginal cost shifters.

In the demand equation, I use the number of affiliated *upscale* hotels in the same sub-city. It serves as a proxy for the chain's loyalty program quality: the more affili-

[13] The HHI reported here is at state level, meaning that market shares are state-wise. The average increase in neighborhood-level HHI, weighted by market size, is 146.62 based on a 5996.75 without-merger level. Although the magnitude is smaller than it at the state level, per *2010 Horizontal Merger Guidelines*, such an increase in the *"HHI of between 100 points and 200 points potentially raise significant competitive concerns and often warrant scrutiny"*.

ated hotels the firm has in a local area, the more attractive it is to its members since it would be easier to accumulate points or to redeem free nights. Additionally, locally affiliated hotels can share the cost of administration or advertising, have higher bargaining power when outsourcing their linen to local services and negotiating commission rate paid to travel agents, etc. Since such network effects do not have to be restricted within hotels of the same class, I use the number of affiliated hotels of *all classes* in a sub-city as a covariate explaining marginal cost variation. Regarding fixed cost, rent is definitely a substantial part. Since the tract-level median rent data in ACS does not have variation across parent companies within a market, I further include a fixed effect for each of the top 3 chains in Texas, i.e. Marriott, Hilton, and IHG.

Table 1.3 shows apparent differences in these variables between markets served versus all markets and between entrants versus all firms, which again justifies their influences on firms' profitability and selection on the observables. Specifically, a served market on average has 1.627 hotels, and firms tend to enter markets with higher social economics level, indicated by higher market size and a higher gross rent. Each entrant has 3.071 affiliated upscale hotels in the same sub-city, more than three times higher than the average level of all firms, and a similar pattern can be found in affiliated hotels of all classes. Additionally, they both vary across firms as well as across markets for the same firm, which makes them qualified demand and marginal cost shifters as described at the beginning of this subsection.

1.3 Model

I now present a 2-stage static oligopoly model of the Texas lodging industry. In the first stage, firms simultaneously make entry decisions based on future profitability. In the second stage, entrants play a Bertrand pricing game, where market shares are generated by utility-maximizing consumers making a discrete choice among the entrants. Figure 1·4 is a timeline that clarifies the information structure and the order in which firms and consumers take action.

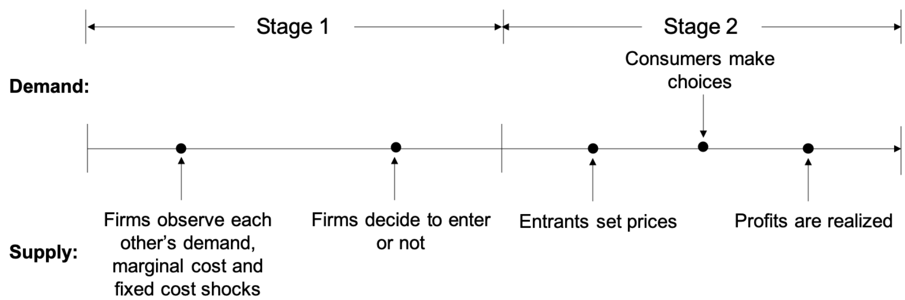


Figure 1·4: Model timeline

In particular, I assume full information in the sense that the fixed cost shocks, along with the demand and marginal cost shocks in the pricing competition stage, are common knowledge when firms consider entry. As a result, firms are self-selected into local markets based on both the observables and the unobservables in both stages, and the associated selection bias is tackled by using the exact profit in the entry decision function. Another source of selection bias is the correlation among the shocks, I explicitly address it by assuming a joint normal distribution and then including the distributional parameters into the estimation. Furthermore, the simultaneous move setting makes it a multi-agent discrete choice model which potentially involves the multiple equilibria issue and thus requires a specialized estimation framework.

In the remainder of this section, I first explain modeling details in a backward order, following the logic of solving a sub-game perfect Nash equilibrium (SPNE), and then discuss the complications associated with the full information and the simultaneous move assumptions, with a focus on how they affect the estimation strategy.

1.3.1 Demand

I model the lodging demand by a multinomial discrete choice. Consumer i chooses the upscale hotel j in market m that generates the highest utility, [14]

$$\begin{aligned} u_{ijm} &= x^{o'}_{jm}\beta + x^{u}_{jm} - \alpha p_{jm} + \epsilon_{ijm}, \quad \epsilon_{ijm} \sim \text{ Type I extreme value, } i.i.d \qquad (1.2) \\ d_{ijm} &= 1[u_{ijm} \geq u_{ikm}, \forall k \in \mathcal{F}_m] \end{aligned}$$

I use superscript o and u to distinguish the observed and unobserved shifters. p_{jm} is the price. The idiosyncratic demand shock ϵ_{ijm} is assumed to follow a Type I extreme value distribution, which makes this a multinomial logit model. d_{ijm} indicates whether consumer i chooses hotel j in market m. In particular, consumers can only choose among the available choice set $\mathcal{F}_m$. Specifically, let y_{km} be a binary indicator for firm k's entry decision into market m, $\mathcal{F}_m$ includes the set of entrants, $\{k : y_{km} = 1\}$, and the outside option of staying in non-hotel rooms. I normalize the deterministic utility of the outside option to 0, then given the multinomial logit setting, the market share of firm j in market m is

$$s_{jm} = \frac{\exp(x^{o'}_{jm}\beta + x^{u}_{jm} - \alpha p_{jm})}{1 + \sum_{\{k:y_{km}=1\}} \exp(x^{o'}_{km}\beta + x^{u}_{km} - \alpha p_{km})} \qquad (1.3)$$

Notably, the equilibrium market shares rely on the demand shifters, prices of all entrants and the entry decisions of all potential entrants. To ease notation for future discussion, I collect the firm-level variables into vectors, e.g. $\boldsymbol{x}^o_m = \{x^o_{jm}\}_j$ and denote the union of the observables and the unobservables by the notation without the superscript, i.e. $\boldsymbol{x}_m = (\boldsymbol{x}^o_m, \boldsymbol{x}^u_m)$. Then the equilibrium market share s_{jm} can be written as function of the choice variable in the same stage $\boldsymbol{p}_m$, and the predetermined

[14] I suppress time subscript t to ease notation.

variables $\boldsymbol{x}_m$, $\boldsymbol{y}_m$, i.e. $s_{jm} = s_{jm}(\boldsymbol{p}_m|\boldsymbol{x}_m, \boldsymbol{y}_m)$.[15]

1.3.2 Supply

Pricing stage After entering into the market, each entrant j sets price to maximize their variable profit

$$\max_{p_{jm}} \pi_{jm} = (p_{jm} - mc_{jm}) M_m s_{jm}(\boldsymbol{p}_m|\boldsymbol{x}_m, \boldsymbol{y}_m)$$

where mc_{jm} is firm j's marginal cost of providing one room in market m, and M_m is the market size. Similar to utility, I allow marginal cost to vary with both observed and unobserved variables and further parameterize it as

$$\log(mc_{jm}) = w_{jm}^{o'}\gamma + w_{jm}^{u}$$

The first-order condition of the profit maximization problem implies the pricing rule

$$p_{jm} = mc_{jm} + \frac{1}{\alpha(1 - s_{jm}(\boldsymbol{p}_m|\boldsymbol{x}_m, \boldsymbol{y}_m))} \tag{1.4}$$

The pricing rules of all entrants in each market m consist a system of equations with the price vector as unknowns. Solving this equation system gives the equilibrium price $p_{jm}(\boldsymbol{x}_m, \boldsymbol{w}_m, \boldsymbol{y}_m)$. Plugging it into the equilibrium market share function, we have $s_{jm} = s_{jm}(\boldsymbol{x}_m, \boldsymbol{w}_m, \boldsymbol{y}_m)$

Entry stage The full information assumption indicates that firms' entry decisions depend on the realized value of variable profit π_{jm}. Then as long as it offsets the fixed cost fc_{jm}, firms will choose to join price competition. [16] Therefore

[15] There is a slight abuse of notation: $\boldsymbol{p}_m$ is not a J-dimension vector as $\boldsymbol{x}_m$ and $\boldsymbol{y}_m$. Since only entrants have a well-defined price, its dimension should be the same as the cardinality of the entrant set $\mathcal{F}_m$.

[16] At the estimation stage, I use panel data and the hotels are assumed to make the entry decisions independently in each period. For the incumbents, the decision can be think of as whether to stay

$$y_{jm} = 1[(p_{jm}(\boldsymbol{x}_m, \boldsymbol{w}_m, \boldsymbol{y}_m) - mc_{jm})M_m s_{jm}(\boldsymbol{x}_m, \boldsymbol{w}_m, \boldsymbol{y}_m) - fc_{jm} \geq 0]. \qquad (1.5)$$

Again fc_{jm} is defined as a log linear function of both the observed and the unobserved variables,

$$\log(fc_{jm}) = z^{o'}_{jm}\delta + z^{u}_{jm}$$

The entry function Eq.(1.5) highlights that when making their entry decisions, firms consider its impact on the price competition outcome, which makes Eq.(1.3)-(1.5) a system of simultaneous equations. Accordingly, $(\boldsymbol{p}_m, \boldsymbol{s}_m, \boldsymbol{y}_m)$ is an equilibrium if it satisfies Eq.(1.3)-(1.5).

1.3.3 Selection and multi-agent discrete choice model

Models with endogenous entry are subject to the canonical selection issue, i.e. entrants may have unobservable advantages in the price competition which might be correlated with other observable product characteristics, mathematically, it means $E[x^u_{jm}|x^o_{jm}, p_{jm}, y_{jm} = 1] \neq 0$ and $E[w^u_{jm}|x^o_{jm}, p_{jm}, y_{jm} = 1] \neq 0$. As mentioned earlier, there are two sources of selection. First, firms are assumed to know the realization of the demand and marginal cost shocks at the entry stage. Second, the unobserved shocks in the three equations are correlated, thus even without the full information assumption, x^u_{jm} and w^u_{jm} should be correlated with $y_{jm} = 1$ through z^u_{jm}.

or not. It is reasonable to concern about the stickiness of the decision across periods, i.e., it is unlikely for a new entrant to exit right in the next period. In that regard, I first assume that most of the dependence has been captured by the explanatory variables. In other words, we would see an entrant stay in the market for a while as long as there is no significant change in those variables, which resembles the reality. Then I abstract from any additional cost related to exit. One may also worry about whether it is possible for a parent company to respond to a negative profit by quitting in any period. Although a franchise agreement usually has a decade-long horizon, the franchiser (namely the parent company has the right to terminate the agreement, effective on the date stated in their written notice, for reason such as the frachisee fails to pay them any fees.)

Consequently, it is wrongful to estimate the price competition stage first and then plug the estimates into the entry inequality for a second-stage estimation, because the estimates from the price competition model should have selection bias as a wage regression only considering the employed labor force. Furthermore, estimating the entry stage first and then addressing the selection problem in the pricing stage using control function method (e.g. Heckman correction in the single-agent setting) is also infeasible, because using a structural demand model to determine the profits makes the impact two-ways. Therefore, regardless of estimation tools, researchers should estimate the two stages simultaneously and consider the vector of the endogenous variables in both stages as the dependent variable, i.e. (p_{jm}, s_{jm}, y_{jm}). Accordingly, I impose a joint normality assumption on $(x^u_{jm}, w^u_{jm}, z^u_{jm})$ to facilitate future estimation, i.e.,

$$(x^u_{jm}, w^u_{jm}, z^u_{jm}) \sim \mathcal{N}(0, \Omega), \quad i.i.d \quad \forall j, m$$

Also, since all firms make entry decisions at the same time, for any firm j, its entry decision y_{jm} depends on the choices of all its rivals, $\boldsymbol{y}_{-jm}$, through its effect on the price competition outcome, as shown in Eq.(1.5). Therefore, I estimate this model as a multi-agent game. Specifically, instead of focusing on each firm, I construct moments based on collections of the endogenous variables across all firms, i.e. $(\boldsymbol{p}_m, \boldsymbol{s}_m, \boldsymbol{y}_m)$. More details are presented in Section 1.4.

Finally, the observations are constructed at the monthly level, which is essentially assuming that the two-stage game is played in every month. This frequency is not unrealistic. By my definition, entry can be fulfilled by an owner company switching to be another brand's franchisee, which has a much lower cost than constructing a new building. [17] Correspondingly, for incumbents, "enter" means stay in the market,

[17] In my data sample, there are 21 new entrants in Texas, 14 of them were from changes of affiliation and the rest 7 were new properties.

while "not enter" could be the termination of a franchise agreement. According to the form of Hyatt Place's franchise agreement,

> The franchiser has the right to terminate the franchise agreement, effective on the date stated in their written notice, for reason such as the franchisee fails to pay them or any of their affiliates any fees.

Therefore, in theory, the franchise agreement can be terminated at any point during its term, for which having a negative profit seems to be a plausible trigger. That said, if the market conditions, as represented by the exogenous variables, have no significant changes, we would rarely see an incumbent exit the market, which is consistent with what a long-term agreement would entail in the real world.

1.4 Estimation strategy

I use moment inequality estimation to accommodate any potential multiple equilibria at the entry stage. Intuitively, suppose there are two heterogeneous hotel chains, either one can make positive profits as a monopolist but not as a duopolist. Therefore, a Nash equilibrium concept would give two equilibria. We observe one of chains enter, but the probability of each entering is unspecified by a model without an assumption on the equilibrium selection mechanism. To circumvent any arbitrary assumptions on such a mechanism, I use moment inequality estimation. Its idea is to restrict the empirical probability within a range generated by the probability of a sufficient condition and the probability of a necessary condition, and then constructing the identified set by collecting all the parameter values that yield these inequality relationships.

Although moment inequality estimation is robust to any equilibrium selection mechanism, it is not widely applied by researchers. One of the main reasons is the computational challenges arising from using numerical method to approximate the probability bounds that have no closed forms. In particular, under the CMT framework, the computational burden increases exponentially with the number of players. To allow all the 11 important hotel chains in the estimation, I propose an innovative lower bound that significantly reduces the computation time. The section below explains the origin of the computational challenge, the construction of the moment inequalities along with the estimation procedure in more details.

1.4.1 Moment inequalities and the dominant strategy equilibrium lower bound

The parameters to be estimated are $\theta = (\beta', \alpha, \gamma', \delta', \omega')'$, where ω is the upper-triangle components of Ω's Choleski factor L, i.e. $LL' = \Omega$. The market subscript m is omitted throughout this section when there is no ambiguity. Given θ and the observed variables $\mathcal{O} = (\boldsymbol{x}^o, \boldsymbol{w}^o, \boldsymbol{z}^o)$, there is a correspondence from any unobserved variables $\mathcal{U} = (\boldsymbol{x}^u, \boldsymbol{w}^u, \boldsymbol{z}^u)$ to the set of market structures $\boldsymbol{y}$ that can be supported as an equilibrium, which I denote as $\mathcal{Y} = C(\mathcal{U}|\mathcal{O}, \theta)$. Notably, although the outcome variables include equilibrium prices $\boldsymbol{p}$, market shares $\boldsymbol{s}$ and market structure $\boldsymbol{y}$, I focus the discussion on $\boldsymbol{y}$ because with any given $(\boldsymbol{y}, \mathcal{U})$, the equilibrium $(\boldsymbol{s}, \boldsymbol{p})$ are determined by Eq.(1.3) and Eq.(1.4).

I define the region of $\mathcal{U}$ where $\boldsymbol{y}$ is the unique equilibrium as $R^u_{\boldsymbol{y}}(\mathcal{O}, \theta) = \{\mathcal{U} : C(\mathcal{U}|\mathcal{O}, \theta) = \boldsymbol{y}\}$ and the region where it is one of multiple equilibria as $R^m_{\boldsymbol{y}}(\mathcal{O}, \theta) = \{\mathcal{U} : \boldsymbol{y} \in C(\mathcal{U}|\mathcal{O}, \theta), C(\mathcal{U}|\mathcal{O}, \theta)/\boldsymbol{y} \neq \emptyset\}$. Then the probability that $\boldsymbol{y}$ is the observed equilibrium is

$$\Pr(\boldsymbol{y}|\mathcal{O}, \theta) = \Pr(\mathcal{U} \in R^u_{\boldsymbol{y}}(\mathcal{O}, \theta)) + \int \Pr(\boldsymbol{y}|\mathcal{U} \in R^m_{\boldsymbol{y}}(\mathcal{O}, \theta)) 1[\mathcal{U} \in R^m_{\boldsymbol{y}}(\mathcal{O}, \theta)] dF_{\mathcal{U}}.$$

where $\Pr(\boldsymbol{y}|\mathcal{U} \in R^m_{\boldsymbol{y}}(\mathcal{O}, \theta))$ is the selection rule. Regardless of what the rule is, the fact that it is a proper probability hence lies in $[0, 1]$ implies

$$\Pr(\mathcal{U} \in R^u_{\boldsymbol{y}}(\mathcal{O},\theta)) \leq \Pr(\boldsymbol{y}|\mathcal{O},\theta) \leq \Pr(\mathcal{U} \in R^u_{\boldsymbol{y}}(\mathcal{O},\theta) \cup R^m_{\boldsymbol{y}}(\mathcal{O},\theta)). \tag{1.6}$$

where the right inequality holds because $R^u_{\boldsymbol{y}}(\mathcal{O},\theta)$ and $R^m_{\boldsymbol{y}}(\mathcal{O},\theta)$ are disjoint by definition.

Because neither the lower probability bound $LB(\boldsymbol{y}|\mathcal{O},\theta) = \Pr(\mathcal{U} \in R^u_{\boldsymbol{y}}(\mathcal{O},\theta))$ nor the upper probability bound $UB(\boldsymbol{y}|\mathcal{O},\theta) = \Pr(\mathcal{U} \in R^u_{\boldsymbol{y}}(\mathcal{O},\theta) \cup R^m_{\boldsymbol{y}}(\mathcal{O},\theta))$ has a closed-form expression, CMT use Monte Carlo approximation. Specifically, for a given parameter vector θ, one need to take many random draws from $\mathcal{N}(0,\Omega)$ and for each draw solve the equilibrium $(\boldsymbol{p},\boldsymbol{s})$ for each possible market structure $\boldsymbol{y}$. If no firm has the incentive to deviate from its entry decision given the net profit, $\boldsymbol{y}$ is a Nash equilibrium. Then $LB(\boldsymbol{y}|\mathcal{O},\theta)$ and $UB(\boldsymbol{y}|\mathcal{O},\theta)$ can be obtained by counting the occurrence of equilibrium $\boldsymbol{y}$ being the unique vs. among the multiple. Recall that $\boldsymbol{y}$ is a J-dimension vector of binary indicators, so the number of possible market structures to check is 2^J, therefore the computational burden increases exponentially with the number of potential entrants.

That said, look at $LB(\boldsymbol{y}|\mathcal{O},\theta)$ and $UB(\boldsymbol{y}|\mathcal{O},\theta)$ separately, the upper probability bound is relatively easy to simulate because most of the possible market structures are irrelevant. In particular, it requires only the observed market structure to be an equilibrium regardless of other possible market structures, which can be guarantee by letting the entry condition in Eq.(1.5) hold for the observed $\boldsymbol{y}$, i.e.,

$$R^u_{\boldsymbol{y}}(\mathcal{O},\theta) \cup R^m_{\boldsymbol{y}}(\mathcal{O},\theta) = \{\mathcal{U} : y_j = 1[\pi_j(Y_j = 1, \boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j}) - fc_j \geq 0], \forall j\}. \tag{1.7}$$

Here, I use upper case Y to distinguish the variable from its specific value y, and suppress the dependence on $(\mathcal{O},\mathcal{U},\theta)$ in the entry condition for simplification of notation. Specifically, for any entrant j, $(Y_j = 1, \boldsymbol{Y}_{-j} = \boldsymbol{y}_{-j})$ is exactly the observed

$\boldsymbol{y}$; while for any non-entrant k, it indicates the counterfactual market structure with k as an entrant in addition to the current ones, which can be rewritten as $\boldsymbol{y}+\mathbf{1}_k$, where $\mathbf{1}_k$ is a vector of the same length as $\boldsymbol{y}$ with k^{th} entry equal to 1 and all other entries equal to 0. Therefore, to calculate this upper bound, the number of equilibria $(\boldsymbol{p}, \boldsymbol{s})$ to solve equals the one observed equilibrium plus a counterfactual market structure for each non-entrant, i.e. $1 + |\{k : y_k = 0\}|$, which is at most J.

However, constructing the lower probability bound requires solving the equilibrium $(\boldsymbol{p}, \boldsymbol{s})$ for every possible market structure. Intuitively, for a random draw $\mathcal{U}$, to make sure that $\boldsymbol{y}$ is the unique equilibrium, one needs to confirm that no other market structure can also be supported an equilibrium, which could induce intimidating computational burden when the number of possible market structures is huge. Since the Texas lodging industry has 11 chain firms therefore there are $2^{11} = 2048$ possible market structures, to keep this estimation framework computationally tractable, I propose to use the probability that $\boldsymbol{y}$ is the strictly dominant strategy equilibrium as the lower probability bound.

Dominant strategy equilibrium probability as lower bound By Definition 8.B.1 in Mas-Colell et al. (1995), an entry strategy is strictly dominant for firm j if the entry condition holds as a strict inequality regardless of the strategies that its rivals might play. It is obvious that a strategy profile consisting of strictly dominant strategies of each firm is a Nash equilibrium, and such a strategy profile is what is called a dominant strategy (DS) equilibrium.

Let $R^{ds}_{\boldsymbol{y}}(\mathcal{O}, \theta)$ denote the region of $\mathcal{U}$ where $\boldsymbol{y}$ is a DS equilibrium, i.e.,

$$R^{ds}_{\boldsymbol{y}}(\mathcal{O}, \theta) = \{\mathcal{U} : y_j = 1[\pi_j(Y_j = 1, \boldsymbol{Y}_{-j}) - fc_j \geq 0], \forall j, \forall \boldsymbol{Y}_{-j}\}.$$

Since strictly dominant strategy is unique for each firm if existing, DS equilibrium is sufficient for being unique. [18,19] Accordingly,

$$\Pr(\mathcal{U} \in R^{ds}_{\boldsymbol{y}}(\mathcal{O},\theta)) \leq \Pr(\mathcal{U} \in R^{u}_{\boldsymbol{y}}(\mathcal{O},\theta)),$$

therefore we can use $\Pr(\mathcal{U} \in R^{ds}_{\boldsymbol{y}}(\mathcal{O},\theta))$ to replace the lower bound in Eq.(1.6).

The reason $\Pr(\mathcal{U} \in R^{ds}_{\boldsymbol{y}}(\mathcal{O},\theta))$ is more convenient to calculate is because it by definition does not involve the complicated inter-player interaction. In particular, in the context of an entry game with multinomial logit demand, since all products are assumed to be each other's substitute, a firm's profit is always reduced by having extra entrants. Therefore, entry is dominant if and only if the firm finds it is profitable to enter even if all its rivals also decide to enter; on the contrary, opting out is dominant if a firm cannot survive even with monopoly profit. Formally, the condition in the definition of $R^{ds}_{\boldsymbol{y}}(\mathcal{O},\theta)$ can be rewritten as

$$y_j = 1[\pi_j(Y_j = 1, \boldsymbol{Y}_{-j} = y_j) - fc_j \geq 0]. \tag{1.8}$$

As a result, similar to the upper bound, the number of equilibrium $(\boldsymbol{p}, \boldsymbol{s})$ to solve now is at most J: for any entrant j, $(Y_j = 1, \boldsymbol{Y}_{-j} = y_j)$ indicates $\boldsymbol{Y}_j = \mathbf{1}$, i.e., all firms decide to enter; while for any non-entrant k, $(Y_k = 1, \boldsymbol{Y}_{-k} = \mathbf{0}) = \mathbf{1}^k$, a monopoly market with firm k as the only entrant. Eventually, in this case where $J = 11$, the number of $(\boldsymbol{p}, \boldsymbol{s})$ to solve for each market m and each θ decreases from $2^J = 2{,}048$ to at most $2 \times J = 22$.

Although the numerical gap between the two lower bounds may cause the loss of sharpness in identification, i.e., the identified set in theory could be wider if using the less restrictive DS lower bound. However, a simulation exercise in Appendix A.4

[18] See `https://homepages.cwi.nl/~apt/stra/ch3.pdf` for a proof.

[19] The existence of a strictly dominant strategy is guaranteed by the infinite domain of $\mathcal{U}$. Intuitively, given any $(\mathcal{O},\theta)$, there should be some very high demand shock x^u_j, and very low marginal and fixed cost shocks w^u_j, z^u_j that make entry always profitable for firm j no matter what the rivals' strategies are.

shows that it is not necessarily true in practice.

A two-player example Now I use an abstract 2-player entry game to fix the implications of the key concepts in moment inequality estimation such as the multiple equilibria issue and probability bounds. To do that, let us ignore price competition for a moment and consider that the two firms decide whether to enter a market based on a reduced-form linear profit function,

$$\begin{cases} y_1 &= 1[\alpha - \beta_2 y_2 - c_1 \geq 0] \\ y_2 &= 1[\alpha - \beta_1 y_1 - c_2 \geq 0], \end{cases} \tag{1.9}$$

where α is a constant, β_1 and β_2 are assumed negative to mimic the competition effect in multinomial discrete choice model.

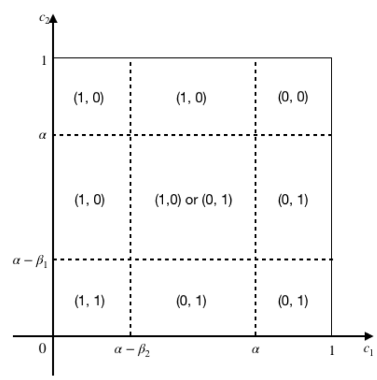
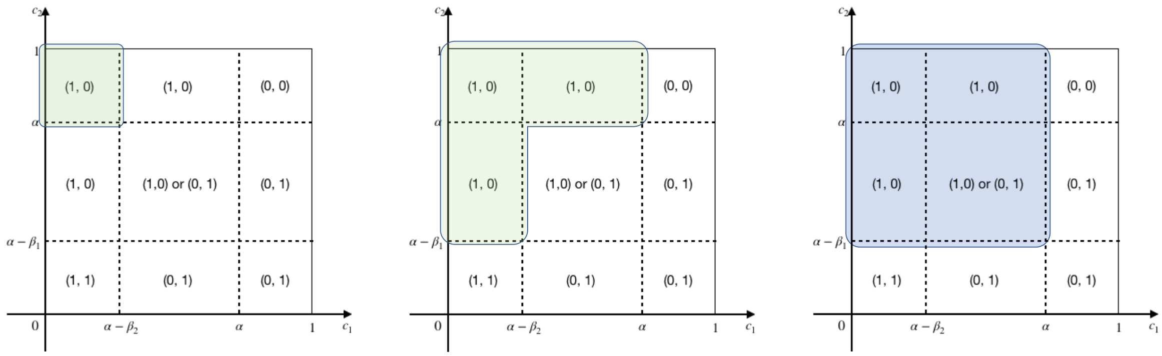


Figure 1·5: 2×2 market structure correspondence

Furthermore, I assume that the cost shock $c_j \in [0, 1]$ and follows some distribution $\mathcal{G}$, *i.i.d* for $j = 1, 2$. Then the equilibrium market structure is a correspondence from $[0, 1] \times [0, 1]$, as shown in Figure 1·5.



(a) Lower bound: DS Eq. **(b)** Lower bound: Unique **(c)** Upper bound

Figure 1·6: Probability bounds in a two-player example

A market structure $\boldsymbol{y} = (y_1, y_2)$ is a Nash equilibrium if Eq.(1.9) holds. When $\boldsymbol{c} = (c_1, c_2) \in [\alpha - \beta_2, \alpha] \times [\alpha - \beta_1, \alpha]$, both $(1, 0)$ and $(0, 1)$ are Nash equilibria, and $\Pr(1, 0)$ is unspecified in a model without equilibrium selection assumption. As a result, MLE is infeasible, for which moment inequality estimation can be remedial.

Figure 1·6 depicts the probability bounds for moment inequality estimation. Specifically, In Figure 1·6(c) the upper bound region is highlighted in blue, formally it is $R^u_{(1,0)} \cup R^m_{(1,0)} = [0, \alpha] \times [\alpha - \beta_1, 1]$. The upper bound only requires that $(1, 0)$ is an equilibrium, therefore it can be easily obtained by checking whether $(1, 0)$ satisfies Eq.(1.9). However, for $R^u_{(1,0)}$, the yellow region in Figure 1·6(b), a researcher has to check Eq.(1.9) for each of $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$, then carefully take out the region where other equilibria are involved. In practice, this region usually does not have a closed form, for which Monte Carlo simulation can be helpful. The procedure usually consists of the following steps: (1) take RS random draws from the assumed distribution of $\boldsymbol{c}$; (2) for each draw $\boldsymbol{c}^r$, check Eq.(1.9) for all possible outcomes and count the number of draws that support only $(1, 0)$ as an equilibrium, denoted as $RS^u_{(1,0)}$; (4) use $RS^u_{(1,0)}/RS$ to approximate $\Pr(\boldsymbol{c} \in R^u_{(1,0)})$. It is not hard to imagine that this procedure will get more complicated when there are more players and the price competition stage is structural therefore numerically solving for the equilibrium prices and market shares is involved.

To tackle the computational challenge, I propose the DS lower probability in Figure 1·6(a). Given $\beta_1, \beta_2 \leq 0$, $y_1 = 1$ and $y_2 = 0$ are strictly dominant strategy indicates that firm 1 wants to enter even if firm 2 also enters, and firm 2 does not want to enter even if it would be the monopoly in the market. By Eq.(1.9), that indicates $\alpha - \beta_2 - c_1 \geq 0$, $\alpha - c_2 \leq 0$; then, $R^{ds}_{(1,0)}$ can be easily solved as $[0, \alpha - \beta_2] \times [\alpha, 1]$.

1.4.2 Objective function in practice

With a little abuse of notation, I let $\Pr(\mathcal{O}, \theta)$, $LB(\mathcal{O}, \theta)$ and $UB(\mathcal{O}, \theta)$ to each denote a vector that collects the empirical probability $\Pr(\boldsymbol{y}|\mathcal{O}, \theta)$ and probability bounds $LB(\boldsymbol{y}|\mathcal{O}, \theta)$, $UB(\boldsymbol{y}|\mathcal{O}, \theta)$ across all observed $\boldsymbol{y}$ respectively, i.e. $\Pr(\mathcal{O}, \theta) = \{\Pr(\boldsymbol{y}|\mathcal{O}, \theta)\}_{\boldsymbol{y}}$. Then any parameter values that confine $\Pr(\mathcal{O}, \theta)$ within $LB(\mathcal{O}, \theta)$ and $UB(\mathcal{O}, \theta)$ therefore make the following deviation function $Q(\theta)$ equal to 0 should be included into the identified set S_θ.

$$Q(\theta) = \int ||\big(LB(\mathcal{O}, \theta) - \Pr(\mathcal{O}, \theta)\big)_+|| + ||\big(UB(\mathcal{O}, \theta) - \Pr(\mathcal{O}, \theta)\big)_-||dF_{\mathcal{O}}, \tag{1.10}$$

where $(a_1, a_2, ..., a_i, ...)_+ = (a_1 1[a_1 \geq 0], a_2 1[a_2 \geq 0], ..., a_i 1[a_i \geq 0], ...)$, $(a_1, a_2, ..., a_i, ...)_- = (a_1 1[a_1 \leq 0], a_2 1[a_2 \leq 0], ..., a_i 1[a_i \leq 0], ...)$ and $|| \cdot ||$ is the Euclidean norm.

Since the probabilities are determined by unobservables, I rewrite Eq.(1.10) with a focus on $(\boldsymbol{x}^u, \boldsymbol{w}^u, \boldsymbol{z}^u)$ in order to outline the procedure of evaluating the objective function at a given θ, detailed estimation steps are discussed in Appendix A.2.

First, I rewrite the empirical probability $\Pr(\boldsymbol{y}|\mathcal{O}, \theta)$ as

$$\Pr(\boldsymbol{x}^u \leq \boldsymbol{t}^x, \boldsymbol{w}^u \leq \boldsymbol{t}^w, \boldsymbol{z}^u(\boldsymbol{y})|\mathcal{O}, \theta). \tag{1.11}$$

The idea again is mapping the observed quantities $(\boldsymbol{p}, \boldsymbol{s}, \boldsymbol{y})$ to the unobservables $\mathcal{U}$.

Conditional on $(\mathcal{O}, \theta)$, I recover $\boldsymbol{x}^u$ and $\boldsymbol{w}^u$ by inverting Eq.(1.3) and Eq.(1.4) based on the observed equilibrium $(\boldsymbol{p}, \boldsymbol{s})$, and then consider the their cumulative densities at some carefully selected thresholds $\boldsymbol{t}^x$ and $\boldsymbol{t}^w$. $\boldsymbol{z}^u(\boldsymbol{y})$ is a simplified way of writing $\{\boldsymbol{z}^u :$ $\boldsymbol{y}$ is the observed equilibrium$|(\boldsymbol{x}^u, \boldsymbol{w}^u)\}$. Then the empirical probability in Eq.(1.11) can be obtained by counting the occurrence of observing $\{\boldsymbol{x}^u \leq \boldsymbol{t}^x, \boldsymbol{w}^u \leq \boldsymbol{t}^w, \boldsymbol{y}\}$. [20]

As analogies to Eq.(1.11), the theoretical probability bounds $LB(\boldsymbol{y}|\mathcal{O}, \theta)$ and $UB(\boldsymbol{y}|\mathcal{O}, \theta)$ are constructed by replacing $\boldsymbol{z}^u(\boldsymbol{y})$ with $\{\boldsymbol{z}^u \in R^{ds}_{\boldsymbol{y}}|(\boldsymbol{x}^u, \boldsymbol{w}^u)\}$ and $\{\boldsymbol{z}^u \in R^u_{\boldsymbol{y}} \cup R^m_{\boldsymbol{y}}|(\boldsymbol{x}^u, \boldsymbol{w}^u)\}$ correspondingly. In particular, since $\Pr(\boldsymbol{z}^u \in R^{ds}_{\boldsymbol{y}}|\boldsymbol{x}^u, \boldsymbol{w}^u)$ and $\Pr(\boldsymbol{z}^u \in R^u_{\boldsymbol{y}} \cup R^m_{\boldsymbol{y}}|\boldsymbol{x}^u, \boldsymbol{w}^u)$ have a closed form under the joint normality assumption, I consider the joint probability of $(\boldsymbol{x}^u, \boldsymbol{w}^u)$ and the conditional probability of $\boldsymbol{z}^u|(\boldsymbol{x}^u, \boldsymbol{w}^u)$ separately to reduce the dimension of simulation. Specifically, by the $i.i.d$ assumption on $z^u_j, \forall j$, the market-level conditional probability can be written as the product of the firm-level ones. Take the lower bound as an example, by Eq.(1.8),

$$\Pr(\boldsymbol{z}^u \in R^{ds}_{\boldsymbol{y}}|\boldsymbol{x}^u, \boldsymbol{w}^u) = \prod_j \Pr(z^u_j \in R^{ds}_{y_j}|\boldsymbol{x}^u, \boldsymbol{w}^u).$$

where $R^{ds}_{y_j}$ is the one-dimension range of z^u_j where y_j is a dominant strategy equilibrium. Specifically, in terms of the firm-level conditional probability, for an entrant $j,\ s.t.\ y_j = 1$, entry is a dominant strategy indicates that firm j makes positive profits even if all its rivals are in the market, i.e.,

$$\begin{aligned} \Pr(z^u_j \in R^{ds}_{y_j}|\boldsymbol{x}^u, \boldsymbol{w}^u) &= \Pr(\pi_j(\mathbf{1}) - exp(z^{o'}_j\delta + z^u_j) \geq 0|\boldsymbol{x}^u, \boldsymbol{w}^u) \\ &= \tilde{\Phi}(log(\pi_j(\mathbf{1})) - z^{o'}_j\delta) \end{aligned}$$

where $\tilde{\Phi}$ is the CDF of $z^u_j|(\boldsymbol{x}^u, \boldsymbol{w}^u)$. On the contrary, for a non-entrant $k,\ s.t.\ y_k = 0$, staying out of the market is a dominant strategy if even the monopoly profit is not positive, i.e.,

[20] Similar to $(\boldsymbol{p}, \boldsymbol{s})$, the dimension of $\boldsymbol{x}^u$, $\boldsymbol{w}^u$, $\boldsymbol{t}^x$, $\boldsymbol{t}^w$ equals the number of entrants in each market.

$$\begin{aligned}\Pr(z_k^u \in R_{y_k}^{ds}|\boldsymbol{x}^u, \boldsymbol{w}^u) &= \Pr(\pi_k(\mathbf{1}^k) - exp(z_k^{o'}\delta + z_k^u) \leq 0|\boldsymbol{x}^u, \boldsymbol{w}^u) \\ &= 1 - \tilde{\Phi}(log(\pi_k(\mathbf{1}^k)) - z_k^{o'}\delta)\end{aligned}$$

In summary, the market-level lower bound probability can be rewritten as

$$\Pr(\boldsymbol{z}^u \in R_{\boldsymbol{y}}^{ds}|\boldsymbol{x}^u, \boldsymbol{w}^u) = \prod_j \tilde{\Phi}(log(\pi_j(\mathbf{1})) - z_j^{o'}\delta)^{y_j}(1 - \tilde{\Phi}(log(\pi_j(\mathbf{1}^j)) - z_j^{o'}\delta))^{1-y_j}$$

Similarly, the market-level upper bound probability has the following form,

$$\Pr(\boldsymbol{z}^u \in R_{\boldsymbol{y}}^u \cup R_{\boldsymbol{y}}^m|\boldsymbol{x}^u, \boldsymbol{w}^u) = \prod_j \tilde{\Phi}(log(\pi_j(\boldsymbol{y})) - z_j^{o'}\delta)^{y_j}(1 - \tilde{\Phi}(log(\pi_j(\boldsymbol{y} + \mathbf{1}^j)) - z_j^{o'}\delta))^{1-y_j}$$

Finally, the theoretical probability bounds can be obtained by integrating $(\boldsymbol{x}^u, \boldsymbol{w}^u)$ over up to $(\boldsymbol{t}^x, \boldsymbol{t}^w)$, i.e.,

$$\begin{aligned}&LB(\boldsymbol{x}^u \leq \boldsymbol{t}^x, \boldsymbol{w}^u \leq \boldsymbol{t}^w, \boldsymbol{z}^u(\boldsymbol{y})|\mathcal{O}, \theta) \\ = &\int^{(\boldsymbol{t}^x, \boldsymbol{t}^w)} \Pr(\boldsymbol{z}^u \in R_{\boldsymbol{y}}^{ds}|\boldsymbol{x}^u, \boldsymbol{w}^u)dF(\boldsymbol{x}^u, \boldsymbol{w}^u)\end{aligned} \tag{1.12}$$

$$\begin{aligned}&UB(\boldsymbol{x}^u \leq \boldsymbol{t}^x, \boldsymbol{w}^u \leq \boldsymbol{t}^w, \boldsymbol{z}^u(\boldsymbol{y})|\mathcal{O}, \theta) \\ = &\int^{(\boldsymbol{t}^x, \boldsymbol{t}^w)} \Pr(\boldsymbol{z}^u \in R_{\boldsymbol{y}}^u \cup R_{\boldsymbol{y}}^m|\boldsymbol{x}^u, \boldsymbol{w}^u)dF(\boldsymbol{x}^u, \boldsymbol{w}^u)\end{aligned} \tag{1.13}$$

In practice, I take random draws from the joint distribution of $(\boldsymbol{x}^u, \boldsymbol{w}^u)$, average $\Pr(\boldsymbol{z}^u \in R_{\boldsymbol{y}}^{ds}|\boldsymbol{x}^u, \boldsymbol{w}^u)$ and $\Pr(\boldsymbol{z}^u \in R_{\boldsymbol{y}}^u \cup R_{\boldsymbol{y}}^m|\boldsymbol{x}^u, \boldsymbol{w}^u)$ over those draws below $(\boldsymbol{t}^x, \boldsymbol{t}^w)$, and then multiply this average by $\Pr(\boldsymbol{x}^u \leq \boldsymbol{t}^x, \boldsymbol{w}^u \leq \boldsymbol{t}^w)$.

