ABSTRACT

Title of dissertation: Essays on Dynamic Pricing and Choice in the Internet and Sharing Economy

Liu Ming, Doctor of Philosophy, 2017

Dissertation directed by: Professor Tunay Tunca
Department of Decision, Operations & Information Technologies

The widespread use of the Internet, social networking, mobile technology and big data has improved people's ability to obtain and use information to an unprecedented level. Influencing consumer behavior and changing concepts of consumption, the Internet and Sharing Economy has carved itself a significant and growing place in the daily life and the economy. The race to commercialize the value of this change has brought about numerous innovations and creative operational solutions in emerging industries worldwide.

The first chapter of my dissertation theoretically and empirically studies consumer equilibrium, pricing, and efficiency of these events. Modeling a continuous time customer arrival and sign-up process, we start by deriving the stochastic dynamic consumer equilibrium. Based on this equilibrium and utilizing sign-up level data from a major Chinese retailer's group buying events, we then structurally estimate consumer arrival rates and utility distributions for 266 events, and empirically

verify the fit and predictive power of the model. Utilizing the estimated arrival rates and consumer utility distributions, we then employ a doubly stochastic Generalized Linear Regression Model to provide empirical evidence for consumer network effects in group buying, and estimate 15.4% increase in consumer demand attributable to the employment of a group buying mechanism. Through counterfactual analysis, we further estimate that employing group buying increased retailer profits by 11.21% on average, corresponding to an annual monetary gain of approximately $4.32M for the 266 events in the data set. We further demonstrate that low deal discounts offered by the retailer for very low and very high consumer arrival rates boost profitability, suggesting that an inverse U-shaped deal discount pattern as a function of consumer arrival rate is recommendable when employing group buying events.

Ride-sharing platforms, such as Uber and Lyft and their Chinese counterpart Didi, set prices dynamically to balance the demand and supply for their services. In the second chapter, we provide an empirical model and analysis of price formation and surplus generation of these services. We first develop a two-sided-market discrete choice model, capturing the formation of mutually dependent demand (consumer) and supply (driver) sides that jointly determine the pricing. Based on this model, we then use a comprehensive data set obtained from Didi to estimate consumer and driver price elasticities as well as other factors that affect market participation. Based on the estimation results and counterfactual analysis, we demonstrate that surge pricing has a significant role in improving the welfare of consumers and Kuaiche

drivers, i.e., by 21.80% and 22.02%, respectively. In terms of government regulations, proposed regulation imposing price caps that match current Taxi rates can decrease consumer surplus by 39.84% while causing a relatively moderate 5.66% decrease in Kuaiche driver surplus. Further, we estimate that restricting driver capacity to equal local Taxi levels would have more severe consequences, resulting in 18.07% and 23.40% reductions in consumer and Kuaiche driver surpluses respectively.

Essays on Dynamic Pricing and Choice in the Internet and Sharing Economy

by

Liu Ming

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Tunay Tunca, Chair/Advisor
Professor John Chao
Professor Zhi-long Chen
Professor Sean Barnes
Professor Kunpeng Zhang

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Tunay Tunca for advising me on interesting and insightful topics and always encourage me to explore the deeper side. His passion and energy on research always inspired me to do better. I could not have done this without his unconditional support. It is one of the lucikest thing in my Ph.D. to have the chance to work with and learn from him.

I would like to thank my family-my parents and my wife Yue Cheng for always supporting me no matter what choices I've made. I certainly couldn't thank them more for their deepest love and encouragement. They are the most important parts in my life and words cannot express the gratitude I owe them.

I thank my best friends, Weiming Zhu and Tianshu Sun, for their support and insightful advices. It is a great pleasure to have them in my Ph.D and hope them best in their future life.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Contents

List of Figures

Chapter 1: Consumer Equilibrium, Pricing, and Efficiency in Group Buying: Theory and Evidence

1.1 Introduction

Rapid development of technology and availability of multiple consumer channels that give many alternatives to consumers have put retailers in an unprecedentedly competitive business environment in recent years. With the proliferation of commerce on the internet, many firms have adopted innovative techniques to attract and retain customers, while struggling to maintain and improve evaporating margins. One of the popular techniques employed recently is Group Buying, which has been pioneered on the Internet more than a decade ago by companies such as *Mobshop* and *Groupon*, and which, in recent years is becoming a fast growing retailing method, especially in emerging economies like China, and being employed in large online consumer commerce platforms like *Taobao*.

The idea behind group buying is simple: in the concept's most basic form, a product or service is made available for purchase during a pre-announced sign-up window, i.e., the group buying event, usually lasting less than a day. If the number of customers who sign up to purchase during the event exceeds a certain threshold,

a *deal*, meaning a discount on the unit retail price, will be available for everyone who signed up. If the number of sign ups does not exceed the pre-specified threshold, then the deal is not offered, and the product or service is available only at a higher *base price*. Even though the concept initially was mostly popular with services such as restaurant meals, activities, and home improvement contractor services, in certain markets it is also recently increasingly becoming popular for deals on goods and products such as electronics, household products, and clothing. In 2013, one of China's most popular Internet commerce platform, Taobao, which is owned by *Alibaba*, hosted more than an estimated 60,000 individual group buying events with revenues totaling more than 15 Billion Chinese Yuan (CNY) (approximately 2.3 Billion USD).

The exact group buying mechanism employed can include a number of variations, such as multiple threshold levels corresponding to multiple discount levels, and mandatory deposits customers have to pay for signing up before the deal becomes certain. For instance, Taobao employs a particular structure for all retailers for whom it hosts on its platform. Taobao's platform has two price thresholds and a mandatory sign-up deposit. The two sign-up thresholds are fixed by Taobao to be 20 and 50, and the deposit is fixed to be 99 CNY – approximately 15 USD. Taobao's platform rules also require a fixed deal price discount amount, albeit the retailer can choose the discount. To illustrate how Taobao's particular mechanism works, consider the following example: A household goods retailer announces a group buying event for one of its products on the platform, posting the date and the start and end times for the event window (usually lasting 12 hours). The retailer also announces

the base price and the discount price it chooses, which determines the first- and second-deal prices. For instance, she can set the base price to 3,000 CNY and the deal discount to 300 CNY, which means that if the total number of consumer sign ups during the event is less than 20, then the unit price will be 3,000 CNY; if it is between 20 and 49, then the unit price will be 2,700 CNY and; and if it is 50 or higher, the unit price will be 2,400 CNY. During the event, customers arrive, observe the number of sign ups, and make a decision to join. If a customer decides to sign up, he has to pay a non-refundable 99 CNY deposit. At the end of the event window the final price is announced according to the number of sign ups during the event. At that point, if a customer decides to stay, he pays the balance of the price net of his deposit. If a customer who signed up decides to leave, he forfeits his deposit and leaves. Finally, the retailer delivers the product to all remaining customers.

Considering the price discounts given by the retailers to all customers who make a final purchase, in order for a group buying mechanism to be successful, it has to attract increased consumer demand. In a group buying event, there are two components for increased purchases: First, with the hope of lower prices when a deal happens, some customers who would not purchase at the base price may now become interested in purchasing the good. This type of increase in sign ups can happen during the event in a dynamic manner before or after a deal threshold is reached. Therefore, customer incentives based on their valuations of the product and their expectations on the sign-up behavior of other customers are critical in determining the final number of units sold, and the resulting equilibrium sign-up behavior should

be studied rigorously to determine the effects of pricing and discounts on purchases. The second type of increase in purchases comes from network effects, namely dissemination of information about the group buying event by customers themselves to attract other customers to sign up, which can ultimately increase the likelihood of a deal happening (cf. Jing and Xie 2011, Chen and Lu 2015). In particular, when there is a group buying event, customers who are aware of the event have incentives to spread the word and inform others to consider signing up, since that increases the probability of reaching a threshold for a deal and lowering the price, in essence, acting like unpaid "sales agents" for the retailer. This behavior has an effect of increasing the demand, i.e., consumer arrival rate during the event, and can boost the number of purchases. An immediate question that arises here is whether one can quantify and test the magnitude of this effect.

Given these dynamics that group buying events introduce and the growing popularity of the concept, a number of interesting research questions emerge. Specifically, what are the equilibrium dynamics of the consumer sign-up process during the group-buying event? How do the consumers' likelihood of signing up and the probability of the deals materializing evolve during the event? How much do the consumers' networking and information dissemination improve demand in group buying events compared to traditional single-price sales? Taking into account the price discounts offered in group buying and the demand increases they induce, do retailers have direct net profit gains from group buying, and if so how much? Finally, what are some empirically verifiable suggestions for patterns of deal discounts in order to improve profitability of group-buying events?

In this paper, in order to address these questions, we theoretically and empirically study group buying utilizing data from events hosted by Taobao. The roots of the answers to our research questions depend on the consumer economic behavior. Therefore, we first build a continuous time model of consumer sign ups in Taobao's group buying events, and study the dynamic stochastic consumer sign-up strategies. We derive the equilibrium of this dynamic game as a recursive differential equation system, and study its properties. Utilizing data from Taobao's platform, we subsequently structurally estimate parameters in our model, determining the consumer arrival rate and the reservation utility distribution for each event. We then verify the fit of our model and utilize our structural estimation to empirically demonstrate and measure consumer network effects and profit gains from group buying, as well as profitable patterns of setting deal discounts.

1.2 Literature Review

Pricing of retail goods and services has long been studied in the literature (cf. Schmalensee and Willig 1989, Wilson 1993, Talluri and Van Ryzin 2006). Studies on pricing a retail product through a group buying mechanism, however, are relatively new, and many aspects of when and how group buying and related consumer discount based mechanisms are profitable for a retailer are still in the process of being disentangled by an increasing number of theoretical and empirical studies. One stream of theoretical studies with early roots employ batch consumer sign ups in analyzing the effectiveness of group buying. Anand and Aron (2003) de-

rive a monopolist's optimal group-buying schedule under different kinds of demand uncertainty and study the impact of production postponement on a group buying strategy. Their results show that the effectiveness of group buying mechanism over traditional single-pricing relies on the nature of uncertainty on the demand curve. Hu et al. (2013) consider the case where consumers make decisions simultaneously in group buying as a batch and show that a sequential sign-up mechanism leads to higher deal success rates. Marinesi et al. (2016) also consider batch customer sign-ups. They show that employing group buying allows firms better utilize their capacity. They further demonstrate that the presence of strategic customers can be advantageous when employing group buying, and the mechanism can help generate significant profit gains. Differing from and complementing this stream of literature, we model a continuous time, stochastic consumer arrival and sign-up process, and find the dynamic consumer equilibrium. Our approach allows us to study the evolution of customer sign-up patterns throughout the event window, and enables detailed structural estimation of the parameters with our group buying data that includes sign-up times for all consumers in each event.

A number of papers study the profitability of group buying. Chen et al. (2002, 2007) explore optimal bidding for risk neutral and symmetric buyers in group-buying events with fixed numbers of goods and customers, and compare the profitability of group-buying events with traditional single-pricing. They find that group buying can outperform the fixed price mechanism only under economies of scale or risk-seeking sellers. Chen and Zhang (2014) find conditions, under which group buying can maximize profits, and show that the profitability of the mechanism depends

on the nature of uncertainty in the market. Chien-Wei and Hsien-Hung (2016) argue that when customers are heterogenous in group buying costs, employing group buying may be preferable to non-discriminated pricing. Deviating from most of the literature that focuses on one retailer, Chen and Roma (2011) study group buying in a two-level distribution channel with one manufacturer and two competing retailers. They show that group buying is beneficial for the smaller of the two retailers but can hurt the larger one, while increasing supplier revenues. In our paper, we empirically demonstrate that combining careful pricing and consumer network effects, a retailer can increase her profits significantly with group buying events compared to single-pricing.

Another stream of literature studies the effects of threshold pricing structure on consumer sign ups in group buying mechanisms. Kauffman and Wang (2001) study group-buying data from one of the earliest group buying companies, Mobshop.com, and find that number of existing sign ups and approaching sign-up thresholds both have positive effects on new orders placed. Subramanian (2012) presents a model where low valuation consumers can wait and benefit from observing previous purchases and finds that profitability of a group buying mechanism can decrease by sending reminders to customers to check the deal's progress. Liang et al. (2014) also study the role of providing information about the number of participating customers, and find that revealing this information can have a positive effect on consumer surplus and the success rate of the mechanism. Wu et al. (2014) explore the possibility of increased consumer sign-up rate effects before and after the thresholds are reached, and find empirical evidence for the first type of effect in all prod-

ucts, while showing that the second one exists only for some products. Our study contributes to this branch by empirically demonstrating that a consumer dynamic strategic behavior can explain the sign-up patterns in group buying events better compared to consumer strategies such as waiting and signing up after thresholds are crossed.

The genesis of group buying is retailers' providing discounts and deals, and given the increased application of discount variants in recent years, a number of studies explore the profitability of such strategies empirically. Wu et al. (2015) analyze daily deals provided by Chinese retailers and find that merchants in fact experience losses from discounts during promotion periods, but they make profits through increased future purchases. Cao et al. (2015) quantify the impact of discount percentage on sales, and conclude that a larger discount percentage may reduce sales by being perceived as a signal of low product quality. Edelman et al. (2016) find that online discount vouchers tend to be more profitable for relatively unknown firms while being not likely to increase profits for better-known ones.

Overall, existing theoretical and empirical studies in the literature paint a mixed picture on the profitability of group buying, and its desirability for sellers. One dimension we aim to contribute to the debate on this front is the consumer network effects and incentives to recruit other customers group buying creates. Jing and Xie (2011) present a theoretical model to study the effect of consumer social interactions, i.e., using a discounted price to motivate consumers to work as "sales agents" to acquire other consumers. They argue that the demand increase brought about by such social interactions can make offering discounts in group buying events

profitable, and efficient interpersonal communication makes the mechanism more profitable to firms. Zhou et al. (2013) empirically study the information diffusion process in group buying, and find that mass media communication and interpersonal communication stimulate the sales at the start of the process while reducing the sales at the end. Zhang and Gu (2015) argue that social interaction affects consumers' purchasing behavior through informational and normative influences on their trust. Chen and Lu (2015) find that social factors including online, media and personal recommendations positively affect consumers' group buying intentions and social influence. On the other side of the argument, Gwee and Chang (2013) claim that purchases at group buying websites are usually impulsive rather than planned, and advocate that firms employing such mechanisms should design their processes to help consumers develop loyalty. Zhang and Tsai (2015) find that consumers' intention to join group buying deals are directly driven by the need for uniqueness and perceived homophily. Hu and Winer (2016) utilize data from Groupon finding that the existence of the deal threshold does not necessarily stimulate customers to inform others, but information about tipping points accelerates customer sign ups.

Our paper contributes to this debate by empirically testing the retailer pricing strategy and the sources of profitability in group buying events. We separate and quantify the strategic sign-up effects and the consumer network and information dissemination effects. We provide evidence that, in fact, firms can make direct profits in group buying events despite giving significant discounts to consumers in many cases, because of the incentives group buying events create for customers to spend effort in networking and recruiting others. In sum, our study not only provides

a theoretical explanation for dynamic consumer behavior in group buying events, but also connects it to novel empirical evidence on the sources of profitability for retailers from these events.

1.3 Theory

Many online retailers and platforms are developing and employing different pricing mechanisms to implement group buying. The one we will study in our paper, is the one employed by Taobao, the largest online retailing platform in China. Our data comes from the events hosted on this platform. In this mechanism, there are two possible deals, and including the base (no deal) unit price, there are three possible prices that can materialize, and a required non-refundable deposit for signing up. Given the platform's deal thresholds and the deposit amount, the retailer chooses the base unit price and a deal discount on the unit price that applies to both deals.

We first present our framework of this model to derive the consumer equilibrium in this setting theoretically, and later apply our findings in structural estimation and counterfactual efficiency analysis.

1.3.1 Model Description

Consider a retailer selling a product to consumers on a third-party platform. The retailer will hold a group buying event on a continuous time window indexed $[0, T]$. During this period, customers arrive following a Poisson process. (For ease of exposition, we will refer to the retailer as "she", and each customer as "he" throughout

the paper.) Each customer has a unit demand for the product, and his reservation value u has a c.d.f. denoted by F with a p.d.f. f. Upon arrival, knowing the pricing pattern announced by the buyer as described below, and the number of customers who have already joined up to that point, each customer decides to join or leave. If the customer joins, he pays a non-refundable deposit $d > 0$, which, following Taobao's process, is set fixed by the platform that hosts the event. Denote the total number of customers who joined by the end of the event window by N.

The pricing of product is in the form of two-threshold group-buying discounts. Specifically, given thresholds M_1 and M_2, $0 < M_1 < M_2$, fixed by the platform, the retailer announces three prices: the base price, p_0, and the first and second deal prices, p_1 and p_2 respectively, where $p_0 \geq p_1 \geq p_2 \geq d$. If the total number of customers who joined, N is less than M_1, no group deal materializes, and the unit price of the product, p, is set to the base price p_0. If $M_1 \leq N < M_2$, then the first deal will be on and the unit price will be set at $p = p_1$. Finally, if $N \geq M_2$, then the second deal is on, and the unit price will be set at $p = p_2$. Following Taobao's process, the price decrement (the deal discount) is constant and set at $\delta \geq 0$. That is $p_0 = p_1 + \delta$ and $p_1 = p_2 + \delta$. At the end of the event window, i.e., at $t = T$, after observing the total number of customers who joined, each customer makes a decision to stay in the deal or drop out. If a customer stays, he pays the balance of the price, $p - d$, and commits to buying the product. If he drops out, however, he forfeits the deposit d. Denote the number of customers that still stay after the drop outs by Q. Finally, the retailer produces Q units to be delivered to the staying customers at unit cost c.

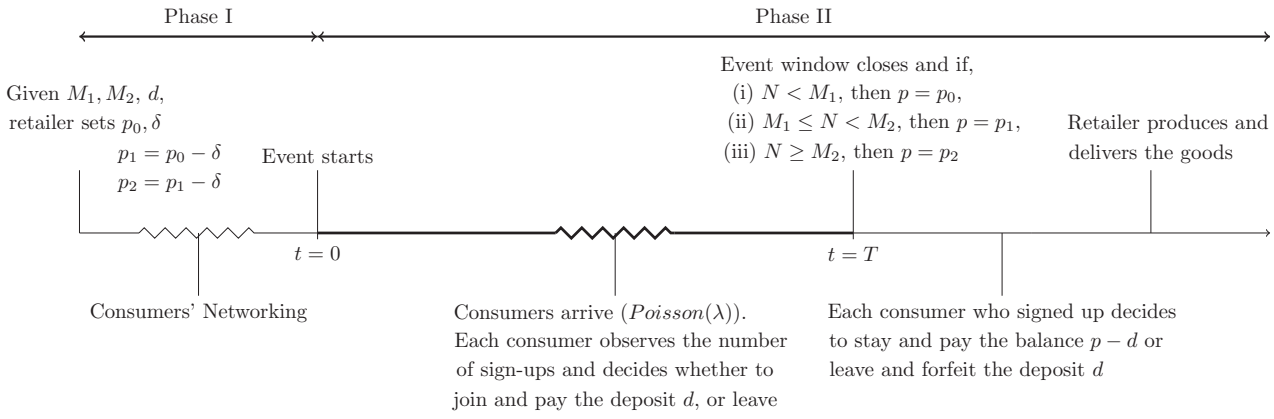

Fig. 1.1: Timeline of Group Buying activity on *Taobao.com*

The event-window itself usually has a length of 12 hours. However, the retailer announces the event and posts the prices and the deal discounts about one-week in advance. During this time before the event, customers may network and recruit others in order to increase participation and hence increase the probability of one of the deals happening. We call this period *Phase I* and the event-window *Phase II*. Let λ_o denote the arrival rate of customers during the event if the product were sold through traditional single-pricing. Due to customer efforts in recruiting other, the arrival rate of customers during the event to increases to $\lambda_g > \lambda_o$. Figure 1.1 summarizes the timeline.

In order to empirically measure demand increase from the network effects, we will need to estimate the consumer arrival rates, λ_g, for the group buying events from the sign up data. To be able to do that, we will first have to study the dynamic consumer equilibrium behavior in Phase II, i.e., during the Group Buying Event. In the rest of this section, we will solve for this equilibrium.

1.3.2 Consumer Equilibrium during the Group Buying Event

We start with the customers' decision after the event. Consider a customer, who arrived and signed up at time $t \in [0, T]$, with utility u. After the event, when the price p is determined, he needs to decide whether to stay or drop out. If the customer drops out he forfeits his deposit, and his overall payoff will be $-d$, while if he stays, his payoff will be $u - p$. The customer will choose the larger at that point and his surplus will be $\max\{u - p, -d\}$.

Given this post-event behavior, each consumer that arrives at time $t \in [0, T]$ observes the price structure, the two deal thresholds, M_1 and M_2, and the total number of sign ups up to that point, (i.e., on $[0, t)$), N_t. Projecting the shaping up of the rest of the event and his decision at the end of the event contingent on the realization of the deals, he makes a decision on whether to sign up or not. Define N_t^+ as the number of sign ups on $[0, t]$, i.e., including the sign-up decision of the customer who arrives at t. That is, if the customer that arrives at time t decides to join, then $N_t^+ = N_t + 1$, otherwise $N_t^+ = N_t$. For any $t \in [0, T]$, define $\pi_k^1(t)$ as the time t probability that *only* the first deal happens given that $N_t^+ = k$, i.e., $\pi_k^1(t) = Pr\{M_1 \leq N < M_2 | N_t^+ = k\}$. Similarly define $\pi_k^2(t)$ as the time t probability that the second deal happens given that $N_t^+ = k$, i.e., $\pi_k^2(t) = Pr\{N \geq M_2 | N_t^+ = k\}$. Note that the following boundary conditions hold:

(i) $\pi_k^1(T) = 0$ for $0 \leq k < M_1$, and $\pi_k^2(T) = 0$ for $0 \leq k < M_2$.

(ii) $\pi_k^1(t) = 1 - \pi_k^2(t)$, for $M_1 \leq k < M_2$.

(iii) $\pi_k^1(t) = 0$, $\pi_k^2(t) = 1$ for $k \geq M_2$.

Finally, for $k \geq 0$, define $H_k(t)$ as the probability that a consumer arriving at time t signing up, given that $N_t = k$. Notice that for all $t \in [0, T]$, and $k \geq 0$, $H_k(t) = 1$ if $u_t \geq p_0$, $H_k(t) = 0$ if $u_t < p_2$. Further, for $k \geq M_2 - 1$ the boundary condition $H_k(t) = 1 - F(p_2)$ holds.

Now, consider the decision of a customer with valuation u who arrives at time t. For $N_t = k \geq 0$ denote his expected utility of signing up by $V_k(u, t)$. Then

$$V_k(u,t) = \pi_{k+1}^1(t)\max\{u-p_1,-d\}+\pi_{k+1}^2(t)\max\{u-p_2,-d\}+\max\{u-p_0,-d\}(1-\pi_{k+1}^1(t)-\pi_{k+1}^2(t)), \tag{1.1}$$

and he would choose to sign up if and only if $V_k(u, t) \geq 0$. This implies

$$H_k(t) = Pr\{V_k(u,t) \geq 0\}, \tag{1.2}$$

where the consumer's utility of not joining is normalized to 0. Utilizing (1.1), we can derive the characterization of a consumer's sign-up decision based on his arrival time, t, and the number of sign ups up to that point N_t. The following lemma states the structure of this decision.