The rest of the estimation procedure closely follows Ciliberto and Tamer (2009) and mainly consist of 2 parts. I first construct the sample analog of $Q(\theta)$, given by

$$Q_M(\theta) = \frac{1}{M}\sum_{m=1}^{M} ||\big(LB(\mathcal{O}_m,\theta) - \Pr(\mathcal{O}_m,\theta)\big)_+|| + ||\big(UB(\mathcal{O}_m,\theta) - \Pr(\mathcal{O}_m,\theta)\big)_-||,$$

where M is the total number of markets. I search for the minimums of $Q_M(\theta)$ by evaluating it at many points, following the MCMC approach in Chernozhukov and Hong (2003).

Second, I construct the 95% confidence sets using the inference method that is robust to partial identification, following Chernozhukov, Hong and Tamer (2007). Specifically, the identified set is defined in a level set fashion, $\hat{S}_\theta = \{\theta : MQ_M(\theta) \leq v_M\}$, where $M \to \infty$ and $v_M \to 0$. The idea of inference is to find a cutoff c such that the associated level set $CS(c) = \{\theta : M(Q_M(\theta) - \min_\theta Q_M(\theta)) \leq c\}$ covers the identified set $\hat{S}_\theta$ with the desired significance level.[21] I approximate the distribution of $MQ_M(\theta)$ by bootstrap, let c be the 95% percentile of the simulated distribution, then construct $CS(c)$ by collecting all the θ's evaluated in the first step that satisfy $M(Q_M(\theta) - \min_\theta Q_M(\theta)) \leq c$. More details regarding the MCMC searching procedure and the inference method can be found in Appendix A.2.

1.5 Results

1.5.1 Descriptive evidences

In this section, I provide reduced-form evidence for the impact of Marriott-Starwood merger on the market equilibrium prices and quantities.

I extend the span of the data to Dec 2018 and employ a difference-in-difference design to identify the merger effect. The regression equation is as follows,

[21] Even though in theory parameter values in the identified set should have $Q(\theta) = 0$, it is not necessarily true in practice due to potential misspecification. See Chernozhukov, Hong and Tamer (2007).

Table 1.5: Difference-in-difference estimates

	All		No rivals entered		Rival entered	
	p_{jm}	q_{jm}	p_{jm}	q_{jm}	p_{jm}	q_{jm}
	(1)	(2)	(3)	(4)	(5)	(6)
M&S hotels in M&S market × Post-merger	2.298***	-15.258***	1.331	-13.673***	10.474***	-17.153*
	(1.009)	(3.590)	(1.097)	(3.732)	(2.415)	(9.164)
Year-month FE	×	×	×	×	×	×
Neighborhood FE	×	×	×	×	×	×
Sub.market-firm FE	×	×	×	×	×	×
Observations	14,341	14,341	13,333	13,333	1,008	1,008
Note:				*p<0.1; **p<0.05; ***p<0.01		

$$
\begin{aligned}
Outcome_{jmt} &= \xi_j + \eta_m + \tau_t + \mu \times D_{jmt} + \nu_{jmt} \\
D_{jmt} &= 1[j \in \{M,S\}, \{M,S\} \in \mathcal{F}_{mt}] \times \textit{Post-merger}_{mt}
\end{aligned}
$$

D_{jmt} is the treatment variable. ξ_j, η_m, τ_t are the firm, market (neighborhood) and time (year-month) fixed effects correspondingly. Additionally, I let the firm effect ξ_j to be sub-city specific. ν_{jmt} is a random shock that is exogenous to the covariates and $i.i.d$ across $\forall j, m, t$. I use M and S as abbreviations for Marriott and Starwood. Specifically, I consider the treatment group as the Marriott and Starwood hotels, i.e., $j \in \{M, S\}$, in markets where they coexist, i.e., $\{M, S\} \in \mathcal{F}_{mt}$, then the control group would be any other hotel or Marriott and Starwood hotels in other types of markets. In this way, I focus on the direct effects of market concentration increase on the merging parties' prices and quantities and take the potential spillover effects in other markets as second order.

The first two columns of Table 1.5 show that overall, Marriott and Starwood hotels increased their with-merger prices by \$2.298, and the number of room-nights sold decreased by -15.258 accordingly. When we separately look at the markets with rivals entered after the merger vs. those without rivals entered, we see that the observational increase in prices is mainly driven by the former type of markets. Regarding the direction of causality, since extra competition, ceteris paribus, would always lower every incumbent's price in the market, it should be the price increase that induces new rivals. Loosely speaking, if the merging parties increase their prices after the merger, a rival firm who previously finds entry unprofitable now can also increase its price by some amount less than the merging parties do so that there would not be much loss in market share, but the extra variable profit gained could overturn its entry decision. Furthermore, the price increase in column (5) should be interpreted as a lower bound for the price effects in these markets, because the new entrants should already offset some of the increased market concentration.

Such differences in these two types of markets are important, because they motivate the endogenous entry setting in two ways. First, it proves that entry is not random. Firms are self-selected into markets based on profitability; therefore, estimation without an entry stage would be biased. Second, rivals can respond to the merger by entry and exit. Specifically, if the extra rivals who would enter when the merging parties increase their price are overlooked, the price increase and thereby the economic damage of the merger would be overestimated.

Table 1.6: Estimation results

	Exogenous Entry		Endogenous Entry
	MLE	Moment Equality	Moment Inequality (95% Confidence Set)
	(1)	(2)	(3)
Panel A: Demand Equation			
Constant	25.670***	25.806***	[8.933, 11.603]
	(0.6126)	(4.0166)	
Price	-0.224***	-0.231***	[-0.134, -0.119]
	(0.0052)	(0.0341)	
No. of affiliated upscale hotels	0.196***	0.195***	[0.127, 0.252]
(N^{uc})	(0.0404)	(0.0563)	
Panel B: Marginal Cost			
Constant	4.665***	4.664***	[4.822, 5.006]
	(0.0040)	(0.0659)	
No. of all affiliated hotels	0.012***	0.021***	[-0.023, -0.001]
(N^{all})	(0.0017)	(0.0103)	
Panel C: Fixed Cost			
Constant			[6.324, 7.388]
Rent			[0.155, 0.290]
Fixed effects			
Hilton			[0.308, 0.407]
IHG			[-0.295, -0.187]
Marriott			[0.337, 0.465]
Panel D: Variance-covariance			
Demand variance	5.072***	4.236***	[1.188, 1.633]
	(0.1282)	(0.9740)	
MC variance	0.090***	0.072***	[0.224, 0.587]
	(0.0023)	(0.0256)	
FC variance			[0.194, 0.631]
Demand-MC covariance	0.190***	0.179***	[0.147, 0.347]
	(0.0024)	(0.0361)	
Demand-FC covariance			[0.080, 0.100]
MC-FC covariance			[-0.206, -0.043]
Panel E: Market Power			
Median elasticity	-17.37	-17.88	[-10.35, -9.20]
Median markup	6.33	6.15	[10.62, 11.95]
No. of obs.	4,430	4,430	6,986
Note:			*p<0.1; **p<0.05; ***p<0.01

1.5.2 Exogenous entry model estimation

As for structural model analysis, I first estimate an exogenous entry model as a baseline. It consists of the pricing competition stage only, as specified by Eq.(1.3) and Eq.(1.4). The differences between this baseline and the endogenous entry model in estimates and counterfactual analysis would once again emphasize the importance of endogenizing the entry stage. In particular, I assume that the unobserved demand

and marginal cost shocks (x^u_{jm}, w^u_{jm}) follows a bivariate normal distribution $\mathcal{N}(0, \Omega')$, so that the setting can resemble the endogenous entry model to the greatest extent.

I estimate the model with two different methods – MLE and method of moment, which I refer to as "Moment Equality" to be an analog with the moment inequality estimation. In particular, because the multiple equilibria issue is eliminated by the observed market structure, MLE is feasible for the exogenous entry model. In particular, the log likelihood function I estimate is

$$LL(\mathcal{O}, \theta) = \sum_{j,m} \log f(x^u_{jm}, w^u_{jm}; \theta)$$

where $x^u_{jm} = \log(s_{jm}) - \log(s_{0m}) - x^{o'}_{jm}\beta + \alpha p_{jm}$ and $w^u_{jm} = \log(p_{jm} - \frac{1}{\alpha(1-s_{jm})}) - w^{o'}_{jm}\gamma$, which are from inverting Eq.(1.3) to Eq.(1.4). Their empirical values at any given parameter θ can be calculated by plugging in other observed quantities. $f(\cdot)$ is the probability density function of the bivariate normal distribution $\mathcal{N}(0, \Omega')$

I then estimate the same model using moment equalities as an informal examination for the idea of performing estimation with cumulative probabilities. Without the multiple equilibria issue, the empirical probability now has an exact model prediction, so the moment inequality estimation essentially degenerates to moment equality estimation. Specifically, the probability in Eq.(1.11) becomes $\Pr(\boldsymbol{x}^u \leq \boldsymbol{t}^x, \boldsymbol{w}^u \leq \boldsymbol{t}^w | \mathcal{O}, \theta)$, and the objective function is

$$\sum_{j,m} \left[F(t^x, t^w; \theta) - \hat{Pr}(x^u_{jm} \leq t^x, w^u_{jm} \leq t^w; \theta) \right]^2$$

where $F(t^x, t^w; \theta)$ is the cumulative density of $\mathcal{N}(0, \Omega')$ up to (t^x, t^w), $\hat{Pr}(x^u_{jm} \leq t^x, w^u_{jm} \leq t^w; \theta)$ is the empirical cumulative probability, which is the sample analog of Eq.(1.11).[22]

[22]Formally, it is defined as $\sum_{j,m} 1[x^u_{jm} \leq t^x, w^u_{jm} \leq t^w; \theta] / \sum_m |\mathcal{F}_m|$

Column (1)-(2) in Table 1.6 present the estimation results using these two methods. Across the two columns, the estimates are very close in magnitude. And not surprisingly, the least square estimates have higher standard errors than MLE, but all parameters achieve the same significance levels, which should dispel any concerns about identification and efficiency when using cumulative density for estimation. Given the similarity in magnitude, the following discussion focuses on the results from the moment equality estimation.

The most critical parameter is the price coefficient, which is -0.231. Combining with the observed market share and price of each firm, I translate the price coefficient into price elasticity and markup, and then report their medians across all firms and markets in Panel E. They together demonstrate a pretty competitive market. In particular, the median markup is \$6.15, which is less than 5% of the average ADR \$116.69. The price elasticity is high at -17.88 – 1% price increase, which is about \$1.16 on average, leads almost 20% market share loss.

The parameter on N^{uc} captures the network effect among the affiliated hotels in the same segment via the loyalty program. Within the upscale class segment, the effect of having an additional affiliated hotel is 0.195. I then divide it by the price coefficient to obtain the willingness to pay. The result shows that travelers are willing to pay \$1.18 more per room-night for one extra member in a firm's loyalty program.

The estimation of the marginal cost equation, unexpectedly, gives a positive estimate for N^{all}, which contradicts the intuition that affiliated hotels, regardless of classes, can share operational costs such as linen cleaning and guest supplies. However, this is likely to be a selection bias that can be corrected by endogenizing entry. More details can be found when we get to the endogenous model results.

Panel D reports the estimated covariance matrix of the unobservables. x^u_{jm} and w^u_{jm} are positively correlated, which is consistent with the intuition that a product with higher quality could be produced at a higher marginal cost.

1.5.3 Endogenous entry

Finally, I estimate the endogenous entry model using the strategy described in Section 1.4 and report the results in Column (3). For each parameter, I report the marginal confidence interval covering the identified set with 95% probability.

The first difference from Column (1)-(2) in Column (3) is the significantly lower constant term in the demand equation. Intuitively, the hotel rooms that could have been provided by non-entrants are supposed to be less appealing in general; therefore , the average utility of staying in a hotel room is taken down when they are incorporated into the framework.

Also, we see that neglecting firm's self-selection induces a downward bias in the price coefficient and an upward bias in the coefficient on N^{all}. That means, conditional on entry, the unobserved demand shock x^u_{jm} is negatively correlated with price p_{jm}, and the unobserved marginal cost shock w^u_{jm} is positively correlated with the number of all affiliated hotels, i.e. $cov(x^u_{jm}, p_{jm}|y_{jm} = 1) < 0$ and $cov(w^u_{jm}, N^{all}_{jm}|y_{jm} = 1) > 0$. Intuitively, given the high competition level indicated by the high price elasticity and the low markup, we would expect the entrants to be those who have a positive unobserved demand shock meanwhile can price lower than their rivals. One may find it counterintuitive, considering that firms with high demand tend to price higher. However, price is determined by both markup and marginal cost. A low marginal cost could dominate the upward pricing pressure from the positive demand shock. In particular, given the negative coefficient on N^{all}_{jm}, entrants with an extensive network, i.e. a large N^{all}_{jm}, could have a lower marginal cost therefore price lower even when facing a high marginal cost shock. That conforms to what the estimated positive correlation between the demand shock and the marginal cost shock and $cov(w^u_{jm}, N^{all}_{jm}|y_{jm} = 1) > 0$ implies, i.e., $cov(x^u_{jm}, N^{all}_{jm}|y_{jm} = 1) > 0$. And the fact that entrants have more affiliated hotels than an average firm can also be considered as supporting evidence, as Table 1.3 shows.

I also calculate the ranges of the median price elasticity and markup based on the confidence intervals. Specifically, for all parameters that are needed for calculating the elasticity and markup, i.e. demand constant, price and N^{uc}, I generate an arithmetic sequence of length ten from the lower bound to the upper bound and then construct a grid by letting these sequences cross with each other. For instance, letting the sequence $a = \{a_1, a_2\}$ cross with $b = \{b_1, b_2\}$, I get a grid with all their possible combinations, i.e. $\{(a_1, b_1), (a_2, b_1), (a_1, b_2), (a_2, b_2)\}$. Next, for each point on the grid, I calculate the price elasticity and the markup for each j, m, take the median over j's and m's, and then report the range of these medians across all evaluated points. As shown in Panel E, Column (3), they are [-10.35, -9.20] and [10.62, 11.95], a lower price elasticity and a higher markup than those in Column (1)-(2). In other words, overlooking the entry stage would underestimate firms' market power.

The endogenous entry model also recovers the fixed cost equation, as shown in Panel C. In general, a firm's fixed entry cost increases with the local rent level. Since all the exogenous variables are standardized in estimation (see Appendix A.2) and the fixed cost is the exponential of a linear index. Combining the summary statistics in Table 1.3, we should interpret the confidence interval of *rent* as one standard deviation, i.e. \$331.9, higher than the average median rent \$983.9 makes the fixed cost $[e^{0.1809}, e^{0.2897}] = [1.20, 1.34]$ times of the average level. In addition to the rent effect, I also allow a level difference in fixed cost for the top 3 firms. Specifically, with other variables being constant, the fixed entry cost is higher for Hilton and Marriott but lower for IHG. That coincides with the directions of their capacity differences. In particular, we would expect a property with a higher fixed cost to be the one with more rooms. That is consistent with what is in the data, the average capacity per property across all firms is 120.72 rooms, while that for Hilton, Marriott and IHG are 129.27, 135.9 and 86.84 rooms respectively.

1.6 Counterfactuals

To measure the welfare effects of Marriott's acquisition of Starwood, I simulate hypothetical markets and solve for the market equilibrium $(\boldsymbol{p}_m, \boldsymbol{s}_m, \boldsymbol{y}_m)$ in both the with- and without-merger scenarios. In particular, instead of comparing the observed reality where Marriott and Starwood have merged with the simulated markets where they operate separately, I simulate with-merger markets as well so that the merger effects from comparing the with- and without-merger scenarios can be isolated from the simulation errors from comparing the simulated without-merger scenario and the observed with-merger scenario.

The markets in this section are hypothetical in the sense that the exogenous variable values are constructed rather than taken directly from the observed markets. Specifically, I consider all the 11 firms in the data as potential entrants to each neighborhood. I introduce variation across these hypothetical markets by setting N_{jm}^{uc} at difference values while fixing N_{jm}^{all} for $\forall j, m$, motivated by the small magnitude of the N_{jm}^{all} coefficient. For the top 3 firms (i.e. Hilton, Marriott, IHG) and Starwood, N_{jm}^{uc} can be at either the 25% percentile or the 75% percentile of each of these firms' empirical distribution. For the other firms, it is set at their medians. *Rent* and *Market Size* are set at the median of all the observed markets in the last period of the data, Dec 2018. This gives 16 combinations of exogenous variable values in total.

Besides the ownership structure change, I also allow the merger to affect the competition outcomes by altering the values of N^{uc} and N^{all}. I refer their effects as *quality improvement* channel and *cost synergy* channel in the following discussion, given that they affect the equilibrium via the demand equation and the marginal cost equation correspondingly. More specifically, recall that N^{uc} is a proxy for the scale of a firm's loyalty program. When Marriott and Starwood merged, the two original loyalty programs of Marriott, *Marriott Rewards*, *Ritz-Carlton Rewards* and the one of Starwood, *Starwood Preferred Guest*, formed a new joint program named *Marriott Bonvoy*. Once travelers link their former accounts to the new one, points earned from any old program can be used interchangeably to redeem free nights at all hotels under the Marriott brands, including those previous Starwood properties. Therefore, the with-merger N^{uc} of Marriott and Starwood hotels should naturally be their sum. Similarly, I sum their N^{all}'s to reflect the fact that an enlarged hotel network could further reduce each hotel's marginal cost.

Furthermore, to investigate the relative importance of quality improvement and cost synergy, for each statistic of interest, I simulate two scenarios where each one of the two channels is muted.

Given partial identification, I report a range of each statistic based on the 95% confidence intervals reported in Column (3), Table 1.6. Taking the mean equilibrium price of firm j, $\bar{p}_j$ as an example, I illustrate the procedure of constructing such a range as follows.

1. Take $R = 100$ random draws from the parameter confidence region, each denoted as θ^r, $r = 1, 2, ..., 100$.

2. With the covariance parameters ω^r, construct the Cholesky decomposition L^r of the variance-covariance matrix Ω^r.

3. For each component in $\mathcal{U}_{jm} = (x^u_{jm}, w^u_{jm}, z^u_{jm})$, take $S = 500$ random draws from a standard normal distribution, and then multiply each three of them with L^r, which gives the $\mathcal{U}^{rs}_{jm}$ that has the desired variances and covariances.

4. Given θ^r and $\mathcal{U}^{rs}_{jm}$, solve for the equilibrium $(\boldsymbol{p}^{rs}_m, \boldsymbol{s}^{rs}_m, \boldsymbol{y}^{rs}_m)$, where $\boldsymbol{p}^{rs}_m = \{p^{rs}_{jm}\}_{j \in \mathcal{F}^{rs}_m}$.

5. Average p_{jm}^{rs} across all m and s that have firm j as an entrant, [23] i.e.,

$$\bar{p}_j^r = \frac{1}{\sum_{s,m} 1[j \in \mathcal{F}_m^{rs}]} \sum_{\{(s,m): j \in \mathcal{F}_m^{rs}\}} p_{jm}^{rs}.$$

6. Collect $\bar{p}_j^r$ for all r's and report $[\min_r \bar{p}_j^r, \max_r \bar{p}_j^r]$

Essentially, I consider each pair of (m, s) as a market, since they denote the variation in the observables and the unobservables respectively. Therefore, I summarize the interested statistics across m and s at each parameter value θ^r and then report the range over all r's.

1.6.1 Equilibrium effects of the merger

Market structure change I start by investigating the probability that the equilibrium market structure changes with the merger. To summarize results, I group all possible market structures into five categories, each denoted as $YCat_n, n = 1, 2, 3, 4, 5$, they are (1) M&S – markets with both a Marriott and a Starwood as entrants; (2) M only – Marriott enters, probably with other firms but not Starwood; (3) S only – Starwood enters, probably with other firms but not Marriott; (4) Others – Neither Marriott nor Starwood is an entrant, but there are other firms in the market; (5) None – the market has been served by any hotel firm yet.

[23]For any fixed r, p_{jm}^{rs} vary with both s and m, therefore I consider market to be at s-m level. Then the number of observations is in general $S \times M = 500 \times 16 = 8000$; However, in the case of prices, this number could be less because the firm may choose not to enter some markets.

Table 1.7: Equilibrium market structures with and without the merger

Panel A: Quality improvement & cost synergy

		Post-merger					
		M&S	M only	S only	Others	None	No. of obs.
Pre-merger	M&S	[0.95, 1.00]		[0.00, 0.05]			[131, 331]
	M only	[0.05, 0.11]	[0.89, 0.95]				[826, 1525]
	S only	[0.01, 0.02]		[0.98, 0.99]			[917, 1351]
	Others	[0.00, 0.01]	[0.01, 0.05]	[0.06, 0.11]	[0.83, 0.93]		[4553, 4702]
	None	[0.00, 0.01]	[0.01, 0.09]	[0.05, 0.15]		[0.76, 0.93]	[184, 1570]
	No. of obs.	[201, 532]	[876, 1625]	[1265, 1872]	[3829, 4293]	[142, 1445]	8000

Panel B: Quality improvement only

		Post-merger					
		M&S	M only	S only	Others	None	No. of obs.
Pre-merger	M&S	[0.92, 1.00]		[0.00, 0.08]			[131, 331]
	M only	[0.03, 0.06]	[0.94, 0.97]				[826,1525]
	S only	[0.00, 0.01]		[0.99, 1.00]			[917, 1351]
	Others		[0.00, 0.01]	[0.04, 0.06]	[0.93, 0.95]		[4553, 4702]
	None		[0.00, 0.02]	[0.03, 0.06]		[0.93, 0.96]	[184, 1570]
	No. of obs.	[171, 364]	[839, 1515]	[1163, 1580]	[4235, 4441]	[172, 1503]	8000

Panel C: Cost synergy only

		Post-merger					
		M&S	M only	S only	Others	None	No. of obs.
Pre-merger	M&S	[0.98, 1.00]		[0.00, 0.02]			[131, 331]
	M only	[0.01, 0.06]	[0.94, 0.99]				[826, 1525]
	S only	[0.00, 0.03]		[0.97, 1.00]			[917,1351]
	Others		[0.01, 0.05]	[0.01, 0.07]	[0.88, 0.98]		[4553,4702]
	None		[0.00, 0.09]	[0.01, 0.11]		[0.81, 0.98]	[184,1570]
	No. of obs.	[152, 464]	[877, 1666]	[986, 1658]	[4062, 4564]	[151, 1529]	8000

Table 1.7 reports the probabilities that the market structure changes across these categories after the merger, each panel is distinguished by the channels through which the merger may alter the equilibrium. Specifically, the probability that a $YCat_n$ market switches to $YCat_{n'}$ is defined as

$$\frac{\sum_{s,m} 1[\boldsymbol{y}_m^{0,rs} \in YCat_n] \times 1[\boldsymbol{y}_m^{1,rs} \in YCat_{n'}]}{\sum_{s,m} 1[\boldsymbol{y}_m^{0,rs} \in YCat_n]},$$

where the superscript 0-1 indicates the without- and with-merger scenarios respectively.

For all the three panels, the diagonal numbers are close to one, meaning that market structure stays the same in most cases. Even though, as long as the market has Marriott or Starwood, it is likely to be affected via price changes, and *entry and* *exit* can be another layer on top of that. In particular, the last row of each panel of Table 1.7 lists the number of markets of each type. In Panel A, the sum of the first three types is [2342, 4029], which implies that [29.3%, 50.4%] of the markets would be affected by this merger through either price changes or entry and exit.

Back to market structure changes, the upper triangle components in Table 1.7 are almost all absent, which indicates that Marriott and Starwood seldom exit markets where they are already in. The only exception is that Marriott may exit an M&S market, and the reason is Marriott's above-average fixed cost – in addition to the benefit from the increased market power, the joint firm may find it more profitable to operate an only Starwood hotel. [24] On the contrary, the lower triangle components are usually positive, mainly driven by the entry of the merging parities, especially Starwood. The last row of Panel A shows that the probability that a Starwood hotel enters a "None" market can be high at 15%. By summing over the cells corresponding to "M&S", "M only" and "S only", we see that the probability of Marriott and Starwood expanding into a new market is between [0.06, 0.24], from which we would expect substantial consumer welfare gains.

Next, I probe into the firm-level entry probability changes with a special focus on whether other firms are likely to be driven out of the market after the merger. I then investigate how Marriott and Starwood price differently without- and with-merger.

Entry and exit probability by firm In Table 1.8, I report the changes in entry probability by market structure and firm. The numbers for Marriott and Starwood are consistent with the implications in Table 1.7. Although the probability that other firms enter decreases with the merger, but the magnitude is less than 2% across all scenarios. In particular, it occurs not only in the markets where Marriott and Starwood coexist and thereby executing the increased power is most likely, but also in the markets with only one merging party. The latter case highlights the network

[24] One caveat is that the cost of exit is not considered in this model.

Table 1.8: Entry & exit probability by firm

Panel A: Quality improvement and cost synergy					
	M&S	M only	S only	Others	None
Marriott	[-0.05, 0.00]		[0.01, 0.02]	[0.02, 0.06]	[0.01, 0.10]
Starwood		[0.05, 0.11]		[0.06, 0.12]	[0.06, 0.16]
Others	[-0.01, 0.00]	[-0.01, 0.00]	[-0.02, 0.00]		
Panel B: Quality improvement only					
	M&S	M only	S only	Others	None
Marriott	[-0.08, 0.00]		[0.00, 0.01]	[0.01, 0.02]	[0.01, 0.02]
Starwood		[0.03, 0.06]		[0.04, 0.06]	[0.03, 0.06]
Others			[-0.01, 0.00]		
Panel C: Cost synergy only					
	M&S	M only	S only	Others	None
Marriott	[-0.02, 0.00]		[0.00, 0.03]	[0.01, 0.05]	[0.00, 0.08]
Starwood		[0.01, 0.06]		[0.01, 0.07]	[0.01, 0.10]
Others	[-0.01, 0.00]	[-0.01, 0.00]	[-0.01, 0.00]		

effects among the affiliated hotels.

Table 1.9: Price changes of the merging party

Panel A: Quality improvement and cost synergy					
	Marriott			Starwood	
	M&S	M only		M&S	S only
M&S	[-0.34, 0.06]		M&S	[-2.36, -0.16]	[0.31, 5.39]
M only	[-3.37, -0.43]	[-2.79, -0.49]	S only	[-4.93, -0.63]	[-2.87, -0.59]
Panel B: Quality improvement only					
	Marriott			Starwood	
	M&S	M only		M&S	S only
M&S	[0.49, 2.71]		M&S	[0.36, 2.28]	[0.85, 4.32]
M only	[0.11, 0.28]	[0.04, 0.37]	S only	[0.14, 0.63]	[0.08, 1.19]
Panel C: Cost synergy only					
	Marriott			Starwood	
	M&S	M only		M&S	S only
M&S	[-0.98, -0.16]		M&S	[-4.19, -0.57]	[0.24, 0.45]
M only	[-3.67, -0.52]	[-3.25, -0.55]	S only	[-5.85, -0.66]	[-5.01, -0.84]

Price changes of Marroiott and Starwood A key criterion for merger evaluation is price change. Table 1.9 shows that for each of the two merging parties by with-merger market structure change. Across the three panels, we see that the price change depends on the net effects of quality improvement and cost synergy. Panel B demonstrates that Marriott properties tend to price higher, presumably leveraging the incomparable scale of its loyalty program; while Panel A and C together show that the upward pricing pressure is overturned by cost synergy in most cases. Starwood's price changes have a similar trend, except for those markets where Marriott properties strategically stay out to maintain Starwood's monopoly profit. In that case, Starwood would be able to lift its price by [$0.31, $5.39]. However, Table 1.7 shows that such strategic exits rarely happen; therefore, we would expect the merger in general benefit consumers by offering better products at lower prices. A more detailed summary of consumer surplus follows later in this section.

Notably, the numbers in Panel B and C do not necessarily add up to their counterparts in Panel A, because the model is highly nonlinear, and even with neither quality improvement nor cost synergy, the changes in ownership structure could make the equilibrium prices different. Also, I acknowledge that in reality, the network effect among the affiliated hotels may have a decreasing margin, therefore the price effects in Table 1.9 should be interpreted as upper bounds.

Table 1.10: Mean across markets served

	Markets served before the merger			New markets
	Price	Profit	Consumer surplus	Consumer surplus
Panel A: Quality improvement and cost synergy				
Pre-merger	[71.90, 135.95]	[4568.90, 65886.51]	[0.20, 9.61]	0
Post-merger	[71.15, 135.77]	[5440.44, 72489.59]	[0.24, 11.35]	[0.05, 0.88]
Difference	[-0.78, -0.04]	[831.42, 6653.34]	[0.04, 1.74]	[0.05, 0.88]
Panel B: Quality improvement only				
Pre-merger	[70.92, 136.12]	[1927.70, 48766.16]	[0.14, 8.46]	0
Post-merger	[71.69, 136.32]	[2235.82, 51453.10]	[0.17, 9.14]	[0.05, 0.61]
Difference	[0.17, 0.77]	[244.94, 3201.34]	[0.02, 0.69]	[0.05, 0.61]
Panel C: Cost synergy only				
Pre-merger	[71.84, 135.84]	[4678.41, 65612,94]	[0.20, 9.73]	0
Post-merger	[70.44, 135.43]	[4909.47, 68911.48]	[0.22, 10.75]	[0.04, 0.66]
Difference	[-1.41, -0.33]	[207.25, 3312.17]	[0.01, 1.02]	[0.04, 0.66]

Market-level average price, consumer surplus and firm profit To have a more comprehensive understanding of the merger's social welfare effects, I report three market-level summary statistics in Table 1.10, including the market-level average price, consumer surplus, and average firm profit.

Specifically, the market-level average price and profit are weighted by market shares, i.e.,

$$\bar{p}_m^{rs} = \frac{\sum\limits_{j \in \mathcal{F}_m^{rs}} p_{jm}^{rs} s_{jm}^{rs}}{\sum\limits_{j \in \mathcal{F}_m^{rs}} s_{jm}^{rs}}, \quad \overline{(\pi - fc)}_m^{rs} = \frac{\sum\limits_{j \in \mathcal{F}_m^{rs}} (\pi_{jm}^{rs} - fc_{jm}^{rs}) s_{jm}^{rs}}{\sum\limits_{j \in \mathcal{F}_m^{rs}} s_{jm}^{rs}}$$

And consumer surplus is defined as the expected maximum utility in dollar terms,

$$\bar{cs}_m^{rs} = -\frac{1}{\alpha^r} E[\max_{j \in \mathcal{F}_m^{rs}} u_{ijm}] = -\frac{1}{\alpha^r} \log(1 + \sum_{j \in \mathcal{F}_m^{rs}} exp(v_{jm}^r)) + C$$

where v_{jm} is the deterministic part of the utility, i.e. $x_{jm}^{o'}\beta - \alpha p_{jm} + x_{jm}^u$, and C is a constant term which is normalized to zero in this exercise. The variation in $\bar{cs}_m^{rs}$ roots in the changes in equilibrium prices and the choice set $\mathcal{F}_m^{rs}$, i.e. market structure. Slightly different from the previous tables, I average these statistics only across the m's and s's that have least one Marriott or Starwood hotel either before or after the merger, which is a necessary condition for a market to be affected. Additionally, I further break them down into markets that have been served without-merger and new markets, since the without-merger price and profit are well-defined only in the former case. I then report the range of the changes formed by r's.

Column 1 of Table 1.10 shows that quality improvement and cost synergy shift prices in opposite directions, but overall, the price drop caused by cost synergy dominates. This is similar to what we observe in Table 1.9. The reason is that other firms would adjust prices in the same direction as the merging parties, i.e. it is either that the merging parties lift their prices, so the rivals would be able to price higher without losing as many customers as previously, or the merging parities lower their prices and thereby initiate a price war.

Across all panels, both consumer surplus and firm profit increase after the merger, even when only quality improvement is accounted for. This indicates that for consumers the utility gain from a better loyalty program prevails the loss from an increased price. [25]

1.6.2 Bias by Overlooking Entry

In this subsection, to establish the accuracy gain from conducting this relatively complicated method, I present how the evaluation results can be biased when market structure is taken as exogenous. The results establish my belief that this framework can be an appealing alternative for competition authorities to evaluate a great variety of merger proposals.

In general, I redo the counterfactual analysis in Section 1.6.1 using the estimates from the exogenous entry model, θ^{exoge} (as shown in Column 2 of Table 1.6), with the same set of simulated markets. To make a fair comparison, I use the same without-merger market structure $\boldsymbol{y}_m^{0,rs}$ but fix it in the with-merger scenario, i.e. $\boldsymbol{y}_m^{0,rs} = \boldsymbol{y}_m^{1,rs}$, and then use θ^{exoge} to solve for the equilibrium prices and market shares. Then, the differences from the results in the previous subsection would be the bias caused by overlooking firms' entry decisions.

Table 1.11: Bias decomposition - price changes

	Marriott		Starwood	
	M&S	M only	M&S	S only
$\theta^{endog.}$, $\boldsymbol{y}_m^{0,rs} \neq \boldsymbol{y}_m^{1,rs}$	[-0.34, 0.06]	[-2.79, -0.49]	[-2.36, -0.16]	[-2.87, -0.59]
$\theta^{exoge.}$, $\boldsymbol{y}_m^{0,rs} = \boldsymbol{y}_m^{1,rs}$	7.41	6.71	11.12	10.29
	[3.89, 13.08]	[2.43, 12.50]	[6.38, 18.74]	[4.45, 18.51]

Bias in price changes Table 1.11 shows the merging parties' price changes. Row 1 copies the diagonal components in Panel A, Table 1.9. Notice that the exogenous model is point identified; therefore, in Row 2, I report the average price as a point, together with a range corresponding to the estimates' 95% confidence intervals. Across the rows, we see that fixing market structure in merger evaluation could draw

[25] I acknowledge that the potential extra cost of managing a gigantic loyalty program is not in the model, otherwise the conclusion might be different.

the opposite conclusion on price changes. Across all types of market, Marriott and Starwood would lift the price by \$7.41 - \$11.12 per room-night. Although it is hard to disentangle this composite effect by covariate, clearly the overestimated constant term in the demand equation and the upward bias in the coefficient on N^{all} contribute to this.

Table 1.12: Bias decomposition - market level means

	Price	Profit	Consumer surplus
Panel A: $\boldsymbol{\theta}^{endog.}$, $\boldsymbol{Y}^0 \neq \boldsymbol{Y}^1$			
Pre-merger	[71.90, 135.95]	[4568.90, 65886.51]	[0.19, 9.51]
Post-merger	[71.39, 136.01]	[5201.18, 71776.32]	[0.23, 11.23]
Difference	[-0.53, 0.35]	[594.29, 5963.66]	[0.04, 1.73]
Panel B: $\boldsymbol{\theta}^{exoge.}$, $\boldsymbol{Y}^0 = \boldsymbol{Y}^1$			
Pre-merger	112.13	27360.74	4.55
	[105.45, 127.00]	[1266.32, 36473.27]	[0.01, 22.81]
Post-merger	113.66	18124.50	3.46
	[107.91, 131.40]	[-3240.76, 65900.53]	[0.01, 21.69]
Difference	1.53	-9236.25	-1.09
	[1.05, 6.37]	[-33143.37, 64427.39]	[-4.37, -0.01]

Similar to the previous subsection, I calculate those market level summary statistics. Results are shown in Table 1.12, and the setting of each panel echos each row in Table 1.11. Since the exogenous entry model does not have the fixed cost function, the profits in Panel B is the variable profit from the exogenous model minus the fixed cost predicted by the endogenous model. It is essentially the same as comparing the variable profits only.