Lemma 1: For each $t \in [0, T]$, given $N_t = k \geq 0$, there exists a threshold $\bar{u}_{k,t} \in [p_2, p_0]$ such that a customer who arrives at time t with reservation utility u signs

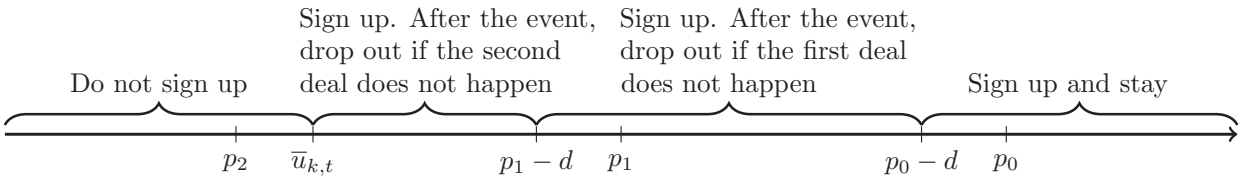

Fig. 1.2: Consumer utility ranges for sign-up and staying decisions at time $t \in [0, T]$ and $k \geq 0$ existing sign ups, when $p_2 \leq \bar{u}_{k,t} < p_1$.

up if and only if $u \geq \bar{u}_{k,t}$. $\bar{u}_{k,t}$ is characterized as

$$\bar{u}_{k,t} = \begin{cases} p_0 - 2\delta - d\left(1 - \dfrac{1}{\pi^2_{k+1}(t)}\right) & \text{if } 0 \leq d \leq \delta\pi^2_{k+1}(t), \\ p_0 - d + \dfrac{d - \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))}{\pi^1_{k+1}(t) + \pi^2_{k+1}(t)} & \text{if } \delta\pi^2_{k+1}(t) < d \leq \delta\pi^1_{k+1}(t) + 2\delta\pi^2_{k+1}(t), \\ p_0 - \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t)) & \text{if } d > \delta\pi^1_{k+1}(t) + 2\delta\pi^2_{k+1}(t). \end{cases} \tag{1.3}$$

Figure 1.2 shows the structure of consumer sign-up behavior on the utility axis (u) for a customer who arrives at time t for the case $p_2 \leq \bar{u}_{k,t} < p_1 - d$. As stated in Lemma 1, the customer will sign up if and only if $u \geq \bar{u}_{k,t}$. However, as we also discussed above, after the event, a customer who signs up will drop out if $u < p - d$. Therefore, as can also be seen in the figure, customers with utility values lower than $p_1 - d$ who signed up will drop out after the event if the first deal does not happen (i.e., $N < M_1$), and customers with utility values lower than $p_2 - d$ who signed up will drop out after the event if the second deal does not happen (i.e., $N < M_2$). The details of the consumer behavior when $\bar{u}_{t,k}$ falls into other intervals follow with similar logic.

Based on Lemma 1, we can now derive the consumer equilibrium by solving

for the continuous time equilibrium evolution of the consumers' sign up probability, $H_k(t)$, and the probabilities of the two deals happening, $\pi_k^1(t)$ and $\pi_k^2(t)$, respectively. Starting with π_k^2, given that at time $t \in [0, T]$, $N_t = k$, in order to calculate the time t probability of second deal happening, we can condition on the arrival time of the next customer. For $x \in (0, T - t]$, suppose that the next customer arrives at time $t + x$. Then since the customer arrival process is Poisson with rate λ_g, the distribution of x is the interarrival distribution for this process, i.e., Exponential with rate λ_g and p.d.f. $\lambda_g e^{-\lambda_g x}$. At the time of his arrival, $t + x$, the next customer observes that $N_{t+x} = k$, and decides to join with probability $H_k(t + x)$. If he joins, then we have $N_{t+x}^+ = k + 1$, and otherwise $N_{t+x}^+ = k$. Consequently, the time $t + x$ probability of the second deal happening is $\pi_{k+1}^2(t + x)$ if he joins, and $\pi_k^2(t + x)$ otherwise. We can then write the a dynamic recursive equation for the probability that the second deal will happen, $\pi_k^2(t)$ by taking the conditional expectation on the arrival and the decision of the next customer as

$$\pi_k^2(t) = \int_0^{T-t} \left(H_k(t + x)\pi_{k+1}^2(t + x) + (1 - H_k(t + x))\pi_k^2(t + x) \right) \lambda_g e^{-\lambda_g x} dx \,, \quad (1.4)$$

with boundary conditions $\pi_k^2(t) = 1$ for $k \geq M_2$, and $H_k^2(t) = 1 - F(p_2)$ for $k \geq M_2 - 1$. In a similar manner, for π_k^1, we can obtain

$$\pi_k^1(t) = \int_0^{T-t} \left(H_k(t + x)\pi_{k+1}^1(t + x) + (1 - H_k(t + x))\pi_k^1(t + x) \right) \lambda_g e^{-\lambda_g x} dx \,, \quad (1.5)$$

with boundary conditions $\pi_k^1(t) = 0$ for $k \geq M_2$, and $\pi_k^1(t) = 1 - \pi_k^2(t)$ for $M_1 \leq$

$k < M_2$. Solving the differential equation system that arises from (1.4) and (1.5) on $t \in [0, T]$ and $k \geq 0$ recursively with the above boundary conditions, we can obtain the consumer equilibrium characterized by (π_k^1, π_k^2, H_k), $k \geq 0$. The following proposition states the result.

Proposition 1: For any $\lambda_g, d > 0$, and $0 < \delta < p_0/2$, there exists a unique dynamic consumer equilibrium. In equilibrium, a customer with utility u arriving at time t with $N_t = k$ signs up if and only if $u \geq \bar{u}_{k,t}$, where $\bar{u}_{k,t}$ is as defined in (1.3). The equilibrium is characterized by the solution to the dynamic recursive equation system

$$\pi_k^i(t) = \lambda_g \int_t^T e^{-\lambda_g \int_t^s H_{k+1}(v)dv} H_{k+1}(s)\pi_{k+1}^i(s)ds, \quad \text{for} \quad 0 \leq k < M_i,\ i = 1, 2, \tag{1.6}$$

with boundary conditions $\pi_k^1(t) = 0$, $\pi_k^2(t) = 1$, for $k \geq M_2$, $\pi_k^1(t) = 1 - \pi_k^2(t)$ for $M_1 \leq k < M_2$, and where

$$H_k(t) = 1 - F(\bar{u}_{k,t}), \text{ for } k \geq 0. \tag{1.7}$$

Figure 1.3 demonstrates the consumer equilibrium outcome as a function of time t for deal thresholds $M_1 = 20$ and $M_2 = 50$. As can be seen in Panel (b), for any given time point t, the probability of the second deal happening $\pi_k^2(t)$ is higher as the number of sign ups on $[0, t]$, i.e., k, becomes higher. Further, for any fixed number of sign ups, as the time progresses, the second deal probability decreases since less time is left for the remaining $M_2 - k$ customers to sign up for the second

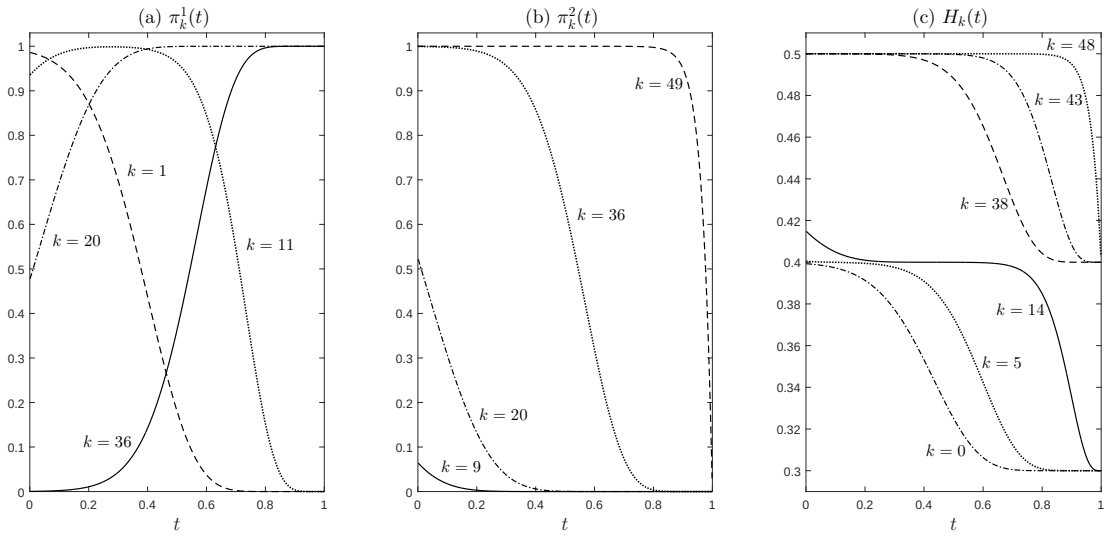

Fig. 1.3: Equilibrium π_k^1, π_k^2 and H_k functions for varying number of existing sign ups. Panels (a) and (b) illustrate the probability of only the first deal materializing, and the probability of only the second deal materializing, $\pi_k^2(t)$, respectively; and panel (c) shows the consumer sign-up probability $H_k(t)$, for selected values of the existing number of consumer sign ups (k). For all panels, $T = 1$, consumer utility distribution is Uniform on $[0, 1]$, $\lambda_g = 30$, $M_1 = 20$, $M_2 = 50$, $p_0 = 0.7$, $p_1 = 0.6$, and $p_2 = 0.5$, and $d = 0.02$.

deal to happen. The patterns for the probability that only the first deal happens, π_k^1, are more subtle as can be seen in Panel (a) of Figure 1.3. First, for a small number of existing sign ups, e.g., $k < 10$, for any given t, as the number of sign ups increases π_k^1 increases, since with a higher number of customers already in, it becomes more likely for the first deal to happen, and π_k^1 is monotonically decreasing in t. However, as k approaches the first deal threshold, 20, additional sign ups do not necessarily increase π_k^1, because they increase the probability of the second deal materializing, and hence reduce the probability that *only* the first deal materializes. Similarly, as

can be seen for $k = 11$, π_k^1 is also no longer monotonic for intermediate k values, since earlier in the time window, as the time progresses without any additional sign ups, the probability of the second deal happening decreases and the overall probability of ending up with only the first deal increases. Finally, when $k \geq 20$, the number of sign ups already exceeds the first deal threshold, so any passing time without new arrivals will reduce the probability that the second deal materializes, and consequently $\pi_k^1(t)$ is monotonically increasing in t.

Panel (c) of Figure 1.3 shows the time evolution of equilibrium consumer sign-up probability H_k for various existing sign-up levels, k. As can be seen from the figure, $H_k(t)$ curves are clustered in two groups, namely for $k < M_1 - 1 = 19$ and $k \geq 19$. For a given $k < 19$, $H_k(t)$ is decreasing in t since having the same number of sign ups k at a larger t means that the probability of any of the two deals materializing is lower, and hence, signing up is less attractive for consumers, i.e., $H_k(t)$ decreases with t. For $k < 20$, as t approaches 1, the probability that any of the deals will materialize vanishes, and almost no customers other than those with utilities greater than the base price, i.e., $u > p_0$, sign up. Hence, in the figure, the customer sign-up probability converges to $1 - F(p_0) = 1 - 0.7 = 0.3$. For $k \geq 19$, each arriving customer knows that the first deal has already materialized or will materialize if he chooses to sign up. Therefore, for $k \geq 19$, for any $t \in [0, 1]$, any arriving customer with $u \geq p_1$ will sign up, and hence $H_k(t) \geq 1 - F(p_1) = 1 - 0.6 = 0.4$. For this larger k value cluster, $H_k(t)$ is also decreasing in t but converges to 0.4 as t approaches 1.

Based on the theoretical analysis we have developed thus far, we will next

empirically explore the effects of group buying on consumer sign-up behavior, and retailer pricing and profits.

1.4 Empirical Analysis

1.4.1 Group Buying Data Description and Empirical Strategy

Our data comes from Group Buying and Single-price events for a major Chinese appliance retailer hosted by Taobao in 2013. The data includes 266 group buying events held on November 11 (proclaimed as "Singles' Day" in China), and 2715 cases of products sold through traditional single-pricing on November 11 and December 24-28 on Taobao's retailing website *Tmall.com*. In this section we will focus on the data description for the group buying events and give further details for the data set, including information for the single-price events and additional product information in the data in Section 1.4.3.1.

In each of the 266 group buying events, a unique product is sold through group buying via the mechanism we had described above. For all events, the first and second sign-up thresholds are $M_1 = 20$ and $M_2 = 50$ respectively, and the required customer deposit for signing up during the event is $d = 99$ CNY. The data for each event includes the product identifier, the three prices, p_0, p_1, and p_2, time length of each event (11 or 12 hours), time for each sign up during the event (hour, minute, and second), and the number of sign ups who stay after the event. In total, we have 41,496 sign-up data observations. In 217 of the 266 events in the data set, the second deal threshold is reached (i.e., $N \geq 50$), in 42 events, the first

deal threshold is reached but not the second ($20 \leq N < 50$), and in the remaining 7 events no deal threshold is reached ($N < 20$). The products sold in the events belong to six major categories: Refrigerators (44 events), Air Conditioners (39), Television Sets (63), Water Heaters (32), Gas Stoves (27), and Washing Machines (61).