Again, the exogenous entry model gives very different results from the endogenous model. The average price is in general upward biased: unlike the uncertain price change direction in the endogenous entry model, the price would definitely increase when entry is exogenous. More importantly, consumer surplus decreases at all parameter values in the confidence region. Intuitively, it is because that the exogenous model does not only have the upward pricing pressure from the increased market power and the quality improvement, but also predict that the merging parties would suffer from extra inefficiency in marginal cost. These two aspects together induce a higher price, which prevails the consumer surplus gain from quality improvement and further results in a net negative effect. For the firms' average profits, the difference interval presents an indecisive direction. For some parameters, the difference is negative. It is mainly due to the reduction in the merging parties' profits, for which the with-merger inefficiency in marginal cost is the direct cause. And another reason is that the properties that cannot make positive profit are not allowed to strategically quit the market in this exogenous entry model.

1.7 Conclusion

This paper shows that the market concentration increase sometimes cannot fully summarize the social welfare implication of a merger, it should be a synthesis of changes in markups, marginal cost, and choice availability. I provide a framework that incorporates all these effects and is computationally tractable when the number of competitors is high. In particular, I estimate a price competition game with endogenous entry, so that the estimates are not contaminated by the self-selection issue and the firms are allowed to respond to the merger by entry and exit.

My results provide evidence that supports the DOJ and FTC's approval for mergers at the size of Marriott's acquisition of Starwood. I find that although the increased market power and product quality improvement tend to increase the prices of the merging parties, such an effect could be predominated by their cost synergies. Additionally, because of the efficiency gains, Marriott and Starwood would serve the markets that were unprofitable to enter before the merger, without much disturbance on other incumbents in the markets. Overall, consumers get lower prices, better quality, and better choice. But under a framework where market structure is fixed throughout the merger, estimates' biases could wrongly conclude a harmful merger.

There are certain limitations in the paper. First, I employ a static model, thereby ignoring the hotel chains' forward-looking behaviors motivated by the franchise contract's long duration. I consider that as a reasonable simplification, because even though in reality a franchise agreement usually has a long term, ranging from 10-30 years, the franchiser has the right to terminate it at any time for multiple reasons. One reason for termination could be that the franchisee fails to pay the franchiser or any of the franchiser's affiliates any fees under their agreement, for which having a negative profit seems to be a plausible trigger. Undeniably, the static setting is also out of tractability concerns. Introducing that dynamics into this already complicated multi-agent game with potential multiple equilibria is well beyond the scope of this paper. Pioneer works along this strand include Ericson and Pakes (1995) and Aguirregabiria and Ho (2012).

Another caveat in the model is that cost synergy is restricted to the network effects from an enlarged affiliation system. There could be, for instance, a with-merger level drop in marginal cost. One way to tackle this issue is to use both without- and with-merger data in the estimation, and add an interaction term between a merging party indicator and a with-merger indicator. This can be the next step of this project, although having more observations and parameters could further aggravate the computational burden.

Chapter 2

Payment Instrument Choice with Scanner Data: An MM algorithm for Fixed Effects in Non-Linear Models (*with Mingli Chen, Marc Rysman and Krzysztof Wozniak*)[1]

2.1 Introduction

Over the past several decades, the U.S. payments system has seen a steady shift away from paper payment instruments, such as cash and check, to digital instruments, such as debit and credit cards. This is important because digital payments are typically regarded as superior in many dimensions: they are faster and cheaper to process, easier for customers to keep track of, and less subject to fraud. Despite this change, however, cash and check still play a larger role in the United States than in many other countries. Anecdotal evidence of young people adopting digital payment while older households persist with cash and check suggests that demographics and heterogeneity between households could be key to explaining the lingering popularity of paper payment instruments.

This paper studies the determinants of payment choice in the short and long

[1]Researchers own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

term. In the short term, we focus on the transaction size as an important determinant. Transaction size has been central to the discussion of payment choice, with households more likely to use noncash payment instruments for larger transactions. Previous papers, such as Klee (2008) and Wang and Wolman (2016), have studied the effect of transaction size on payment choice by using scanner data drawn from retailers. However, since these data sets did not allow the authors to track individuals over time, the resulting estimates were not able to separate the within and between effects. In particular, while previous literature has shown that consumers are more likely to pay with card larger for larger transactions, it is also possible that households that use cards more often tend to have higher transaction sizes on average. A central goal of this paper is to separate these effects.

We also study the drivers of the long-term changes in payment usage. In particular, increased preference for card payments could be a key driver of the observed increase in card usage in the United States. At the same time, it is important to disentangle changes in individual consumers' preference for card payments from other factors could have contributed to the growing card use. These alternative explanations include a shift in the composition of households towards those that prefer cards, a growth in the number of transactions made by households that prefer cards, or an overall growth in transaction sizes. For instance, if older households prefer cash and check and also experience decreases in transactions, while younger households prefer card and experience increases in transactions, we will observe an aggregate increase in card usage although no household has experienced a change in preferences for card relative to cash and check. Our paper aims to decompose the changes in preferences from other factors which contributed to the long-term shift towards digital payment instruments.

Our analysis leverages consumer scanner data from Nielsen to obtain a novel transaction-level data set on payment choice. Unique among data sources used in the payments literature, this data set can track payment choices for specific consumers over time. The resulting panel data structure allows us to use multidimensional fixed effects to disentangle the many factors which impact payment choice, both in the short and the long terms. To our knowledge, no previous academic work has used such data to study payment choice.

In order to fully capture the many factors driving payment choice, we estimate a multinomial discrete choice model with household-quarter-choice fixed effects. However, estimating such a model using data with three payment choices, over 10,000 households, and 20 quarters of data, which in the richest specification translates into more than 1 million fixed effects, creates a significant numerical challenge. Standard procedures for estimating multinomial logit models fail with this many parameters. An important contribution of our paper is to introduce a new method to address this numerical problem. We believe our solution is applicable in a wide variety of setting where a researcher wishes to estimate fixed effects in a non-linear model.

In particular, we estimate our model using the Minorization-Maximization (MM) algorithm. This algorithm, which can be seen as a generalization of the Expectation-Maximization (EM) algorithm, has been developed in the statistics literature (Hunter and Lange, 2004; Lange, 2016). However, to date, it has seen almost no applications in econometrics. We utilize the MM algorithm to linearize the logit model so that we can apply linear techniques, such as demeaning, to the fixed effects estimation. Sequential fixed effects estimation and minorization allow us to find numerically identical estimates to maximum likelihood at a tiny fraction of the computational and memory costs that estimating dummy variables using traditional methods would entail. Finally, because we estimate many fixed effects in a panel setting, we face the incidental parameters problem. As in several previous papers, we address the incidental parameters problem with ex-post bias reduction via the jackknife following Dhaene and Jochmans (2015).

Looking at the short-term payment decision, our results show that transaction size is an important determinant of payment choice in the short term, which confirms an important finding from previous literature. In particular, we find that going from the 1st quartile of the empirical distribution of transaction size, \$11.94, to the 3rd quartile, \$57.06, leads to a 20.7 percentage point increase of probability of using a card. Notably, we find that our model specification with a full set of household-quarter-choice fixed effects results in a lower effect on average, 16.8 percentage point. This finding suggests the impact of transaction size on payment choice is smaller than had been estimated by papers not able to directly account for heterogeneity in unobserved household payment preferences, although the difference is only moderate in magnitude. More importantly, the models with household-specific fixed effects unveil a substantial heterogeneity in transaction size effect across households. In particular, there are households whose probability of using card does not vary with transaction size at all or even slightly decreases with it. In particular, that interquartile effect of transaction size for the most responsive household is 39.9 percentage point higher than that of the least sensitive one. Such a divergence would have been suppressed in the specification with choice fixed effects only.

We then turn to the long term, and in particular to the almost 10 percentage point increase in card usage over the five-year period in our data. We use our model to decompose the factors driving this change into (a) changes in household preferences, (b) changes in the number and value of transactions, and (c) entry and exit of households from the sample. Our results show that only about a third of the growth in popularity of card payments is due to changes in individual households' preferences. This finding suggests that household preferences change relatively slowly, and that public policy efforts to shift consumers to digital payments may take time to yield substantial results.

Overall, our paper makes several contributions. We demonstrate consumer scanner data can be a powerful tool for studying payment choice. We provide an attractive new approach to estimating multinomial discrete choice models with fixed effects. Within the MM literature, we provide a new formalization of the MM algorithm and a new minorization for the multinomial logit model, which could be extended to a number of linear-index likelihood models. We present new results on the importance of transaction size in determining short-term payment choice, and show that accounting for persistent unobserved household heterogeneity reduces the magnitude of that effect. Finally, we decompose long-term trends in payment choice and show the relatively limited role that changes in consumer preferences have in driving these trends.

2.2 Literature Review

There are many studies whose aim is to identify the determinants of payment choice, with the majority focusing on the decision in the short term. However, many of them are hampered by data constraints. In particular, it is difficult to track the payments of individual households, especially for payments made with cash. One method for tracking payment choice is to survey consumers retrospectively, as used in Schuh and Stavins (2010) and Koulayev et al. (2016). These papers rely on a survey that asks consumers about payment use over the previous month. However, since shopping trip details are not captured alongside payment choice, data from such surveys make it difficult to study the determinants of each individual choice, or why choice varies across shopping trips. Another method is to ask survey participants to fill out a diary of payment behavior, as used in Rysman (2007), Arango, Huynh and Sabetti (2015) and Wakamori and Welte (2017). While such diaries are an important data source, Jonker and Kosse (2009) raises questions about their accuracy. In particular, the authors show that the daily number of transactions in seven-day surveys is significantly less than in one-day surveys, suggesting data from payment diaries may suffer from "diary fatigue." A third widely-used method is to obtain data directly from consumer bank accounts, as do White (1975), Stango and Zinman (2014), and Dutkowsky and Fusaro (2011). While data thus obtained does not suffer from diary fatigue, it typically provides no information on cash usage. Moreover, individual consumers may have multiple transaction accounts, some of which may not show up in the available transaction record.

Consumer scanner data has important advantages over these data sources. In particular, in our data set we observe payment choice decisions for individual household continuously over a period of five years, something that no existing diary data set can come close to matching. At the same time, our data has certain limitations. First, the Nielsen data does not capture every transaction a household makes. Nonetheless, the data is probably most complete with regard to grocery trips, a significant touchpoint for payment choice, and an important focus of the payments industry. Second, the method that Nielsen uses to track payments does not allow us to distinguish between debit and credit card payments, a common issue in payment literature. Importantly, though, we are able to distinguish between the three most common retail payment instruments: cash, check and payment card.

A paper closely related to ours is Klee (2008). Klee also uses scanner data from grocery purchases to study payment choice. However, since her data set is drawn from the cash register of a grocery chain, she is not able to track consumers over time. Moreover, since the data set does not capture consumer demographics directly, the author accounts for it by using census data for store locations. This contrasts with our paper, where we observe consumer demographics directly, and importantly can use household identifiers to account for unobserved heterogeneity using panel techniques such as fixed effects. In addition, our study covers packaged food shopping from a wide array of retail channels, not just a single store. Like us, Klee cannot distinguish between debit and credit, although she distinguishes between signature and PIN-based card transactions. Wang and Wolman (2016) follows a similar approach. Ultimately, most of the papers we discuss here rely on data sets that cover relatively short time periods. We are not aware of another paper that attempts to decompose long-term changes in payment instrument use the way we do.

In addition to adding to payment literature, our paper also makes a significant contribution in the area of estimating fixed effects in non-linear models. Several other papers precede us in this regard. A classic contribution is Chamberlain's conditional logit model (Chamberlain, 1980). Unfortunately, typical implementations handle only the binary outcome case, and extending the model to multinomial outcomes creates significant combinatoric complexities. Furthermore, the model does not naturally deliver fixed effects estimates, which our approach does.

In contrast with Chamberlain's approach, we combine a computational approach to estimating the fixed effects model with a jackknife bias correction. Papers such as Hospido (2012) and D'Haultfœuille and Iaria (2016) introduce efficient methods for computing the dummy variables model and, like us, rely on ex-post bias correction. In particular, Hospido (2012) exploits the sparsity of the fixed effects, while D'Haultfœuille and Iaria (2016) rely on simulation of the choice set to cheaply compute the Hessian of the objective function, at the expense of introducing integration error. Unfortunately, in our application, the Hessian is larger than can be addressed by either of these aforementioned approaches.

The paper whose approach to estimating fixed effects in non-linear models is probably Stammann, Heiß and McFadden (2016). In particular, the authors advocate for both iterative demeaning to obtain estimates as well as ex-post bias correction, in their case based on Hahn and Newey (2004). Another approach relies on concentrating out fixed effects and maximizing over the remaining parameters, as used in Hinz, Hudle and Wanner (2019) and Stamann (2018). Our understanding is that the approaches in all three papers have been developed only for binary outcomes and do not easily expand to multinomial settings. A final approach relies on differencing out fixed effects in a way that leads to estimation with moment inequalities, as used in Ho and Pakes (2014b) and Shum, Song and Shi (2018). Our approach differs from moment inequality estimation in that it generates point identification and directly estimates fixed effects. In addition, we view our approach as computationally less challenging than estimation with moment inequalities.

Our estimation approach uses the Minorization-Maximation (MM) algorithm (sometimes called the Majorization-Minorization algorithm). The MM algorithm has a long history in statistics and dates back to around the time of the introduction of the Expectation-maximization (EM) algorithm, which can be regarded as a special case of the MM algorithm. In general, the MM algorithm expands the set of functions that can be used in the E-step of the EM algorithm, and has appeared under many names in different papers, often depending on what function was used. Böhning and Lindsay (1988) is an important early citation, while Hunter and Lange (2004) and Lange (2016) provide a helpful overview and history. We are aware of only one other paper in the econometrics literature which uses the MM algorithm. In particular, James (2017) shows that the MM algorithm can be advantageous in the context of the mixed multinomial logit model of McFadden and Train (2000), but does not discuss the application to fixed effects or dynamic models.

Finally, our approach is similar to that of Chen (2019), who uses the EM algorithm in the context of the binary probit to estimate a model with interactive fixed effects. In particular, she uses the EM algorithm to obtain a linear form of the model and then applies known techniques for handling interactive fixed effects in the linear case. In addition, she also uses ex-post bias correction, in her case a known analytic form. In fact, her implementation of the EM algorithm resembles the MM algorithm. She does not consider multinomial models. Following her ideas, our model could be extended to handle interactive fixed effects in a multinomial logit model.

2.3 The Minorization - Maximization (MM) Estimation Procedure

The central idea in our paper is that we use the MM algorithm to estimate a nonlinear discrete choice model in an iterative series of two steps. First, in each iteration, we construct a simpler concave surrogate function that minorizes the complicated log-likelihood function (that is, the surrogate function is less than the log-likelihood function everywhere but at the current best guess of the parameters, where it is equal). Second, we maximize the surrogate function instead of the log-likelihood function. The ascent of the log-likelihood is guaranteed by the property of minorization. By alternating between these steps of minorization and maximization, the MM algorithm finds the parameters that maximize the original log-likelihood function.

Before diving into details of our multinomial logit model, we first discuss the definition of the MM algorithm in general, and provide conditions required for convergence.

2.3.1 The Transfer Minorization

Relative to the standard mathematical definition of a minorizing function, we add an extra condition that makes the function suitable for use in a maximization problem. We refer to our minorization as a transfer minorization. Our name is based on the terminology of Lange, Hunter and Yang (2000), which refers to the minorization as the transfer function. That is, we transfer optimization from the function of interest to the minorization of the function.

Definition 1. *Suppose $\mathcal{L}$ is a real-valued function on R^p that is twice differentiable and S is a real-valued function on $R^p \otimes R^p$, we say that S is a transfer minorization of $\mathcal{L}$ if:*

(a) $S(\boldsymbol{\theta};\boldsymbol{\theta}') \leq \mathcal{L}(\boldsymbol{\theta})$ *for all* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}'$*;*

(b) $S(\boldsymbol{\theta}';\boldsymbol{\theta}') = \mathcal{L}(\boldsymbol{\theta}')$ *for all* $\boldsymbol{\theta}'$*;*

(c) $\nabla^{20} S(\boldsymbol{\theta};\boldsymbol{\theta}')$ *exists, and is negative definite at* $\boldsymbol{\theta}$*.*

where $\nabla^{mn} S(\boldsymbol{\theta};\boldsymbol{\theta}')$ is the m^{th} order derivative w.r.t. $\boldsymbol{\theta}$ and n^{th} order derivate w.r.t to $\boldsymbol{\theta}'$.

Analogously, S is a transfer majorization of $\mathcal{L}$ if $-S$ is a transfer minorization of $-\mathcal{L}$. The first two conditions of Definition 1 are from de Leeuw and Lange (2009) and are standard for defining a minorization. The third condition ensures that the minorization is well-behaved around the focal point. Arguably, we could use a less strict condition, such as that the minorization has the same sign as $\mathcal{L}$ in some region around $\boldsymbol{\theta}$, but in practice, we are not aware of any implementations of the MM algorithm that do not satisfy the third condition. As shown in de Leeuw and Lange (2009), an implication of Definition 1 is:

Corallary 1. *If S is a transfer minorization of $\mathcal{L}$, then for all $\boldsymbol{\theta}$:*

$$\nabla^{10} S(\boldsymbol{\theta};\boldsymbol{\theta}) = \nabla \mathcal{L}(\boldsymbol{\theta})$$

This is basically the necessary condition that $\boldsymbol{\theta}$ minimizes the distance between $S(\cdot;\boldsymbol{\theta})$ and $\mathcal{L}(\cdot)$.

Intuitively, when faced with a likelihood function $\mathcal{L}$ that is difficult to maximize, we instead choose to maximize another function S. The function S is chosen to be

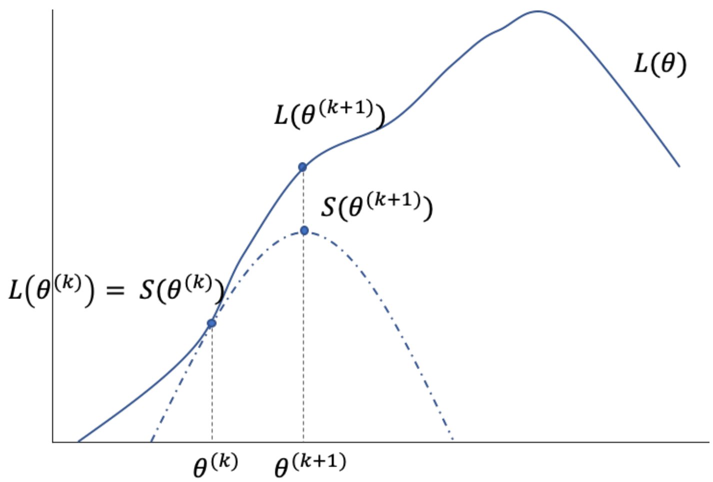


Figure 2·1: MM algorithm

easy to maximize so that its maximand is always closer to a local optima than the current guess. Figure 2·1 provides an example. In the figure, we would like to find the optimum of $\mathcal{L}$ and we start with a guess $\theta^{(k)}$. Rather than seek to optimize $\mathcal{L}$ directly, we construct a transfer minorization $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$. As a minorization, it is always below $\mathcal{L}$ and is equal to $\mathcal{L}$ at $S(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})$. As a transfer minorization, S is well-behaved around $\boldsymbol{\theta}^{(k)}$, i.e. it is differentiable and concave. It is optimized at the point $\boldsymbol{\theta}^{(k+1)}$. At this point, we will construct a new transfer minorization $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k+1)})$ (not shown in the figure). Iterative application of this process leads to the maximum of $\mathcal{L}$, as shown in the next proposition. First, we define the MM algorithm:

Definition 2. *Let $\boldsymbol{\theta}^{(0)}$ be the initial guess of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(k)}$ be the guess after k cycles of the algorithm. The Minorization-Maximization (MM) Algorithm iteratively applies the following two-step procedure:*

1. *Minorization step: Compute $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$.*
2. *Maximization step: Choose $\boldsymbol{\theta}^{(k+1)}$ to be a value of $\boldsymbol{\theta} \in R^p$ that maximizes $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$.*

At each step, let $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(k+1)}$. Repeat steps 1 and 2 until $\boldsymbol{\theta}^{(k)}$ converges.

To show convergence, we generalize the Theorem 4 in Dempster, Laird and Rubin (1977) from the EM algorithm context to MM algorithm. In Theorem 1, we list conditions for the sequence $\boldsymbol{\theta}^{(k)}, k = 0, 1, 2, ...$ to converge to a point where $\nabla\mathcal{L}(\cdot) = 0$ in the context of maximum likelihood estimation (MLE). Specifically, a likelihood function $f(x;\boldsymbol{\theta})$ is the density of the true DGP $f(x;\boldsymbol{\theta}_0)$ with the true parameter vector $\boldsymbol{\theta}_0$ replaced with its hypothetic value $\boldsymbol{\theta}$.[2]

Theorem 1. *Suppose that $\mathcal{L}(\boldsymbol{\theta})$ is the log likelihood function and $S(\boldsymbol{\theta};\boldsymbol{\theta}')$ is a minorization of $\mathcal{L}(\boldsymbol{\theta})$ for maximization, and $\boldsymbol{\theta}^{(k+1)} = argmax_{\boldsymbol{\theta}}\, S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}), k = 0, 1, 2, ...$ is an instance of an MM algorithm, then:*

1. *$\boldsymbol{\theta}^{(k)}$ converges to a $\boldsymbol{\theta}^*$ in the closure of Ω.*
2. *$\nabla\mathcal{L}(\boldsymbol{\theta}^*) = 0$, $\nabla^{20}S(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*)$ is negative definite with eigenvalues bounded away from zero.*

Proof. See Appendix B.1.1.[3] □

Thus, any function that satisfies Definition 1 for some function $\mathcal{L}$ can be used in the MM algorithm to find an optimum to $\mathcal{L}$.[4] This approach allows substantial

[2]More formally, we consider the log likelihood function as defined in Hayashi (2000, p.448). In particular, let $\{\boldsymbol{x}_n\}$ be an i.i.d. sequence where the density of $\boldsymbol{x}_n$ can be indexed by a finite-dimensional vector $\boldsymbol{\theta}_0$: $f(\boldsymbol{x}_n;\boldsymbol{\theta}_0)$, $\boldsymbol{\theta}_0 \in \Omega$. Because $\{\boldsymbol{x}_n\}$ is independently distributed, the joint density of the data $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$ at a hypothetical value $\boldsymbol{\theta}$ is

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N;\boldsymbol{\theta}) = \prod_{n=1}^{N} f(\boldsymbol{x}_n;\boldsymbol{\theta}).$$

This density, viewed as a function of $\boldsymbol{\theta}$ is called the *likelihood function.* The maximum likelihood (ML) estimator of $\boldsymbol{\theta}_0$ is the $\boldsymbol{\theta}$ that maximizes the likelihood function. Because the log transformation is a monotone transformation, maximizing the likelihood function is equivalent to maximizing the log likelihood function.

$$\mathcal{L}(\boldsymbol{\theta}) = \log\ f(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N;\boldsymbol{\theta}) = \log\left[\prod_{n=1}^{N} f(\boldsymbol{x}_n;\boldsymbol{\theta})\right].$$

[3]We provide Theorem 1 and the rest of the formalism in this section because we could not find the mathematical statement we required in the existing literature. But to be clear, our approach relies heavily on the cited literature and for sure, the existing literature seems to operate as if it is well-known that the MM algorithm converges to a local optima.

[4]The papers that offer the closest version of what we present in this section define a minorization based only on parts (a) and (b) of Definition 1 and then include something like part (c) in the supposition of the analog to Theorem 1. For example, see Böhning and Lindsay (1988). We prefer to have all of the requirements in Definition 1 so we can simply check if a candidate function satisfies the definition to know that it can be used in the MM algorithm. Because our definition differs in

freedom in selecting the transfer minorization. From the perspective of the MM algorithm, the EM algorithm is a special case and it works because the conditional expectation of $\mathcal{L}$ used in the EM algorithm is a transfer minorization. If the E-step of the EM algorithm does not deliver a function that is easy to optimize, as in our case, the researcher is free to use some other minorizing function. Hunter and Lange (2004) discuss several approaches. In the next section, we focus on a Taylor expansion, which delivers a least-squares optimization problem in our context.

2.4 A Minorization for the Multinomial Logit

In this section, we first present the multinomial logit model, then develop a transfer minorization, and finally discuss our estimation method.

2.4.1 Model

In the multinomial logit model, an agent makes a discrete choice among several options, each of which draws an Extreme Value error. The model is distinguished by a closed-form logistic function for the probability of each choice. In our presentation, we emphasize a fixed effect that varies by consumer, quarter and product.

We observe N consumers make a discrete choice among J products in each of T time periods. Consumer i in period t who chooses product j obtains utility u_{ijt}. Utility is defined as:

$$u_{ijt} = x_{it}\beta_j + \xi_{ijq(t)} + \varepsilon_{ijt}, \tag{2.1}$$

where $q(t)$ is the quarter of the year that t falls in where $q \in \{1, \ldots, Q\}$, x_{it} is a vector observable characteristics that varies by consumer and time, ξ_{ijq} is a consumer-quarter-product fixed effect and ε_{ijt} is a scalar *i.i.d* idiosyncratic shock distributed according to a Type I Extreme Value distribution. The variable y_{ijt} is a binary

this way from the previous research, we coin a new name for our version: the transfer minorization. But to be clear, we rely heavily on existing contributions in our approach.

indicator for the product that consumer i chooses in t, where:

$$y_{ijt} = \mathbb{1}[u_{ijt} \geq u_{ikt}, \forall k \neq j], \ \forall i, t.$$

The parameter vectors β_j and ξ_{ijq} are to be estimated. We collect these parameters as $\boldsymbol{\theta} = (\{\beta_j\}_{j=1,\ldots,J}, \{\xi_{ijq}\}_{i=1,\ldots,N;j=1,\ldots,J;q=1,\ldots,Q})$. Then the log-likelihood function can be written as:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i,t} \log \ l(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}; \boldsymbol{y}_{it}) \tag{2.2}$$

where $x_{it}\boldsymbol{\beta}+\boldsymbol{\xi}_{iq(t)} = (x_{it}\beta_1+\xi_{i1q(t)}, x_{it}\beta_2+\xi_{i2q(t)}, \ldots, x_{it}\beta_J+\xi_{iJq(t)})$, $\boldsymbol{y}_{it} = (y_{i1t}, y_{i2t}, \ldots, y_{iJt})$ and:

$$\begin{aligned} l(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}; \boldsymbol{y}_{it}) &= \textstyle\prod_{j=1}^{J} \left(p_{ijt}(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)})\right)^{y_{ijt}}, \\ p_{ijt}(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}) &= \frac{\exp(x_{it}\beta_j + \xi_{ijq(t)})}{\sum_{k=1}^{J} \exp(x_{it}\beta_k + \xi_{ikq(t)})}. \end{aligned}$$

In practice, we must normalize the mean utility of one of the choices to be zero, as is standard in the multinomial logit.

2.4.2 The transfer minorization

This subsection derives a function S that satisfies the conditions of Definition 1 to be a transfer minorization of $\mathcal{L}$.

Theorem 2. *Let $\mathcal{L}(\boldsymbol{\theta})$ be the log likelihood function for a multinomial logit model defined as in Eq.(2.2). Let $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ be defined as:*

$$\begin{aligned} S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) = \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2} \sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}; \boldsymbol{y}_{it})^2 \\ - \frac{1}{2} \sum_{i,j,t} \left(x_{it}\beta_j^{(k)} + \xi_{ijq(t)}^{(k)} - h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}; \boldsymbol{y}_{it}) - x_{it}\beta_j - \xi_{ijq(t)}\right)^2, \end{aligned} \tag{2.3}$$

where

$$
\begin{aligned}
h_j(\boldsymbol{\psi}_{it}^{(k)}; \boldsymbol{y}_{it}) &= \frac{\partial \log l(\boldsymbol{\psi}_{it}^{(k)}; \boldsymbol{y}_{it})}{\partial \psi_{ijt}^{(k)}} = y_{ijt} - p_{ijt}(\boldsymbol{\psi}_{it}^{(k)}) \\
\boldsymbol{\psi}_{it}^{(k)} &= x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}
\end{aligned}
$$

then $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ *is a transfer minorization of* $\mathcal{L}(\boldsymbol{\theta})$*.*

Proof. See Appendix B.1.2. □

By Theorem 1, $\boldsymbol{\theta}^{(k+1)} = argmax_{\boldsymbol{\theta}}\, S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}), k = 0, 1, 2, ...$ converges to a local, if not global, maximum of $\mathcal{L}(\boldsymbol{\theta})$.

Eq.(2.3) is a first-order Taylor expansion of the likelihood function. Note that the parameters that we search over in the iterative MM algorithm, i.e. $\boldsymbol{\theta}$ rather than $\boldsymbol{\theta}^{(k)}$, appear only in the third part of the right-hand side of Eq.(2.3). Focusing on this third part, we can think of optimization of $S(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ as a linear regression of $x_{it}\beta_j^{(k)} + \xi_{ijq(t)}^{(k)} + h_j\left(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}; \boldsymbol{y}_{it}\right)$ on x_{jt} (with separate coefficients for each j) and $\xi_{ijq(t)}$. This is the central benefit of our MM approach to the multinomial logit. We convert non-linear optimization to sequential linear optimization, which is particularly attractive when there are many regressors. Rather than use OLS directly, we use linear panel data methods to address the large number of parameters represented by ξ_{ijq}.

Thus, the functional form of $h_j(\boldsymbol{\psi}_{it}^{(k)}; \boldsymbol{y}_{it})$ is clearly important to our technique. For the multinomial logit, this function takes on a particularly simple form: $y_{ijt} - p_{ijt}(\boldsymbol{\psi}_{it}^{(k)})$. Thinking of $x_{it}\beta_j^{(k)} + \xi_{ijq(t)}^{(k)}$ as the expectation of u_{ijt} at iteration k, our iterative linear regression uses a dependent variable above the expectation for observations with $y_{ijt} = 1$ and below the expectation for observations with $y_{ijt} = 0$.

Our full algorithm is as follows. We begin with a guess of the parameters $\boldsymbol{\theta}^{(0)}$. We iterate on the following sequence of steps, which updates the parameters $\boldsymbol{\theta}^{(k)}$ in iteration k:

1. Minorization step: Given $\boldsymbol{\theta}^{(k)}$, we calculate:

$$v_{ijt}^{(k)} = x_{it}\beta_j^{(k)} + \xi_{ijq(t)}^{(k)} + h_j\left(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}; \boldsymbol{y}_{it}\right). \tag{2.4}$$

2. Maximization step:

   (a) Update $\boldsymbol{\beta}^{(k)}$ by demeaned OLS:

   $$\beta_j^{(k+1)} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}'\tilde{x}_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}'\tilde{v}_{ijt}^{(k)} \;\; \forall j,$$

   where $\tilde{x}$ indicates demeaning at the consumer-quarter-product level.

   (b) Update $\{\boldsymbol{\xi}_{iq}^{(k)}\}_{i=1,\ldots,N,q=1,\ldots,Q}$ by computing:

   $$\xi_{ijq}^{(k+1)} = \frac{1}{T_q}\sum_{t\in\mathcal{T}_q}\left(v_{ijt}^{(k)} - x_{it}\beta_j^{(k+1)}\right) \;\; \forall i, j, q,$$

   where T_q is the number of time periods in quarter q and $\mathcal{T}_q$ is the set of these time periods.

3. Return to Step 1 for iteration $k+1$ as long as the difference between $\boldsymbol{\theta}^{(k+1)}$ and $\boldsymbol{\theta}^{(k)}$ is above some tolerance.

2.4.3 Connection to the EM Algorithm for Binary Probit

If we could observe u_{ijt}, we could estimate our coefficients and fixed effects directly by linear techniques rather than relying on discrete outcome methods such as the multinomial logit. In this sense, Eq.(2.4) in Step 1 of the MM algorithm above has the feel of data augmentation Tanner (1996). That is, we calculate v_{ijt} as an approximation of the unobserved u_{ijt}. That is the intuition behind many applications of the EM algorithm.

In the multinomial logit, v_{ijt} does not coincide with the expectation of u_{ijt}. That is, $v_{ijt}^{(k)} \neq E[u_{ijt}|x_{it}, y_{ijt}, \theta_j^{(k)}]$. Indeed, substituting $v_{ijt}^{(k)}$ with $E[u_{ijt}|x_{it}, y_{ijt}, \theta_j^{(k)}]$ would lead to biased results because the likelihood of the expectation is not equal to the expectation of the likelihood. Generating the expectation of the likelihood function for the case of the multinomial logit is more complicated than our minorization approach, so the EM algorithm is relatively unattractive in our context.

However, the MM and EM algorithm coincide in the case of the binary probit. Indeed, Chen (2019) estimates the binary probit by the EM algorithm and derives a functional form equivalent to our MM algorithm. See also Greene (2018).

In the case of binary Probit,

$$l(x_{it}\beta + \xi_{iq(t)}; y_{it}) = \begin{cases} \Phi(x_{it}\beta + \xi_{iq(t)}) & y_{it} = 1 \\ 1 - \Phi(x_{it}\beta + \xi_{iq(t)}) & y_{it} = 0 \end{cases},$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, $\phi(\cdot)$ is its PDF.[5]

Accordingly,

$$\begin{aligned} h(x_{it}\beta + \xi_{iq(t)}; y_{it}) &= \frac{\partial \log\left(x_{it}\beta + \xi_{iq(t)}; y_{it}\right)}{\partial(x_{it}\beta + \xi_{iq(t)})} = \begin{cases} \dfrac{\phi\left(x_{it}\beta + \xi_{iq(t)}\right)}{\Phi\left(x_{it}\beta + \xi_{i}\right)}, & y_{it} = 1 \\ \dfrac{\phi\left(x_{it}\beta + \xi_{iq(t)}\right)}{1 - \Phi(x_{it}\beta + \xi_{iq(t)})}, & y_{it} = 0 \end{cases} \\ &= \frac{\left(y_{it} - \Phi\left(x_{it}\beta + \xi_{iq(t)}\right)\right)\phi\left(x_{it}\beta + \xi_{iq(t)}\right)}{\Phi\left(x_{it}\beta + \xi_{iq(t)}\right)\left(1 - \Phi\left(x_{it}\beta + \xi_{iq(t)}\right)\right)} \end{aligned}$$

Thus, we see that $h(x_{it}\beta + \xi_{iq(t)}; y_{it})$ corresponds to the well-known Mills ratio and, as a result, $h(x_{it}\beta + \xi_{iq(t)}; y_{it}) = E[\epsilon_{it}|y_{it}]$, where ϵ_{it} corresponds to the normally distributed error from the probit model. In this sense, the MM and EM algorithms correspond in the case of the binary probit.

	Binary probit	Binary logit	Multinomial logit	
	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}_2$	$\hat{\beta}_3$
MLE	0.9697	1.0184	0.9983	0.4775
MM-algorithm	0.9697	1.0184	0.9983	0.4775

Table 2.1: Comparison of $\hat{\beta}$ between MLE and MM-algorithm

2.4.4 Simulation Results

In this section, we show that our iterative algorithm generates results for $\hat{\beta}$ that are numerically identical to those obtained using standard MLE techniques. We generate data from three nonlinear models with a linear index: (1) a binary probit model, (2) a binary logit model and (3) a multinomial logit model. For each model, we perform the estimation using both traditional gradient-based techniques[6] and our MM algorithm. We generate simulated data using $N = 10$, $T = 1000$, $J = 2$ for binary choice models and $J = 3$ for the multinomial logit. We assume there are consumer-level fixed effects in all models, generated independently from the standard normal distribution. We do not add a time element to the fixed effects in this exercise. We let x_{it} be a scalar, also drawn independently from the standard normal distribution. For the two binary choice models, we set the true parameter $\beta = 1$; for the multinomial logit model, we normalize $\beta_1 = 0$ and $\xi_{i1} = 0,\ \forall i$, and set the true parameters $\beta_2 = 1$ and $\beta_3 = 0.5$. The results are summarized in Table 2.1.

2.5 Data

In our empirical investigation, we take advantage on the Nielsen Consumer Panel Dataset, available through the Kilts Center for Marketing at the Chicago Booth

[5]Notation looks slightly different here because $J = 2$. It is no longer necessary to treat β, $\xi_{iq(t)}$ and y_{it} as vectors with only two choices.

[6]For the two binary choice models, we use the R function “glm” in package “stats”, with its default method iterative reweighted least square (IWLS) ; for multinomial logit, we use the R function “mlogit” in package “mlogit”, with its default method Broyden–Fletcher–Goldfarb–Shanno (BFGS)

School of Business. The data set provides detailed coverage of purchase choices at household level, including the quantity bought and price paid for each product. The data includes detailed information on products purchased characteristics based on UPC codes scanned by the panelists. Crucially for our study, consumers indicate how they paid for each shopping trip. Finally, Nielsen verifies the information using receipts submitted by panelists.

We focus our study on three payment choices: cash, check, and card. Following the approach adopted in most payment literature (for example, Klee, 2008), the "card" category pools purchases categorized as either debit or credit card. This approach reflects concerns about panelists not distinguishing accurately between the two types of payment card.[7]

Our data set includes the five-year time period between 2013 and 2017, and captures over 31 million shopping trips made by 77,657 households.[8] Reflecting turnover in the data set, on average we observe a household for 2.53 years. Nonetheless, there is a considerable number of households which remain longer than the average: 44.3% of households say for longer than three years, 31.0% stay longer than four years, and 19.9% of households stay for the entire five years.