The outline of our strategy for empirical analysis is as follows: Utilizing our theoretical analysis and the continuous time consumer equilibrium expressions derived in Section 1.3, and given the price levels at each event and the sign-up data, we first perform a structural maximum likelihood estimation for each group buying event, and jointly estimate (i) the consumer arrival rate (λ_g), and (ii) the consumer utility distribution function (F) and its parameters. Based on this estimation methodology, we then test the model fit and assumptions, and check for the predictive power of our model. Next, utilizing our estimates for group buying events and utilizing the extended data set that includes observations on traditional single-price sales, we estimate the increase in demand brought about by group buying due to network effects, and based on this estimate, perform a counterfactual analysis to determine the profit gains from employing group buying over a single-price strategy. Finally, using the estimation results, we empirically demonstrate recommendable pricing patterns that help improve retailer profits.

1.4.2 Estimation

1.4.2.1 Estimation of Model Parameters

We start with the estimation of model parameters. For each event and at any given time point $t \in [0, T]$, given the base price and the discount chosen by the retailer (p_0, δ), and the number of arrivals, $k \geq 0$, up to that point, the instantaneous sign-up rate for the next customer is

$$\lambda_{k+1}(t) = \lambda_g H_k(t), \tag{1.8}$$

where $H_k(t)$ is as defined in Proposition 1, and since the consumer arrival process is Poisson, conditional on the number of existing sign ups k, the sign-up rate for the $k+1^{st}$ customer is independent of the history of the process on $[0, t)$. Hence for each t and k, the next sign-up follows a process that is distributionally equivalent to the first sign-up of a non-homogenous Poisson process with instantaneous arrival rate as given in (1.8). Therefore, at time t, with k existing sign ups, the appearance time for the $k+1^{st}$ sign-up is exponentially distributed with density

$$\varphi_k(t, s|H) \triangleq \int_t^s \lambda_k(\tau)d\tau \cdot e^{-\int_t^s \lambda_k(\tau)d\tau} = \int_t^s \lambda_g H_k(\tau)d\tau \cdot e^{-\int_t^s \lambda_g H_k(\tau)d\tau}. \tag{1.9}$$

In order to estimate the consumer utility distribution, we will determine the best parameter fit for a variety of family of distributions, namely Uniform, Normal and Log Normal, Beta, Gamma, Weibull and Gumbel. Let $\boldsymbol{\xi}$ be the parameter vector

for the consumer utility distribution for a given type of distribution. For instance, for Normal distribution $\boldsymbol{\xi}$ will be (μ, σ), i.e., the mean and standard deviation of the distribution. Thus, for a given distribution type, we will find the best fitting parameter vector $\boldsymbol{\theta} = (\lambda, \boldsymbol{\xi})$ through Maximum Likelihood Estimation. For a given event, let $(t_1, t_2, ..., t_N)$ be the sign-up times observed in the data. In order to perform the estimation, for any candidate parameter vector $\boldsymbol{\theta}$, we first calculate the consumer equilibrium outcome (π^1, π^2, H) utilizing Proposition 1. Denote the equilibrium consumer sign-up probability function sequence $H : \mathbb{N} \times [0, T] \to [0, 1]$, derived for the parameter vector $\boldsymbol{\theta}$ as $H^{\boldsymbol{\theta}}$. For instance, again for Normally distributed consumer utility, we have $\boldsymbol{\theta} = (\lambda, \mu, \sigma)$ and for $k \geq 0$ and $t \in [0, T]$,

$$H_k^{\boldsymbol{\theta}}(t) = Pr\{u \geq \bar{u}_{k,t}\} = \frac{1}{2}[1 - \text{erf}(\frac{\bar{u}_{k,t} - \mu}{\sigma\sqrt{2}})], \tag{1.10}$$

where $\bar{u}_{k,t}$ is as given in Proposition 1 for the parameter vector $\boldsymbol{\theta}$. Then we can write the likelihood function as

$$\mathcal{L}(\boldsymbol{\theta}; t_1, t_2, ..., t_N) = \prod_{k=1}^{N} \varphi_{k-1}(t_{k-1}, t_k | H^{\boldsymbol{\theta}}), \tag{1.11}$$

where $t_0 = 0$. The maximum likelihood estimate for the given event and distribution type then is

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{argmax}}\ \mathcal{L}(\boldsymbol{\theta}; t_1, t_2, ..., t_N) = \underset{\boldsymbol{\theta}=(\lambda_g, \boldsymbol{\xi})}{\text{argmax}} \prod_{k=1}^{N} \left\{ \int_{t_{k-1}}^{t_k} \lambda_g H_{k-1}^{\boldsymbol{\theta}}(\tau) d\tau \cdot e^{-\int_{t_{k-1}}^{t_k} \lambda_g H_{k-1}^{\boldsymbol{\theta}}(\tau) d\tau} \right\}. \tag{1.12}$$

Tab. 1.1: Estimation Results

Arrival Rate (λ_g)	**No. of Events**	**Min**	**Average**	**Max**	**Standard Deviation**		
	266	22.33	201.04	411.28	96.08		
		Mean			**Standard Deviation**		
Utility Distribution	**No. of Events**	**Min**	**Average**	**Max**	**Min**	**Average**	**Max**
Beta	169	1002.42	3616.79	6511.37	165.16	367.22	485.03
Log-Normal	49	892.35	3392.41	7231.87	128.91	343.86	544.22
Normal	48	1311.9	3587.62	5835.86	201.49	416.13	587.58
McFadden Pseudo R^2:		Min	Average	Max			
		0.35	0.64	0.89			

For a given distribution type, at each iteration of the estimation, for the corresponding parameter vector $\boldsymbol{\theta}$, the equilibrium sign-up probability function sequence $H^{\boldsymbol{\theta}}$ is calculated by solving the dynamic equation system (1.6)-(1.7), and the objective (1.12) is calculated. Iterations continue until the optimal parameter vector $\boldsymbol{\theta} = (\lambda_g, \boldsymbol{\xi})$ is obtained. Repeating the process for each of the distribution types we listed above, and choosing the one that yields the highest likelihood score, we can determine the best fitting distribution with its parameters and the corresponding best estimate for the arrival rate for each event.

Table 1.1 presents the results of the estimation. Category-based breakdown of these estimation results are given in Appendix 2.8. As can be seen from Table 1.1, among the 266 events in our data set, for approximately two thirds of them (169) Beta distribution is the best fit for the consumer utility distribution. The

rest of the events are split approximately evenly between Log-Normal and Normal as the best fit. The distributional estimates indicate that average unit reservation value for customers is about 3,500 CNY (approximately 500 USD) with a standard deviation at nearly one tenth of the mean. We can also observe from the table the estimated arrival rate of customers vary significantly across events ranging from 22.3 to 411.3. Similarly, the estimated means of the consumer utility distributions have significant variability, ranging from about 1,000 to approximately 7,000 CNY. However, standard deviations for estimated consumer utility distributions vary less relative to their means.

1.4.2.2 Model Fit and Predictive Power

In this section, we provide evidence for the close fit and predictive power of our model, demonstrating that it provides a very good approximation for the consumer equilibrium behavior that emerges in group buying events.

Testing the Model Fit and Distributional Assumptions

Table 1.1 states that the average McFadden Pseudo-R^2 value of the model parameter estimation for the 266 events is 64%, with a minimum of 35%, and can be as high as 89%. Further, 224 out of 266 events (84.21%) have Pseudo-R^2 values greater than 50%, indicating that the model fit to the data, in general, is very good.

We further test the fit of the Poission inter-arrival time distributional assumption of the model. For this, we need to take into account the modulation of the arrival process with equilibrium consumer sign-up process. Recall from the discus-

sion in Section 1.4.2 that conditional on the previous consumer sign-up, each sign-up has a distribution, equivalent to the first sign-up of a non-homogenous Poisson process with instantaneous arrival rate, given in (1.8), and hence, the interarrival times for sign-ups are exponential with density given in 1.9. Therefore, given the sign-up time vector $(t_1, t_2, ..., t_N)$ for event k, $1 \leq k \leq N$,

$$\frac{t_k - t_{k-1}}{\int_{t_{k-1}}^{t_k} \lambda_g H_{k-1}(\tau) d\tau \cdot e^{-\int_{t_{k-1}}^{t_k} \lambda_g H_{k-1}(\tau) d\tau}} \tag{1.13}$$

has an exponential distribution with mean 1, where $t_0 = 0$, and $H_k(t)$, $k = 0, \ldots, N$ is as defined in Proposition 1. For a given event, we can then test the distributional fitting of the observed data with the Poisson arrival rates implied by the consumer sign-up process from the model, by applying a Kolmogorov-Smirnov distributional fitting test to the statistic vector given in (1.13) for each event. Running this test for each of the 266 group buying events in the data set, we find that 257 of the events (96.6%) pass the test, implying that Poisson inter-arrival time distributional assumption of the model is supported.

Testing the Predictive Power of the Model

Next, we will test how good our model is in predicting the evolution of the consumers' reaction to threshold crossing. In particular, one can argue that some customers may wait for the price reduction thresholds to be crossed and lower prices to be guaranteed before signing up and paying the deposit. If that is the case, we should see a significant increase in sign ups after the second threshold is crossed compared to

the number that would be predicted by the pre-threshold crossing sign-up observations. On the other hand, if our model is a good approximation for the consumer sign up behavior, the sign-up behavior before the second threshold is crossed should have strong predictive power on the number of consumer that will sign up after this threshold is crossed.

In order to test this, first let us define $\tau_1 = \inf\{t | N_t^+ \geq M_1 - 1,\, 0 \leq t \leq T\}$, and $\tau_2 = \inf\{t | N_t^+ \geq M_2 - 1,\, 0 \leq t \leq T\}$. That is, τ_1 and τ_2 are the time points on the event window $[0, T]$, when the first and second deal sign-up thresholds are crossed with one more sign up, provided that the corresponding thresholds are crossed during the event window. For the purposes of our tests, we will focus on the 217 out of the 266 events that are in our data set, where there were sufficient sign ups that the second deal materialized, i.e., where $N \geq M_2$.[1] In each of these events, according to our model, the arrival process for $t > \tau_2$ would be a Poisson process with arrival rate $\lambda_g(1 - F(p_2))$, and hence the expected number of arrivals on $(\tau_2, T]$ is

$$\nu(\lambda, F, \tau_2) \triangleq \mathbb{E}[N_T - N_{\tau_2}] = \lambda_g(1 - F(p_2))(T - \tau_2). \tag{1.14}$$

Therefore, we can test our model's power of predicting the number of arrivals

[1] For this test, we are focusing on the second threshold rather than the first threshold, and hence setting our "training" data set to $[0, \tau_2]$ and not $[0, \tau_1]$, for several reasons. First, for each event, before the first threshold is crossed, there are only 20 observations, which is a limited amount of data to construct a reliable estimation from. Second, the estimation "learns" from our model's dynamic evolution of consumer equilibrium on the entirety of $[0, \tau_2]$, until the second threshold is crossed, and after which, there is no longer any consumer strategic sign-up behavior exists. Limiting the estimation to $[0, \tau_1]$ would mean limiting the training data set to a subset of what is available, and throwing away all the observations on $[\tau_1, \tau_2]$, that can adjust the estimation with further information on observations of consumer strategic behavior.

Tab. 1.2: Test results for $E[N_T - N_{\tau_2}]$

Mean $E[N_T - N_{\tau_2} \mid t_1, t_2, ..., t_{M_2-1}]$	82.16
Mean difference with observed $N_T - N_{\tau_2}$	-0.44
Standard Deviation	15.85
p-value	0.6875

after τ_2 as follows: First, for each event i, $1 \leq i \leq n$, using Maximum Likelihood Estimation as described in Section 1.4.2 on $[0, \tau_{1i}]$, we can estimate $\hat{\lambda}_{gi}$, and $\hat{F}_i$. That is,

$$(\hat{\lambda}_g, \hat{F}) = \underset{\lambda_g, F}{\operatorname{argmax}} \ \mathcal{L}(\lambda_g, F; t_1, t_2, ..., t_{M_2-1}) \,, \tag{1.15}$$

where $\mathcal{L}$ is as described in Section 1.4.2. Then, we can calculate $\nu(\hat{\lambda}_{gi}, \hat{F}_i, \tau_{2i})$, $1 \leq i \leq n$ as defined in (1.14), and comparing to the actually observed $N_T^i - N_{\tau_2}^i$, perform a t-test to determine whether the model's predictions are in line with the data or can be rejected. Table 1.2 summarizes the test results for the 217 events where the second deal materialized. The mean sign ups predicted by model is 82.16, which is very close to the mean of the observed number of sign ups and the difference is not significant with a p-value of 0.6875. Therefore, we can conclude that our model matches very closely with the observed data.

Figure 1.4, further illustrates this point. The figure normalizes time by setting τ_2 as the reference point (i.e., $\tau_2 = 0$), and adjusting all time points, t, relative to this threshold for the $n = 217$ events in our data. With this labeling, for each time point t, the figure tracks the average cumulative number of arrivals for all events that have data for that time point t, normalized by each event's consumer arrival

rate, $\hat{\lambda}_g T$, estimated in Section 1.4.2.1. That is, for any t, denoting the number of events that are active at time t by $n(t)$, the figure plots

$$\eta(t) = \sum_{i=1}^{n(t)} \frac{N_{ti}}{\hat{\lambda}_{gi} T_i n(t)} \,. \tag{1.16}$$

In addition, the figure plots the normalized projection line for our model specification, based on how it would project the expected growth of cumulative sign ups after the second deal threshold is effectively crossed (i.e., for the time interval $(\tau_2, T]$). Specifically, for $t \geq 0$,

$$Y(t) = \sum_{i=1}^{n} \frac{N_{0i} + \hat{\lambda}_{gi}(1 - \hat{F}_i(p_{2i}))(t - \tau_{2i})}{\hat{\lambda}_{gi} T_i n} \,. \tag{1.17}$$

Thus, $Y(t)$ denotes the expected number of total sign ups at time t according to our model, if the event ended at time t, normalized by the expected number of arrivals during the event window according to the model estimate on $[0, \tau_2]$, $\hat{\lambda}_g T$.

As can be seen from the figure, the model projection aligns very well with the data (with a slope of 0.033), closely mimicking the approximate average Observed Sign-up Rate (OSR, slope 0.031), as it was also confirmed by the t-test in Table 1.2. Overall, we can conclude that our model provides a good estimate for the consumer behavior resulting in predictions very closely fitting with the actual data. The data suggests that if there is any consumer behavior not captured in our model (e.g., threshold waiting behavior), its effect on the equilibrium outcome is very small and statistically insignificant. In the rest of the paper, we will use our model's parameter

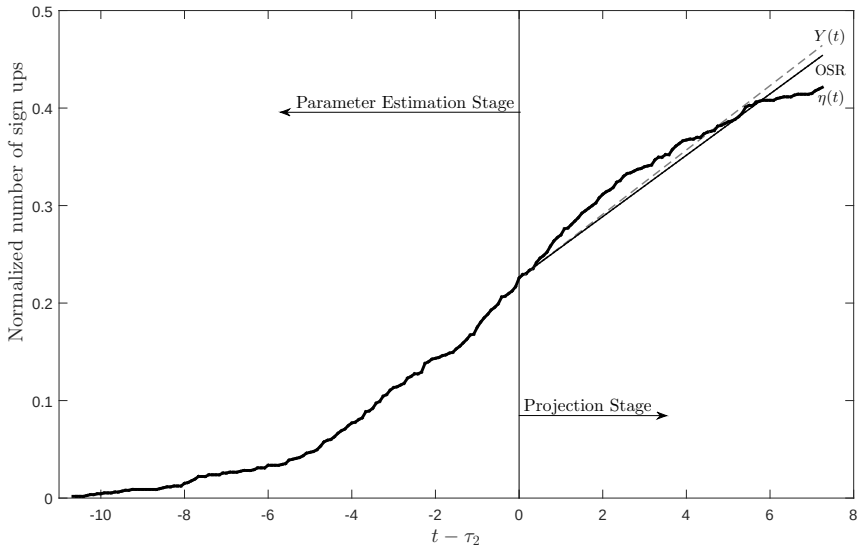

Fig. 1.4: Normalized cumulative customer arrivals and projections from the alternative models for the 217 events where the second deal threshold is reached. $\eta(t)$ represents the average cumulative number of arrivals, OSR is the linear regression line for the growth trend of the observed data for $t > \tau_2$, and $Y(t)$ is the average estimated linear growth trend for $t > \tau_2$ for the estimated model based on consumer sign ups on $[0, \tau_2]$. All quantities for each of the 217 events are normalized by the total expected customer arrivals for that event, estimated by the model.

estimates to empirically study consumer network effects and net profit gains from employing group buying events.

1.4.3 Network Effects and Gains from Group Buying

There are two components of the effect of employing a group buying mechanism on retailer profits: First, the price discounts when the number of sign ups exceeds the threshold levels tend to reduce retailer profits. On the other hand, group buying

gives incentives to customers to spread the word about the event and recruit others through networking, leading to an increase in the customer arrival rate during the event. The net effect of employing a group buying mechanism is determined by the interaction between these two opposing factors. In this section, utilizing data on the products sold through both group buying and traditional single-price mechanisms, we first estimate the magnitude of the demand increase due to network effects in group buying events. Then, through a counterfactual analysis, we estimate the retailer profits had the retailer sold the product through traditional single-price events, and study the combined effects of group buying discounts and increased demand due to customer networking to determine the net gains on retailer profits brought about by employing group buying.

1.4.3.1 Description for the Extended Data Set

We start by describing the extended data set we use for our analysis, which, in addition to the data on 266 Group Buying Events we deacribed in Section 1.4.1 covers 2715 observations of sales numbers for products sold under traditional single-price sales over 6 days. To determine the date effect for each category, we use the control data set of 260 products including all six categories that were sold on Single's day, Christmas day and four days following Christmas day through single-price events at the same price together with the sales data of the 231 group buying products on these dates. So for each day we have the same 491 products sold on the same online store and in total we have 2946 observations of the sales data together with other characteristics such as the review score. The review score is also an important

factor reflecting how satisfied consumers feel about the products. Everyone who has bought the product can rate it from 0 to 5, where 5 is the highest, as their feedback. The weight of each consumer's review score is the same. We observe the average review score for each product at the end of each day. Interestingly, we find that the lowest score is 4.6 and the highest 4.9. It is not surprising that no product can achieve the highest 5 while obviously there is also no large-scale low-rate behavior. This is very normal on Taobao.com since the review score is of great account to the retailer. On one side, consumers need to give a solid explanation for a low rate, otherwise the retailer can appeal to Taobao to withdraw it and on the other side retailers will try their best in the upfront and aftermarket to eliminate low rates. We also treat the deal discount, i.e., the price difference of group buying events on Single's Day as one of our covariates as it demonstrates how the magnitude of deal discount in group buying pricing strategy affects its consumer arrival rates.

1.4.3.2 Quantifying the Increase in Customer Demand for Group Buying due to Customer Network Effects

Utilizing the extended data set, we can now estimate the determinants of the consumer arrival rate for purchasing a product in both group buying and single-price events. For this, we employ a logarithmic doubly stochastic arrival process model, or a Generalized Linear Regression Model with log-link function (Nelder and Baker 1972), which we will estimate. Specifically, the arrival process for a given event i is Poisson with rate λ_i, which is also stochastic, and for the corresponding regressor

variable vector $\mathbf{X}_i \in \mathcal{R}^n$,

$$\mathbb{E}(\lambda_i|\mathbf{X}_i) = e^{\beta \mathbf{X}_i}, \tag{1.18}$$

where $\beta \in \mathcal{R}^n$ is a vector of coefficients. For our estimation, we employ a doubly Poisson arrival process, meaning that the distribution of λ_i conditional on $\mathbf{X}_i$ is Poisson with mean as given in (1.18) (Colin Cameron and Trivedi 1998, Coxe et al. 2009).

The determinants of the magnitude of the potential consumer demand for an event include various factors, such as the characteristics of the product that is sold during the event, e.g., product category, quality, and color, in addition to the date of the event and whether group buying was employed in the event, and if so, the discount offered. We include these factors in our regressor vector, and in order to separate the date and category effects, we also include date and category interaction terms. In particular, taking the logarithm of both sides of (1.18), we have

$$\begin{aligned} \log(\mathbb{E}(\lambda_i|X_i)) &= \beta \mathbf{X}_i \\ &= \sum_{j=1}^{6} \beta_{1j} I_{ij} + \beta_2 \frac{\delta_i}{p_{si}} + \sum_{j=1}^{6} \beta_{3j} T_i I_{ij} \\ &\quad + \sum_{t=1}^{5} \sum_{j=1}^{6} \beta_{4jt} D_{it} I_{ij} + \sum_{j=1}^{6} \beta_{5j} D_{i1} T_i I_{ij} + \sum_{j=1}^{6} \beta_{6j} R_i I_{ij} \end{aligned} \tag{1.19}$$

In (1.19), I_{ij} is a dummy variable that indicates whether product sold in event i was in category j. δ_i/p_{si} is the deal discount for a given event, normalized by the estimated optimal single-price corresponding to an event (please see below for further details). T_i (or *Treatment* indicator) is a dummy variable that indicates if

the product in event i is one of those that were put on for sale group buying on November 11, 2013. Note that if a product was put up for a group buying event on Singles' Day 2013, if later on that product is sold in a single-price event i, T_i will still be one, i.e., this variable controls for the effect of being put on a Group Buying event for the potential demand for the product. D_{it} is the day indicator for the event. In particular, if event i was held on Day t, then $D_{it} = 1$, otherwise $D_{it} = 0$, where $t = 1$ for Single's Day (November 11), and $t = 2, \ldots, 5$ correspond to the four days after the Christmas Eve, December 25-28, 2013, respectively, i.e., Christmas Eve is the base. Finally, R_i is the average consumer review score for the product sold in event i, measured on a scale of 1 to 5.

For category $j = 1, \ldots, 6$, β_{3j}, measure the difference in logarithm of the potential demand between the products that were chosen to be sold by group buying events and those that were never included in group buying events, and β_{4jt} measure the date effects, reflecting the increase in logarithm of the potential demand for day $t \in \{1, \ldots, 5\}$ compared to Christmas Eve. Importantly, the coefficients β_{5j}, $j = 1, \ldots, 6$, capture *the consumer network effects* associated with group buying events. Specifically, for an event i, and category j, $D_{i1}T_iI_{ij} = 1$ if and only if product sold in the event is in category j and was one of the products that were included in a group buying event on November 11 (i.e., $T_i = 1$), and the date of the event was November 11, i.e., if and only if the event was a group buying event (note that if the same product was sold on any other day in the data set, T_i would still be one but D_{i1} would be zero, making the whole term zero). As such, the coefficient β_{5j} captures the increase in the logarithm of the expected arrival rate for an event

in category j due to the event being a group buying event.

The GLM estimation proceeds with a Maximum Likelihood Estimation of (1.19) by fitting it to the observed λ_i values under Poisson Distribution. Note that λ_i in (1.19) captures the magnitude of the *potential* consumer demand. Consumers who are arriving with rate λ_i make purchasing decisions based on the pricing of the product. For the 266 group buying events, the λ_i calculated in Section 1.4.2.1 correspond to this potential demand arrival rate. For the remaining 260 events that were priced through traditional single-pricing, we have to calculate the corresponding rates from the observed sign up numbers and other information we have.

Estimation of Consumer Arrival Rates and Utility Distributions for Single-Price Events

As we mentioned above, for each single-price event in the data set, we have the number of purchases, N_{oi}, for that event over a 12 hour period (i.e., $T = 12$). Note that given the c.d.f., F_i of the consumer utility distribution for the product in event i, and the single-price for the event p_{ci}, the consumer purchase process in event i is Poisson with rate $\lambda_i(1 - F_i(p_{ci}))$. Using this information, for event i, the Maximum Likelihood Estimate[2] for the customer arrival rate λ_i is

$$\lambda_i = \frac{N_{oi}}{T(1 - F_i(p_{ci}))}. \tag{1.20}$$

Therefore, in order to estimate the customer arrival rate for a product sold through

[2] Please see Section **??** in the Online Supplement for a proof.

a single-price event, we need to first estimate the customer utility distribution for that product. First, if the product sold in a single-price event on December 24-28 is the same as one of those products sold under group-buying, we already have the consumer utility distribution estimate for the product calculated in Section 1.4.2.1. For those products that do not appear in any of the 266 group buying events in our data set, we use the existing utility distribution estimates for each product category and product characteristics to estimate the parameters of the utility distribution. Note that as presented in Table 1.1, best distribution estimates for all 266 products are Beta, Log-normal, or Normal distributions. All three of these distribution types are two parameter distributions and can be uniquely identified by their mean and standard deviations. Therefore, using the 266 estimated utility distributions as our training data, we first estimate the determinants of the mean and standard deviation of consumer utility in each category.