The average household in the data set makes 403.6 shopping trips in total. This translates into 153.7 trips per year on average, or 36.4 trips per quarter.[9] The number of shopping trips per quarter varies between households, with households at the 25th

[7]Debit cards may be authorized by signature or PIN (Personal Identification Number, consisting of 4 to 6 digits), while credit cards are typically authorized by signature. Industry studies and previous literature suggest that many U.S. consumers do not understand the difference between signature debit and credit cards. Past versions of the Nielsen panel appeared to give consumers contradictory instructions on this issue, for instance, instructing consumers to indicate "credit" if they used a signature. We found it difficult to verify the instructions for the current data set. See Cohen, Rysman and Wozniak (2017).

[8]Although the Nielsen panel runs over a decade, the version of the data available through Kilts Center for Marketing only includes payment choice information beginning in 2013.

[9]Among these trips, 34.6% involve grocery stores, other top types of retailers include discount stores (15.7%), drug store (6.0%), dollar store (4.9%), warehouse club (4.9%), quick serve restaurants (3.1%), etc.

percentile making 18 trips on average, while those at the 75th percentile making an average of 50 trips. As our most granular specification uses two choice fixed effects per household-quarter in a non-linear model, the incidental parameters problem is potentially an issue with these numbers of observations.

Over our entire data set, the market share for transactions for cash, check, and card is 30.9%, 2.4%, and 66.8%, respectively.[10] These shares vary substantially across different types of transactions and across households. Figure 2·2 shows the steady increase in the total number of transactions over time, as well as how the payment choice market shares changed over time. In particular, the figure shows a substantial increase in card use over time, as well as a falling market share for both cash and check.

Table 2.2: Transaction size distribution ($)

Mean	Std. Err.	10%	[25%, 75%]	90%
46.84	65.10	5.22	[11.94, 57.06]	107.51

Following previous literature, we examine transaction size as a key driver of payment choice. Table 2.2 illustrates the distribution of transaction size in our data set. While the average transaction size is $46.84, the variation in transaction size is large. In particular, the 10th percentile in the distribution is just $5.22, the interquartile range is [$11.94, $57.06], and the 90th percentile is $107.51.

Figure 2·3 illustrates how important transaction size is in determining payment choice. In particular, the figure shows that the market share of cash falls from above 60% to below 20% as transaction size moves from $5 to $150, with most of the remaining share absorbed by card. Similar to card, the market share for check also increases with transaction size, although it only rises to around 4% for the largest

[10] Two types of payment instruments in the raw data are excluded from our estimation, they are "Scanner does not collect Method of Payment" and "Other Payment" which accounts for 40.3% and 1.71% of the trips in the raw data.

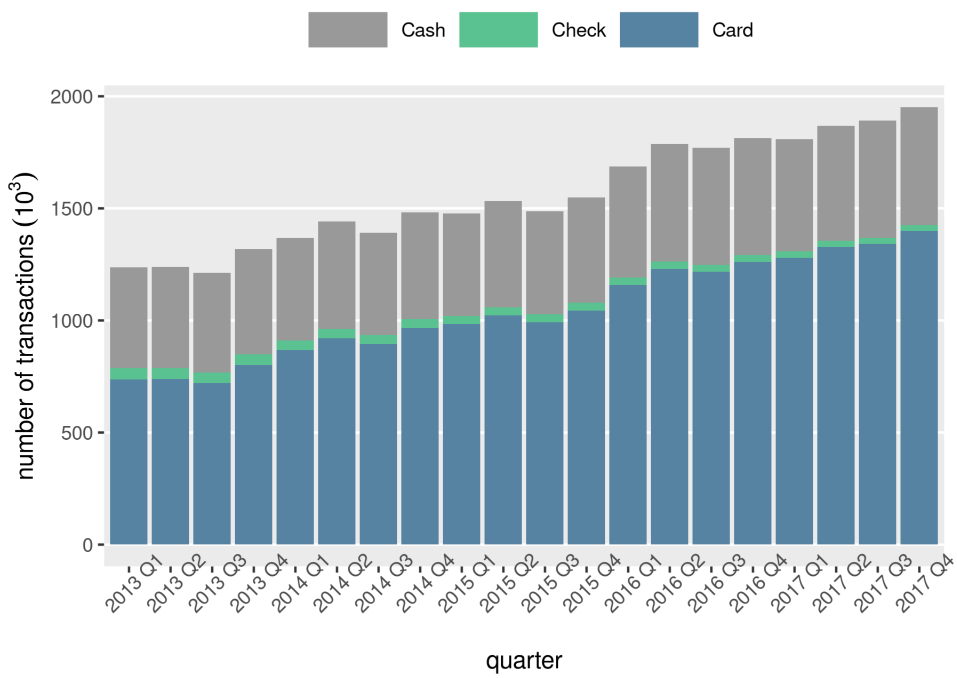


Figure 2·2: Transactions over time

transactions. It is important to recognize, however, that while check has a low market share overall, 37.6% of the households in our data used check at least once.

2.6 Results

We now turn to specifying the multinomial logit model for use in our analysis of households' payment choice. In the specification used in our paper, households are faced with an exogenously-determined set of shopping trips with predetermined transaction sizes, for which they must choose a payment instrument. In particular, household i on shopping trip t where it pays with instrument $j \in \{cash, check, card\}$, receives utility:

$$u_{ijt} = \beta_j \log(x_{it}) + \xi_{ijq(i,t)} + \varepsilon_{ijt}.$$

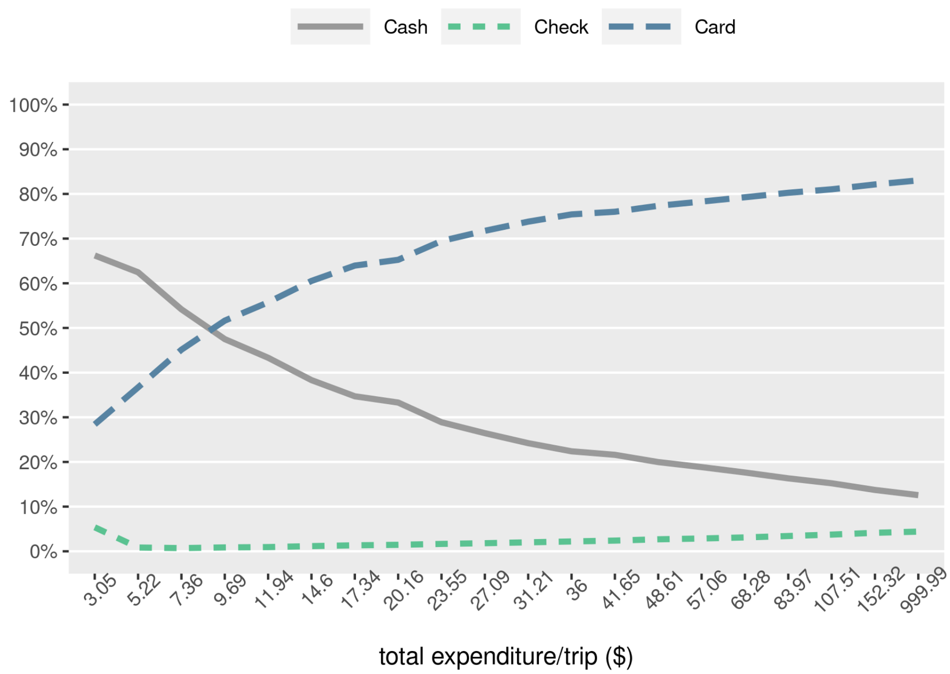


Figure 2·3: Market share in transactions by transaction size

where $q(i,t)$ is the quarter when the shopping trip takes place, and x_{it}, a scalar, is the transaction size in dollars. Note that in our approach, t indexes shopping trips rather than calendar time, so two households may be in different quarters for the same shopping trip number. Thus, rather than write $q(t)$ as in Eq.(2.1), we write $q(i,t)$. As above, ε_{ijt} is distributed Extreme Value.

As is standard, we normalize the mean utility of one choice to zero. In particular, we normalize the utility of $j = cash$ to zero, so $\beta_{cash} = 0$ and $\xi_{i,cash,q(i,t)} = 0$ for all i and q. We interpret the rest of the coefficients as the value relative to the value for cash.

Looking to determine the importance of accounting for unobserved consumer preferences in the payment choice context, we specify an alternative version of our model which does not account for household fixed effects. In particular, for the alternative version of the model we specify $\xi_{ijq(i,t)} = \xi_j$ for all i and q. Finally, we also consider a model with a more limited set of household-payment choice fixed effects, where $\xi_{ijq(i,t)} = \xi_{ij}$ for all q.

Results appear in Table 2.3. Standard errors in this table are conventional maximum likelihood standard errors derived from the inverse of the Hessian matrix. As the Hessian is very large, we exploit the sparsity of the matrix in order to invert it, as presented in Appendix B.2.

Before discussing parameters, it is worth considering how long it takes to estimate these models. For the case of only payment choice fixed effects (the first panel), we estimate the likelihood model by both our MM algorithm and by a traditional gradient optimization routine, BFGS. Results are numerically very close, especially for β_{card}, most likely because of the considerably larger number of shopping trips paid for with card. Moreover, our experience suggests the remaining differences could be further reduced by lowering the tolerance levels in the optimizers. The last row of the table shows that estimating the model using the MM algorithm takes about half the time it takes using the BFGS algorithm. For the case with household-payment choice fixed effects, the MM algorithm takes about 13 hours to estimate, while for the case with household-quarter-choice fixed effects, it takes about 24 hours. In contrast, we ran the BFGS algorithm using a dummy variable approach to implementing the fixed effect estimation for these two cases, and never reached convergence for either case.[11] We do not report BFGS results for these cases.

[11] We write “greater than 24 hours” in the table, but in practice, we let the program run for at least twice this.

	ξ_j			ξ_{ij}			ξ_{iqj}	
	MM	BFGS		MM	BFGS		MM	BFGS
β_{check}	0.7188	0.7278	β_{check}	1.0730		β_{check}	1.1595	
	(0.0012)	(0.0012)		(0.0017)			(0.0029)	
β_{card}	0.7085	0.7085	β_{card}	1.0563		β_{card}	1.1570	
	(0.0004)	(0.0004)		(0.0006)			(0.0011)	
ξ_{check}	-4.8441	-4.8760	$\bar{\xi}_{check}$	-7.7561		$\bar{\xi}_{check}$	-7.9493	
			$[min, max]$	[-15.2902, 6.5288]		$[min, max]$	[-16.0147, 7.4957]	
ξ_{card}	-1.4628	-1.4627	$\bar{\xi}_{card}$	-2.1049		$\bar{\xi}_{card}$	-2.0048	
			$[min, max]$	[-15.1744, 6.6956]		$[min, max]$	[-15.9977, 8.0039]	
Number of FE's estimated	N = 2			N = 155,314			N = 1,631,808	
time	$\sim$ 30 min	$\sim$ 1 hr		$\sim$ 13 hr	$>$ 24 hr		$\sim$ 24 hr	$>$ 24 hr

Notes: All estimations were run on a CPU with 8 8G-memory processors

Table 2.3: Estimates & Computational Time: MM vs. BFGS

We now turn to analyzing the MM estimates for the three alternative model specifications. First, as expected, we find in all specifications that the estimated coefficients on transaction size are positive for both check and card. This agrees with findings in previous papers, as well as the trends presented in Figure 2·3 – namely, that the likelihood that consumers pay with check or card increases significantly for larger transactions. Second, we look to compare the estimated coefficients for the three model specifications. However, although we find that richer fixed effects in the second and third model specifications result in higher transaction value coefficients, the absolute values of the estimated fixed effects also grow considerably. To get a better sense of the impact of accounting for unobserved household heterogeneity, we turn to computing average marginal effects.

Marginal effect of transaction size: Given our multinomial logit assumption, we define the marginal effect (ME) of x_{it} on the probability of choosing a payment method, $p_{ijt}(\theta)$, as:

$$\text{ME}_{ijt} = \frac{\partial p_{ijt}(\theta)}{\partial x_{it}} = p_{ijt}(\theta) \left(\beta_j - \sum_{k=1}^{3} \beta_k p_{ikt}(\theta) \right) / x_{it}. \tag{2.5}$$

To summarize, we report the average marginal effect (AME) of transaction size on each payment method by averaging Eq.(2.5) across all households i and trips t.

Results appear in Table 2.4. Focusing on the first row of results, the table shows

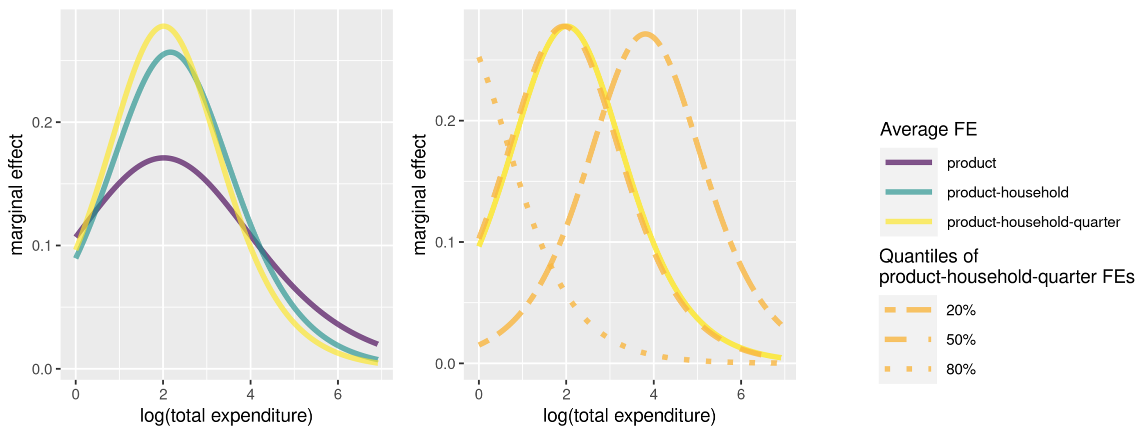


Figure 2·4: Marginal effect of log transaction size) on card usage

	ξ_j			ξ_{ij}			ξ_{ijq}		
	cash	check	card	cash	check	card	cash	check	card
$\hat{AME}$	-0.0107	0.0004	0.0103	-0.0101	0.0005	0.0096	-0.0100	0.0005	0.0095
$\tilde{AME}$				-0.0099	0.0005	0.0094	-0.0097	0.0005	0.0092
Bias: $\hat{AME}$ - $\tilde{AME}$				-0.0002	0.0000	0.0002	-0.0003	0.0000	0.0003

Table 2.4: Average marginal effects of transaction size

that the specifications with household-payment choice and household-choice-quarter fixed effects (panels 2 and 3) result in very similar AME estimates. Comparing these results to those obtained using the specification with payment choice fixed effects only (panel 1), we find that accounting for unobserved household preferences moves the AME towards zero, though the magnitude is under 10%. To understand the intuition for this finding, consider how the different model specifications seek to explain behavior of households with large transaction sizes that predominantly use check or card. The first specification is not able to account for this household heterogeneity, and instead seeks to explain the data by increasing the estimated effect of transaction size on likelihood of paying with check and card.

To unveil the heterogeneity across households that is masked by the relatively small difference on average, we graph how the marginal effects of transaction size on card usage vary for different fixed effect specifications in Figure 2·4. In the left panel,

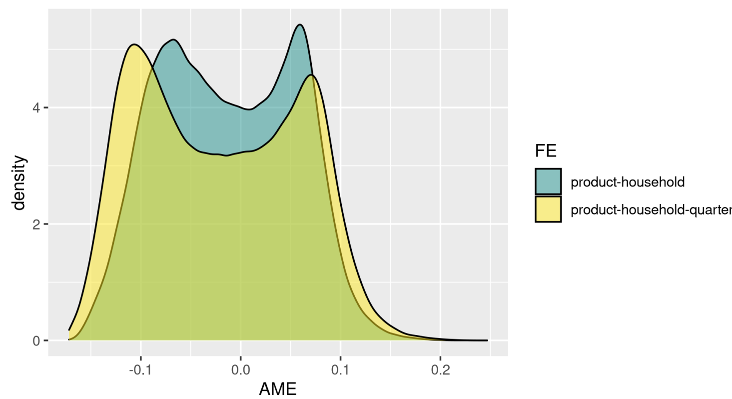


Figure 2·5: Distribution of difference from the product FE model in marginal effect

we see that the household fixed effect specifications generate higher marginal effects for moderate transaction sizes but lower marginal effects for very low and very high transaction sizes, therefore leading to a small average difference. Additionally, the lines with richer fixed effects have more variation than the line with only choice fixed effect, echoing their higher coefficients on transaction size in Table 2.3.

The left panel of Figure 2·4 averages over the estimated fixed effects for any given transaction size. In order to better understand how the model with richer fixed effects changes the marginal effect, we hold the cash fixed effect constant at the average value and graph the marginal effect of transaction size on card usage for three different levels, the 20th, 50th and 80th percentiles of the distributions of the estimated household-quarter check and card fixed effects. The result is in the right panel. The different levels of fixed effects produce parallel shifts of the marginal effect function. Averaging these lines together over the realized fixed effects leads to the household-quarter-choice fixed effect line in the left panel. Overall, Figure 2·4 reveals significant heterogeneity in marginal effects under the specifications with expanded fixed effects.

To further explore this heterogeneity, we calculate the difference in AME between the baseline specification (only choice fixed effects) and the two other specifications with richer fixed effects for each household-quarter. Figure 2·5 presents the distribution of this difference.[12] In Figure 2·4, the marginal effects based on the baseline specification are typically between 0.1 and 0.2, except for very high transaction sizes; meanwhile, we see in Figure 2·5 that the differences between specifications are often at a similar level, which indicates that a difference greater than 50% is not uncommon. Also, the dip in the middle of the distributions suggests that there are relatively few households for which the baseline specification is accurate. Thus, bias from leaving out household-quarter fixed effects has a substantial impact on measured marginal effects.

The incidental parameter problem: The parameters presented in the first row of Table 2.3 are subject to potential bias due to the incidental parameters problem. To correct for such potential bias, we adopt the split-panel jackknife correction developed in Dhaene and Jochmans (2015). In particular, we divide the set of shopping trips taken by each household into two halves, and re-estimate the model on each of the two subsets. Let $\hat{\boldsymbol{\beta}}$ denote the estimates from the full sample, and $\hat{\boldsymbol{\beta}}^1$ and $\hat{\boldsymbol{\beta}}^2$ be the estimates from each half, such that:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^1 &= \mathit{argmax} \textstyle\sum_{i,t\in[1,\lfloor T_i/2\rfloor]} \log l\left(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(i,t)}(\boldsymbol{\beta}); \boldsymbol{y}_{iqt}\right) \\ \hat{\boldsymbol{\beta}}^2 &= \mathit{argmax} \textstyle\sum_{i,t\in[\lfloor T_i/2\rfloor+1,T_i]} \log l\left(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(i,t)}(\boldsymbol{\beta}); \boldsymbol{y}_{it}\right) \end{aligned}$$

Here, we write $\boldsymbol{\xi}_{iq(i,t)}(\boldsymbol{\beta})$ to emphasize that the estimates of the fixed effects will change in both estimations. With these results, the bias-corrected estimates are:

$$\tilde{\boldsymbol{\beta}} = 2\hat{\boldsymbol{\beta}} - \frac{\hat{\boldsymbol{\beta}}^1 + \hat{\boldsymbol{\beta}}^2}{2}.$$

The effect of the bias correction on the estimated coefficients is shown in Table 2.5.

[12] To ease the comparison between Figure 2·4 and Figure 2·5, the AME's in Figure 2·5 are with respect to the log of transaction size.

The bias correction reduces the parameter estimate of both β_{check} and β_{card}, although the change is relatively small in magnitude. As expected, the bias correction is larger in the third specification with the richest fixed effects. Nonetheless, the relatively small magnitude of the bias corrections suggests that, even in the richest specification, the average number of trips per quarter is 36.4 is sufficiently large to significantly mitigate the incidental parameters problem. Looking at the second row in Table 2.4, the bias correction results in a similarly small reduction in the absolute values of the estimated AMEs.

	ξ_{ij}		ξ_{ijq}	
	β_{check}	β_{card}	β_{check}	β_{card}
$\hat{\beta}$	1.0730	1.0563	1.1595	1.1570
$\tilde{\beta}$	1.0519	1.0319	1.1247	1.1146
Bias: $\hat{\beta} - \tilde{\beta}$	0.0211	0.0244	0.0348	0.0425

Table 2.5: Split-panel Jackknife Correction

2.7 Long-term decomposition

Figure 2·2 illustrates the extent to which the share of card usage has grown over time. The goal of this section is to estimate the extent to which each of the many different potential factors contributed to this growth. In particular, one of the key factors that could have contributed to this growth is a gradual increase in household preferences for card payments. At the same time, changes in the composition of transactions or transaction sizes across households, as well as entry and exit of households could have also resulted in a shift of payments towards card. Intuitively, consider a young household that always pays using a card and an older household that always uses cash or check. If the young household has children, its average number of shopping trips and average transaction size will both likely increase. Similarly, once the older household reaches retirement age its average number of shopping trips and average transaction size will both likely shrink. In this example, the market share of card would increase purely due to changes in the composition of transaction number and size, without any changes in preferences of individual households. Similarly, the older household leaving the sample and being replaced by a another young household which favors card over cash/check would result in further growth in card's market share, this time due to changes in the composition of households in the sample.

To facilitate discussion, we introduce new notation. First, we use $\mathcal{I}_q$ to denote the set of households in quarter q and $\mathcal{T}_{iq}$ to denote the set of trips household i took in quarter q. The transactions market share of payment choice j in quarter q is thus:

$$s_{jq} = \frac{1}{\sum_{i \in \mathcal{I}_q} |\mathcal{T}_{iq}|} \sum_{i \in \mathcal{I}_q, t \in \mathcal{T}_{iq}} \frac{\exp\left(x_{it}\beta_j + \xi_{ijq(i,t)}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ikq(i,t)}\right)}.$$

The market share s_{jq} can change over time for a number of reasons: (a) the number of transactions $|\mathcal{T}_{iq}|$ can change, (b) the average size for those transactions x_{it} can change, (c) household preferences $\xi_{ijq(i,t)}$ can change, or (d) the set of households $\mathcal{I}_q$ can change, which can be further broken down into entry and exit. We proceed by sequentially fixing each of these values at their realization in the first quarter each household i is observed in the data, denoted as $\underline{q}(i)$, or in the case of exit the last quarter denoted as $\overline{q}(i)$, and then computing market shares for the last quarter.

Transaction size distribution within households: For each household, we fix the number of trips and the transaction size on each trip at the level of their first quarter, but let their fixed effects evolve with time. We calculate the household-level choice probabilities and then aggregate them to market shares with the number of trips in the current quarter as weights. So the counterfactual last quarter market share is:

$$s^1_{jQ} = \frac{1}{\sum_{i \in \mathcal{I}_Q} |\mathcal{T}_{iQ}|} \sum_{i \in \mathcal{I}_Q} \frac{|\mathcal{T}_{iQ}|}{|\mathcal{T}_{i\underline{q}(i)}|} \sum_{t \in \mathcal{T}_{i\underline{q}(i)}} \frac{\exp\left(x_{it}\beta_j + \xi_{ijQ}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ikQ}\right)}. \tag{2.6}$$

Consider the case in which the set of transactions sizes realized in $\underline{q}(i)$ was the same as in Q. That would imply that the number of transactions in each period was the same, so $|\mathcal{T}_{iQ}| = |\mathcal{T}_{i\underline{q}(i)}|$, and the set of x_{it} was the same for the first and last period that i was in the data. In this case, $s_{jQ} = s^1_{jQ}$. The difference $s_{jQ} - s^1_{jQ}$ provides a measure of how changes in the distribution of transactions contributes to the change in market share $s_{jQ} - s_{j1}$.

Household-quarter-choice fixed effects: We capture the change in preferences within s with our -quarter-choice fixed effects. In order to mute the effect of preferences, we fix -quarter-choice fixed effects at the level of the first quarter the is observed and then calculate the market share in the final quarter Q as:

$$s^2_{jQ} = \frac{1}{\sum_{i \in \mathcal{I}_Q} |\mathcal{T}_{iQ}|} \sum_{i \in \mathcal{I}_Q, t \in \mathcal{T}_{iQ}} \frac{\exp\left(x_{it}\beta_j + \xi_{ij\underline{q}(i)}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ik\underline{q}(i)}\right)}. \tag{2.7}$$

In this case, $s_{jQ} - s^2_{jQ}$ provides a measure of the contribution of changes in -quarter-choice fixed effects, and this term equals zero only if fixed effects are the same in the first and last period.

Number of transactions across households: As in the earlier young vs. older household example, the growth of card usage in this case could also be due to shifts in transactions from non-card to card users. To isolate this effect, we first calculate the household level choice probabilities, and when aggregating them to compute market share, we weight by the number of trips in the household's first quarter rather than the number of trips in the current quarter. Then, the last-quarter market share becomes:

$$s^3_{jQ} = \frac{1}{\sum_{i \in \mathcal{I}_Q} |\mathcal{T}_{i\underline{q}(i)}|} \sum_{i \in \mathcal{I}_Q} \frac{|\mathcal{T}_{i\underline{q}(i)}|}{|\mathcal{T}_{iQ}|} \sum_{t \in \mathcal{T}_{iQ}} \frac{\exp\left(x_{it}\beta_j + \xi_{ijQ}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ikQ}\right)}. \tag{2.8}$$

Entry: In this scenario, we focus on those households that remain in the data set all the way from the first to the last quarter. They are allowed to exit as observed in the data, in order to be distinguished from the exit channel. Specifically, starting with the 31,178 households in 2013 Q1, 15,477 of them stay until the last quarter 2017 Q4, which is 32.6% of all households at that time. The market share in the final quarter when fixed $\mathcal{I}_Q = \mathcal{I}_1$:

$$s^4_{jQ} = \frac{1}{\sum_{i \in \mathcal{I}_1} |\mathcal{T}_{iQ}|} \sum_{i \in \mathcal{I}_1, t \in \mathcal{T}_{iQ}} \frac{\exp\left(x_{it}\beta_j + \xi_{ijQ}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ikQ}\right)} \tag{2.9}$$

Exit: We consider a counterfactual scenario where no households leave the sample. Therefore, all households that ever show up in the sample stay until the last quarter 2017 Q4, i.e. $\mathcal{I}_Q = \bigcup_{q=1}^{Q} \mathcal{I}_q$, which gives 77,656 households. For those households that leave before 2017 Q4, we assume that their number of trips, the transaction size of each trip and fixed effects are the same as in the last quarter that they are observed in the data, i.e. $\mathcal{T}_{iq} = \mathcal{T}_{i\overline{q}(i)}, \forall q > \overline{q}(i)$. In particular, $\overline{q}(i) = Q$ for household i that is observed in 2017 Q4. In this scenario, the market share in the final quarter is:

$$s^5_{jQ} = \frac{1}{\sum_{i \in \bigcup_{q=1}^{Q} \mathcal{I}_q} |\mathcal{T}_{i\overline{q}(i)}|} \sum_{i \in \bigcup_{q=1}^{Q} \mathcal{I}_q, t \in \mathcal{T}_{i\overline{q}(i)}} \frac{\exp\left(x_{it}\beta_j + \xi_{ij\overline{q}(i)}\right)}{\sum_{k=1}^{J} \exp\left(x_{it}\beta_k + \xi_{ik\overline{q}(i)}\right)} \tag{2.10}$$

Then the contribution of each channel is the difference $s_{jQ} - s^k_{jQ}$ for each $k = \{1, 2, 3, 4, 5\}$. Note that the sum of these differences does not exactly equal $s_{jQ} - s_{j1}$, in part because of joint effects. By isolating each effect separately, we do not capture the role of simultaneous changes in channels, for instance because in practice, ξ_{ijq} and x_{it} change jointly. Still, these differences give a first-order approximation of how much each type of change contributes to the overall change. Therefore, for demonstration purpose, we rescale these differences so that the sum of them equals to $s_{jQ} - s_{j1}$.[13]

The results of the decomposition are shown in Figure 2·6. In particular, the figure

[13] In this sense, our measure is similar to Variance Partition Coefficients, as in Goldstein, Browne and Rasbash (2002). See also Grömping (2007).

illustrates that while changes in household payment preferences are the single biggest driver of long-term change in market share, on their own they account for only about a third of the overall change. Entry of household with stronger preferences for card, as well as exit of households with a relatively higher preference for card and check, are the other two factors which contributed significantly to the growth in card's market share between 2013 and 2017. By contrast, we find that the growth in transaction sizes within households with a stronger preference for card barely contributed to the growth in card's popularity over time. What's more, we find that changes in the composition of the number of transactions across households actually contributed against the growth in card's popularity, suggesting over time household with a stronger preference for card reduced the number of trips they make relative to other households. Overall, our results show that changes in consumer payment preferences were one of the key factors that contributed to growing popularity of card, accounting for about a third of the growth observed over the time period 2013-17. At the same time, we find that the largest contribution came from a change in the composition of the Nielsen sample. This finding suggests that changes in payments usage over time is driven in large part by ingrained preferences of young consumers entering the economy, which the consumers will then only slowly change over the course of their lives.

2.8 Conclusion

Although the transition to digital payments has been one of the most significant developments in the payment industry in recent years, the continued prevalence of cash and check raise important policy questions. This paper studies the determinants of payment choice in the short and long term. Decomposing the drivers of long-term payment shifts allows us to contributing significantly to payment literature, which typically focuses only on short-term payment decisions. Key to this is our ability

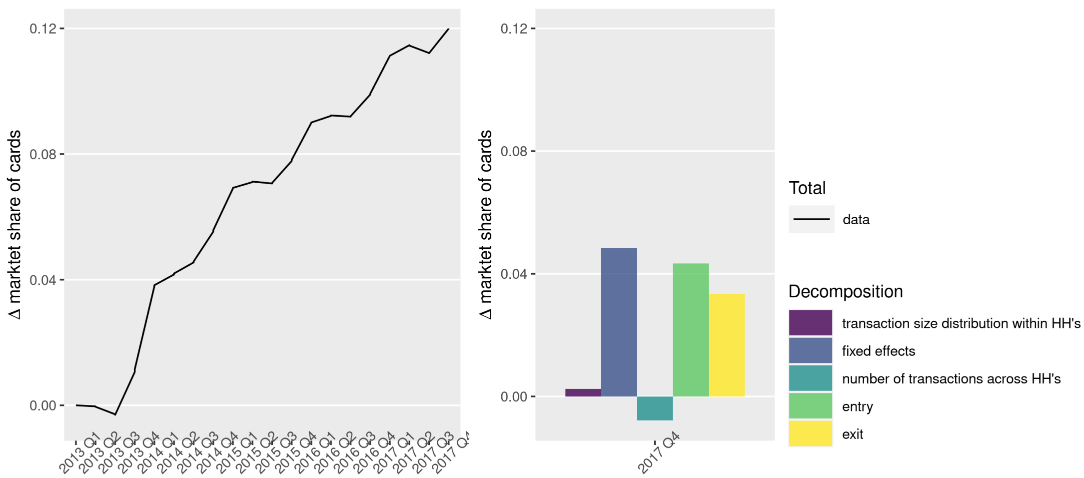


Figure 2·6: Long-term decomposition

to observe, and properly capture in our model, unobserved household preferences for payments.

In our paper, we use a novel source of data on payment behavior: consumer panel surveys. Although these sources are typically used to study consumer shopping behavior and responses to advertising, we show that these data can be usefully employed to study payment behavior. Doing so allows us to keep track of individual households' payment behavior over multiple year through the lens of high-frequency shopping trip data.

Crucially for our analysis, we adapt novel statistical methods to account for unobserved household heterogeneity using panel data techniques. Doing so allows us to estimate a multinomial models of payment choice with over 1 million fixed effects, which would not have been possible using traditional techniques used to estimate non-linear models. In particular, we present a new method for addressing fixed effects in multinomial models based on the Minorization-Maximization (MM) algorithm, which can be seen as a generalization of the Expectation-Maximization (EM) algorithm. While the MM algorithm has a significant history in the statistics literature, we are aware of almost no presence in econometrics. We discuss the application of our method not only to the multinomial logit model, but also several other well-known models.

The results of our estimation shine new light both on the short and long-term payment decisions. First, our results suggest that while transaction size is an important determinant of payment choice in the short term, its effect is smaller than previously estimated in papers not able to directly account for unobserved household payment preferences. Looking to the long term, we use our model to study the key factors driving the increase in card usage observed in our data. In particular, we decompose the changes in payment instruments' market shares into those (a) driven directly by changes in household preferences, (b) driven by changes in the number and average value of transactions made by individual households, and (c) driven indirectly by entry of younger households into and exit of older households. Our results suggest that while changes in household payment preferences are an important factor, they explain only about a third of the observed growth in card usage. Instead, the model suggests that the primary driver of long-term changes in payments has been the entry of young households with stronger preferences for card payments, as well as exit of older households with stronger preferences for cash and check payments.

Chapter 3

Spatial Competition and Missing Data: an Application to Cloud Computing

3.1 Introduction

For a variety of reasons, firms historically care about physical location when making decisions about where to invest in physical capital.[1] Because the internet lowers costs of communication over space, it was reasonable to suspect internet adoption could mitigate the importance of physical location for investment decisions. According to this theory, the internet allows rural firms to have access to similar resources as urban firms without moving to urban locations, allows faster rural growth and ameliorates incidence issues of agglomeration economies (Forman, Goldfarb and Greenstein, 2008; Glaeser and Gottlieb, 2009).

Evidence of the internet leading to the "death of distance" is scant, however, for regional economic outcomes. At the regional level, Forman, Goldfarb and Greenstein (2012) shows that advanced internet is not sufficient, in and of itself, to enable wage growth in a city. Rather, they find advanced internet is a complement to a city's existing human capital and stock of firms. At the firm level, Giroud (2013) shows that between 1977-2005, decreased travel times to plants through new plane routes or roads led to increased investment and productivity. Hence, traditional investment decisions

[1]Monitoring and information acquisition is one explanation (Giroud, 2013). Transportation costs, despite long run declines, are another (Glaeser and Kohlhase, 2004). Agglomeration economies are a third (Glaeser and Gottlieb, 2009).

vary positively with proximity. A sharp test for the death of distance hypothesis, then, is how the internet impacts the affinity for proximity in firm investment decisions.

In this paper, we ask whether firm investment decisions enabled solely by the internet, cloud computing, systematically shows a predilection for proximity. Cloud computing providers, like Amazon Web Services (AWS) and Microsoft's Azure, rent compute resources called "virtual machines" to customers who connect to them via the internet. Once connected to a virtual machine (VM) via the internet, cloud users can perform functions previously required to be hosted by on-premise servers like compute operations, read and write data operations or web application hosting.[2]

There is a large and growing demand for cloud computing: according to Gartner, the worldwide public cloud services market is projected to grow 17.3 percent in 2019 to total $206.2 billion, up from a $175.8 billion forecast in 2018.[3] Cloud computing is also important for general economic growth and productivity because renting cloud compute resources lowers fixed hardware capital costs for start-up firms and turns them into marginal operating expenses. Despite this fast growth and economic importance, there is very little empirical work on understanding the cloud computing market and welfare derived from it.

When a firm deploys a cloud computing instance, they must choose a physical location where to deploy it. Cloud providers have multiple data center locations and prices vary by location. Working with a far-away data center can impact latency. Latency is the time between a task request and task execution. As a rule of thumb, distance related latency is on the order of one millisecond per 100 miles.[4] To put

[2]What was formally a fixed investment costs of server ownership into variable costs. With some configurations or "stacks" cloud users can outsource IT personnel to experts employed by the cloud providers.

[3]See "Gartner Forecasts Worldwide Public Cloud Revenue to Grow 17.3 Percent in 2019", accessed on 09/30/2018, `https://www.gartner.com/en/newsroom/press-releases/2018-09-12-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2019`

[4]See `https://www.365datacenters.com/portfolio-items/beyond-bandwidth-distance-matters-choosing-data-center/`. Data packets travel at the speed of light in a vacuum but in

that in context, a 2017 research by Google finds that "average time to first byte" (a measure of web server responsiveness) was roughly 2,000 milliseconds when averaged across all websites in the U.S.[5] Thus, an increase of roughly 1,500 miles of distance between a VM user and the data center location corresponds to a latency increase of ∼15 milliseconds (less than a 1% increase in time to first byte for the average U.S. web server). Hence, over relatively small distances like those faced by customers in our dataset (e.g., choosing between a data center 1,000 versus 1,500 miles away) latency might not always be the sole disutility of selecting far away data centers outside of niche use cases like high frequency trading.

We test two hypotheses for how distance impacts firm investment decisions in this paper. First, we investigate the strength of firm preferences for proximity when investing in the cloud. Whereas some firms, such as high frequency traders, care a great deal about 5 milliseconds of latency, many users do not. As a result, we view this as a strong test for the importance of distance and firm investment decisions: if firms are willing to pay more to use the nearest data center when only marginally closer than another data center, it is evidence against the "death of distance" hypothesis.

Second, we estimate how competition impacts the "death of distance" hypothesis. Public cloud providers like AWS, Azure and Google Cloud Platform (GCP) are fiercely competing in the rapidly growing cloud computing market. Market competition can impact both which cloud provider customers choose and, for their chosen provider, which specific data center location customers choose. We develop and estimate a structural model of cloud demand to investigate how strategic firm level decisions interact with preferences for proximity.

---

actual fiber which powers the internet, the speed is a bit less. Further, bends in fiber cables can slow down data speeds. Lastly, there are other factors not related to distance which can increase latency like congestion and reading data packets are not directly related to distance.

[5]See `https://think.storage.googleapis.com/docs/mobile-page-speed-new-industry-benchmarks.pdf`.

Having the type of data we have, detailed data for a single firm and aggregate data for another, is a common problem with both developing business strategy and in competition policy. We introduce a novel mixed logit demand model of spatial competition that is estimable with detailed data of a single firm but only aggregate sales data of a second and apply it to the cloud computing industry. The model lets us perform counterfactual analysis over 1) how spatial competition between cloud providers impacts optimal price setting behavior and 2) optimal data center locating decisions. We can thereby show how firms' cloud investment patterns change with competition upstream in the cloud computing provider industry.