Specifically, for each category, $j = 1, \ldots, 6$, we first determine a specific set of factors that affect the value of a product in that category. For example, the Television Sets are generally homogeneous and their value for the customers are mainly determined by the screen size and display resolution. Similarly, the category-specific value determinants we use are capacity for Refrigerators, capacity and energy level for Air Conditioners, capacity and power for Water Heaters, power and panel for Gas Stoves, and capacity and energy level for Washing Machines. We obtain the corresponding data for all 491 products in our data set from the product design department of the appliance company and through online public sources. Then, for each category j, using the products in that category for which the estimates of the

mean, $\{\hat{\mu}_{gij}\}$, and standard deviation, $\{\hat{\sigma}_{gij}\}$, of consumer utility distribution exist from our analysis in Section 1.4.2.1, and given the category-specific factor matrix, denoted as Z_{gj}, we run the set of regressions

$$\log(\hat{\mu}_{gij}) = \alpha_0 + \alpha_1 Z_{gij} + \epsilon^{\mu}_{ij} \tag{1.21}$$

$$\log(\hat{\sigma}_{gij}) = \gamma_0 + \gamma_1 Z_{gij} + \epsilon^{\sigma}_{ij} \tag{1.22}$$

where Z_{gij} is the row of Z_{gj} that corresponds to product i in category j, and ϵ^{μ}_{ij} and ϵ^{σ}_{ij} are the affiliated error terms. We choose a logarithmic regression structure as it is commonly used in estimating consumer utility, and since our robustness checks with other regression structures show that the fit of the logarithmic regression is the best. The estimation results are given in Table 2.1 in the Online Supplement. The fit of the model is very good for the distribution mean regressions, as the F-Ratio for each model is highly significant, with p-values for the F-test less than 0.01%, indicating the validity of the model, and the adjusted R^2 values vary from a minimum of 0.40 and reaching as high as 0.92. As could be expected, the fit of the model for the estimation of standard deviation is somewhat worse, with adjusted R^2 values ranging from 0.21 to 0.71, but again the F-ratio for each model is highly significant, with corresponding p-values less than 1%.

Taking the regression results estimated from the training data-set, we can then estimate the consumer utility distribution for the products that were only sold through single-price events. Specifically, for product i of category j in this group,

denote the factor vector by as Z_{cij}. The estimated mean and standard deviation, $\hat{\mu}_{cij}$ and $\hat{\sigma}_{cij}$ for the utility distribution for this product can then be calculated as

$$\hat{\mu}_{cij} = e^{\alpha_0 + \alpha_1 Z_{cij}} \quad \text{and} \quad \hat{\sigma}_{cij} = e^{\gamma_0 + \gamma_1 Z_{cij}} . \tag{1.23}$$

For the consumer utility distribution for these products, we use the Beta distribution with mean and standard deviation as specified in (1.23), since in our utility estimations in Section 1.4.2.1, as given in Table 1.1, for approximately two-thirds of product utility distributions, a Beta distribution is the best fit.[3] For the product in single-pricing event i, calculating $1 - F_i(p_{ci})$ using these estimated consumer utility distributions and the product price in that single-price event, p_{ci}, we can then estimate the consumer arrival rate for event i, using (1.20).

Determining the optimal single price for each product

As we mentioned above, we would like to control for the size of the deal discount when determining the network effects, as we expect larger deal discounts to have greater effect on average in boosting demand through network effects. However, each product is on a different price scale, and in order to have a fair comparison among deal discounts, we need to normalize the deal discounts across products to obtain the relative magnitude of the deal discounts to the product's "value" or a benchmark price. For this we will use the estimated optimal single price for each.

[3] In order to perform a robustness check on this assumption, we have also performed our regression analysis below with Log-Normal and Normal Distributions. The results show that our analysis is very robust to this assumption, with all our results being preserved and with only negligible changes in regression coefficients.

product

For a given product with consumer arrival rate λ, and the marginal production cost c, the optimal traditional non-group-buying single price solution is

$$p_s^* \triangleq \underset{p\geq 0}{\operatorname{argmax}}\,(p-c)\lambda T(1-F(p)) = \underset{p\geq 0}{\operatorname{argmax}}\,(p-c)(1-F(p)). \tag{1.24}$$

p_s^* is the expected profit maximizing price for the Poisson consumer arrival process with rate λ, if the retailer simply sets a single price, p, and each arriving customer with reservation price u purchases the good when $u \geq p$. For convenience in expression, we will refer to p_s^* as the *single price optimum* in the rest of the paper. Denote the corresponding optimal profit $(p_s^* - c)\lambda T(1 - F(p_s^*))$ by Π_S.

For product i in category j, denote the marginal unit production cost as c_{ij} and define the date-effect corrected estimated single-price November 11 2013 customer arrival rate as λ_{sij}. Then utilizing the customer utility distribution estimated in Section 1.4.2, F_{ij}, and by (1.24), we can calculate the retailer single-price profit maximizer, $p_{sij} = \operatorname{argmax}_{p\geq 0}\,(p - c_{ij})T(1 - F_{ij}(p))$. For this, we need to have an estimate of the unit production cost, c_{ij}. In the data set, we do not have the unit production cost at the product level. However, for each of the six categories of products in the data set, we have the average margin, based on the group buying event final price. The number of products in each category and the corresponding margins are given in Table 1.5. Thus, utilizing these percentage margins and the group buying event final price for each product, we can calculate the estimated production cost, c_i for each product. Subsequently, we calculate the p_{si}^* as above.

Estimation Results and Measuring Demand Increase from Group Buying

Finally, by using the estimated consumer arrival rates for all events, we can perform the GLM estimation with the specification given in (1.19). The estimation results are given in Table 1.3. Note that there are 55 independent variables on the right hand side of the regression equation (1.19). For conciseness, the date effects for December 25-28 are omitted in the table in order to focus the variables most relevant to the determination of the Group Buying effects. For most date-category combinations for December 25-28, the estimated regression coefficients are not significant, indicating no significant consumer demand difference between those dates and Christmas Eve.

As can be seen from the table, the consumer review score has a significant and positive effect on the estimated consumer arrival rate for all categories, other than Washing Machines, for which the effect is not significant. The product being one of those selected to be sold by group buying on November 11th (i.e., treatment) has a mixed effect on the demand depending on the category. This is because in some cases the retailer sells products that are not very popular through group buying to boost their sales, i.e., being in the treatment group may be an indicator of inherent low demand. In other cases, the retailer chooses more popular products to include in group buying to promote its own brand. In either case, this variable controls for and separates product popularity from group buying effects we are aiming to extract. As can also be seen from the table, the positive effect of Single's Day in increasing consumer demand is strong and highly significant across all categories. In particular, compared to Christmas Eve, the estimated average demand increases

Tab. 1.3: Regression of date and network effects

	Estimate	Std. Error	p-Value
Refrigerators	0.2231	0.7382	0.7625
Air Conditioners	2.0455***	0.7045	0.0037
Television Sets	1.9872***	0.6752	0.0032
Water Heaters	1.5931**	0.6665	0.0168
Gas Stoves	1.9548***	0.7784	0.0095
Washing Machines	3.0157***	0.5980	<0.0001
$\boldsymbol{\delta/p_s}$	**0.2804*****	0.1026	0.0063
Treatment: Refrigerators	0.1219***	0.0083	<0.0001
Treatment: Air Conditioners	0.0093	0.0081	0.2489
Treatment: Television Sets	-0.1063***	0.0083	<0.0001
Treatment: Water Heaters	0.1709***	0.0075	<0.0001
Treatment: Gas Stoves	0.0656***	0.0089	<0.0001
Treatment: Washing Machines	-0.1296***	0.0068	<0.0001
Single's Day: Refrigerators	0.2289***	0.0158	<0.0001
Single's Day: Air Conditioners	0.1365***	0.0151	<0.0001
Single's Day: Television Sets	0.0845***	0.0141	<0.0001
Single's Day: Water Heaters	0.1991***	0.0151	<0.0001
Single's Day: Gas Stoves	0.1301***	0.0148	<0.0001
Single's Day: Washing Machines	0.1594***	0.0128	<0.0001
Group Buying: Refrigerators	**0.1025*****	0.0206	<0.0001
Group Buying: Air Conditioners	**0.1116*****	0.0225	<0.0001
Group Buying: Television Sets	**0.1477*****	0.0216	<0.0001
Group Buying: Water Heaters	**0.0688*****	0.0209	0.0024
Group Buying: Gas Stoves	**0.1012*****	0.0216	<0.0001
Group Buying: Washing Machines	**0.1149*****	0.0175	<0.0001
Review Score: Refrigerators	0.4553***	0.0446	<0.0001
Review Score: Air Conditioners	0.1262**	0.0424	0.0188
Review Score: Television Sets	0.1582***	0.0407	0.0006
Review Score: Water Heaters	0.1936***	0.0400	<0.0001
Review Score: Gas Stoves	0.1322**	0.0465	0.0145
Review Score: Washing Machines	-0.0991	0.0360	0.1193

*:$p < 0.1$, **:$p < 0.05$, ***:$p < 0.01$

Adjusted R^2: 0.6524

from approximately 8.45%, for Television Sets, to 22.89%, for Refrigerators.

Having controlled for the above factors, we can now observe the net effects of group buying on consumer demand. First, for each category, the group buying effects are positive and highly significant, providing evidence that group buying events increase the consumer interest in a product by generating a special promotional effect, or giving incentives to customers to spread the information about the event and recruit other potential customers for the event. According to the regression results, the additional constant demand boost with group buying ranges from around 7% to more than 15%. In addition, the coefficient of the normalized group buying deal discount, δ/p_s is positive and significant at the 5% level, indicating that the higher the deal discount relative to the product's optimal monopoly price, the higher the increase in the expected consumer demand. This is because, in general, the higher the deal discount, the more the customers would have incentives to spend effort to act as voluntary sales agents and recruit other customers to join the event. The regression estimates that offering a deal discount at the size of the product's monopoly price helps increase the consumer demand by about 24%.

We further perform a robustness check of the double Poisson process assumption by calculating heteroskedasticity adjusted robust standard errors and coefficient estimates (Cameron and Trivedi 2009). The results are given in Table 1.4. As can be seen from the table, all constant group buying effects as well as the deal discount size effect are still significant, providing support for the robustness of our findings under the model assumptions. Finally, by using the discount-to-monopoly price ratio (δ/p_s) for each event, we can calculate the total estimated demand boosting

Tab. 1.4: Robust SE of Network Effects

	Network Effects				
	Avg.(%)	Robust SE	p-value	Average	
Normalized Price Diff	0.2804**	0.1105	0.0112	Price Diff	**Total Effects**
Category based fixed effects					
Refrigerators	10.79%***	0.0317	0.0012	0.1188	14.56%
Air Conditioners	11.81%***	0.0308	0.0003	0.1178	15.57%
Television Sets	15.92%***	0.0368	<0.0001	0.0958	19.08%
Water Heaters	7.12%**	0.0300	0.0335	0.0932	9.97%
Gas Stoves	10.65%***	0.0299	0.0007	0.1011	13.84%
Washing Machines	12.18%***	0.0263	<0.0001	0.1106	15.72%
All Products				0.1065	15.40%

*:$p < 0.1$, **:$p < 0.05$, ***:$p < 0.01$

effect from group buying. As can be seen from Table 1.4, the average normalized deal discount in each category ranges from 9.32% to 11.88% of the estimated optimal monopoly price. For each event i in category $j = 1, \ldots, 6$, calculating $e^{\beta_{5j}+\beta_{2j}(\delta_i/p_{si})}$, we can calculate the total percentage increase in customer demand through group buying for each event i using robust estimates. As can be seen in Table 1.4, the estimated average demand increase ranges from 9.97% to 19.08% across the six categories, and overall average demand increase due to group buying network effects is 15.40%. In sum, we can conclude that there is evidence of significant network effects in increasing consumer demand through group buying events.

1.4.3.3 *Counterfactual Analysis for Profit Gains through Group Buying*

Having determined the percentage increase in the consumer arrival rate coming from employment of group buying, we can estimate the single-price event customer arrival rates for the 266 products in the data set to perform a counterfactual analysis to estimate the base customer arrival rates for these products and calculate expected optimal retailer profit *if* they had been sold on November 11 through single-price events instead of group buying events. Then, we can calculate the realized profits from the group buying events, and estimate the net percentage gains in retailer profits from employing group buying.

For a product i that is in category j, the estimated consumer arrival rate if the product were sold through single price instead of group buying, λ_{oi} can be calculated by taking out the group buying effects in increasing the consumer arrival rate utilizing the estimated regression equation (1.19), which implies $\lambda_{oi} = \lambda_i e^{-(\beta_{5j}+\beta_{2j}(\delta_i/p_{si}))}$. In order to calculate the profits for the counterfactual scenario for selling a product at a single price, recall that in Section 1.4.3.2, we obtained the estimated single-price optimum p^*_{si} for each product. Utilizing this price together with the c.d.f. of the estimated consumer utility distribution for the product, F_i, its estimated marginal cost c_i and the estimated single price arrival rate λ_{oi}, and obtain the counterfactual single-price mechanism profit $\Pi^*_{Si} = (p^*_{si} - c_i)\lambda_{oi}T(1 - F_i(p_{si}))$.

Having calculated the retailer's counterfactual single-price expected profits, we can next calculate the estimated realized profits of the retailer from each one of the group buying events in order to determine the net percentage profit increase

for each event. The retailer's realized profit from a group-buying event depends on the total number of customers who signed up during the event window and stayed in after the event, Q. Whether a customer who signed up stays or drops out, in turn, depends on the number of sign ups, N. The firm's profit equals to the profits from the number of units sold, i.e., $(p-c)Q$, plus the deposits collected from the customers who drop out, i.e., $d(N-Q)$. Therefore, for product i, the firm's total realized profit from the group buying event is $\Pi_{Gi} = (p^i - c_i)Q_i + d(N_i - Q_i)$, where p^i denotes the realized group-buying price for product i. The estimated gains vary from -50% to 100% with an average of 11.21% and standard deviation of 25.23%. Utilizing a t-test for significance, yields a t-value of 6.75, which is significant at the 1% level (p-value $1.16 \cdot 10^{-10}$). Finally, once again applying a Wilcoxon rank sum test for robustness we obtain a z-value of 4.73 with a p-value of $2.22 \cdot 10^{-6}$, i.e., it is again significant at 1% level. The breakdown of the profit gains according to product categories are given in Table 1.5.

As can be seen from the table, estimated profit gains for all categories are statistically significant, except for Refrigerators and Water Heaters which are 5.25% and 5.51% respectively. The category of Television sets obtains the highest profit gain of 22.57% by implementing Group Buying and Air Conditioners and Gas Stoves also have a profit increase higher than 10%. Washing Machines, however, has the third least profit gain among all categories. Overall, in our data set, group buying events bring a significant boost of 11.21% to profits.

Finally, we can calculate the approximate estimated annual monetary gains by employing group buying events for the retailer in our data set. For the 231 events

Tab. 1.5: Counterfactual Analysis Results

Category	**No. of Events**	**Avg. Margin**	**Profit Gain**		
			Avg.(%)	Std.Dev	p-value
Refrigerators	39	6.6%	5.25%	0.2511	0.1994
Air Conditioners	33	21.4%	12.38%**	0.2776	0.0153
Television Sets	54	16.7%	22.57%***	0.2743	<0.0001
Water Heaters	28	18.5%	5.51%	0.2057	0.1675
Gas Stoves	23	5.3%	11.07%**	0.2398	0.0375
Washing Machines	54	27.2%	6.46%**	0.2098	0.0277
All Products	231	17.0%	11.21%***	0.2523	<0.0001

*:$p < 0.1$, **:$p < 0.05$, ***:$p < 0.01$

we analyzed, the average monetary profit per event is 45,400 CNY. The retailer in our data set ran an estimated 5,775 individual group buying events in 2013. Therefore the total annual profits from the group buying events for the retailer is about 262 Million CNY. With a 11.21% average gain over the optimum single-price, we can then calculate the average annual monetary gain as 29.37 Million CNY, which is approximately 4.32 Million USD. Further, with 266 events in the batch in our data set, retailer profit for each such batch is 12.08 Million CNY. Scaling up to more than 60,000 estimated group buying events hosted by Taobao each year, we can estimate the annual profits generated through group buying events on the platform to be about 2.72 Billion CNY, or 419.08 Million USD. Again applying the 11.21% average gain, we can then find that the annual profit increase through the employment of group buying events hosted by Taobao to be about 279.66 Million CNY, or approximately 41.16 Million USD.

1.4.3.4 The Effect of Retailer Pricing Patterns on Profit Gains

As we have observed from Tables 1.4 and 1.5, despite the estimated network effects from group buying is positive and statistically significant, not every product category is estimated to have had significant profit gains from employing the mechanism. For Refrigerators and Water Heaters, the estimated profit gains are substantially lower than those for Air Conditioners and Television Sets, and the gains from the former two categories are not statistically significant. What causes these differences among different categories? For this, let us examine the tradeoffs underlying retailer's decision to select the deal discount, δ. In determining the deal discount, δ, the retailer considers two opposing factors: First, discounts yield losses for the dealer compared to the base price (Cao et al. 2015), and they need to be kept as small as possible. On the other hand, as we have also empirically demonstrated in Section 1.4.3.2, the larger the deal discounts, the higher the consumer arrival rate, λ, since larger discounts give customers increased incentives to spend effort to disseminate information and recruit other customers (Jing and Xie 2011), and an increase in customer arrival rate on average increases profits. Therefore, the retailer has to set the discount, δ, to the right level, to balance between losses caused by the discounts and the gains from incentivizing customers to network and bring in others.

Given this trade-off, let us first consider the case when the existing customer arrival rate without any networking and promotion by the customers (base arrival rate) is very low. In that case, the probability of crossing even the first threshold is very low and it is very likely that no deal will happen. Therefore, there is very

little incentive for customers to spend effort recruiting other customers through networking. As a result, the retailer has very little incentive to give discounts, and she should be setting a very low discount level, δ. On the other end of the spectrum, for very high base consumer arrival rates it is very likely that even without any networking by the customers, the second deal threshold will be crossed. Hence, in this case the upside for the customers to spend effort in recruiting others is again very low, and it is again very difficult to incentivize them to network. Therefore, for high base customer arrival rates, it should similarly be in the best interest of the retailer to minimize her losses by setting a very small discount. We then conclude that, for profitable pricing, for very low and very high consumer arrival rates, the discount level, δ, should be set close to zero. On the other hand, for intermediate consumer arrival rate levels, there are likely benefits from setting positive deal discounts. Therefore, an inverse U-shaped pattern of (normalized) deal discounts as a function of the arrival rate would help maximize the profitability from group buying. On the flip side, if the deal discounts are not showing such an inverse U-shaped pattern, then it is likely that the retailer gave too much discounts for low and/or high consumer arrival rate products and the retailer profits will be reduced.

To formally demonstrate this role of pricing patterns on profits, we focus on the pricing pattern of each category separately. The deal discounts are set by separate managers who are independently responsible for each category and the pattern for each category reflects the pricing choices of those decision makers. For each event i of the 266 group buying events in the data set, we first start by estimating the base consumer arrival rate, λ_{oi}, by using the estimated by the regression (1.19) as

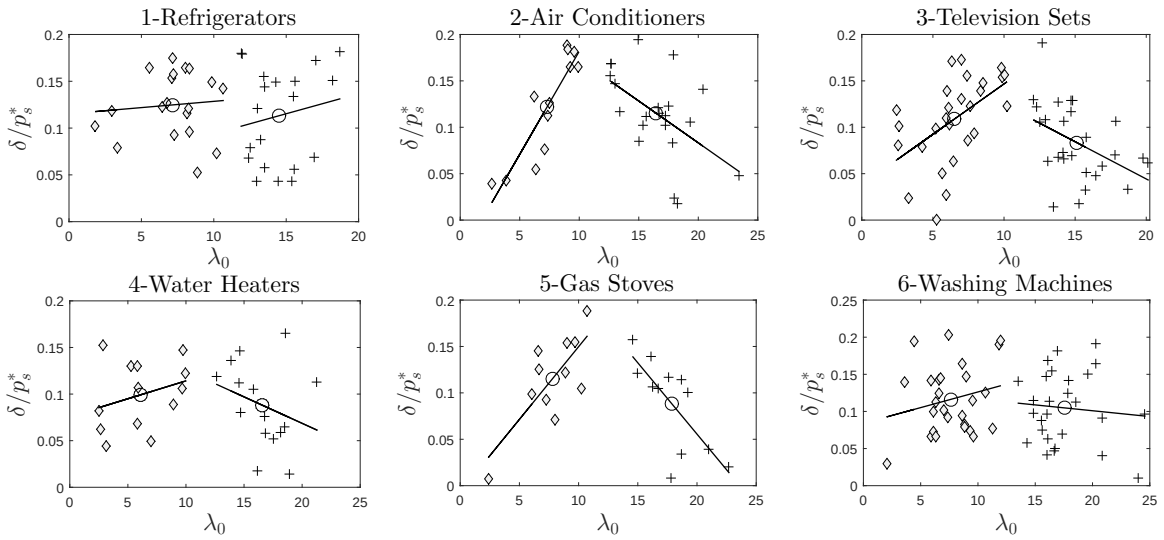

Fig. 1.5: Clustering of normalized deal discounts based on product categories and the corresponding cluster based fitted regression lines. In all panels, the circles indicate the cluster centroids.

discussed in Section 2.5. Then for each category, we plot δ_i/p^*_{si}, as calculated in Section 1.4.3 versus the estimated λ_{oi}. In order to detect the shape of the pricing patterns, we then proceed to formally identify the individual segments of the data by clustering for each category. We perform the clustering by employing K-means, K-medoids and GMM methods separately (Bouman et al. 1997; Vassilvitskii 2007), and choosing the method with a lower Euclidean total distance of the centroids to every point in their respective cluster. In order to determine the number of clusters, we evaluate the performance of different numbers of clusters using the Gap criterion (Pujari 2001), which gives us a suggested optimal number of clusters by looking for the highest decrease in error measurement, and by testing it again using the Davies-Bouldin index method (Aggarwal and Reddy 2013). The data plots and clustering results are demonstrated in Figure 1.5. The clustering process for each category in

optimality results in the two clusters seen for each product category – one for small and one for large λ_o values (left and right clusters). For a given category, if the deal-discount pricing demonstrates an inverse U-shaped pattern as discussed above, then δ/p_s^* values will be increasing for low λ_o values (left cluster) and decreasing for high λ_o values. To measure these patterns formally, we then group the data points in all the left and right clusters for all categories in two groups, $k = 1$ and $k = 2$ respectively, and for $k \in \{1, 2\}$ run the regression

$$\frac{\delta_i^k}{p_{si}^k} = \sum_{j=1}^{6} \alpha_j^k I_{ij}^k + \sum_{j=1}^{6} \beta_j^k I_{ij}^k \lambda_i + \epsilon_i^k. \tag{1.25}$$

In (1.25), the subscript i indicates the i^{th} event in group k, I_{ij}^k is the indicator function for event i in group k being in category j, $j \in \{1, \ldots, 6\}$, and ϵ_i^k is the corresponding error term. The full regression results are given in Table 2.1 in the Online Supplement. For each category $j \in \{1, \ldots, 6\}$, we plot the regression line corresponding to that category, $\delta_i^k/p_{si}^k = \alpha_j^k + \beta_j^k \lambda_i$ in Figure 1.5 on the corresponding cluster for visual demonstration.