We use a proprietary dataset with anonymized customer level zip codes linked to the location of data centers they choose. The dataset consists of all customers who deployed one popular type of VM on Microsoft's Azure in 2016. At the time, Azure was the second largest public cloud provider in the world behind AWS. We restrict the dataset to focus on location decisions of US and Canadian firms to locate in US and Canadian data centers. We leverage the rollout of new data centers in the U.S. and Canada over our time period to provide variation in the choice set of data centers and identify key demand parameters, allowing a subset of demand parameters to vary by a cloud user's industry. Due to large lead times in data center construction and 2016 being early in the public cloud sector, we argue data center location is plausibly exogenous.[6]

We have detailed data of a single firm, Azure, but only aggregate sales data of a second, AWS. Specifically, we use quarterly cloud revenue data from AWS available in their 10-Ks. We treat absent customer level data from the AWS as a missing data problem and leverage the structural model, detailed Azure data and Expecta-

[6]We argue below that over the planning period for these data centers, reliability for servicing internal workloads was the primary reason for data center construction. For example, a two year lead time for data center construction implies that 2014 was the planning period. In 2014, share prices for Microsoft hadn't yet responded to increased cloud revenue reported in 10-Ks.

tion Maximization (EM) algorithm to back out AWS customer location. The EM algorithm addresses the missing data problem iteratively: we first construct an expectation of the likelihood by integrating over the latent consumer locations based on their posterior distribution, and then maximize the likelihood function over demand parameters.

Identifying key demand parameters relies on the rollout of new data centers by both AWS and Azure and 2016 price changes. By observing the rate at which new customers begin purchasing Azure when new AWS or Azure data centers open and how those rates vary over space, we can identify preferences for proximity to data centers. We argue that using data from early days of the cloud computing market and the long lead times to construct new data centers as a source of identification is adequate to identify preference parameters. We project demand parameters identified from the granular Azure data to the observed AWS data center characteristics and sum across AWS data centers. The gap between the projection and the observed AWS market share is attributed to the fixed effects of cloud providers. Lastly, the population distribution of consumers can be inferred by the choice probabilities calculated from the identified demand model and the observed market shares of Azure, accounting for the presence of an outside good (on-premise servers). We show via simulation that the model successfully recovers the demand parameters and unobserved consumer spatial distribution then take the model to the data given Azure's market share.

Our core empirical finding is that cloud users have a preference for nearby data centers. As a result, the spatial layout of DCs relative to customer location induces a significant variation in local market power. We have no identifying variation to test whether this preference for proximity is driven by latency concerns or other factors like a secular preference for proximity. It is hard to imagine latency is driving the magnitude of preference for proximity we find in the data. Our data covers North America and Canada only. Introduction of new data centers in our sample often change distance to nearest data center by only a few hundred miles or latency decreases of a few milliseconds. However, we find that cloud customers are willing to pay roughly 60% premiums for a reduction in distance of roughly 600 miles (i.e., 1000 kilometers).

We use estimated parameters to perform counterfactual exercises to determine how market structure would change if new data centers are introduced in different locations. Among the six possible counterfactual Microsoft Azure data center locations, the most profitable one could generate a market share gain roughly 25% higher than the least. Thus, the revenue reductions to cloud providers of placing a data center sub-optimally are large. The model lets us decompose increases in market share across customers purchasing the outside good (on-premise servers) versus purchasing from a competitor and we find that much of the increase in market share is from the outside good although a meaningful share is from the competitor.

We also perform a counterfactual where we decrease price of all Azure data centers by 15% and assess changes on market shares. Consistent with economic theory of spatial competition, we find that the benefits of price competition are greatest where both Azure and AWS have a data center. Thus, our results provide evidence that spatial competition is important in the early stages of the cloud computing industry. Comparing the two counterfactuals, opening a new data center increases consumer surplus by roughly 75% of the consumer surplus from the price decrease. This is large since the new data center would impact only ∼10% of all Azure customers (e.g., surplus increases only for cloud users that deploy there) but it is plausible given the implied willingness to pay for proximity.[7]

[7] As a back of the envelope calculation if all customers receive a 15% price decrease their customer surplus increases by 15%. A new data center impacts those customers that deploy in it and there

There are three main lessons from this research. First, we find evidence that cloud customers display a material preference for proximity in deploying VMs that is hard to explain with latency issues. Indeed, we show that a large fraction of cloud users do not deploy in the nearest DC implying that latency is often not a major hurdle to cloud deployments. Data center age, for example, correlates with where cloud customers deploy. By focusing on the North American market, we ignore data sovereignty issues but highlight that those are likely to also be important. We view a preference for proximity as being inconsistent with internet enabling the "death of distance" at least over our sample in the early stages of cloud adoption. We acknowledge that latency could be an important issue for some cloud applications, but the magnitude and scale over which preferences for proximity manifest and the observed distances in our dataset (we observe only North American customers) makes it difficult for latency to be plausibly responsible.

Second, our findings imply that market competition could help mitigate incidence issues from spatial allocation of capital. The "death of distance" narrative promised increase growth in rural areas attributable to better access to information and freer flow of goods and services. Our results imply that increased strategic spatial competition as the cloud market matures would reduce equilibrium prices and also increase incentives to invest in additional data centers. Although we don't endogenous DC location decisions in this paper to address it formally, intense competition among the major cloud providers (e.g., AWS, Azure and Google Cloud Platform, Alibaba, etc.) is likely increasing access and surplus to the cloud for all potential cloud customers

were 10 DCs at the end of our sample so roughly 10% of customers benefit from the new DC. Recall that aggregate consumer surplus from the new DC is 75% of the welfare increase from a 15% price decrease for all newly deploying customers. If N are the total number of cloud customers then $.1 * N * \Delta CS_{newDC} = .75 * N \Delta CS_{PriceChange} = .75 * N * .15$ and solving for ΔCS_{newDC} yields the increase in consumer surplus for customers deploying in the new DC in our counterfactual. Hence we must observe an increase in consumer surplus of (.15/.1)*.75 = 112.5%. This is plausible: a new proximate DC could be worth roughly twice as much to cloud users as distant DC based on our parameter estimates.

across both the price and distance margins.

Third, more generally our results show that product managers for goods characterized by spatial competition can effectively estimate demand for their goods using detailed data of only a single firm so long as market data for the competitor firm exists and there is variation in the number of stores over time. In addition to benefits to managers, we highlight how this technique can also be used by economists to perform welfare analysis. While our use case is cloud computing, the technique could be useful for managers and researchers interested in questions regarding the impacts of opening and closing of brick and mortar stores faced with increasing online competition.

This paper contributes to three strands of literature. First, understanding the economic geography of the internet has important incidence implications. Despite higher adoption rates for early internet in rural areas (Forman, Goldfarb and Greenstein, 2005), it appears that the benefits enabled by the internet accrue in only a subset of cities (Forman, Goldfarb and Greenstein, 2012, 2008). Further, recent research suggests that proximity to data centers could cause increased growth (Jin and McElheran, 2019). Our work pushes these findings by investigating how spatial competition could impact the economic geography of internet enabled economic gains.

Second, in the field of discrete choice modeling, applications of EM algorithm date back at least to Bhat (1997), Train (2007) and Train (2008). Many of these applications use EM to address missing data on consumer attributes. In what might be the most closely related EM based approach to ours, Conlon and Mortimer (2013) addresses missing data on product availability. At a high level, competitor sales are similar to missing data regarding any product generally. Unlike these previous papers, though, the data structure in our case has two problems: the aggregate level competitors' data makes both their consumer's attributes and disaggregated (e.g., store level) sales unobservable. Because this is a spatial model of competition, the consumer-store level attributes of AWS are of added importance.

While we view the EM algorithm as the most appropriate remedy for our missing data problem for both efficiency and computational feasibility, there are other related techniques in the literature. Other demand frameworks for a similar data structure include those in Berry, Levinsohn and Pakes (2004), a Berry, Levinsohn and Pakes (1995) inspired model leveraging micro moments of consumer characteristics. However, these "Micro-BLP" models are less efficient than maximum likelihood estimation (MLE) by attenuating the information on choices at individual level. The marketing literature often uses Bayesian techniques in the sense that demand parameters are also treated as latent variables. Examples include but are not restricted to Chen and Yang (2007), Musalem, Bradlow and Raju (2008), Jiang, Manchanda and Rossi (2009), Musalem et al. (2010) and Zheng, Fader and Padmanabhan (2012). Specifically, Feit et al. (2013) is probably the most related work to ours. They use a mixture of individual level usage data for digital platforms and aggregate data on usage for traditional platforms to estimate the multi-platform media consumption, albeit in a context of a multivariate model and computationally more burdensome because of the inevitable Markov Chain Monte Carlo simulation.

Third, this paper expands the existing literature on spatial competition broadly in addition to our application regarding the cloud computing industry. In terms of data structure, previous works on spatial competition usually use either aggregate or disaggregate data only. For instance, Davis (2006) estimated a model of spatial competition in the movie theater industry with market share data. Davis (2006) aggregates consumer heterogeneity with an observed geographic consumer distribution from census data and then focuses on identifying the functional form of travel cost. Smith (2004) estimates a two-stage discrete-continuous model for the supermarket industry, and the complexity of unobserved consumer attributes is circumvented by consumer level data from a survey. While spatial competition and firm entry decisions are important economic questions (Seim, 2006), our novel demand estimation approach to combine micro and macro data can be applied to estimating demand elasticities as well.

The remainder of the article proceeds as follows. In Section 2 and 3, we give a brief introduction of the IaaS public cloud industry and describe the general framework of the model. Section 4 describes the model, describes how EM algorithm can be employed to address the missing data problem, and identification. Section 5 shows the performance of a Monte Carlo experiment. In Section 6, gives the estimation results from the data. Section 7 performs two counterfactual exercises highlighting the spatial competition aspects of cloud implied by estimated preference parameters. We conclude this paper in Section 8.

3.2 Industrial Background

According to the beginner's guide on the website of Microsoft Azure,

"Cloud computing is the delivery of computing services–servers, storage, databases, networking, software, analytics, and more–over the Internet ('the cloud'). Companies offering these computing services are called cloud providers and typically charge for cloud computing services based on usage, similar to how you are billed for water or electricity at home."

Most cloud computing services fall into one of three broad categories: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). In this paper, we focus on IaaS and model consumer's problem as a discrete choice among data centers. Focusing on IaaS over PaaS and SaaS is ideal in our setting because the user must specify a specific DC to deploy their cloud resources. Alternative PaaS and SaaS offerings often have a more curated experience in which the firm

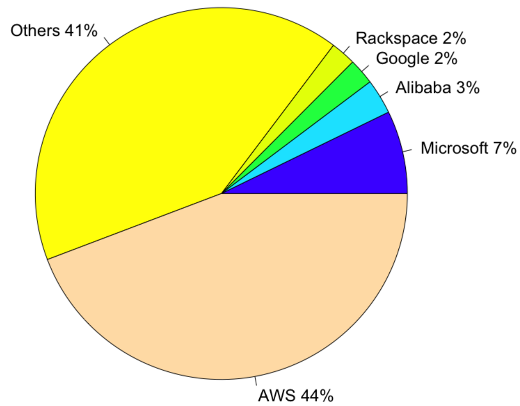
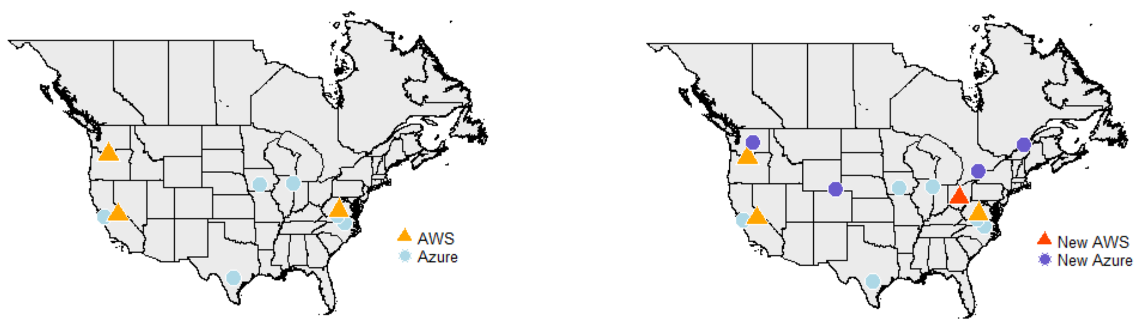


Figure 3·1: 2016 IaaS Public Cloud Computing Market Share from Gartner. Gartner data from survey results of firms. Market shares materially changed over the past five years so that Azure and Google's GCP now have much larger shares.

makes deployment decisions.

DCs are facilities that house computer systems and associated components, such as telecommunications and storage systems. Consumers rent virtual machines (VMs) at DCs as complements to local machines on a pay-as-you-go basis. The value proposition to customers is driven capacity management, cloud providers' economies of scale and management of hardware and security. Put another way, replacing lumpy capital expenditures on wholly-owned servers with smoother operating expenses in the cloud, being able to scale up and down demand for compute resources but not always provision for max demand as with own servers, and outsourcing hardware security concerns all are valuable. Some use cases include housing large datasets, serving website, web App or Application content and performing period machine learning model training.

Amazon Web Services (AWS) and Microsoft Azure are the two firms that have the largest market shares in global IaaS public cloud market. In 2016, the year our data spans, the total value of this market reached $22 billion U.S. dollars, of which AWS had 44%, followed by Microsoft Azure at 7.1% [8], as shown in Figure 3·1. Total cloud demand has increased significantly since 2016 to $44.4 Billion in 2019 with AWS's market share staying roughly constant but Azure's growing to 18%.[9]



(a) Jan 2016 **(b)** Dec 2016

Figure 3·2: North American Data Center (DC) layout in 2016. During the calendar year both AWS and Azure added DCs in different parts of the U.S. and Canada. We leverage how new DC introduction differentially impacts customers in different locations to estimate preferences fro DC proximity.

The core distinction between on-premise servers and cloud computing is that cloud customers rent compute resources from a public cloud provider. When purchasing an on-premise server, a firm puts the server in their compute facility, normally in their office building. When a customer decides to rent compute resources and configure a VM, they select a physical location for that VM to be located. Figure 3·2 shows the location for all U.S. and Canadian data centers of both AWS and Azure. A shorter physical distance between a VM and its users is correlated with lower latency (e.g., shorter wait times for webpages to load). Each firm had a footprint in Canada by the end of 2016. We estimate demand of a single popular SKU for all U.S. and Canadian

[8]See "Gartner Says Worldwide IaaS Public Cloud Services Market Grew 31 Percent in 2016", accessed on 11/01/2018, `https://www.gartner.com/en/newsroom/press-releases/2017-09-27-gartner-says-worldwide-iaas-public-cloud-services-market-grew-31-percent-in-2016`.

[9]See `https://www.gartner.com/en/newsroom/press-releases/2020-08-10-gartner-says-worldwide-iaas-public-cloud-services-market-grew-37-point-3-percent-in-2019`.

consumers with workloads in any DC in either the U.S. or Canada.

Spatial proximity is likely an important aspect of DC differentiation for speed-sensitive users. Data transfer takes time, and the resulted latency could be further amplified due to security protocols. For example, this is likely to be a real concern when considering leverage cloud servers across on the other side of the globe. Our dataset, though, includes only US and Canadian customer demand for VMs within US data center locations. It is plausibly less likely to be an issue for domestic data center location decisions where the U.S. is roughly 3000 miles across and as a rule of thumb, distance related latency is on the order of 100 miles per millisecond. Observing a preference for proximity could be related to latency preferences or non-performance related preferences to be physically close to data centers (sometimes called "server hugging").

Spatial proximity is determined by both DC location and consumer location and consumer heterogeneity along this margin could play a critical role in this demand system. Therefore, the estimation for demand parameters overlooking consumer heterogeneity in location could miss an important consumer preference. This motivates our mixed logit framework. Although consumer location is observable only for Microsoft customers, the EM algorithm we detail below enables a simultaneous estimation of both demand parameters and consumer spatial distribution, which is needed for any policy analysis respecting spatial demand preferences.

3.3 Data

We merge several datasets together for our analysis. These datasets include actual purchase data from Azure customers including customer locations for a single popular general-purpose product or shop keeping unit (SKU), aggregate sales for AWS, data center locations, pricing data for AWS and Azure, census data on business locations

Table 3.1: Comparison between Microsoft *basic A1* and AWS *t2.small*

Name	Brand	vCPUs	RAM(GiB)
basic A1	Mircosoft	1	1.75
t2.small	Amazon	1	2

Note: We compare demand for customers' first deployment of basic A1 for Microsoft and estimate first deployments of AWS' t2.small. These products are similar in terms of performance. Differences in product fixed effects will be covered by AWS fixed effects in the empirical model.

for the U.S. and the analog for Canadian businesses.

The most novel attribute of our data is a random subset of Microsoft customer level choice data for the a general-purpose cloud computing SKU: *basic A1* SKU. The analog of the *basic A1* SKU for AWS we consider is the *t2.small* SKU. A detailed technical comparison between *basic A1* and *t2.small* can be found in Table 3.1. The VMs are similar across CPU and RAM. One difference between the virtual machines SKUs is across product quality: whereas *basic A1* is a dedicated core, *t2.small* is a burstable VM. That means scaling up *t2.small* cores due to a "burst" in compute demand might not always be available if deployed whereas a dedicated core would be. This will be picked up in the brand/product fixed effects we estimate in the empirical model.[10]

There was modest price variation in our sample period. Figure 3·3 shows region level prices across DCs for Azure's *basic A1* and AWS's *t2.small* in 2016. Prices are quoted in the hourly price of deploying a one core Azure *basic A1* or AWS *t2.small* VM. To put these prices into context, at $0.08/hour a one core VM would cost $700.80 if deployed for 24 hours a day for all 365 days in a year, which is more than a one core personal computer would have cost in 2016. The premium accounts for the ability to only pay for what is used (e.g., used the VM for 20 hours then shut it off), in addition to outsourcing security and IT.

Figure 3·3 shows when the first Canadian DCs of Microsoft and AWS were in-

[10]Because we only evaluate one product from both AWS and Azure, brand and product fixed effects are operationally identical in this paper.

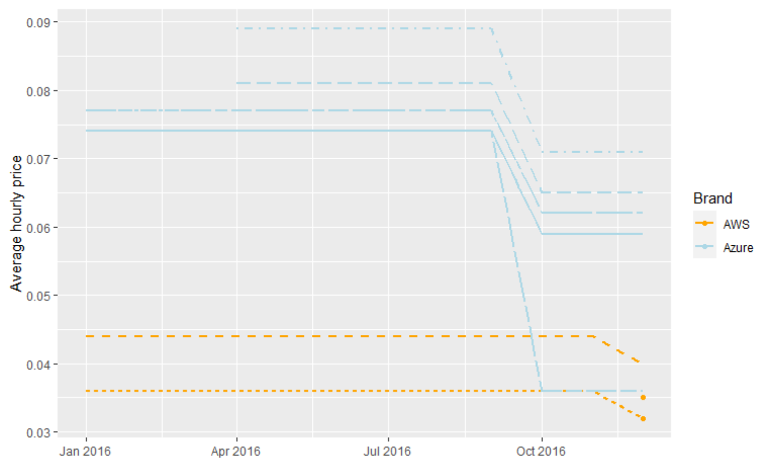


Figure 3·3: Prices of Microsoft *basic A1* and AWS *t2.small* Across Regions.

Note: Figure shows price dispersion for AWS and Azure over time and space. One large price drop for Azure and one small drop for AWS is responsible for identifying price coefficients. New region prices are shown as appearing midway through the year.

troduced in April, 2016 and Dec, 2016, those prices were slightly higher than prices in other regions for Azure. The price drop in Microsoft DCs in October 2016 and price decrease in December 2016 from AWS helps us identify price coefficients. Finally, prices were higher for Azures *basic A1* SKU relative to AWS's *t2.small* reflecting some time invariant differences in attributes which will be picked up by the brand/product fixed effects. Generally speaking, cloud providers tended to have a few large data centers at low prices then smaller data centers more geographically dispersed at slightly higher prices over our sample period.

Related to pricing, we focus on the location of initial VM deployment decision of customers in the paper. When a cloud user deploys a VM they pick a DC where the VM will be deployed. One advantage of the cloud relative to wholly owned servers is cloud user can turn off their VM at any time and stop paying in a "pay as you go" cloud computing business model. Usage decisions are second order for our question of willingness to pay for proximity for the cloud. In order for usage decisions to matter, there would need to be a substitution margin along which cloud users choose a location as a function of both their physical location and the expected duration of the deployment. We view this as unlikely although we discuss how adding a usage decision could impact the model and findings below. Also, by focusing on the initial deployment, we circumvent the complications of multiple DC users.

The Azure customer purchase data includes date of initial purchase, the specific DC location where the *basic A1* VM is deployed, the zip code for the customer and the industry of the customer when available. Table 3.2 shows a summary of observables for both the anonymized Azure data and the pricing data for both AWS and Azure. Table 3.2 also highlights the increase in data centers across both Azure (four to ten) and AWS (three to five) in 2016. The table also shows explicitly that we only observe the location of Azure customers. For this reason we leverage the EM algorithm to infer the location of AWS customers.

Most customers do not have an industry associated with them so we classify them as "unknown". Roughly 25% of customers, however, do have an industry noted in our data. Observing industries is very likely non-random so the industry composition in Table 3.2 likely isn't representative of the overall customer industry of Azure. Based upon conversations with Microsoft employees, industry is often reported when a cloud user leverages an intermediary to deploy their cloud workloads (e.g., when the end customer uses a vendor cloud service provider to manage their cloud resources). Hence observing a reported industry could be a proxy for leveraging a vendor to operate cloud IT.

We use variation in the number of DCs over time to identify taste parameters for proximity to data centers. The intuition is as follows: consider two sets of cloud users all from Ohio. Azure has no DC in Ohio over our sample but AWS opens a DC in mid-2016. The first set of customers need to deploy in January 2016 before AWS opened a DC in Ohio. Hence, we would observe some customers with Ohio zip codes

Table 3.2: Summary Statistics

Panel A: Consumer Characteristics		
Consumer Characteristics	**Microsoft Azure**	**AWS**
Locations	observable	unobservable
Industry	observable	unobservbale
Discrete Manufacturing	4.5%	
Education	1.1%	
Health	1.3%	
Hospitality & Transportation	1.0%	
Insurance	0.6%	
Media / Telecome and Utilities	1.3%	
Nonprofit	0.5%	
Professional Services	13.2%	
Choice	DC level	Brand level
Panel B: DC Characteristics		
DC Characteristics	**Microsoft Azure**	**AWS**
Locations	observable	observbale
Changes in number of DCs in 2016	$4 \rightarrow 10$	$3 \rightarrow 5$
Start date of Canadian DC	Apr, 2016	Dec, 2016
Average hourly price	*basic A1*	*t2.small*
	\$0.0708 (0.0120)	\$0.0381 (0.0040)

Note: Table summarizes data used in the analysis. Industry only reported for roughly 30% of observations in our data and professional services and discrete manufacturing appear overly represented for those observations reporting industry. Both AWS and Azure saw and increase in the number of DCs over the time period.

as signing up in Azure DCs in January 2016. AWS then opens a DC in Ohio. Assume the second set of customers, also from Ohio, want to deploy in December 2016. If Ohio cloud customers value proximity then we would observe few Ohio customers deploying in Azure DCs in December. The same logic applies to opening new Azure DCs for the spatial distribution of customers in other Azure DCs.

Figure 3·4 shows average distance between the zip codes of customers making new deployments and the zip codes of the DCs they deploy to in the Azure data by month. Vertical lines indicate the dates would new DCs are available to customers. If the new DCs had no impact on deployment decisions the lines would be flat over time. The line shows some month on month variation in addition to a decreasing trend over time. Hence, there is some reduced form evidence in the Azure data for a preference for proximity.

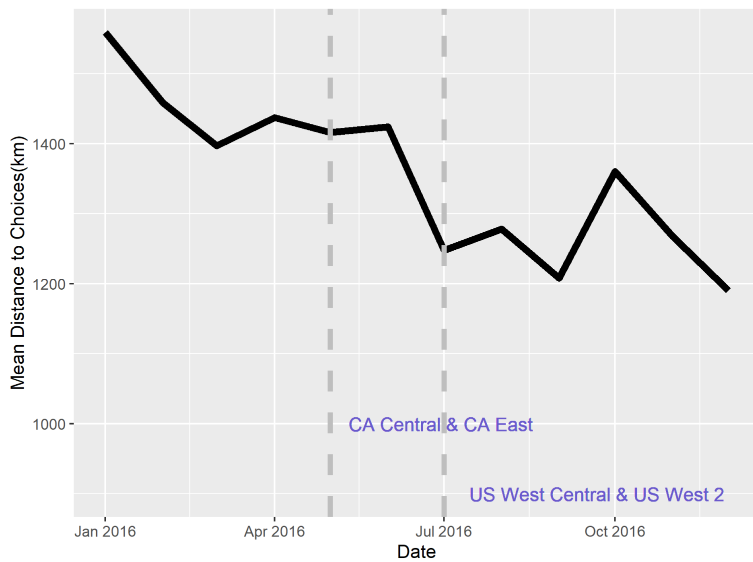


Figure 3·4: Average distance of deployments to customer zip code
Note: Figure shows drop in average distance between a deploying customer's billing zip code and their choice of data center over time. Sharpest one month drop occurs in same time period as new Azure DC locations.

Figure 3·5 shows a density of customer location relative to deployed DC for customers that choose the closest DC relative to those that do not choose the closest DC at the time of deployment. The green shaded density shows the location in kilometers for customers that choose the nearest and the grey density customers who do not choose the nearest. The Figure is meant to highlight that when customers choose the nearest DC location they are often moving from something like 1000 to 4000 kilometers to being within 1000 kilometers. The average distance distance is roughly 1000 kilometers (vertical dashed lines). If 100 miles maps to one millisecond of latency then the gain in latency between the two densities is roughly six milliseconds. Aside from niche use cases like high frequency traders, six milliseconds is not likely to be material for most cloud VMs.

Our method relies on having detailed customer sales for a single firm (Azure) and aggregate customer sales for the second (AWS). While we have very good data on Azure customers, we have no customer level data for AWS customers. However,

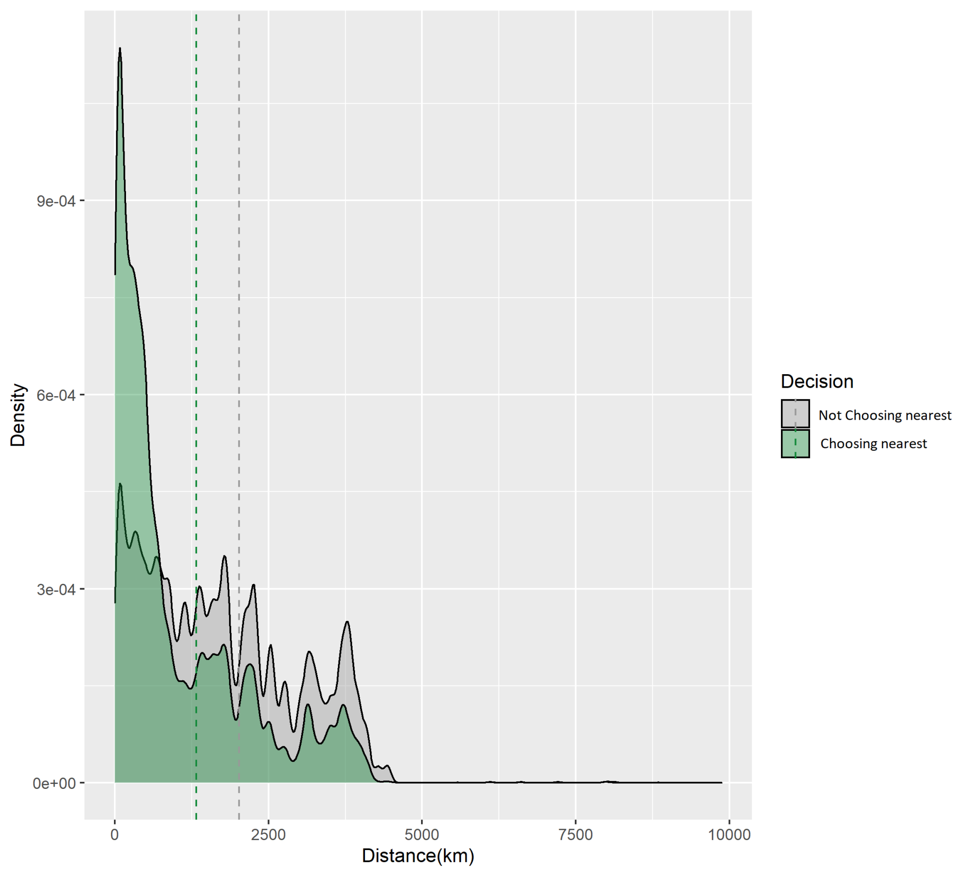


Figure 3·5: Average distance conditional on proximity decision
Note: Figure plots smoothed distance distribution for customers choosing the closest DC to them at time of deployment (shaded dark green) versus those not choosing the closest DC at time of deployment (shaded light grey). Average distance difference across each group is on the over of 600 kilometers and supports broadly overlap.

we observe aggregate global cloud sales from AWS from 10-K SEC filings in 2016.[11] SEC reported sales are worldwide, not restricted to North American market, but we observe U.S. sales as a percent of worldwide sales for Azure.

We make three strong but plausible assumptions to back out US AWS sales for *t2.small* in 2016 from their 10-K leveraging insights from Azure data. First, we apply the global revenue share of North America for Azure to AWS. While this is likely to be imperfect, it is hard to imagine that geographical revenue shares are significantly different across the providers. Second, we calculate the revenue share of the Azure SKU, *basic A1*, within North America relative to all other cloud products. We then apply that product revenue share to AWS. Third, we calculate the average sales of *basic A1* customers and apply it to the inferred AWS *t2.small* customers to get a customer count for *t2.small* for AWS customers.

This inferred approach is appealing because it permits us to get a plausible customer base for AWS customers. A simpler version in the same spirit would be to multiply the number of observed Azure customers in our sample by the market share ratio of AWS to Azure show in Figure 3·1. In practice the two approaches give qualitatively similar customer count numbers, which we don't report due to confidentiality clauses in our data sharing agreements since it would provide customer count data for Azure customers. We discuss in our results section how results could be impacted by getting inferred AWS customer counts wrong.

Having both detailed data for Azure and aggregate data for AWS, the final piece of data is aggregate market size data for the cloud computing market in North America by state or province. The vast majority of cloud computing resources are used by firms as opposed to sold directly to consumers and the cloud is a substitute for on-premise compute resources. We thus assume the market for cloud computing is defined by the

[11]More details can be found on `http://phx.corporate-ir.net/phoenix.zhtml?c=97664&p=irol-reportsother`

total number of private sector firms in the U.S. and Canada. For the U.S. market we take the total number of businesses by state from Henry J Kaiser Family Foundation (KFF).[12] KFF tracks data on number of private sector firms by size.

For the Canadian market, we take data from Statistics Canada, a Canadian government agency which can be considered as the counterpart of the U.S. Census Bureau[13]. The data is from the Business Register (BR), a continuously-maintained central repository of baseline information on businesses and institutions operating in Canada. The variable is referred as "Canadian Business Counts" in the repository, including all active Canadian locations with employees. The number we use was collected in December, 2016.

We trim the market level data in two ways. First, because the data is at yearly level we take the numbers in 2015 and 2016 as they were collected at the end of each year, and then extrapolate them into each month in 2016 based on a constant growth rate assumption for both US and Canada data. Furthermore, we only consider firms with more than 50 employees as potential cloud users, they are 24.43% of all private firms in the U.S. in 2016 and 4.7% for Canada. We only look at larger firms since in 2016 cloud computing was more likely to be utilized by larger, tech savvy firms.[14]

3.4 Model

This section introduces our structural demand model for cloud deployments. We model all Canadian and U.S. consumers' utility to take the standard random utility model (RUM) form. In addition to allowing price and firm fixed effects to impact

[12]See `https://www.kff.org/other/state-indicator/number-of-firms-by-size`.

[13]More detailed information can be found on the following website:`https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3310003401`.

[14]Of course, many smaller tech savvy start-ups also leverage the cloud. We discuss robustness around this trimming decision in the results section. We make a final technical assumption to multiply the market size by *basic A1*'s demand share within Azure, so that the patterns in market shares are kept consistent with *basic A1* and *t2.small* demand. As we discuss below, this final resizing only changes the magnitude of the outside option relative to shares for AWS and Azure.

utility, we explicitly include distance between a consumer and data center and a shifter for if the data center is domestic. Finally, we leverage the EM algorithm to get around the missing data problem of unobserved AWS customer locations.

Cloud computing is a classic discrete-continuous good because consumers first decide to rent cloud computing resources, then decide how much to rent (Hanemann (1984)). For simplicity in what is already a non-trivial problem, we focus only on the initial purchase decision for two of the most popular general compute cloud products during this time: *t2.small* for AWS and *basic A1* for Azure. We do not model continued deployment decisions in this paper and focus on the location of new VM deployments.

We assume the utility of customer i choosing DC j in period t is

$$\begin{aligned} u_{ijt} &= \gamma_{m_i} \times d(\mathbf{l}_i, \mathbf{l}_j) + \beta \times price_{jt} + \psi \times \mathbb{1}_{ij}\{domestic\} \\ &\quad + \xi \times DCAge_{jt} + \zeta \times \mathbb{1}_j^{AWS} + \epsilon_{ijt}, \forall j \in \mathcal{F}^t \end{aligned} \tag{3.1}$$

where

- $i = 1, 2, ..., I$ is the index for customers, $j = 1, 2, ..., J$ is the index for DCs and $t = 1, 2, ..., T$ is the time index.
- $\mathbf{l}$ is a 2-dimensional vector indicating locations, with the first component as longitude and the second as latitude.
- $d(\mathbf{l}_i, \mathbf{l}_j)$ is a function returning the distance between consumer i and DC j, i.e. $d(\mathbf{l}_j, \mathbf{l}_i) = ||\mathbf{l}_i - \mathbf{l}_j||$, where $|| \cdot ||$ is the great-circle distance. We allow preferences for proximity to vary based on consumer i's industry.
- m_i indictes consumer i's industry, we allow industry-specific distance coefficient to reflect the fact that different industries may have distinguished degrees of latency aversion.

- $price_{jt}$ is the price of DC j in period t.
- $\mathbb{1}_{ij}\{domestic\}$ is an indicator variable for if customer i is in the same country as the data center j.
- $DCAge_{jt}$ is the age of DC j in period t.
- $\mathbb{1}_j^{AWS}$ is an indicator for AWS DCs.
- ϵ_{ijt} is a type I extreme value that is $i.i.d$ across $\forall i,\ j,\ t$.

Equation (3.1) has a standard form but a couple of attributes merit discussion. First is the distance metric. We determine a customer's location based upon their observed billing address zip code and the approximate location of different data centers (nearest city). This introduces some measurement error: cloud customers care about latency between their deployment and the user of that deployment. For example, Netflix, a streaming video on demand provider, might prefer to put their cloud workloads close to their customers' locations rather than their corporate headquarters. While there is correlation between cloud customer's location and the location of their customers, that correlation is not perfect. This introduces measurement error and thus attenuation bias. As a result, the impacts of distance we estimate are likely a lower bound. Also, by including an indicator for domestic DCs, $\mathbb{1}_{ij}\{domestic\}$, we allow a general preference for domestic DCs due to concerns about information security or logistic convenience.

Second, since consumers' utility of different DCs vary with their locations, it is possible in principle to model utility function in a "random coefficient" fashion. Specifically, whereas we can calculate distance explicitly for Azure consumers, distances for AWS or non-cloud users are unknown. Consumers' heterogeneous tastes across DCs could be thought of as determined by their unobserved attributes, therefore similar to a "random coefficient" model. We put more structure on the problem by making assumptions about the spatial distribution of all possible cloud consumers because the counterfactual exercise we want to perform are the welfare implications of changing the location of DCs. Thus our modeling assumptions are driven by the nature of problem we seek to solve.

Third, we allow for the utility of data centers to vary by the age of the data center measured in months. This allows for cloud customers to learn about new data centers over time. It also allows for growth in complementary services: our analysis examines only a single cloud computing product but there are complementarities between products (e.g., VMs and data storage). Allowing for DC age to impact utility is a reduced form way of allowing complementarities to manifest.

We model the outside option as on-premise IT infrastructure. We assume that all consumers have one such option in their choice set, denoted as $j = O$ with characteristics $d(\mathbf{l}_i, \mathbf{l}_O) = 0, \forall i$, $price_{Ot} = 0, \forall t$, $\mathbb{1}_{ij}\{domestic\} = 1, \forall i$. Since all consumers had been using in-house infrastructure before cloud, it is unnecessary to model learning effects with a time-variant variable such as $DCAge_{jt}$, thus $\xi \times DCAge_O$ can be normalized up to a constant. Instead, we assume there is a time-variant fixed effect for the outside option, $\alpha + \tau ln(t)$, which can be interpreted as the general time trend of cloud computing. A negative coefficient on τ would reflect the general increase in market share of cloud computing relative to on-premise offerings. Therefore, the utility of the on-premise option available to all possible cloud customers is:

$$u_{iOt} = \alpha + \tau \times ln(t) + \epsilon_{iOt} \tag{3.2}$$

where α includes the domestic effect as well as the constant term in time trend.

3.4.1 The Likelihood Function

Since we assume ϵ_{ijt} are from type I extreme value distribution, the probability for customer i to choose DC j in period t takes the familiar logit form[15]:

$$P(y_{ijt} = 1|m_i, \mathbf{l}_i, \mathbf{l}_t^{DC}, \mathbf{z}_t, \boldsymbol{\theta}_1) = \frac{exp(v_{ijt})}{exp(v_{iOt}) + \sum_{k \in \mathcal{F}_t} exp(v_{ikt})} \tag{3.3}$$

where

- v denotes the deterministic part of the utility function, i.e. $v_{ijt} = u_{ijt} - \epsilon_{ijt}$; $v_{iOt} = u_{iOt} - \epsilon_{iOt}$
- y_{ijt} is a 0-1 binary variable indicates whether consumer i signs up for DC j in period t.
- $\mathcal{F}_t$ is the product set in period t, including the product set of Microsoft's Azure, $\mathcal{F}_t^M$, and that of Amazon's AWS, $\mathcal{F}_t^A$, i.e. $\mathcal{F}_t = \mathcal{F}_t^M \cup \mathcal{F}_t^A$
- $\mathbf{l}_t^{DC} = \{\mathbf{l}_j, \forall j \in \mathcal{F}_t\}$ is the set collecting the locations of all available DCs in period t
- $\mathbf{z}_{jt} = (price_{jt},\ DCAge_{jt},\ \mathbb{1}_j^{AWS})$ is the product characteristics vector, and $\mathbf{z}_t = \{\mathbf{z}_{jt}, \forall j \in \mathcal{F}_t\}$ collects $\mathbf{z}_{jt}$ across all DC's.
- $\boldsymbol{\theta}_1 = (\gamma_m,\ \beta,\ \psi,\ \xi,\ \zeta,\ \alpha,\ \tau)$ is the set of utility parameters.

The probability of not signing up for Microsoft Azure or AWS is

$$P(y_{iOt} = 1|m_i, \mathbf{l}_i, \mathbf{l}_t^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1) = \frac{exp(v_{iOt})}{exp(v_{iOt}) + \sum_{k \in \mathcal{F}_t} exp(v_{ikt})} \tag{3.4}$$

[15]In our Azure data, consumers are from different purchase channels. In this estimation, we focus on two of them, web direct and volumn license. The reason is that consumers from other channels such as "Benefits" may have a different pricing scheme.

Unobserved AWS demand Although we can directly use Eq.(3.3) to denote the probability that a Azure customer chooses any specific DC, the industries and locations AWS customers as well as their DC-level choices are unobservable in our dataset. We leverage our inferred AWS product revenue, conditional probabilities and the EM algorithm to get around this problem.

First, we write the likelihood as the probability of choosing AWS as a brand, which is the sum of probabilities of choosing any of their DCs:

$$P(y_{iAt} = 1 | m_i, \mathbf{l}_i, \mathbf{l}_t^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1) = \frac{\sum_{j \in \mathcal{F}_t^A} exp(v_{ijt})}{exp(v_{iOt}) + \sum_{k \in \mathcal{F}_t} exp(v_{ikt})} \tag{3.5}$$

where y_{iAt} indicates whether consumer i chooses AWS in period t.

Missing consumer locations The industries and locations of non-Microsoft customers are unobservable in our data which makes the calculation of the conditional choice probabilities infeasible. To circumvent this problem, we will take consumer's industry and location as two random variables, get the joint probability of industry, location and choice, then integrate out its uncertainty in industry and location for AWS and the outside option consumers. Particularly, the likelihood function in period t can be written as

$$\begin{aligned} L_t(\boldsymbol{\theta}) \quad = \quad & \prod_{i \in C_t^M} \prod_{j \in \mathcal{F}_t^M} (P(y_{ijt} = 1 | m_i, \mathbf{l}_i, \mathbf{l}^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1) f_t(m_i, \mathbf{l}_i | \boldsymbol{\theta}_2))^{y_{ijt}} \\ & \times \prod_{i \in C_t^A} \int_{m_i, \mathbf{l}_i} P(y_{iAt} = 1 | m_i, \mathbf{l}_i, \mathbf{l}^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1) f_t(m_i, \mathbf{l}_i | \boldsymbol{\theta}_2) dm_i d\mathbf{l}_i \\ & \times \prod_{i \in C_t^O} \int_{m_i, \mathbf{l}_i} P(y_{iOt} = 1 | m_i, \mathbf{l}_i, \mathbf{l}^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1) f_t(m_i, \mathbf{l}_i | \boldsymbol{\theta}_2) dm_i d\mathbf{l}_i \end{aligned} \tag{3.6}$$

where $C_t^f, f = M, A, O$ are the sets of consumers for Microsoft's Azure, Amazon's

AWS and non-cloud users respectively. The key attribute of equation (3.6) is the distribution of location for AWS and outside option purchasers on the second and third lines. The density $f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)$ can be viewed as a industry-specific spatial distribution of consumers of all options in the market. Although we observe the industries and locations of Azure customers, we write their joint probabilities of industry, location and choice separately to keep the format consistent across brands. This will also enable us to infer $f_t(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)$ based on the observed Azure customer locations. More details can be found in Section 6. In practice we take the industry-specific spatial distribution of consumers in the market to be the that of medium and large firms across U.S. states and Canadian provinces as described in the Data section above.

Taking logs for the function above and compacting $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ as $\boldsymbol{\theta}$ then summing over time gives:

$$\begin{aligned} LL(\boldsymbol{\theta}) &= \sum_t LL_t(\boldsymbol{\theta}) \\ &= \sum_t \Big(\sum_{i \in C_t^M} \sum_{j \in \mathcal{F}_t^M} y_{ijt} log(P_{it}^j(\boldsymbol{\theta}_1) f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)) \\ &\qquad + Q_t^A log(\int\limits_{m_i, \mathbf{l}_i} P_{it}^A(\boldsymbol{\theta}_1) f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2) dm_i d\mathbf{l}_i) \\ &\qquad\qquad + Q_t^O log(\int\limits_{m_i, \mathbf{l}_i} P_{it}^O(\boldsymbol{\theta}_1) f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2) dm_i d\mathbf{l}_i)\Big) \end{aligned} \tag{3.7}$$

where $P_{it}^j(\boldsymbol{\theta}_1)$, $P_{it}^A(\boldsymbol{\theta}_1)$ and $P_{it}^O(\boldsymbol{\theta}_1)$ simplifies $P(y_{ijt} = 1|m_i, \mathbf{l}_i, \mathbf{l}_t^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1)$, $P(y_{iAt} = 1|m_i, \mathbf{l}_i \in C_b, \mathbf{l}_t^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1)$ and $P(y_{ijt} = 1|m_i, \mathbf{l}_i, \mathbf{l}_t^{DC}, \mathbf{z}_{\cdot t}, \boldsymbol{\theta}_1)$ correspondingly. Since we take expectation over the unknown consumer's industry and location, the expected choice probability is same for every AWS consumer or any potential cloud consumer. Therefore, we multiply them by the total quantities Q_t^A and Q_t^O.[16]

[16]Recall we make some assumptions on AWS revenue composition to get Q_t^A and the market size broadly to get Q_t^O.

3.4.2 EM Algorithm

Maximizing the log likelihood function above with the usual Newton or quasi-Newton routines can be numerically difficult and computationally unstable. This is a key motivation for leveraging the EM algorithm. The EM algorithm is a two-stage iterative method which involves calculating an expectation of the log likelihood function weighted by the Bayes' probabilities at some initial values and then updating the parameters by maximization.

Following Bhat (1997), it can be shown that with a given distribution of firms in North American and a set of preference parameters (θ_1), maximizing Eq.(3.7) is mathematically equivalent to maximizing the alternative log likelihood function in Eq.(3.8), where $f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)$ are replaced by its Bayesian posterior counterparts, i.e. the probabilities that an AWS customer or a non-cloud user is from industry m_i and located at $\mathbf{l}_i$ which we denote as $h^A_{m_i,\mathbf{l}_i,t}$ and $h^O_{m_i,\mathbf{l}_i,t}$ respectively.

$$\begin{aligned}\sum_t \Big(\sum_{i \in C_t^M} \sum_{j \in \mathcal{F}_t^M} & y_{ijt} log(P_{it}^j(\boldsymbol{\theta}_1) f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)) \\ & + Q_t^A \int_{m_i,\mathbf{l}_i} h^A_{m_i,\mathbf{l}_i,t}(\boldsymbol{\theta}) log(P^A_{it}(\boldsymbol{\theta}_1) f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)) dm_i d\mathbf{l}_i \\ & \quad + Q_t^O \int_{m_i,\mathbf{l}_i} h^O_{m_i,\mathbf{l}_i,t}(\boldsymbol{\theta}) log(P^O_{bt}(\boldsymbol{\theta}_1) f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)) dm_i d\mathbf{l}_i \end{aligned} \quad (3.8)$$

Then this maximization problem can be solved iteratively: starting from some initial values $\boldsymbol{\theta}^s$, we first update the Bayesian posterior probabilities, and then maximize Eq.(3.8) for $\boldsymbol{\theta}^{s+1}$ conditional on the Bayesian posteriors. Details of the approach are carefully described in the Appendix.

Lastly, due to the property of log operation, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be separately updated. Specifically , we iteratively maximize the following two objective functions,

$$\begin{aligned}
\boldsymbol{\varepsilon}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^s) &= \sum_t (\sum_{i\in C_t^M}\sum_{j\in \mathcal{F}_t^M} y_{ijt} log P_{it}^j(\boldsymbol{\theta}_1) \\
&\quad + Q_t^A \int_{m_i,\mathbf{l}_i} h^A_{m_i,\mathbf{l}_i,t}(\boldsymbol{\theta}^s) log P^A_{it}(\boldsymbol{\theta}_1) dm_i d\mathbf{l}_i \\
&\quad\quad + Q_t^O \int_{m_i,\mathbf{l}_i} h^O_{m_i,\mathbf{l}_i,t}(\boldsymbol{\theta}^s) log P^O_{bt}(\boldsymbol{\theta}_1) dm_i d\mathbf{l}_i) \\
\boldsymbol{\varepsilon}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^s) &= \sum_t (\sum_{i\in C_t^M}\sum_{j\in \mathcal{F}_t^M} y_{ijt} log f(m_i,\mathbf{l}_i|\boldsymbol{\theta}_2) \\
&\quad + Q_t^A \int_{m_i,\mathbf{l}_i} h^A_{m_i,\mathbf{l}_i,t}(\boldsymbol{\theta}^s) log f(m_i,\mathbf{l}_i|\boldsymbol{\theta}_2) dm_i d\mathbf{l}_i \\
&\quad\quad + Q_t^O \int_{m_i,\mathbf{l}_i} h^O_{m_i,\mathbf{l}_i,t}(\boldsymbol{\theta}^s) log f(m_i,\mathbf{l}_i|\boldsymbol{\theta}_2) dm_i d\mathbf{l}_i)
\end{aligned}$$

In practice, instead of assuming a parametric distribution for $f(m_i, \mathbf{l}_i|\boldsymbol{\theta}_2)$, we assume a discrete distribution of consumer's industry and location, or say a discrete industry-specific spatial distribution. The discrete distribution can approximate any arbitrary distribution when discretization is fine enough. Specifically, we take each U.S. state and Canadian province as a bin b, and then the probability that a consumer (including non-cloud users) from industry m belongs to a certain bin b in period t is q_{mbt}, and these q_{mbt}'s are treated as parameters to estimate. Using states and provinces is both convenient and appropriate since data on market size (medium and large firms) is available at the state level and provides good variation in distance from newly introduced DCs. Details of this approach are again in the Appendix.

In sum there are a few departures from normal log likelihood maximization we make in our approach. First, we replace location probabilities with Bayesian posteriors. Second, we iteratively solve for parameters governing the distribution of consumers for each product and preference for the product. Third, we discretize the spatial distribution of North America. This final step is an advantage for us since variation in Azure demand in geographical bins over time in response to new Azure and AWS DCs help us identify the model's parameters. The iterative maximization process across geographical and preference parameters continues until convergence as we describe in detail in the next section.

3.4.3 Identification

In this section, we show that the parameters can be identified in the following order: (1) $\boldsymbol{\theta}_{1,1} = (\gamma_m, \beta, \psi, \rho, \xi)$; (2) $\boldsymbol{\theta}_{1,2} = (\zeta, \alpha, \tau)$; (3) $\boldsymbol{\theta}_2 = \{q_{mbt}\}_{mb}, t = 1, 2, ..., T$.

First, $(\gamma_m, \beta, \psi, \xi)$ are identified from the substitution pattern of Microsoft customers among Microsoft DCs. Since our Microsoft data is at individual level, including the industry and location of each customer, m_i, $d(\mathbf{l}_i, \mathbf{l}_j)$'s and $\mathbb{1}_{ij}\{domestic\}$'s are deterministic, i.e. there is no unknown interaction between individual attributes and product characteristics. Therefore, the *Independence of Irrelevant Alternatives* (IIA) property of logit model makes it possible to focus on only a subset of products (Train (2009)).

Next, if we consider ζ and $\{v_{Ot}\}_t$ [17] as the general preference for all AWS DCs and the outside option over Microsoft, with the product characteristics and $\boldsymbol{\theta}_{1,1}$ as given, the unexplained part of market share ratios should be attributed to that "general preference", which gives the identification of $\boldsymbol{\theta}_{1,2} = (\zeta, \alpha, \tau)$.

Specifically, for $\forall j \in \mathcal{F}_t$, we write the AWS fixed effect separately from other components in the utility index, i.e.

$$v_{ijt} = \mu_i(l_j, \mathbf{z}^1_{jt}, \boldsymbol{\theta}_{1,1}) + \zeta \mathbb{1}^{AWS}_j,$$

where

[17] With a slight abuse of notation, we suppress the subscription i in v_{iOt} since the deterministic utility from the outside option is individual-invariant.

$$\begin{aligned} \mathbf{z}_{jt}^1 &= (price_{jt},\ DCAge_{jt}) \\ \mu_i(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}) &= \beta price_{jt} + \gamma_{m_i} d(\mathbf{l}_i, \mathbf{l}_j) + \psi \mathbb{1}_{ij}\{domestic\} + \xi DCAge_{jt} \end{aligned}$$

Note that μ_i is individual-specific due to the consumer's heterogeneous industry and location.

Then, within each combination of m and b, the model gives the market share ratio of AWS to Microsoft as the fraction of the exponentials of their inclusive values,

$$\frac{Q_{mbt}^A}{Q_{mbt}^M} = \frac{\sum_{j \in \mathcal{F}_t^A} exp(\zeta + \mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))}{\sum_{j \in \mathcal{F}_t^M} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))} = exp(\zeta) \frac{\sum_{j \in \mathcal{F}_t^A} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))}{\sum_{j \in \mathcal{F}_t^M} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))}$$

Here with a little abuse of notation, we use subscript m, b to emphasize that function $\mu_i(\cdot)$ is the same for consumers from the same industry m and located in bin b. Also, for Azure consumers, even though $\mathbf{l}_i$ is observed, we lower the granularity to bin b level in this section just to illustrate the concept.

Relate this to the observed market level AWS demand by $Q_t^A = \sum_{m,b} Q_{mbt}^A$, we have

$$Q_t^A = exp(\zeta) \sum_{m,b} \frac{\sum_{j \in \mathcal{F}_t^A} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))}{\sum_{j \in \mathcal{F}_t^M} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))} Q_{mbt}^M$$

which gives

$$\zeta = log(Q_t^A / \sum_{m,b} \frac{\sum_{j \in \mathcal{F}_t^A} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))}{\sum_{j \in \mathcal{F}_t^M} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))} Q_{mbt}^M)$$

Similarly,

$$v_{Ot} = log(Q_t^O / \sum_{m,b} \frac{1}{\sum_{j \in \mathcal{F}_t^M} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))} Q_{mbt}^M)$$

Then α and τ are identified by the linear relation $v_{Ot} = \alpha + \tau ln(t)$.

Finally, given $\boldsymbol{\theta}_1$, the model could infer the local market size based on the observed local demand of Microsoft, i.e.

$$\begin{aligned} q_{mbt} &= \frac{Q_{mbt}^M + Q_{mbt}^A + Q_{mbt}^O}{M_t} \\ &= (Q_{mbt}^M \\ &\quad + exp(\zeta) \frac{\sum_{j \in \mathcal{F}_t^A} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))}{\sum_{j \in \mathcal{F}_t^M} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))} Q_{mbt}^M \\ &\quad + exp(v_{Ot}) \frac{1}{\sum_{j \in \mathcal{F}_t^M} exp(\mu_{mb}(l_j, \mathbf{z}_{jt}^1, \boldsymbol{\theta}_{1,1}))} Q_{mbt}^M)/M_t \end{aligned}$$

where M_t is the market size in period t.

3.5 Monte Carlo Experiment

To test the model's identification, we performed a Monte Carlo experiment. It is important to assess whether the model's parameters are recoverable with only detailed Azure data because Azure had only a 7% market share over our sample. Accordingly, the basic structure of the simulated data sets used in the Monte Carlo borrows from the true data in two ways.[18] First, the number of consumers in each industry-state/province is generated based on the distribution (e.g., $\{q_{mbt}\}_{m,b,t}$) that we recover from estimation. Second, the taste parameters that we use to generate each consumer's DC choices are the same as the estimates from the actual data.

[18] Of course, the DC layout as well as their prices are also consistent with our observed data in each period.

The main variation across these simulated data sets are the idiosyncratic random utility shocks ϵ_{ijt}. We simulated 100 data sets. For each data set, we let each consumer chooses the DC that gives the highest utility. We then keep the individual choices of Azure customers while aggregating AWS customers and those who choose the outside option up to market shares at period level. With the spatial distribution of consumers masked so that they must be estimated as when we estimate the model with our actual data, we estimate each simulated data set with the EM-algorithm described above.

The results of the Monte Carlo are shown in Appendix Table C.1. The model performs reasonably well. For all 18 parameters except one the true simulated parameter is within the 95% confidence interval of the parameters estimated from the simulated data. The one parameter that is outside of the 95% confidence interval is the indicator variable for a DC being domestic. The domestic indicator is only marginally outside the confidence interval (CI): true value 1.58 and 95% CI of [1.415,1.516]. Thus, there is some evidence we estimate a domestic indicator that makes cloud customers look slightly less interested (less than 10%) in deploying their VM in country.

That the coefficient on the domestic indicator variable is somewhat imprecise is not surprising given the nature of our data. It is identified almost entirely by Canadian customers choosing to deploy VMs in the Canadian Azure DCs after they open halfway through 2016. However, the number of unique Canadian firms in our sample is an order of magnitude lower than the number of U.S. firms in the sample. As a result, the indicator variable is likely to be measured imprecisely and, perhaps, modestly downward biased. The downward bias could be due in part to the domestic preference loading onto the estimated distance preference for Canadian customers. That said, the Monte Carlo shows the true value of the distance coefficient (γ for unknown industries) is exactly in the center of the 95% CI.

3.6 Estimation Results

Table 3.3 shows results from estimating the model with the data. We performed estimation in R and convergence times on a single PC were on the order of 10 hours. We do not report the number of observations so as to not reveal information on the number of unique customers for this Azure SKU over our sample, per the confidentiality agreement with Microsoft.

Table 3.3 shows that all parameters are precisely estimated and have the expected sign. The price coefficient is negative. The coefficients on DC age, the domestic indicator, AWS fixed effect, outside option (OO) fixed effects are all positive. The positive AWS and OO fixed effects reflect market share sizes over the sample (e.g., Azure $<$ AWS $<$ OO based upon our assumption of outside good market size).[19]

Table 3.3: Estimates

	Estimates	Std. Err.
Distance (in 1000km)		
Discrete Manufacturing	-1.439^{***}	0.012
Education	-1.439^{***}	0.013
Health	-1.430^{***}	0.015
Hospitality & Transportation	-1.439^{***}	0.013
Insurance	-1.439^{***}	0.014
Media / Telecom and Utilities	-1.750^{***}	0.015
Nonprofit	-1.442^{***}	0.016
Professional Services	-1.334^{***}	0.011
Unknown	-0.517^{***}	0.007
Price	-12.124^{***}	0.001
Domestic	1.872^{***}	0.019
DC Age	0.905^{***}	0.026
AWS FE	2.162^{***}	0.017
OO FE	1.461^{***}	0.009
OO trend	2.295^{***}	0.001

Note: All parameters statistically significant. All coefficients have expected sign with distance and price both negative and highly significant. The model includes AWS and OO fixed effects in the first and seventh month of our data where there was some backfilled reporting from previous months due to a Microsoft reporting delays. Those coefficients are statistically significant and an order of magnitude lower than AWS and OO FEs; we consider them as nuisance parameters therefore do not report them here. We normalized DC age so that DC age is measured with respect to the oldest DC in the sample.

[19] The model includes AWS and OO fixed effects in the first and seventh month of our data where there was some backfilled reporting from previous months due to a Microsoft reporting delays. Those coefficients are statistically significant and an order of magnitude lower than AWS and OO FEs; we don't report them as consider them nuisance parameters. Due to this reporting issue in the timing of some of the Azure data, we don't put much stock in the sign of the coefficient on the logarithmic time trend of the outside good.

The key coefficient of interest is the coefficient on distance where we take the baseline to be preference for distance in unknown industries. The coefficient is negative and highly significant. Coefficients for other industries don't exhibit much variation and are roughly twice the magnitude of the estimated coefficient of distance for observations without a recorded industry. Thus, there is first order correlation between observing industry and preference for proximity. As mentioned above, based upon internal conversations it could be that observing industry is correlated with using a vendor to operate cloud resources. Because we don't fully observe the data generating process for that field in the data, we instead focus on cloud users with unknown industries, which make up the vast majority of the sample.

It is more informative to evaluate the ratio of the coefficient on distance to the coefficient on price rather than each coefficient in isolation. The ratio of distance to price is the willingness to pay for one kilometer. Hence, the willingness to pay to be 1,000 kilometers closer to a data center in our sample is 4.2 cents (per hour) for unknown industries (the majority of the sample). Recall that the average price over the sample for Azure's basic A1 product is 7.1 cents (per hour). Hence we estimate a price premium of roughly 60% of the average Azure hourly price. As a point of comparison DC level prices often varied by 20-50% within a public cloud provider over our sample. Hence, the implied point estimate for unknown industries seems modestly large but not out of the question where are the point estimate for cloud users that reported their industry seems larger than we expect. For customers in known industries, the estimated disutility of distance is stronger but, as mentioned above, we don't view those point estimates as reliable due to how industry data is recorded in our sample: many of those customers go through a third party to deploy their workloads.

There are two important caveats worth noting relating the negative and significant impact of distance on utility. First, the positive coefficient on DC age reflects that old DCs tend attract more deployments than newer DCs all else being equal. This could reflect some amount of inertia: as a customer deploys a new type of VM they are likely to put it in the same DC as where they might have older deployments. If so, this would reduce the likelihood of finding a strong negative utility for distance: identification of the distance parameter is driven by new DCs opening and evaluating how many customers proximate to its location start deploying workloads there. If older DCs have a stronger attraction, the likelihood of deploying in a new, proximate DC would be lower.

Second, the coefficient on domestic is strong and positive. We believe this could cause some modest downward bias (e.g., more negative) in the distance parameter so that we estimate a stronger dis-utility of distance than the true effect. The reason is that Azure opened two data centers in Canada which are both more proximate to many Canadians and also domestic. Hence, some preference for deploying in a domestic DC could be loaded on to the distance parameter. That said, there is significant variation in proximity for nearby but not domestic data centers for Canadians since the Canadian DCs are located in or east of Ontario. Thus, when Canadian DCs are opened halfway through our sample, cloud users in Vancouver, British Columbia can choose between a nearby DC in Washington state that is not domestic and a far away Canadian DC that is domestic. Therefore, any downward bias on the distance parameter is likely modest.

It is somewhat surprising that the data doesn't show any modest variation at the industry level. 2016 was still the early days of cloud usage. Some industries like health and education could have been later adopters and are now displaying similar distance preferences as discrete manufacturing and professional services did over our sample. For example, it might have been that more sophisticated cloud users display

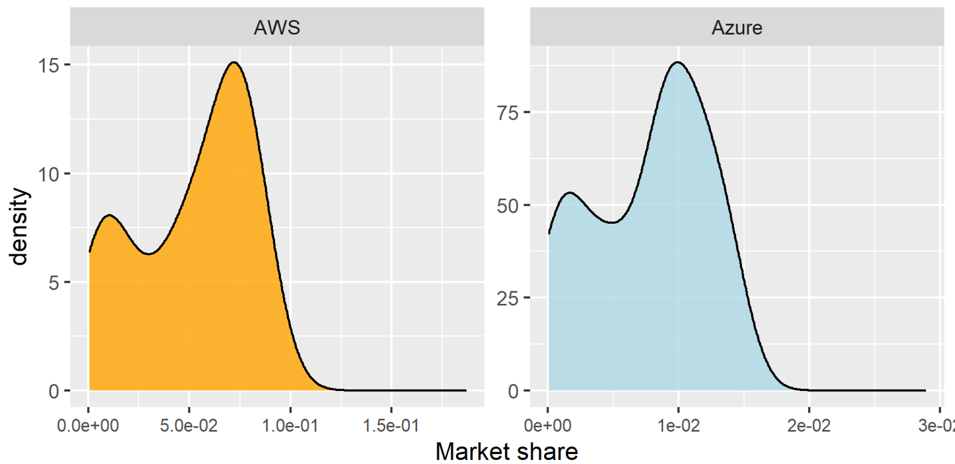


Figure 3·6: AWS vs Azure Market Share Distribution
Note: Market share distributions show material variation over states and provinces. Although not clear from this figure, it should be unsurprising that market shares also compliment each other: where AWS has a larger share Azure tends to have smaller share and vice versa.

weaker preferences for proximity meaning that we estimate a "short run" effect in this paper. As noted above, though, it is possible that observing industry is really a proxy for using a third party to deploy and manage virtual machines. As a result, any underlying differences in preference for proximity could be second order to using third parties for IT management.

Figure 3·6 takes the estimates and aggregates consumers from all industries in a "bin" (e.g., U.S. state or Canadian province) to display heterogeneity in market shares over space for Azure and AWS. Figure 3·6 gives the densities of each firm's estimated market shares. The scale of the market share distribution on the left (AWS) is higher than that of the distribution on the right (Azure) but both are market shares accounting for the outside good. Hence a "bin" with 10% share for AWS and 2% share for Azure implies an 88% share for the outside good where the outside good is the number of firms with more than 50 employees in the geography.

The important aspect of Figure 3·6 is the non-trivial heterogeneity in market share across locations. Both firms exhibit bimodal market shares over space: they

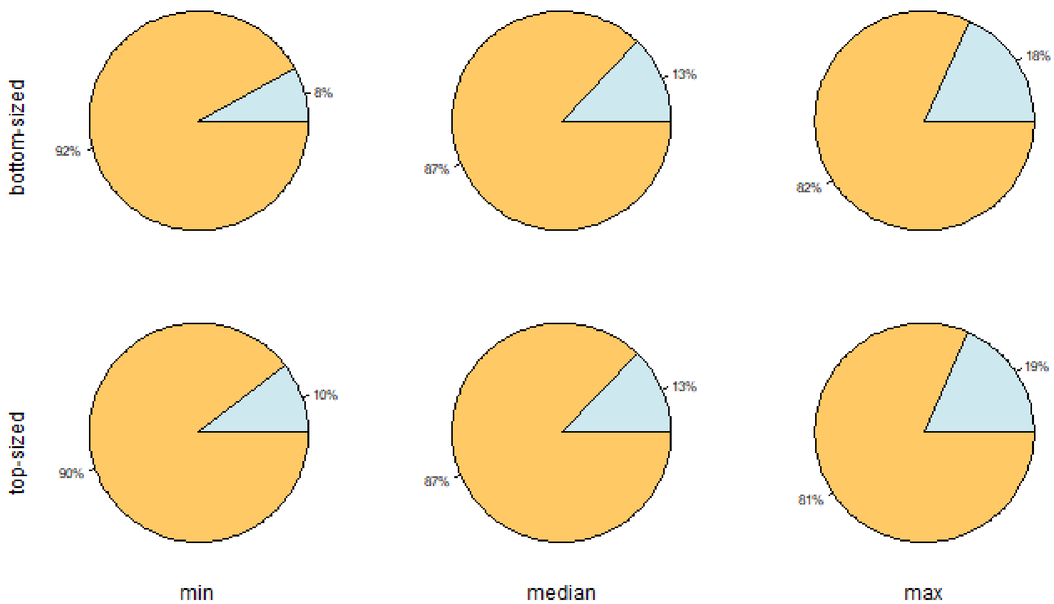


Figure 3·7: Microsoft vs AWS Market Share by Market Size
Note: Consistent with Figure 6, market shares vary across states. This is true for both large and small states where Azure market share can range from up to 19% of AWS market share down to less than 10%. Recall this figure does not report market share of the outside option nor other cloud providers so it is not directly comparable to Figure 1.

have some regions above the firm-specific mean market share and some regions below, and the distribution is not single-peaked. Even though AWS was the market leader during 2016, there are some areas where Azure has a market share in the low single digit percents and some nearly 20% of AWS's share.

Microsoft had more data centers than AWS during this time period, possibly earning higher market share in the regions where AWS did not have a data center. This finding is consistent with cloud customers having preferences for proximity in our sample. It is also consistent with competition being important for welfare in this market insofar as competition leads to more DCs being built in different locations.

While Figure 3·6 shows estimated aggregate variation in market share within firms over space, Figure 3·7 shows estimated variation in market share across Azure (blue) and AWS (orange) at the state/province level. We only select six regions for clarity and don't report precise locations associated with each state/province per our data sharing agreement. We instead show variation in the minimum, median and maximum Azure market share across markets sized below the median (bottom-sized markets) and those sized above (top-sized markets). Note that Figure 3·7 represents a relatively small level of aggregate market penetration relative to the outside option for both AWS and Azure which highlights that the cloud computing industry is still young and rapidly growing.

Figure 3·7 shows that we estimate changes in market share across regions of more than 100% for relatively small markets (8% to 18%) and roughly 100% for relatively large markets (10% to 19%). The model estimates a right tail as well: median Azure market share was slight less than half of the difference between the minimum and maximum market share. Qualitatively, we do estimate relatively larger market shares in some states where Azure has a DC but AWS does not, and vice versa. Finally, these market shares are from 2016 data and since then Azure has grown in market share. Thus these numbers do not reflect current market shares nor do they necessarily represent what would happen if new DCs were built today since more DCs have been constructed between 2016 and 2020.

3.7 Counterfactuals

With the estimated taste parameters, we move on to counterfactual analysis. The strength of this modeling approach is the ability to estimate heterogeneous market shares over space using disaggregate data for one firm but aggregate data for another. Our counterfactuals focus on using the model to optimize data center location and examine the interplay of price competition and spatial competition in the cloud industry. All the data used in counterfactual analysis is the December 2016 data so that the counterfactuals reflect the most recent view of the data we observe.

First, we propose six states in southern U.S. where Microsoft currently has no DC, and ask which one would bring the most market share increase if Microsoft put one more DC there. These examples are chosen for their relevance. Microsoft Azure introduced four new DCs in North America in 2016 which increased its total number to ten, twice that of AWS. Therefore, it is reasonable to quantify the impact of a denser product space.

Second, we condition on the current DC layout in North America, and predict the market share responses to a counterfactual 15% price change for all Azure DCs. We then investigate how counterfactual changes in market shares vary based upon how vigorous spatial competition is. Put another way, we simulate a price decrease and evaluate how it impacts market shares in locations where both Azure and AWS have a DC, where only Azure has a DC and where neither Azure nor AWS have a DC.

Lastly, while we calculate changes in consumer surplus, fully capturing strategic supply side equilibrium responses is beyond the scope of these exercises. Neither AWS nor Azure alters DC layout or adjust prices of existing DCs in response to our counterfactual exercises. Accounting for equilibrium competition best responses is beyond the scope of our paper as our contribution highlights spatial competition for cloud computing rather than equilibrium competitive behavior.

3.7.1 New DC Location

The six proposed states for which we simulate Azure building a DC are spread evenly in southern U.S. where there was no DC in 2016 from Arizona to Florida. We assume the price for the newly constructed DC is set at the average Azure price level in December 2016 so that the different demand responses could be attributed to the differences in local market size and DC layout. All changes in Azure market share are normalized to changes from introducing a DC in Arizona, which Azure actually did enter in 2018. We decompose increases in market share by the "market stealing"

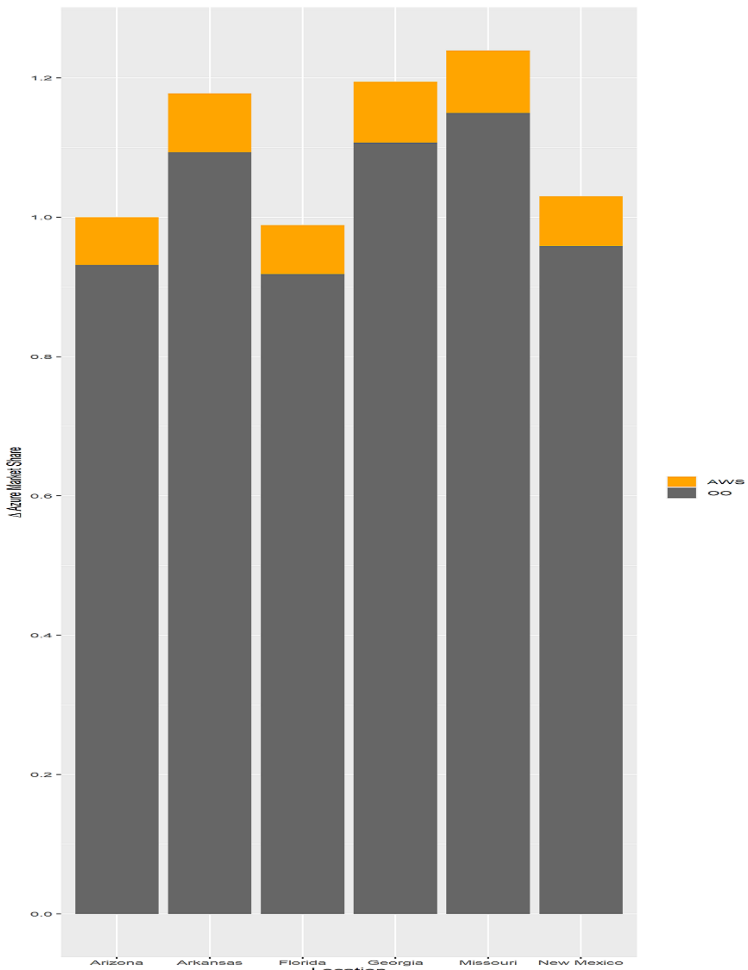


Figure 3·8: Introducing a New DC

Note: Figure reports the change in Azure market share from new customers in the counterfactual where Azure builds a new DC in one of six states. The increase in market share is reported relative to the percentage change in market share of a new DC introduced in Arizona. The Figure also reports where the increase in Azure market share comes from: the outside option over AWS. The southeast U.S. seem to indicate the largest percentage increase in share for AWS due in part due to a relatively large increase in share from acquired from AWS.

effect of taking share from AWS and the cloud "market growing" effect of moving customers off their own premise and onto the cloud.

Figure 3·8 shows the results from the new DC counterfactual measured in percentage increase in market share relative to Arizona. There are a couple of important findings. First, it shows that introducing a new DC in Missouri generates highest market share gains for Microsoft Azure, which is around 25% higher than the "numeraire" state of Arizona. This result is consistent with estimation results since during this time period the DC density in Missouri is comparatively lower than the others. Therefore, a newly-introduced DC would provide greater utility increase relative to the outside good and AWS by reducing distance.[20]

It is useful to put the 25% number into perspective of the overall costs of running a DC to assess whether something like differences in wholesale electricity costs could drive location decisions. According to a report from the U.S. Chamber of Commerce roughly annual operating expenses are less than 10% of the capital costs of a data center and roughly 50-75% of operating expenses are electricity.[21] Hence electricity is on the order of 5-7.5% of annualized amortized DC costs. According to the U.S. Energy Information Administration, in 2016 average wholesale electricity prices in low cost Texas were \$27.16/MWh versus \$34.54 in high cost PJM for a 25% difference[22] The implication is that electricity cost differences on the order of 2% of annualized costs could explain DC location decisions that results in a 25% difference in market share changes. This seems unlikely.

Second, there is some modest variation in the size of the "market stealing" versus the "market growing". Building a DC in Georgia, Missouri and Arkansas leads to

[20] Of course, because we measure changes in within state market shares this says nothing of the aggregate increase in revenue for putting a new DC in Missouri relative to Arizona.

[21] See https://www.uschamber.com/sites/default/files/ctec_datacenterrpt_lowres.pdf.

[22] See https://www.eia.gov/electricity/wholesale/#history.

a larger market share increase than Arizona, Florida and New Mexico. In Georgia, Missouri and Arkansas there is a larger proportional increase in the "market stealing" versus the "market growing" effect driving the increase in market share. The implication is that appropriately siting DCs can lead to increased local market shares driven disproportionately by the market stealing effect. It is perhaps for this reason that all public cloud providers have dramatically increased their geographical footprint in the last five years, all roughly doubling the number of unique DC locations globally. This clearly is beneficial to consumers who, based on our estimates, appear to non-trivially value proximity. However, this strategic effect seems second order to the market growing effect based upon our sample.

Finally, we calculate the consumer surplus gain generated by a new Azure DCs. We define consumer surplus as the expected maximum money metric utility for new customers registered in Dec 2016, i.e. $t = T$. Put another way, we don't account for gains to existing customers since we only model the initial deployment decision. Because we don't account for the differences in usage intensity among consumers, consumer surplus estimates measured in dollars should be thought of as gains in the first hour of a single deployment of a one core VM. Since the lifespan of a VM is often many cores and many hours, the level of the surplus gains reported here are extreme lower bounds and as such we focus on percentage changes across counterfactuals. Specifically, for cloud users in industry m at location b at time T,

$$E(CS_{mbT}) = -\frac{1}{\beta}E[\max_{j\in\mathcal{F}_T} u_{mbjT}] = -\frac{1}{\beta}log(1+\sum_{j\in\mathcal{F}_T}(exp(u_{mbjT}))) + C$$

The subscripts m and b emphasize that utility depends on the hetergeneous cloud user industry m and location b. C is a constant term which is negligible when calculating the surplus differences.

The expectation value is defined relative the set of available data centers ($\mathcal{F}_T$)

plus the outside option whose utility is normalized to 0. Thus when a new Azure DC j' opens, there will be one more element in the choice set thus makes the joint set $\mathcal{F}_T \cup j'$, and the exception is supposed to increase since maximization function weakly increases with the number of choices. The strength of this approach is comparing by how much consumer surplus increases when DCs are placed in better versus worse locations. Finally, we aggregate this individual level expectation to the North American market level by summing over all industries and locations, i.e.

$$E(CS_T) = \sum_{sb} E(CS_{mbT}) q_{mbT} \times M_T,$$

where M_T is the market size, i.e. the total number of firms with 50 or more employees in North America. As in the exercise above we compare the percentage increase in consumer surplus to a single baseline state, Arizona.

Table 3.4: Consumer surplus effects of new DC locations (AZ baseline)

Location	Arizona	Arkansas	Florida	Georgia	Missouri	New Mexico
% $\Delta E(CS_T)$	100%	117.9%	98.9%	119.6%	124.1%	103.1%

The consumer surplus effect of each new DC location are summarized in the Table 3.4. Table 3.4 shows that percentage changes in consumer surplus by state are almost identical to changes in market share for Azure. This is not surprising: increases in market share indicate increases in consumer surplus as more cloud users begin to consume Azure.

3.7.2 Price Drop

Figure 3·9 provides demand responses to an overall 15% price drop of Microsoft Azure across all regions. We show results of the price impact in three representative U.S. states with different market structures: both a AWS and Azure DC (Virginia), neither a AWS nor an Azure DC (Georgia) or only an Azure DC (Texas). Each bar shows

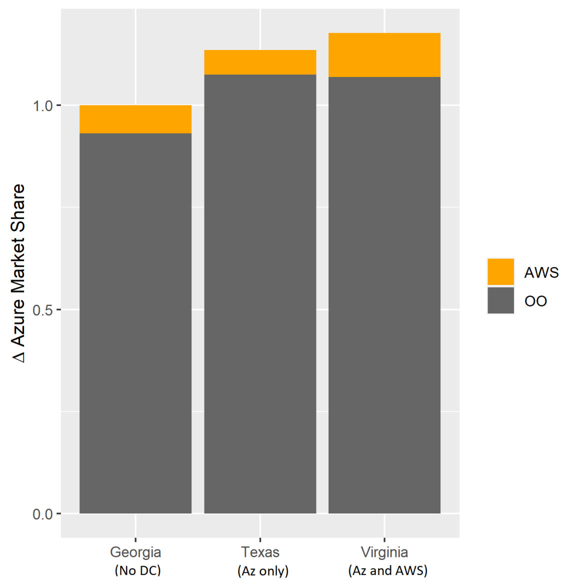


Figure 3·9: Price Competiton

Note: Figure reports the change in Azure market share from new customers in the counterfactual where Azure decreases the price of all Azure DCs by 15%. The increase in market share is reported relative to the percentage change in market share in Georgia. The Figure highlights that increases in market share from the price drop will vary based upon how much spatial competition there is in region. For example, both AWS and Azure have DCs in Virginia and we observe larger market share changes there due to a price decrease.

the percentage of switchers from AWS to Azure. All market share changes are pegged to Georgia in this counterfactual.

Figure 3·9 shows that the incremental change in market share varies by local market structure. Intuitively, in areas where both AWS and Microsoft DCs are available like Virginia, we estimate that price plays a relatively more important competitive role and that price cuts have a significant impact on market shares. On the contrary, the potential gain is less pronounced in states like Texas where Microsoft is the only cloud provider. In other words, the loss from a price rise would also be limited, a straightforward implication of local market power.

Another implication from this counterfactual is the amount of switchers across AWS and Azure are generally small in all three scenarios of price competition. Recall that during this time period AWS was both the early market leader and much larger in publicly reported revenue numbers. Put another way, this doesn't appear to be a Bertrand, winner take all market. This is consistent with AWS's early leadership in the cloud market but also spatial competition as being a material driver of increase Azure market share over this sample. It contradicts the idea of fully location agnostic demand in the cloud computing industry and the internet being the "death of distance".

3.7.3 Counterfactual Comparison

We calculate changes in consumer surplus effect in the same way for the price drop counterfactual as for the new DC location counterfactual. Thus we can compare the change in consumer surplus from building a single new DC and compare it to the change in consumer surplus from a 15% price decrease from all Azure DCs. This serves as a sanity check for our estimates and the counterfactuals built upon them.

We find that the increase in consumer surplus from building a new DC in the six states from our counterfactual was 77% of the increase in consumer surplus from a 15% across the board price decrease. Recall that the 15% price decrease impacts all new customers in North America in a single month (the entire set of new customers in a month) whereas the new DC will cause a change in behavior of only a fraction of the monthly extensive margin (e.g., just the customers induced to move to Azure based upon the new DC). Assessing orders of magnitude, this makes sense at a high level: assume roughly 10% of total new customers deploy in the new DC in any given month (i.e., there were 10 Azure DCs at the end of the sample) those customers have a large reduction in distance between the previously closest DC and the new proximate one. Recalling that the implied willingness to pay for 1000kms (∼600 miles) in proximity was 60% of the average price of Azure, those 10% of new cloud customers' benefit implies an average decrease in distance of roughly 750 miles, which seems plausible since not every Azure customer deploys in the closest DC (see Figures 4 and 5 above).

There is another intuitive way to perform a back of the envelope calculation: all customers receiving a 15% price decrease in the price counterfactual have a 15% increase in their customer surplus to a first order approximation. Alternatively, in the new DC counterfactual, only those customers deploying in the new DC have their consumer surplus are impacted by it. There were 10 DCs at the end of our sample so roughly 10% of customers benefit from a new DC. Recall that aggregate consumer surplus from the new DC is 75% of the welfare increase from a 15% price decrease for all newly deploying customers. Hence, for customers deploying in the new DC in our counterfactual, we must observe an increase in consumer surplus of (.15/.1)*.75 = 112.5%.[23] This is again plausible: a new proximate DC could be worth roughly twice as much to cloud users as distant DC based on our parameter estimates.

Given the staggering growth in cloud adoption in the last five years by firms, it is hard to imagine latency concerns being the sole driver of this barrier. For example, the distance decrease of a customer in Atlanta, Georgia to the nearest Azure DC at the time of our study was around 600 miles, or about 6 milliseconds of latency. While we cannot rule out latency as a driver with our data, these results indicate the presence of a secular preference for proximity consistent with "server hugging". If preference based server hugging does explain this result, our evidence suggests an alternative preference based rationale for why the internet may not lead to the "death of distance" in the case of cloud computing.

3.8 Conclusion

We find that cloud compute customers care about proximity to a surprising degree even within the US where latency difference across data centers are often separated in the single digit milliseconds. Our result is consistent with a growing body of work that

[23]If N are the total number of cloud customers then $.1 * N * \Delta CS_{newDC} = .75 * N \Delta CS_{PriceChange} = .75 * N * .15$ and solve for ΔCS_{newDC}.

finds that the internet has not in fact been the “death of distance” although we can’t fully rule out strong preferences for reduced latency with our data. Because customers do care about distance, vigorous spatial competition of public cloud providers like AWS, Azure, GCP and Alibaba in the quickly maturing cloud market are likely to benefit cloud users a great deal and more quickly move firms from wholly owned on-premise servers to remote rented cloud based compute resources. The number of data centers of each cloud provider has roughly doubled in the past five years.

While we do not model an equilibrium entry decision in this paper, there is clear room to expand this line of research in that dimension. Such work could have particular importance given that cloud computing resources lower barriers to entry of new firms and therein enable more productivity from the global labor force. Such models could be used by policy makers to encourage more competition in industries where spatial competition is important and enables aggregate productivity of the economy.

Our methodology could also be useful to other economists. We estimate our demand system when the dataset contains disaggregate consumer level choice data of one firm and aggregate market share data of another. We show that both the taste parameters and a discrete distribution of unobserved consumer attributes can be recovered with EM algorithm under the framework of mixed logit. It enables the identification of demand parameters up to brand level fixed effects which could be further pinned down by the observed market shares. Given demand parameters, the consumer spatial distribution, i.e. the local market sizes, is identified by the inverse of model-predicted local Microsoft market share. A Monte Carlo exercise supports identification.

Finally, our data is from 2016 which are the early days of the cloud computing industry. In 2016 and even in 2020 cloud revenue is growing rapidly. Cloud computing is not a product in long run equilibrium and preferences for cloud attributes are likely to change as cloud users learn and experiment with cloud resources. Hence, our results might not be externally valid in a fully mature cloud computing market.

Appendix A

Appendix for Chapter 1

A.1 Market size

Using the subsample of served markets, I regress the number of upscale hotel rooms (monthly) on population/m^2, number of employees and median income level. They are from or constructed using the following variables in the ACS.

Population/m^2: *B01003 Total Population* gives the population in each census tract. The U.S. *Gazetteer* files, also from the Census Bureau, provides the area measure of each census. With these two variables, I calculate Population/m^2, which is effectively the population density.

Number of employees: *C24050 Industry by Occupation for the civilian employed population 16 years and over*. In particular, I use the category *Arts entertainment and recreation and accommodation and food services*, which is the finest one that is related to the lodging industry.

Median income level: *B19013: Median Household Income.*

The U.S. Gazetteer files also have the representative latitude and longitude coordinates of each census tract, which is defined as *inner point*, with which I calculate the distance between the neighborhoods and census tracts. For each census tract, I pick the geographically closest census tract, and consider the demographics of that census tract as the proxies for the demographics of the neighborhood. I use yearly data from 2010-2018. The regression results are shown in the follow table.

	Estimates
population/m^2	0.698***
	(0.0545)
employees	2.6624***
	(0.7477)
median income	0.056***
	(0.0034)
Year FE	×

Table A.1: OLS regression of market size on demographics

Next, I use these estimates to predict the "number of upscale hotel rooms" of the potential markets. Then I divide this number by the ratio of hotel rooms to all lodging rooms in the corresponding year for all markets, and consider the results as the *market size*. These ratios are nation-wide. Its original source is *Second Measure*, and I found them in an article by Rani Molla published on *vox.com*.[1] Except for hotels room, other lodging options include *Airbnb* and *HomeAway*.

A.2 Estimation procedure

A.2.1 Estimation

Evaluate the objective function at a given θ

1. To calculate empirical probability $\hat{Pr}(\boldsymbol{x}^u_m \leq \boldsymbol{t}^x, \boldsymbol{w}^u_m \leq \boldsymbol{t}^w, z^u(\boldsymbol{y}_m)|\mathcal{O}_m \in \mathcal{O}^g; \theta)$,

    (a) Pick thresholds $(\boldsymbol{t}^x, \boldsymbol{t}^w)$. Specifically, I use the median and maximum of each unobserved shock's empirical distribution. For example, $t^x_j = \{t^{x,1}_j, t^{x,2}_j\} = \{median(x^u_{jm}), max(x^u_{jm})\}$ is a two-component sequence. These thresholds will group $(\boldsymbol{x}^u_m, \boldsymbol{w}^u_m)$ into difference cells. Consider a simpler case where $J = 2$, one cell could be $x^u_{1m} \in (-\infty, t^{x,1}_j]$, $x^u_{2m} \in (t^{x,1}_j, t^{x,2}_j]$, $w^u_{1m} \in (t^{w,1}_j, t^{w,2}_j]$, $w^u_{2m} \in (-\infty, t^{w,1}_j]$. I denote each cell of the grid by $\mathcal{O}^g$;

    (b) Categorize each market m into a grid $\mathcal{O}^g$ based on $\mathcal{O}_m$;

[1] https://www.vox.com/2019/3/25/18276296/airbnb-hotels-hilton-marriott-us-spending

(c) Given observed $(\boldsymbol{p}_m, \boldsymbol{s}_m)$ and $(\boldsymbol{x}_m^o, \boldsymbol{w}_m^o)$, calculate $(\boldsymbol{x}_m^u, \boldsymbol{w}_m^u)$ by inverting the market share function Eq.(1.3) and the pricing rule Eq.(1.4), i.e., for $\forall j, m$ such that $y_{jm} = 1$,

$$\begin{aligned} x_{jm}^u &= ln(s_{jm}) - ln(s_{0m}) - x_{jm}^{o'}\beta - \alpha \\ w_{jm}^u &= exp(p_{jm} + \frac{1}{\alpha(1 - s_{jm})}) - w_{jm'}^o\gamma \end{aligned}$$

(d) calculate

$$\hat{Pr}(\boldsymbol{x}_m^u \le \boldsymbol{t}^x, \boldsymbol{w}_m^u \le \boldsymbol{t}^w, \boldsymbol{y}_m | \mathcal{O}_m \in \mathcal{O}^g; \theta) = \frac{\sum_m 1[\boldsymbol{x}_m^u \le \boldsymbol{t}^x, \boldsymbol{w}_m^u \le \boldsymbol{t}^w, \boldsymbol{y}_m, \mathcal{O}_m \in \mathcal{O}^g; \theta]}{\sum_m 1[\mathcal{O}_m \in \mathcal{O}^g]}$$

2. Simulate the probability bounds $LB(\boldsymbol{x}_m^u \le \boldsymbol{t}^x, \boldsymbol{w}_m^u \le \boldsymbol{t}^w, z^u(\boldsymbol{y}_m) | \mathcal{O}_m \in \mathcal{O}^g; \theta)$ and $UB(\boldsymbol{x}_m^u \le \boldsymbol{t}^x, \boldsymbol{w}_m^u \le \boldsymbol{t}^w, z^u(\boldsymbol{y}_m) | \mathcal{O}_m \in \mathcal{O}^g; \theta)$

(a) take $RS = 100$ random draws from a bi-variate standard normal $\mathcal{N}(0, I_2)$ where I_2 is a 2×2 identity matrix, multiply each draw by $\Omega_{1:2,1:2}$ and denote it as $(x_{jm}^{u,r}, w_{jm}^{u,r})$. [2]

(b) For each market m ,

i. solve for equilibrium $(\boldsymbol{p}_m, \boldsymbol{s}_m)$ given each random draw $(x_{jm}^{u,r}, w_{jm}^{u,r})$ for the following market structures for each bound:

A. upper bound: $\boldsymbol{y}_m$ and $\boldsymbol{y}_m + \boldsymbol{1}^j$;

B. lower bound: $\boldsymbol{1}$ and $\boldsymbol{1}^j$.

ii. calculate the mean $\Omega_{1:2,3}\Omega_{3,3}^{-1}$ and the variance $\Omega_{1:2,1:2} - \Omega_{1:2,3}\Omega_{3,3}^{-1}\Omega_{3,1:2}$ of the conditional distribution $z_{jm}^u | x_{jm}^{u,r}, w_{jm}^{u,r}$

iii. approximate Eq.(1.12) and Eq.(1.13) by

$$\frac{\sum_r (\prod_j \tilde{\Phi}(\pi_j(\boldsymbol{1}) - \boldsymbol{z}_{jm}^o\delta)^{y_{jm}} (1 - \tilde{\Phi}(\pi_j(\boldsymbol{1}^j) - \boldsymbol{z}_{jm}^o\delta))^{1-y_{jm}}) 1[\boldsymbol{x}_m^{u,r} \le \boldsymbol{t}^x, \boldsymbol{w}_m^{u,r} \le \boldsymbol{t}^w, \mathcal{O}_m; \theta]}{R}$$

[2] $\Omega_{a:b,c:d}$ is the subset of Ω including $a^{th} - b^{th}$ rows and $c^{th} - d^{th}$ columns. Given the joint normal assumption, $z^u | (x^u, w^u) \sim \mathcal{N}(\Omega_{1:2,3}\Omega_{3,3}^{-1}(x^u, w^u), \Omega_{1:2,1:2} - \Omega_{1:2,3}\Omega_{3,3}^{-1}\Omega_{3,1:2})$

and

$$\frac{\sum_r (\prod_j \tilde{\Phi}(\pi_j(\boldsymbol{y}_m) - \boldsymbol{z}^o_{jm}\delta))^{y_{jm}}(1 - \tilde{\Phi}(\pi_j(\boldsymbol{y}_m + \mathbf{1}^j) - \boldsymbol{z}^o_{jm}\delta)^{1-y_{jm}})1[\boldsymbol{x}^{u,r}_m \leq \boldsymbol{t}^x, \boldsymbol{w}^{u,r}_m \leq \boldsymbol{t}^w, \mathcal{O}_m; \theta]}{R}$$

(c) summarize over all markets

$$\begin{aligned} & LB(\boldsymbol{x}^u_m \leq \boldsymbol{t}^x, \boldsymbol{w}^u_m \leq \boldsymbol{t}^w, z^u(\boldsymbol{y}_m)|\mathcal{O}_m; \theta) \\ = & \frac{\sum_m LB(\boldsymbol{x}^u_m \leq \boldsymbol{t}^x, \boldsymbol{w}^u_m \leq \boldsymbol{t}^w, z^u(\boldsymbol{y}_m)|\mathcal{O}_m; \theta)1[\mathcal{O}_m \leq \mathcal{O}^g]}{\sum_m 1[\mathcal{O}_m \leq \mathcal{O}^g]} \\ & UB(\boldsymbol{x}^u_m \leq \boldsymbol{t}^x, \boldsymbol{w}^u_m \leq \boldsymbol{t}^w, z^u(\boldsymbol{y}_m)|\mathcal{O}_m; \theta) \\ = & \frac{\sum_m LB(\boldsymbol{x}^u_m \leq \boldsymbol{t}^x, \boldsymbol{w}^u_m \leq \boldsymbol{t}^w, z^u(\boldsymbol{y}_m)|\mathcal{O}_m; \theta)1[\mathcal{O}_m \leq \mathcal{O}^g]}{\sum_m 1[\mathcal{O}_m \leq \mathcal{O}^g]} \end{aligned}$$

3. calculate the objective function in Eq.(1.10).

Search for θ that minimizes the objective function I follow the MCMC procedure in Chernozhukov and Hong (2003) to search for the θ that minimizes the objective function. Notably, moment inequality estimation may result in a identified set instead of a point, which is referred as *partial identification.* So in theory, there could be multiple θ's such that $Q(\theta) = 0$, but in practice there could be misspecification and therefore the objective function may have a unique minimizer or even $\min_\theta Q(\theta) \geq 0$. Following Chernozhukov, Hong and Tamer (2007), I address this issue in the inference stage. Given this potential for misspecification, finding the optimum(s) is not the only goal of this search, I am also looking for candidate values that should be included in the confidence set. Therefore, I keep track of all the θ's that are evaluated during the search. The main steps of this procedure includes

1. Find good start values.

   (a) Construct a range for start values using the estimates and standard errors from the exogenous entry model (Column 2, Table 1.6). For example, the estimate for price coefficient $\hat{\alpha} = -0.232$, and the standard error is 0.027, so I consider the interval -0.232 ± 0.027 as a reasonable range to start the search.

(b) Take $RS = 5000$ random draws from the constructed ranges, each denoted as θ^r. For the parameters in the fixed cost function, I first infer each firm's variable profit π_{jm} using the point estimates in Column2, and then take logs and run an OLS by regressing π_{jm} on the fixed cost covariates, then use the estimates and standard errors from the OLS to construct the range.

(c) Evaluate the objective function $Q(\theta)$ at each $\theta^r, r = 1, 2, ..., 500$.

(d) Pick 10 θ^r's that gives the least objective function, each denoted as $\theta^s, s = 1, 2, ..., 10$.

2. Do the search separately for each θ^s. For each strand of search, I update the each parameter in θ^s iterative via a Gibbs-Hastings procedure.

(a) Dropping the subscript for start value s for ease of notation, for each component in θ, say the price coefficient α, take a draw ψ from the univariate normal density $\mathcal{N}(\alpha^{(j)}, \sigma_\alpha^{(j)})$, where (j) denotes the j-th iteration, and the variance is specified based on the scale of the current value, $\sigma_\alpha^{(j)} = 10^{\lfloor \log(|\alpha^{(j)}|) \rfloor - 1}$.

(b) Let $\theta_{[\psi,\alpha]}$ denote the parameter vector θ with the component α replaced by ψ, calculate $Q(\theta^{(j)}_{[\psi,\alpha]})$.

(c) Update $\alpha^{(j+1)}$ from $\alpha^{(j)}$ using

$$\alpha^{(j+1)} = \begin{cases} \psi & \text{with probability } \rho(\theta^{(j)}, \psi) \\ \alpha^{(j)} & \text{with probability } 1 - \rho(\theta^{(j)}, \psi) \end{cases}$$

where

$$\rho(\theta^j, \psi) = inf(\frac{exp(Q(\theta^{(j)}_{[\psi,\alpha]}))}{exp(Q(\theta^{(j)}))}, 1)$$

(d) Stop searching when $\dfrac{|Q(\theta^{(j+1)}) - Q(\theta^{(j)})|}{1 + |Q(\theta^{(j)})|} \leq 10^{-4}$

A.2.2 Inference

With the collection of θ's given by the MCMC search, I use the methodology of Chernozhukov, Hong and Tamer (2007) to construct the confidence set, same as in Ciliberto and Tamer (2009), Ciliberto, Murry and Tamer (2018).

A.3 Identification details

This section closely follows Ciliberto and Tamer (2009) and Ciliberto, Murry and Tamer (2018).

In CMT, the point identification of the fixed cost function parameters relies on a firm heterogeneous covariate that is excluded from the variable profit function and has a wide range, say $z^o_{jm} \in (-\infty, +\infty)$. Intuitively, the inter-dependence among firms' entry decisions is the main threat to identification. If z^o_{jm} affects a firm's entry decision in a positive way, suppose that in some markets it goes to infinity for $\forall j \neq 1$, then $y_{jm} = 1, \forall j \neq 1$ for sure. In that case, the parameters in the entry equation can be identified by how firm 1's entry probability varies with each covariates in these markets. In my fixed cost function, even though $rent_m$ is not firm-specific, there is a fixed effect for each top 3 firm, and estimates indicate that Marriott has a significant higher average fixed cost. Therefore, if in some low rent markets all firms would definitely enter except for Marriott, the *rent* coefficient can be identified by how Marriott's entry choices vary with rent across these markets.

Furthermore, when $y_{jm} = 1$ almost for sure, there will be no selection based on the demand shock x^u_{jm} and the marginal cost shock w^u_{jm}. Therefore the identification of the parameters in the price competition stage follows that for a classical demand system, which requires at least $2 \times J$ instruments for both the endogenous prices and the endogenous market shares (Berry and Haile, 2016). As argued in Section 4.2 of Ciliberto, Murry and Tamer (2018), the instruments here are similar to the "BLP instruments" (Berry, Levinsohn and Pakes, 1995), only that since the moments are conditional on the exogenous variables of each potential entrant, I essentially use each firm's exogenous variable separately as instruments instead of summing or averaging them within a market. Notably, all potential entrants, not only the actual incumbents, are included in the condition, otherwise the instruments themselves would have the selection problem.

In particular, consider the simultaneous equation system formed by the inverse market share function and pricing rule

$$\begin{cases} ln(s_{jm}) - ln(s_{0m}) = \beta N_{jm}^{uc} + \alpha p_{jm} + x_{jm}^{u} \\ p_{jm} = mc_{jm} + \dfrac{1}{\alpha(1 - s_{jm})} = exp(\gamma N_{jm}^{all} + w_{jm}^{u}) + \dfrac{1}{\alpha(1 - s_{jm})}. \end{cases}$$

The pricing rule indicates that p_{jm} is correlated with x_{jm}^{u} through the markup, more specifically through the market share function. In that sense, the excluded variable in the marginal cost function N_{jm}^{all} can be a valid instrument, since it alters price but is not correlated with x_{jm}^{u}. Once the demand side is identified, the role of aff_uc_{km} and aff_all_{km}, $\forall k \neq j$ as instruments in identifying the supply side can been seen by first recalling that the s_{jm} is a function of the utility indices of all firms[3], therefore aff_uc_{km}, $\forall k \neq j$ shifts s_{jm} directly and aff_all_{km}, $\forall k \neq j$ does it through p_{km}, $\forall k \neq j$. However, it is unlikely that firm j's unobserved marginal cost shock in neighborhood m, w_{jm}^{u}, would be correlated with the scale of firm k's loyalty program in the entire MSA.

[3]Non-entrants' exogenous variables affect s_{jm} by determining the identities of the actual incumbents.

The identification theorems in Ciliberto and Tamer (2009) and Ciliberto, Murry and Tamer (2018) are required for point identification. Since the inference method I use is robust to partial identification, slight looseness in the condition that the exogenous variables have infinite ranges is not catastrophic.

A.4 Simulation

I use simulation to show that compared to using unique equilibrium lower bound, using DSE lower bound can dramatically reduce the computation time and at the same time keep the confidence interval similarly tight. The model I use in this simulation exercise follows the structure of the two-player example in Section 1.4.1. I generalize it so that more players are involved, and impose a distributional assumption on the cost shock to facilitate simulation. Specifically,

$$\begin{aligned} y_{jm} &= 1[\alpha - \textstyle\sum_k \beta_{km} y_{km} - c_{km} \geq 0] \\ c_{jm} &\sim U[0,1],\ i.i.d\ \forall j, m. \end{aligned}$$

Similar to the model in Section 1.3, $j, k = 1, 2, ..., J$ is the index for firms, and $m = 1, 2, ..., M$ indicates markets. For each J, I estimate the model using the DSE lower bound (DSE hereafter) and the unique equilibrium. In particular, since there is no structural price competition stage, the estimation using the unique equilibrium follows the procedure in Ciliberto and Tamer (2009). Thus I refer it as CT hereafter.

I compare the performance of DS and CT for $J = 2, 3, ..., 10$. Throughout all J's, I specify the true parameter values $\alpha = 0.75$, $\beta_j = 0.75/J$. And I simulate $M = 10,000$ markets for each J The objective function is the same as Eq.(1.10), and the probability bounds are constructed as follows.

DS In this simple example, the upper bound has a closed-form formula, so does the lower bound when it is based on the DS. Specifically,

Table A.2: Computation Time Comparison - 95% Confidence Set

Number of firms	2	3	4	5
$\alpha =$		0.75		
CT [2009]	[0.748, 0.759]	[0.749, 0.764]	[0.735, 0.804]	[0.750, 0.786]
DS	[0.747, 0.753]	[0.747, 0.755]	[0.739, 0.762]	[0.726, 0.772]
$\beta_1 =$	0.375	0.25	0.1875	0.15
CT [2009]	[0.328, 0.468]	[0.220, 0.284]	[0.164, 0.235]	[0.149, 0.182]
DS	[0.337, 0.438]	[0.231, 0.285]	[0.166, 0.240]	[0.120, 0.178]
Computation time				
CT [2009]	5.08 hrs	12.13 hrs	21.78 hrs	55.46 hrs
DS	0.61 mins	1.29 mins	3.59 mins	7.85 mins

Computational setting: M = 10000, nBS = 100, nCores = 4; for CT[2009], R = 10000

$$\begin{aligned} LB(\boldsymbol{y}|\theta) &= \prod_j (1 - \sum_{k\neq j} b_k)^{y_j} (1-a)^{1-y_j} \\ UB(\boldsymbol{y}|\theta) &= \prod_j (1 - \sum_{k\neq j} y_k b_k)^{y_j} (1-(a - \sum_{k\neq j} y_k b_k)))^{1-y_j} \end{aligned}$$

CT When using the unique equilibrium lower bound, there is no closed-form. So following Ciliberto and Tamer (2009), I use Monte Carlo simulation to approximate both of the probability bounds.[4]. The estimation procedure has been briefly discussed in the two-player example in Section 1.4.1, for more details, the interested readers are referred to the supplementary materials of Ciliberto and Tamer (2009). In particular, I let $RS = 10,000$.

The rest of the procedure is essentially the same as described in Appendix A.2. A slight difference is that I use the R-package *GenSA* (Yang Xiang et al. (2013)) to search for the optimum θ, therefore the search algorithm is simulated annealing instead of MCMC.

The results are shown in Table A.2. All the jobs were run on 4-core CPU's, and parallel computation is used whenever applicable.

[4]Note that even though the upper bound has a closed-form, it is more convenient to use simulation in practice because the equilibrium $\boldsymbol{y}_m$ has been solved for the lower bound anyways.

Appendix B

Appendix for Chapter 2

B.1 Proofs

B.1.1 Proof of Theorem 1

(1) Given that $\boldsymbol{\theta}^{(k+1)} = argmax_{\boldsymbol{\theta}} S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ and by definition of minorization for maximization,

$$\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) \geq S(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \geq S(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) = \mathcal{L}(\boldsymbol{\theta}^{(k)}) \tag{B.1}$$

Because $\mathcal{L}(\boldsymbol{\theta})$ is a log likelihood function, we have that $\mathcal{L}(\boldsymbol{\theta}) \leq 0, \forall \boldsymbol{\theta}$. Then the fact that $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ is an increasing sequence bounded above implies its convergence to some $\mathcal{L}^* \leq 0$. Hence for any $\delta > 0$, there exists a $p(\delta)$ such that for all $p \geq p(\delta)$ and all $r \geq 1$,

$$\sum_{s=1}^{r} \{\mathcal{L}(\boldsymbol{\theta}^{(p+s)}) - \mathcal{L}(\boldsymbol{\theta}^{(p+s-1)})\} = \mathcal{L}(\boldsymbol{\theta}^{(p+r)}) - \mathcal{L}(\boldsymbol{\theta}^{(p)}) < \delta \tag{B.2}$$

From Eq.(B.1), we have

$$S(\boldsymbol{\theta}^{(p+s)}; \boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)}; \boldsymbol{\theta}^{(p+s-1)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(p+s)}) - \mathcal{L}(\boldsymbol{\theta}^{(p+s-1)}), \ \forall s \geq 1, \tag{B.3}$$

and by Taylor expansion,

$$\begin{aligned}
& S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)};\boldsymbol{\theta}^{(p+s-1)}) \\
= & -(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'\nabla^{10} S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) \\
& \quad -(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'\nabla^{20} S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)}) \\
= & -(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'\nabla^{20} S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})
\end{aligned} \tag{B.4}$$

where $\boldsymbol{\theta}_0^{(p+s)}$ is some point on the line segment joining $\boldsymbol{\theta}^{(p+s-1)}$ and $\boldsymbol{\theta}^{(p+s)}$. The second equality holds because $\nabla^{10} S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) = 0$ is the necessary condition for $\boldsymbol{\theta}^{(p+s)} = \mathit{argmax}_{\boldsymbol{\theta}}\, S(\boldsymbol{\theta};\boldsymbol{\theta}^{(p+s-1)})$ given that $S(\boldsymbol{\theta};\boldsymbol{\theta}')$ is twice differentiable by definition.

Furthermore, Definition 1 (2) indicates that $\nabla^{20} S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})$ is negative definite, i.e.. $-\nabla^{20} S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)})$ is positive definite. Therefore, let I be an identity matrix, $\exists \lambda > 0$ such that $-\nabla^{20} S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) > \lambda I$, i.e. $\nabla^{20} S(\boldsymbol{\theta}_0^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) + \lambda I$ is positive definite.

Then Eq.(B.4) can be rewritten as

$$S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)};\boldsymbol{\theta}^{(p+s-1)}) > \lambda(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)}), \forall s \geq 1 \tag{B.5}$$

Combining Eq.(B.2) , Eq.(B.3), Eq.(B.4) and Eq.(B.5), we have

$$\begin{aligned}
& \lambda \sum_{s=1}^{r} (\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)})'(\boldsymbol{\theta}^{(p+s-1)} - \boldsymbol{\theta}^{(p+s)}) \\
< & \sum_{s=1}^{r} S(\boldsymbol{\theta}^{(p+s)};\boldsymbol{\theta}^{(p+s-1)}) - S(\boldsymbol{\theta}^{(p+s-1)};\boldsymbol{\theta}^{(p+s-1)}) \\
< & \ \delta
\end{aligned} \tag{B.6}$$

for all $p \geq p(\delta)$ and all $r \geq 1$, which proves $\boldsymbol{\theta}^{(k)}$ converges to some $\boldsymbol{\theta}^*$ in the closure of Ω.

(2) Since $\boldsymbol{\theta}^{(k)}, k = 0, 1, 2, ...$ converges to $\boldsymbol{\theta}^*$,

$$\boldsymbol{\theta}^* = \mathit{argmax}\, S(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$$

Then by Corollary 1,

$$\mathcal{L}(\boldsymbol{\theta}^*) = \nabla^{10} S(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*) = 0$$

Similarly, $\nabla^{20} S(\boldsymbol{\theta}^*; \boldsymbol{\theta}^*)$ is negative definite. Q.E.D

B.1.2 Proof of Theorem 2

We show that $S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ is a minorization of $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ at $\boldsymbol{\theta}^{(k)}$ for maximization by checking the three requirements in Definition 1.

(1) $\underline{S(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \leq \mathcal{L}(\boldsymbol{\theta}^{(k)})}$

We start from a representative consumer i. Recall that $\boldsymbol{\phi}_{it} = x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)}$. Then, in the case of multinomial logit,

$$l(\boldsymbol{\phi}_{it}; \boldsymbol{y}_{it}) = \prod_j \Big(\frac{exp(\phi_{ijt})}{1 + \sum_{k \neq j} exp(\phi_{ikt})} \Big)^{y_{ijt}}$$

$$\begin{aligned} h_j(\boldsymbol{\phi}_{it}; \boldsymbol{y}_{it}) &= \frac{\partial \log\ l(\boldsymbol{\phi}_{it}; \boldsymbol{y}_{it})}{\partial \phi_{ijt}} \\ &= y_{ijt} - p_{ijt} \end{aligned}$$

$$\begin{aligned} h_{jk}(\boldsymbol{\phi}_{it}; \boldsymbol{y}_{it}) &= \frac{\partial^2 \log\ l(\boldsymbol{\phi}_{it}; \boldsymbol{y}_{it})}{\partial \phi_{ijt} \partial \phi_{ikt}} \\ &= \begin{cases} -p_{ijk}(1 - p_{ikt}), & j = k \\ p_{ijt} p_{ikt}, & j \neq k \end{cases} \end{aligned}$$

The Taylor expansion of $log\ l(\boldsymbol{\phi}_{it})$ at $\tilde{\boldsymbol{\phi}}_{it}$ is ,

$$\begin{aligned} \log\ l(\boldsymbol{\phi}_{it}; \boldsymbol{y}_{it}) &= \log\ l(\tilde{\boldsymbol{\phi}}_{it}; \boldsymbol{y}_{it}) + (\boldsymbol{\phi}_{it} - \tilde{\boldsymbol{\phi}}_{it})' \nabla \log\ l(\tilde{\boldsymbol{\phi}}_{it}; \boldsymbol{y}_{it}) \\ &\quad + \frac{1}{2} (\boldsymbol{\phi}_{it} - \tilde{\boldsymbol{\phi}}_{it})' \nabla^2 \log\ l(\boldsymbol{\phi}^*_{it}; \boldsymbol{y}_{it}) (\boldsymbol{\phi}_{it} - \tilde{\boldsymbol{\phi}}_{it}) \end{aligned} \tag{B.7}$$

Given that

$$\nabla^2 \log\ l(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) = \begin{bmatrix} h_{11}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) & h_{12}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) & \dots & h_{1J}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) \\ h_{21}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) & h_{22}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) & \dots & h_{2J}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) \\ \vdots & \vdots & & \vdots \\ h_{J1}(\boldsymbol{\phi}_{it}^*) & h_{J2}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) & \dots & h_{JJ}(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) \end{bmatrix}$$

$$= \begin{bmatrix} -p_1(1-p_1) & p_1p_2 & \dots & p_1p_J \\ p_2p_1 & -p_2(1-p_2) & \dots & p_2p_J \\ \vdots & \vdots & & \vdots \\ p_Jp_1 & p_1p_2 & \dots & -p_J(1-p_J) \end{bmatrix}, \tag{B.8}$$

we have $\nabla^2 \log\ l(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) \geq -I$, i.e. $\nabla^2 \log\ l(\boldsymbol{\phi}_{it}^*;\boldsymbol{y}_{it}) + I$ is semi-positive definite matrix.

Therefore,

$$\begin{aligned} \log\ l(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it}) &\geq \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + (\boldsymbol{\phi}_{it} - \tilde{\boldsymbol{\phi}}_{it})'\nabla \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) - \frac{1}{2}(\boldsymbol{\phi}_{it} - \tilde{\boldsymbol{\phi}}_{it})'(\boldsymbol{\phi}_{it} - \tilde{\boldsymbol{\phi}}_{it}) \\ &= \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + \sum_j h_j(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it})(\phi_{ijt} - \tilde{\phi}_{ijt}) - \frac{1}{2}\sum_j(\phi_{ijt} - \tilde{\phi}_{ijt})^2 \\ &= \log\ l(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) + \frac{1}{2}\sum_j h_j(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it})^2 - \frac{1}{2}\sum_j(\tilde{\phi}_{ijt} - h_j(\tilde{\boldsymbol{\phi}}_{it};\boldsymbol{y}_{it}) - \phi_{ijt})^2 \end{aligned} \tag{B.9}$$

Recall that $l(\boldsymbol{\phi}_{it};\boldsymbol{y}_{it})$ is the individual log-likelihood, the log-likelihood function is its sum over i and t. and we substitute $\boldsymbol{\phi}_{it}$ and $\tilde{\boldsymbol{\phi}}_{it}$ in inequality (B.9) with $x_{it}\boldsymbol{\beta}+\boldsymbol{\xi}_{iq(t)}$ and $x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)}$ respectively. It gives

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) \geq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(t)};\boldsymbol{y}_{it})^2 - \\ \frac{1}{2}\sum_{i,j,t}(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)} - h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}_{iq(t)}^{(k)};\boldsymbol{y}_{it}) - x_{it}\boldsymbol{\beta} - \boldsymbol{\xi}_{iq(t)})^2 \end{aligned} \tag{B.10}$$

where the RHS of the above inequality as $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$.

(2) $\underline{S(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}) = \mathcal{L}(\boldsymbol{\theta}^{(k)})}$

By definition,

$$\begin{aligned} S(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)}) &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)};\boldsymbol{y}_{it})^2 - \\ &\quad \frac{1}{2}\sum_{i,j,t}(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)} - h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)};\boldsymbol{y}_{it}) - x_{it}\boldsymbol{\beta}^{(k)} - \boldsymbol{\xi}^{(k)}_{iq(t)})^2 \\ &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)};\boldsymbol{y}_{it})^2 - \frac{1}{2}\sum_{i,j,t} h_j(x_{it}\boldsymbol{\beta}^{(k)} + \boldsymbol{\xi}^{(k)}_{iq(t)};\boldsymbol{y}_{it})^2 \\ &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) \end{aligned}$$

(3) $\nabla^{20}S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ exists, and $\nabla^{20}S(\boldsymbol{\theta}^{(k+1)};\boldsymbol{\theta}^{(k)})$ is negative definite

To ease notation, we consider $\boldsymbol{\xi}_{iq(t)}$ as coefficients on indicator variables, combine these indicator variables with x_{it} and denote the combined vector as z_{ijt}. Then by the definition of $S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ in Eq.(B.10),

$$\nabla^{20}S(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = -2\boldsymbol{Z}'\boldsymbol{Z},\ \forall\boldsymbol{\theta}$$

where $\boldsymbol{Z}$ is the matrix with $z_{ijt},\ \forall i,j,t$ stacked by rows.

Q.E.D

B.2 Computing Standard Errors

We rewrite the likelihood function as $l(x_{it}\boldsymbol{\beta} + \boldsymbol{\xi}_{iq(i,t)};\boldsymbol{y}_{it}) = l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta});\boldsymbol{y}_{it})$. For each household i and each trip t,

$$\begin{aligned} \frac{\partial \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta});\boldsymbol{y}_{it})}{\partial\boldsymbol{\theta}} &= \frac{\partial \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta});\boldsymbol{y}_{it})}{\partial\boldsymbol{\phi}_{it}} \times \frac{\partial\boldsymbol{\phi}_{it}}{\partial\boldsymbol{\theta}} \\ &= (\boldsymbol{y}_{it} - \boldsymbol{p}_{it})' \times \frac{\partial\boldsymbol{\phi}_{it}}{\partial\boldsymbol{\theta}} \end{aligned} \tag{B.11}$$

where $\dfrac{\partial\boldsymbol{\phi}_{it}}{\partial\boldsymbol{\theta}}$ is a $J \times (J + J \times I \times Q)$ matrix.

Let $\text{diag}_J(a)$ be a $J \times J$ matrix with diagonal elements as a and off-diagonal elements as 0,

$$\left[\frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\theta}}\right]_{.;1:J} = \frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\beta}} = \mathrm{diag}_J(x_{it})$$
$$\left[\frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\theta}}\right]_{.;J\times i+1:J} = \frac{\partial \boldsymbol{\phi}_{it}}{\partial \boldsymbol{\xi}_{iq(i,t)}} = \mathrm{diag}_J(1)$$

while the other elements are zero because $\boldsymbol{\xi}_{hq(h,t)}$ is not in $l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})$ if $h \neq i$. Therefore, Equation B.11 can be rewritten as:

$$((y_{i1t} - p_{i1t})x_{it}, \cdots, (y_{iJt} - p_{iJt})x_{it}, 0, \cdots, 0, y_{i1t} - p_{i1t}, \cdots, y_{iJt} - p_{iJt}, 0, \cdots, 0)$$

Then $\dfrac{\partial^2 \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ matrix can be written as

$$\begin{bmatrix}
H_{\boldsymbol{\phi}_{it}} x_{it}^2, & \mathbf{0}, & \cdots, & \mathbf{0} & H_{\boldsymbol{\phi}_{it}} x_{it} & \mathbf{0}, & \cdots, & \mathbf{0} \\
\mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} \\
H_{\boldsymbol{\phi}_{it}} x_{it} & \mathbf{0}, & \cdots, & \mathbf{0} & H_{\boldsymbol{\phi}_{it}} & \mathbf{0}, & \cdots, & \mathbf{0} \\
\mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0} & \mathbf{0} & \mathbf{0}, & \cdots, & \mathbf{0}
\end{bmatrix}$$

where $H_{\boldsymbol{\phi}_{it}}$ is as defined in Eq.(B.8)

And then the Hessian matrix is $\dfrac{1}{I \times T} \sum_{i,t} \dfrac{\partial^2 \log l(\boldsymbol{\phi}_{it}(\boldsymbol{\theta}); \boldsymbol{y}_{it})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$

Appendix C

Appendix for Chapter 3

C.1 Detailed treatment of the EM algorithm

Following Bhat (1997), it can be shown that maximizing Eq.(3.7) is mathematically equivalent to maximizing

$$\sum_t \Big(\sum_{i \in C_t^M} \sum_{j \in \mathcal{F}_t^M} y_{ijt} log(P_{it}^j(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)) + Q_t^A \int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta}) log(P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)) d\mathbf{l}_i + Q_t^O \int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^O(\boldsymbol{\theta}) log(P_{bt}^O(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)) d\mathbf{l}_i$$

if $h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta})$ and $h_{\mathbf{l}_i,t}^O(\boldsymbol{\theta})$ are taken as given.[1] Here, $h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta})$ and $h_{\mathbf{l}_i,t}^O(\boldsymbol{\theta})$ are the

[1] Take the second term in Eq.(3.7) as an example, the necessary first-order conditions for maximizing it is

$$\begin{aligned}
\frac{\partial log(\int_{\mathbf{l}_i} P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2) d\mathbf{l}_i)}{\partial \boldsymbol{\theta}} &= \frac{1}{\int_{\mathbf{l}_i} P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2) d\mathbf{l}_i} \frac{\partial \int_{\mathbf{l}_i} P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2) d\mathbf{l}_i}{\partial \boldsymbol{\theta}} \\
&= \int_{\mathbf{l}_i} \frac{1}{\int_{\mathbf{l}_i} P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2) d\mathbf{l}_i} \frac{\partial P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}} d\mathbf{l}_i \\
&= \int_{\mathbf{l}_i} \frac{P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\int_{\mathbf{l}_i} P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2) d\mathbf{l}_i} \frac{\partial P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}} / P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2) d\mathbf{l}_i \\
&= \int_{\mathbf{l}_i} \frac{P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\int_{\mathbf{l}_i} P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2) d\mathbf{l}_i} \frac{\partial log P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}} d\mathbf{l}_i \\
&= \int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta}) \frac{\partial log P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}} d\mathbf{l}_i
\end{aligned}$$

It is equivalent to maximizing $\int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta}) log(P_{it}^A(\boldsymbol{\theta}_1) f(\mathbf{l}_i|\boldsymbol{\theta}_2)) d\mathbf{l}_i$ with $h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta})$ as given.

Bayesian posterior probabilities that an AWS customer or a non-cloud user is located at $\mathbf{l}_i$, i.e.

$$h^{A}_{\mathbf{l}_i,t}(\boldsymbol{\theta}) = \frac{P^{A}_{it}(\boldsymbol{\theta}_1)f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\int\limits_{\mathbf{l}_i} P^{A}_{it}(\boldsymbol{\theta}_1)f(\mathbf{l}_i|\boldsymbol{\theta}_2)d\mathbf{l}_i} \tag{C.1}$$

$$h^{O}_{\mathbf{l}_i,t}(\boldsymbol{\theta}) = \frac{P^{O}_{it}(\boldsymbol{\theta}_1)f(\mathbf{l}_i|\boldsymbol{\theta}_2)}{\int\limits_{\mathbf{l}_i} P^{O}_{it}(\boldsymbol{\theta}_1)f(\mathbf{l}_i|\boldsymbol{\theta}_2)d\mathbf{l}_i} \tag{C.2}$$

C.1.1 Maximization

Equation (3.8) can be maximized iteratively: starting from some initial values $\boldsymbol{\theta}^s$, we update $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^{s+1}$ which maximizes Eq.(3.8) conditional on $h^{A}_{\mathbf{l}_i t}(\boldsymbol{\theta}^s)$ and $h^{O}_{\mathbf{l}_i t}(\boldsymbol{\theta}^s)$. Formally,

$$\begin{aligned}
\boldsymbol{\varepsilon}(\boldsymbol{\theta}|\boldsymbol{\theta}^s) &= \sum_t \Big(\sum_{i\in C^M_t} \sum_{j\in \mathcal{F}^M_t} y_{ijt} log(P^{j}_{it}(\boldsymbol{\theta}_1)f(\mathbf{l}_i|\boldsymbol{\theta}_2)) \\
&\quad + Q^A_t \int\limits_{\mathbf{l}_i} h^{A}_{\mathbf{l}_i,t}(\boldsymbol{\theta}^s) log(P^{A}_{it}(\boldsymbol{\theta}_1)f(\mathbf{l}_i|\boldsymbol{\theta}_2))d\mathbf{l}_i \\
&\quad\quad + Q^O_t \int\limits_{\mathbf{l}_i} h^{O}_{\mathbf{l}_i,t}(\boldsymbol{\theta}^s) log(P^{O}_{bt}(\boldsymbol{\theta}_1)f(\mathbf{l}_i|\boldsymbol{\theta}_2))d\mathbf{l}_i \Big) \qquad \text{(C.3)} \\
\boldsymbol{\theta}^{s+1} &= argmax_{\boldsymbol{\theta}} \quad \boldsymbol{\varepsilon}(\boldsymbol{\theta}|\boldsymbol{\theta}^s)
\end{aligned}$$

Furthermore, due to the property of log operation, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can be separately updated by maximizing the following two objective functions,

$$\begin{aligned}\boldsymbol{\varepsilon}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^s) &= \sum_t(\sum_{i\in C_t^M}\sum_{j\in\mathcal{F}_t^M} y_{ijt}logP_{it}^j(\boldsymbol{\theta}_1)\\ &\quad +Q_t^A\int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta}^s)logP_{it}^A(\boldsymbol{\theta}_1)d\mathbf{l}_i\\ &\quad +Q_t^O\int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^O(\boldsymbol{\theta}^s)logP_{bt}^O(\boldsymbol{\theta}_1)d\mathbf{l}_i)\\ \boldsymbol{\varepsilon}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^s) &= \sum_t(\sum_{i\in C_t^M}\sum_{j\in\mathcal{F}_t^M} y_{ijt}logf(\mathbf{l}_i|\boldsymbol{\theta}_2)\\ &\quad +Q_t^A\int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^A(\boldsymbol{\theta}^s)logf(\mathbf{l}_i|\boldsymbol{\theta}_2)d\mathbf{l}_i\\ &\quad +Q_t^O\int_{\mathbf{l}_i} h_{\mathbf{l}_i,t}^O(\boldsymbol{\theta}^s)logf(\mathbf{l}_i|\boldsymbol{\theta}_2)d\mathbf{l}_i)\end{aligned}$$

C.1.2 A Discrete Spatial Distribution

Instead of assuming a parametric distribution for $f(\mathbf{l}_i|\boldsymbol{\theta}_2)$, we assume a discrete spatial distribution of consumer locations, so in theory it can approximate any arbitrary distribution when the discretization is fine enough. Specifically, we take each U.S. state and Canadian province as a bin B, and then the probability that a consumer (including non-cloud users) belongs to a certain bin b in period t is q_{bt}, and these q_{bt}'s are treated as parameters to estimate. So the objective functions is given as [2]

[2]For Microsoft customers,

$$\begin{aligned}\sum_{i\in C_t^M}\sum_{j\in\mathcal{F}_t^M} y_{ijt}logf(\mathbf{l}_i|\boldsymbol{\theta}_2) &= \sum_b\sum_{i\in B_b}\sum_{j\in\mathcal{F}_t^M} y_{ijt}logq_{bt}\\ &= \sum_b\sum_{i\in B_b} logq_{bt}\sum_{j\in\mathcal{F}_t^M} y_{ijt}\\ &= \sum_b\sum_{i\in B_b} logq_{bt}\\ &= \sum_b Q_t^M q_{bt}^M logq_{bt}\end{aligned}$$

The third equation holds because these Microsoft customers must choose one of the Microsoft DCs. And Q_t^M is the demand for Microsoft in period t, and q_{bt}^M is the spatial distribution specific for Microsoft consumers. Therefore $Q_t^M q_{bt}^M$ is the number of Microsoft customers in bin B_b, which is observable in our dataset .

$$\begin{aligned}\boldsymbol{\varepsilon}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^s) &= \sum_t (\sum_{i\in C_t^M}\sum_{j\in\mathcal{F}_t^M} y_{ijt} log P_{it}^j(\boldsymbol{\theta}_1) \\ &\quad + Q_t^A \sum_b h_{bt}^A(\boldsymbol{\theta}^s) log P_{bt}^A(\boldsymbol{\theta}_1) \\ &\quad + Q_t^O \sum_b h_{bt}^O(\boldsymbol{\theta}^s) log P_{bt}^O(\boldsymbol{\theta}_1)) \\ \boldsymbol{\varepsilon}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^s) &= \sum_t (Q_t^M \sum_b q_{bt}^M log q_{bt} + Q_t^A \sum_b h_{bt}^A(\boldsymbol{\theta}^s) log q_{bt} + Q_t^O \sum_b h_{bt}^O(\boldsymbol{\theta}^s) log q_{bt})\end{aligned}$$

where

$$h_{bt}^A(\boldsymbol{\theta}^s) = \frac{P_{bt}^A(\boldsymbol{\theta}_1^s)q_{bt}^s}{\sum_b P_{bt}^A(\boldsymbol{\theta}_1^s)q_{bt}^s} \tag{C.4}$$

$$h_{bt}^O(\boldsymbol{\theta}^s) = \frac{P_{bt}^O(\boldsymbol{\theta}_1^s)q_{bt}^s}{\sum_b P_{bt}^O(\boldsymbol{\theta}_1^s)q_{bt}^s} \tag{C.5}$$

Intuitively, $\boldsymbol{\varepsilon}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^s)$ can be considered as a variant of an ordinary multinomial logit model: since AWS customers in bin b share the same log likelihood $log P_{bt}^A(\boldsymbol{\theta}_1)$, it is multiplied by $Q_t^A h_{bt}^A(\boldsymbol{\theta}^s)$, the "posterior" number of AWS customers in bin b. Parallely, $log P_{bt}^O(\boldsymbol{\theta}_1)$ is multiplied by the "posterior" number of people who choose the outside option, $Q_t^O h_{bt}^O(\boldsymbol{\theta}^s)$. Therefore, we are essentially matching the predicted choice probabilities, or say market shares, with the "observed" ones given by $\boldsymbol{\theta}^s$.

For $\boldsymbol{\varepsilon}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^s)$, if we rewrite it as

$$\boldsymbol{\varepsilon}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^s) = \sum_t \sum_b (Q_t^M q_{bt}^M + Q_t^A h_{bt}^A(\boldsymbol{\theta}^s) + Q_t^O h_{bt}^O(\boldsymbol{\theta}^s)) log q_{bt},$$

it can be interpreted as pairing each q_{bt} with the "observed" total probability that a consumer belongs bin b. Moreover, it has a closed-form optimizer, i.e.

$$q_{bt}^{s+1} = \frac{Q_t^M q_{bt}^M + Q_t^A h_{bt}^A(\boldsymbol{\theta}^s) + Q_t^O h_{bt}^O(\boldsymbol{\theta}^s)}{M_t},$$

where $M_t = Q_t^M + Q_t^A + Q_t^O$ is used to denote the market size in period t. This closed-form solution would significantly ease the computation.

Henceforth, we repeat the procedure in Eq.(C.3) until parameters converge.

C.2 Monte Carlo Results

Table C.1: Monte Carlo Experiment

	True value	Mean absolute error	Median absolute error	95% confidence interval
Panel A: Taste Parameters				
γ (Distance in 1000km)				
Discrete Manufacturing	-0.920	0.028	0.023	$[-0.985, -0.872]$
Education	-1.920	0.065	0.054	$[-2.082, -1.783]$
Health	-1.593	0.049	0.036	$[-1.752, -1.523]$
Hospitality & Transportation	-2.265	0.073	0.059	$[-2.480, -2.148]$
Insurance	-4.550	0.120	0.083	$[-4.758, -4.184]$
Media / Telecome and Utilities	-1.749	0.053	0.038	$[-1.887, -1.621]$
Nonprofit	-4.438	0.133	0.106	$[-4.475, -4.084]$
Professional Services	-0.709	0.016	0.014	$[-0.751, -0.675]$
Unknwon	-0.490	0.011	0.008	$[-0.522, -0.466]$
β	-0.018	0.039	0.039	$[-0.216, -0.090]$
ψ	1.584	0.086	0.086	[1.415, 1.561]
ξ	1.068	0.044	0.036	[0.950,1.123]
ζ	2.229	0.037	0.035	[2.194, 2.328]
ζ_7	-0.001	0.001	0.003	$[-0.001, -0.001]$
α	6.327	0.074	0.064	[6.127, 6.458]
α_7	-0.897	0.012	0.010	$[-0.912, -0.866]$
τ	0.221	0.005	0.005	[0.199, 0.222]
Panel B: Consumer Spatial Distribution Parameters				
Std Err. q_{mbt}	4.342×10^{-3}	1.4870×10^{-5}	1.5021×10^{-5}	$[4.120 \times 10^{-3}, 4.315 \times 10^{-3}]$

This Figure shows implied cloud market size by industry for different regions in the US with dark colors being large market sizes. Put another way, this Figure shows the total demand by industry of AWS plus Azure for different regions. It recovers sensible patterns such as discrete manufacturing is prominent in the upper midwest and west coast and professional services being largest on the east coast and and west coast. We take this as evidence the model is recovering sensible market level patterns.

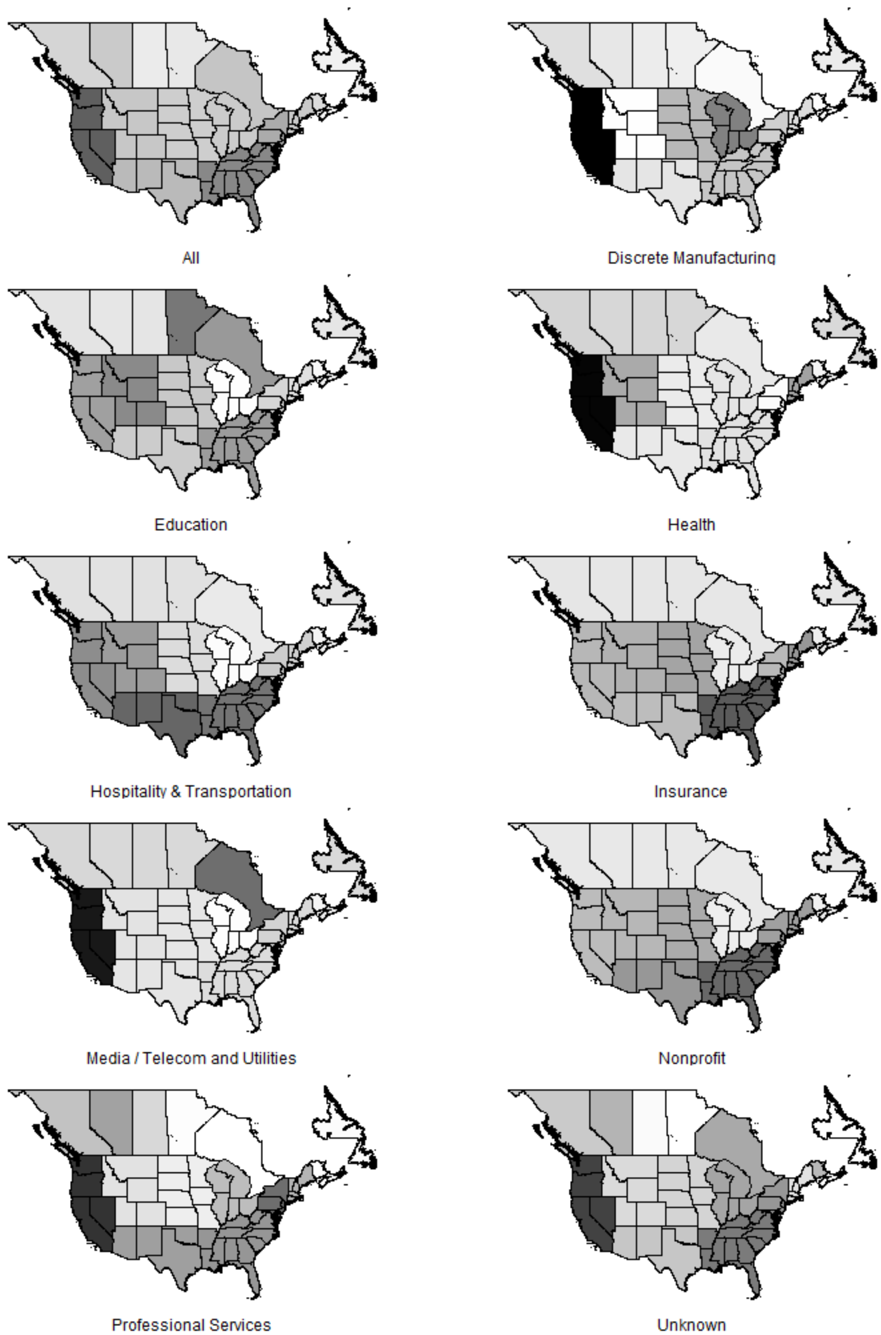


Figure C·1: Market size map

References

Aguirregabiria, Victor, and Chun-Yu Ho. 2012. "A dynamic oligopoly game of the US airline industry: Estimation and policy experiments." Journal of Econometrics, 168(1): 156–173.

Aradillas-Lopez, Andres. 2010. "Semiparametric estimation of a simultaneous game with incomplete information." Journal of Econometrics, 157(2): 409–431.

Arango, Carlos, Kim Huynh, and Leonard Sabetti. 2015. "Consumer payment choice: Merchant card acceptance versus pricing incentives." Journal of Banking and Finance, 55: 130–141.

Bajari, Patrick, Han Hong, and Stephen P Ryan. 2010. "Identification and estimation of a discrete game of complete information." Econometrica, 78(5): 1529–1568.

Berry, Steven, and Philip Haile. 2016. "Identification in differentiated products markets." Annual review of Economics, 8: 27–52.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile prices in market equilibrium." Econometrica: Journal of the Econometric Society, 841–890.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated products demand systems from a combination of micro and macro data: The new car market." Journal of political Economy, 112(1): 68–105.

Berry, Steven T. 1992. "Estimation of a Model of Entry in the Airline Industry." Econometrica: Journal of the Econometric Society, 889–917.

Bhat, Chandra R. 1997. "An endogenous segmentation mode choice model with an application to intercity travel." Transportation science, 31(1): 34–48.

Björnerstedt, Jonas, and Frank Verboven. 2016. "Does merger simulation work? Evidence from the Swedish analgesics market." American Economic Journal: Applied Economics, 8(3): 125–64.

Bjorn, Paul A, and Quang H Vuong. 1984. "Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation."

Böhning, Dankmar, and Bruce G Lindsay. 1988. “Monotonicity of Quadratic-Approximation Algorithms.” Annals of the Institute of Statical Mathematics, 40: 641–663.

Bougette, Patrice, Kai Hüschelrath, and Kathrin Müller. 2014. “Do horizontal mergers induce entry? Evidence from the US airline industry.” Applied Economics Letters, 21(1): 31–34.

Bresnahan, Timothy F, and Peter C Reiss. 1990. “Entry in monopoly market.” The Review of Economic Studies, 57(4): 531–553.

Bresnahan, Timothy F, and Peter C Reiss. 1991. “Entry and competition in concentrated markets.” Journal of political economy, 99(5): 977–1009.

Chamberlain, G. 1980. “Analysis of Covariance with Qualitative Data.” Review of Economic Studies, 47: 225–238.

Chen, Mingli. 2019. “Estimation of Nonlinear Panel Models with Multiple Unobserved Effects.” Unpublished manuscript, University of Warwick.

Chen, Yuxin, and Sha Yang. 2007. “Estimating disaggregate models using aggregate data through augmentation of individual choice.” Journal of Marketing Research, 44(4): 613–621.

Chernozhukov, Victor, and Han Hong. 2003. “An MCMC approach to classical estimation.” Journal of Econometrics, 115(2): 293–346.

Chernozhukov, Victor, Han Hong, and Elie Tamer. 2007. “Estimation and confidence regions for parameter sets in econometric models.” Econometrica, 75(5): 1243–1284.

Ciliberto, Federico, and Elie Tamer. 2009. “Market structure and multiple equilibria in airline markets.” Econometrica, 77(6): 1791–1828.

Ciliberto, Federico, Charles Murry, and Elie T Tamer. 2018. “Market structure and competition in airline markets.” Available at SSRN 2777820.

Cohen, Andrew M, and Michael J Mazzeo. 2007. “Market structure and competition among retail depository institutions.” The Review of Economics and Statistics, 89(1): 60–74.

Cohen, Michael, Marc Rysman, and Krzysztof Wozniak. 2017. “Payment Choice with Consumer Panel Data.” Unpublished Manuscript.

Conlon, Christopher T, and Julie Holland Mortimer. 2013. “Demand estimation under incomplete product availability.” American Economic Journal: Microeconomics, 5(4): 1–30.

Davis, Peter. 2006. "Spatial competition in retail markets: movie theaters." The RAND Journal of Economics, 37(4): 964–982.

de Leeuw, Jan, and Kenneth Lange. 2009. "Sharp Quadratic Majorization in One Dimension." Computational Statistics & Data Analysis, 53: 2471–2484.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statisical Society, Series B, 39: 1–38.

Dhaene, Geert, and Koen Jochmans. 2015. "Split-panel Jackknife Estimation of Fixed-effect Models." Review of Economic Studies, 82: 991–1030.

Dutkowsky, Don, and Marc Fusaro. 2011. "What Explains Consumption in the Very Short Run? Evidence from Checking Account Data." Journal of Macroeconomics, 33: 542–552.

D'Haultfœuille, Xavier, and Alessandro Iaria. 2016. "A Convenient Method for the Estimation of the Multinomial Logit Model with Fixed Effects." Economics Letters, 141: 77–70.

Eizenberg, Alon. 2014. "Upstream innovation and product variety in the us home pc market." The Review of Economic Studies, 81(3): 1003–1045.

Ericson, Richard, and Ariel Pakes. 1995. "Markov-perfect industry dynamics: A framework for empirical work." The Review of economic studies, 62(1): 53–82.

Fan, Ying, and Chenyu Yang. 2020. "Merger, Product Repositioning and Firm Entry: The Retail Craft Beer Market in California." Available at SSRN 3681556.

Feit, Eleanor McDonnell, Pengyuan Wang, Eric T Bradlow, and Peter S Fader. 2013. "Fusing aggregate and disaggregate data with an application to multiplatform media consumption." Journal of Marketing Research, 50(3): 348–364.

Fernandez, Nerea, and Pedro L Marin. 1998. "Market power and multimarket contact: Some evidence from the Spanish hotel industry." The Journal of Industrial Economics, 46(3): 301–315.

Focarelli, Dario, and Fabio Panetta. 2003. "Are mergers beneficial to consumers? Evidence from the market for bank deposits." American Economic Review, 93(4): 1152–1172.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2005. "How did location affect adoption of the commercial Internet? Global village vs. urban leadership." Journal of urban Economics, 58(3): 389–420.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2008. "Understanding the inputs into innovation: Do cities substitute for internal firm resources?" Journal of Economics & Management Strategy, 17(2): 295–316.

Forman, Chris, Avi Goldfarb, and Shane Greenstein. 2012. "The Internet and local wages: A puzzle." American Economic Review, 102(1): 556–75.

Giroud, Xavier. 2013. "Proximity and investment: Evidence from plant-level data." The Quarterly Journal of Economics, 128(2): 861–915.

Glaeser, Edward L, and Janet E Kohlhase. 2004. "Cities, regions and the decline of transport costs." In Fifty Years of Regional Science. 197–228. Springer.

Glaeser, Edward L, and Joshua D Gottlieb. 2009. "The wealth of cities: Agglomeration economies and spatial equilibrium in the United States." Journal of economic literature, 47(4): 983–1028.

Goldstein, Harvey, William Browne, and Jon Rasbash. 2002. "Partitioning variation in multilevel models." Understanding statistics: statistical issues in psychology, education, and the social sciences, 1(4): 223–231.

Greene, William H. 2018. Econometric Analysis. . 8th ed., Prentice Hall.

Grömping, Ulrike. 2007. "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition." The American Statistician, 61: 139–147.

Gugler, Klaus, and Ralph Siebert. 2007. "Market power versus efficiency effects of mergers and research joint ventures: evidence from the semiconductor industry." The Review of Economics and Statistics, 89(4): 645–659.

Hahn, Jinyong, and Whitney Newey. 2004. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models." Econometrica, 72: 1295–1319.

Hanemann, W Michael. 1984. "Discrete/continuous models of consumer demand." Econometrica: Journal of the Econometric Society, 541–561.

Hayashi, Fumio. 2000. Econometrics. Princeton University Press.

Hinz, Julian, Alexander Hudle, and Joschka Wanner. 2019. "Separating the Wheat from the Chaff: Fast Estimation of GLMs with High-Dimensional Fixed Effects." Unpublished manuscript, European University Institute.

Ho, Kate, and Ariel Pakes. 2014a. "Hospital choices, hospital prices, and financial incentives to physicians." The American Economic Review, 104(12): 3841–3884.

Ho, Kate, and Ariel Pakes. 2014b. "Hospital Choices, Hospital Prices, and Financial Incentives to Physicians." American Economic Review, 104: 3841–3884.

Ho, Katherine. 2009. "Insurer-provider networks in the medical care market." The American Economic Review, 99(1): 393–430.

Hosken, Daniel S, Luke M Olson, and Loren K Smith. 2016. "Can entry or exit event studies inform horizontal merger analysis? Evidence from grocery retailing." Economic Inquiry, 54(1): 342–360.

Hospido, Laura. 2012. "Modelling heterogeneity and dynamics in the volatility of individual wages." Journal of Applied Econometrics, 27: 386–414.

Hunter, David R, and Kenneth Lange. 2004. "A Tutorial on MM Algorithms." The American Statistician, 58: 30–37.

Ishii, Joy. 2005. "Compatibility, competition, and investment in network industries: ATM networks in the banking industry." Unpublished working paper.

James, Jonathan. 2017. "MM Algorithm for General Mixed Multinomial Logit Models." Journal of Applied Econometrics, 32: 841–857.

Jiang, Renna, Puneet Manchanda, and Peter E Rossi. 2009. "Bayesian analysis of random coefficient logit models using aggregate data." Journal of Econometrics, 149(2): 136–148.

Jin, Wang, and Kristina McElheran. 2019. "Economies before Scale: Learning, Survival, and Performance of Young Plants in the Age of Cloud Computing." University of Toronto Working Paper.

Jonker, Nicole, and Anneka Kosse. 2009. "The Impact of Survey Design on Research Outcomes: A Case Study of Seven Pilots Measuring Cash Usage in the Netherlands." Working Paper 221/2009 Bank of Netherlands.

Kahle, David, and Hadley Wickham. 2013. "ggmap: Spatial Visualization with ggplot2." The R Journal, 5(1): 144–161.

Kalnins, Arturs. 2006. "Markets: The US lodging industry." Journal of Economic Perspectives, 20(4): 203–218.

Kalnins, Arturs, and Wilbur Chung. 2004. "Resource-seeking agglomeration: a study of market entry in the lodging industry." Strategic Management Journal, 25(7): 689–699.

Kalnins, Arturs, Luke Froeb, and Steven Tschantz. 2017. "Can mergers increase output? Evidence from the lodging industry." The RAND Journal of Economics, 48(1): 178–202.

Klee, Elizabeth. 2008. "How People Pay: Evidence from Grocery Store Data." Journal of Monetary Economics, 55: 526–541.

Kooreman, Peter. 1994. “Estimation of econometric models of some discrete games.” Journal of Applied Econometrics, 9(3): 255–268.

Koulayev, Sergei, Marc Rysman, Scott Schuh, and Joanna Stavins. 2016. “Explaining Adoption and Use of Payment Instruments by US Consumers.” RAND Journal of Economics, 47: 293–325.

Lange, Kenneth. 2016. MM Optimization Algorithms. SIAM.

Lange, Kenneth, David R Hunter, and Ilsoon Yang. 2000. “Optimization transfer using surrogate objective functions.” Journal of computational and graphical statistics, 9: 1–20.

Leisten, Matthew. 2020. “Information, Managerial Incentives, and Scale: Evidence from Hotel Pricing.”

Lewis, Gregory, and Georgios Zervas. 2016. “The welfare impact of consumer reviews: A case study of the hotel industry.” Unpublished manuscript.

Li, Sophia Ying, Joe Mazur, Yongjoon Park, James W Roberts, Andrew Sweeting, and Jun Zhang. 2018. “Endogenous and selective service choices after airline mergers.” National Bureau of Economic Research.

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

Mazzeo, Michael J. 2002a. “Competitive outcomes in product-differentiated oligopoly.” Review of Economics and Statistics, 84(4): 716–728.

Mazzeo, Michael J. 2002b. “Product choice and oligopoly market structure.” RAND Journal of Economics, 221–242.

McFadden, D., and K. Train. 2000. “Mixed MNL models for discrete response.” Journal of Applied Econometrics, 15: 447–470.

Musalem, Andres, Eric T Bradlow, and Jagmohan S Raju. 2008. “Who’s got the coupon? Estimating consumer preferences and coupon usage from aggregate information.” Journal of Marketing Research, 45(6): 715–730.

Musalem, Andrés, Marcelo Olivares, Eric T Bradlow, Christian Terwiesch, and Daniel Corsten. 2010. “Structural estimation of the effect of out-of-stocks.” Management Science, 56(7): 1180–1197.

Nevo, Aviv. 2000. “Mergers with differentiated products: The case of the ready-to-eat cereal industry.” The RAND Journal of Economics, 395–421.

Pakes, Ariel. 2010. "Alternative models for moment inequalities." Econometrica, 78(6): 1783–1822.

Pakes, Ariel, Jack Porter, Kate Ho, and Joy Ishii. 2015. "Moment inequalities and their application." Econometrica, 83(1): 315–334.

Peters, Craig. 2006. "Evaluating the performance of merger simulation: Evidence from the US airline industry." The Journal of law and economics, 49(2): 627–649.

Rushmore, Stephen, Dana Michael Ciraldo, and John Tarras. 2000. Hotel Investments Handbook 2000. West Group.

Rysman, Marc. 2007. "Empirical Analysis of Payment Card Usage." Journal of Industrial Economics, 60: 1–36.

Schuh, Scott, and Joanna Stavins. 2010. "Why Are (Some) Consumers (Finally) Writing Fewer Checks? The Role of Payment Characteristics." Journal of Banking and Finance, 34: 1745 – 1758.

Seim, Katja. 2006. "An empirical model of firm entry with endogenous product-type choices." The RAND Journal of Economics, 37(3): 619–640.

Shum, Matthew, Wei Song, and Xiaoxia Shi. 2018. "Estimating Semi-parametric Panel Multinomial Choice Models using Cyclic Monotonicity." Econometrica, 86: 737–761.

Smith, Howard. 2004. "Supermarket choice and supermarket competition in market equilibrium." The Review of Economic Studies, 71(1): 235–263.

Stamann, Amrei. 2018. "Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects." Unpublished manuscript, arXiv:1707.01815.

Stammann, Amrei, Florian Heiß, and Daniel McFadden. 2016. "Estimating Fixed Effects Logit Models with Large Panel Data." Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics, No. G01-V3.

Stango, Victor, and Jonathan Zinman. 2014. "Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Bank Overdraft Fees." Review of Financial Studies, 27: 990–1030.

Suzuki, Junichi. 2013. "Land use regulation as a barrier to entry: evidence from the Texas lodging industry." International Economic Review, 54(2): 495–523.

Sweeting, Andrew. 2009. "The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria." The RAND Journal of Economics, 40(4): 710–742.

Tamer, Elie. 2003. "Incomplete simultaneous discrete response model with multiple equilibria." The Review of Economic Studies, 70(1): 147–165.

Tanner, Martin. 1996. Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions. Springer.

Train, Kenneth. 2007. "A recursive estimator for random coefficient models." University of California, Berkeley.

Train, Kenneth E. 2008. "EM algorithms for nonparametric estimation of mixing distributions." Journal of Choice Modelling, 1(1): 40–69.

Train, Kenneth E. 2009. Discrete choice methods with simulation. Cambridge university press.

Wakamori, Naoki, and Angelika Welte. 2017. "Why Do Shoppers Use Cash? Evidence from Shopping Diary Data." Journal of Money, Credit and Banking, 2017: 115–169.

Wang, Zhu, and Alexander F. Wolman. 2016. "Payment Choice and Currency Use: Insights from Two Billion Retail Transactions." Journal of Monetary Economics, 84: 94–115.

White, Kenneth J. 1975. "Consumer Choice and Use of Bank Credit Cards: A Model and Cross-Section Results." Journal of Consumer Research, 2: 10–18.

Williamson, Oliver E. 1968. "Economies as an antitrust defense: The welfare tradeoffs." The American Economic Review, 58(1): 18–36.

Yang Xiang, Sylvain Gubian, Brian Suomela, and Julia Hoeng. 2013. "Generalized Simulated Annealing for Efficient Global Optimization: the GenSA Package for R." The R Journal Volume 5/1, June 2013.

Zheng, Zhiqiang, Peter Fader, and Balaji Padmanabhan. 2012. "From business intelligence to competitive intelligence: Inferring competitive measures using augmented site-centric data." Information Systems Research, 23(3-part-1): 698–720.

CURRICULUM VITAE

[redacted]

[redacted]