The relationship of pricing patterns based estimated from regression (1.25) and their relationship to profitability is summarized in Table 1.6. It can be seen from the table that for Television Sets, Air Conditioners and Gas Stoves, the normalized deal discounts are increasing with the consumer arrival rate and decreasing for larger ones, and the slopes are highly statistically significant. That is, these three categories sharply demonstrate the recommended inverse U-shaped pricing pattern. For Washing Machines, Water Heaters, and Refrigerators on the other hand, the

Tab. 1.6: Effect of pricing patterns on the profit gains

Category	**Left Clusters**			**Right Clusters**			**Profit Gain**
	Estimate	Std. Error	p-Value	Estimate	Std. Error	p-Value	
Television Sets	0.0009^{***}	0.0003	0.0053	-0.0007^{**}	0.0003	0.0236	$22.57\%^{***}$
Air Conditioners	0.0019^{***}	0.0003	<0.0001	-0.0008^{**}	0.0003	0.0127	$12.38\%^{**}$
Gas Stoves	0.0013^{***}	0.0004	0.0065	-0.0013^{***}	0.0004	0.0066	$11.07\%^{**}$
Washing Machines	0.0004	0.0003	0.2712	-0.0001	0.0003	0.6573	$6.46\%^{**}$
Water Heaters	0.0003	0.0003	0.3248	-0.0005	0.0004	0.2978	5.51%
Refrigerators	0.0004	0.0005	0.4785	0.0001	0.0003	0.7149	5.25%

*:$p < 0.1$, **:$p < 0.05$, ***:$p < 0.01$

slopes are not significant there is no statistical support for the existence of the recommended price pattern. These observations are also clearly visible in Figure 1.5. Remarkably, as can be seen from Table 1.6, there is a clear correspondence between the average estimated profit gains due to group buying and the pricing patterns. In particular, the three categories that demonstrate the recommended inverse U-shaped pattern have much higher profit gains compared to the three categories that failed to employ this pricing pattern. Further, a two-sample t-test confirms the difference in estimated profit gains is statistically significant at 1% level with a t-value of 3.4668 and p-value of 0.0006. In addition, for robustness check, we also perform the non-parametric Wilcoxon Rank Sum test for the difference, again finding is significant at 1% level with a z-value of 3.2458 and a corresponding p-Value of 0.0012. These observations demonstrate the profit boosting effect of an inverse U-shaped pricing pattern in group buying, with low deal discounts for least and most popular products and higher deal discounts for products of intermediate popularity. Man-

agers who employ group buying can improve profitability by following such pricing patterns.

1.5 Concluding Remarks

In this study, we examined consumer behavior and retailer pricing strategy in online group buying events, and demand and profit gains induced by the employment of this retail strategy. Replicating Taobao's group buying event format, we presented a game-theoretical model of group buying. We considered consumer sign-up behavior as a continuous time dynamic problem, and derived the stochastic consumer equilibrium as a solution to a recursive differential equation system. The basic idea is that after observing existing sign ups, the decision of whether to pay a non-refundable deposit and sign-up depends on a consumer's belief about the success rate of the deal, i.e., the decisions of subsequent arrivals. Through our equilibrium, we analyzed the evolution of the likelihood of number of sign ups reaching the posted deal thresholds and the consumer propensity of signing up.

Given the availability of prices, thresholds, and consumer sign-up times for each event, we were able to structurally estimate our model's parameters such as the consumer arrival rates and reservation price distributions through the equilibrium expressions we derived in our theoretical model. Note that based on the consumer sign-up behavior from our analysis, the sign-up process for a given event is not a Poisson process (homogenous or non-homogenous), but the rate and the distribution of sign ups can be calculated, and that allows us to perform the estimation on model

parameters. Our results also suggest that most consumers do not systematically employ strategies that involve waiting up to a late time point in the event window and signing up only after observing that the deal is guaranteed to materialize. Utilizing the estimation and calibrating the posted deal prices and discounts by the corresponding estimated single-price optimum values, our results show that when the consumer arrival rate is very small or very large, the retailer should set the deal discount amount relatively very low, almost replicating a single-price sales scenario. Therefore, our results suggest that the bulk of the gains from group buying events come from not highly popular products or those with very low consumer interest, but rather for the products with *medium* levels of consumer demand, since in this region, the retailer can harvest the benefits of customers' networking activities by offering discounts.

Our study can be considered as a first step into a broader future stream of research that analyzes group buying, and provides insights into this novel method of retailing that is growing in popularity. There are many future directions of research to deepen our understanding of the subject. One such research avenue can be analyzing other formats of implementing the concept, such as increased number of deal thresholds or varying deposits levels. The analysis of such more complicated settings, though, can be challenging, especially in a dynamic, multi-agent environment as we aimed to tackle in this paper. However, the insights provided from our study as well as follow up studies can be helpful in future design and efficient use of group buying mechanisms, and ultimately better harvesting of the value generated from this innovative channel strategy.

Chapter 2: An Empirical Analysis of Price Formation, Utilization, and Value Generation in Ride Sharing Services

2.1 Introduction

The establishment of mobile Internet technologies have led to the modernization of people's traveling choices. Compared to the traditional street hail approach, the birth of ride-sharing platform, which shakes the taxi industry by utilizing the mobile internet features and integrating online payment and offline service, links passengers and drivers in a more efficient manner and enhances the traveling experiences for both parties. Ride-sharing platforms not only reduce the mismatch between customers and drivers but also provide a new solution to our everyday transportation problem.

Ride-sharing platforms typically connect passengers (demand) and drivers (supply) through mobile applications. On such mobile applications, passengers can choose a range of car services to get matched with different type of cars. Passengers have access to the number and locations of nearby cars and can place orders by typing or leaving a voice message of the destination. Information about the driver license plate and contact information becomes available immediately after the order

is taken by a driver.

Ride-sharing platforms also adopt a dynamic pricing approach to coordinate market supply with demand. Such pricing scheme used by ride-sharing services providers such as Uber and Didi is often referred to as Surge Pricing. During times of high demand, the fare surges to price out customers with lower willingness to pay or better outside options and would provide incentive for drivers to work. However, the effectiveness of surge price has generated heated controversy among practitioners and regulators, as it's not yet clear if surge pricing could help the ride-sharing service provider grasp a larger share of the market and improve the social welfare in general, or it's just a tool serving to increase the platform's revenue.

Another feature that ride-sharing platforms has in common is that the capacity of the driver is not centrally controlled by the platform, as studied by (Gurvich, Lariviere, and Moreno 2015). Instead, the number of drivers that are working at a specific time is indirectly determined by the appealingness of the surge price relative to the outside options of each driver. It's argued that a capacity cap should be imposed by regulators, since one of potential consequences of the indirect scheduling approach is that there could be too much supply when the price is set inadequately high, and the traditional Taxi industry would be harmed thereby. Previous studies show mixed results: (Gurvich, Lariviere, and Moreno 2015) illustrate that the ride-sharing firm should limit agent flexibility and restrict the number of drivers that can work in some time intervals, while (Cachon, Daniels, and Lobel 2015) indicate that all stakeholders can benefit from the platform with self-scheduling capacity that is indirectly determined by surge pricing.

One example of a ride-sharing platform who employs surge pricing is Didi Chuxing (Didi for short in the rest of the paper), which has acquired Uber China recently. Didi is currently the largest Taxi-hailing platform in China with a market share of more than 90%. The company is present in more than 400 Chinese cities and completes 16 million rides on a daily basis in Q2 2016 with a total of 1.43 billion rides completed in 2015[1]. Didi provides taxis as well as basic and premium car-hailing services (known as Kuaiche and Zhuanche respectively) while dynamically setting the prices to better match the supply with the demand. The company thus serves as an ideal ground for analyzing the effectiveness of surge pricing as well as the necessity for regulators to impose price and capacity regulations. The number of Didi Kuaiche drivers in Beijing as of June 2015 climbed to around 95,000 while the number of Taxi remained at 66,000. This expansion of ride-sharing service brings about conflict between Taxi industry and Didi, as the latter has been eating Taxi's market share. The government is alarmed too by the newly-emerged ride-sharing service. In the past a few months local governments in China have promulgated administrative regulations towards Didi on both capacity and price fronts. Similar to surge price scheme, the effectiveness and the necessity to regulate ride-sharing market need to be studied under scrutiny as well. Does the price regulation increase customer's welfare as the customers will be paying less? what is the capacity regulation on customers welfare? To answer these questions, we conduct empirical analysis on the impact of price and capacity constraints on customers' as well as drivers' welfare using a comprehensive dataset obtained from Didi. We hope to contribute to the

[1] https://en.wikipedia.org/wiki/Didi_Chuxing

understanding of surge pricing and the impact of government regulation on car-sharing platform from our empirical evidence.

In this paper, we first employ a discrete choice model for both Kuaiche and Taxi to describe customers' travel choices and driver's decision on whether to work at a given time. The number of passengers and available drivers affect supply and demand respectively at the same time together with the current price and other characteristics. We then estimate the consumer shares for Kuaiche and Taxi and driver's capacity shares for these two services over time to capture the effect of price and other factors on demand and supply. Utilizing a dataset including order details and driver information for Kuaiche and Taxi from Didi, we estimate the price per order for each service in different time periods. Because of the simultaneity between demand and supply, we address the endogeneity by using the exclusive exogenous parameter in each side (demand/supply) as an instrumental variable for the regression of the other side (supply/demand). Estimated price elasticities and effects of the number of consumers and drivers enables us to run a counterfactual analysis with regards to price and capacity regulations. By deriving consumer's and driver's welfare over time from their utility functions, we demonstrate the dynamic change of welfare for both consumers and drivers by imposing a price cap and capacity limit.

The rest of the paper is organized as follows: We review the related literature in Section 2.2 and Section 2.3 proposes a two-sided market discrete choice model for two services on Didi platform to describe the dynamics between demand and supply. Section 2.4 includes data description and parameter estimations as well as

our empirical results illustrating the effect of a set of parameters on both demand and supply. Section 2.5 presents the counterfactual analysis of price and capacity regulations on Kuaiche service. Section 2.6 is the concluding remark.

2.2 Literature Review

Our research is primarily related to three streams of existing literature: research on two-sided markets, recent research on peer-to-peer sharing platforms, especially ride sharing platforms, and empirical research on government regulations.

There is a vast body of literature on two-sided market building on the seminal work by Rochet and Tirole (2003). In a two-sided market, two groups of agents interact, and there are both same-side and cross-side network effects in the market. A majority of the literature focuses on the decisions (e.g., pricing) of the markets/platforms in the presence of the two network effects. Rochet and Tirole (2006) develop a model of two-sided markets by integrating usage and membership externalities. They also study a pricing model in a two-sided market including usage and membership fees. Weyl (2009) studies the price scheme in a two-sided market by introducing the vulnerability of demand. He also studies the importance of externalities and its impact on socially optimal pricing. There are some empirical studies on various two-sided markets that are more closely related to our research. Argentesi and Filistrucchi (2007) studies the Italian newspaper industry by estimating the market power in a structural model. They find that the competition on advertising prices among newspapers mitigates the market power on the reader side. Song

(2013) develops a structural model of platform demand in two-sided markets and estimate the market power using magazine advertising data. The paper shows that mergers in the magazine market are less harmful to readers and advertisers than a one-side market model predicts. The ride-sharing platforms we are studying in this paper can be viewed as a two-sided market where riders and drivers interact and both same-side and cross-side network effects present. We focus on providing insight on how government regulations impact the surplus for both riders and drivers and the social welfare as a whole.

There is growing academic interest on peer-to-peer platforms in the sharing economy. The closest to our paper are empirical papers on peer-to-peer platforms. Cohen, Hahn, Hall, Levitt, and Metcalfe (2016) estimates demand elasticities using individual-level observations of UberX in four cities in the U.S. Based on the elasticities, they estimate that in 2015 the UberX service generated about $2.9 billion in consumer surplus in the four U.S. cities included in our analysis and that the overall consumer surplus generated by the UberX service in the United States in 2015 was $6.8 billion. Li, Moreno, and Zhang (2016) study the differences between behavior of non-professionals hosts and professional hosts in Airbnb, a room rental/sharing platform. They provide empirical evidence that professional hosts earns more than nonprofessional hosts who are less likely to offer different rates based on demand across different times. Zervas, Proserpio, and Byers (2016) investigate Airbnb's entry in Texas and its impact on the incumbent hotel industry. They show empirically that the entry of Airbnb has a 8-10% impact on the local hotel revenue. They also find that local hotels responded to the entry of Airbnb with less aggressive pricing,

which benefits consumers, not just participants in the sharing economy. Fraiberger and Sundararajan (2015) develops a dynamic model of a peer-to-peer rental market for durable goods where consumers may trade in their durable assets such as cars. They calibrate the model with a data from Getaround, a peer-to-peer car rental marketplace to assess the welfare implications of sharing economy. They find that peer-to-peer platforms lead to significant welfare improvements through better allocation of goods, lower use-good prices and more efficient ownership. Buchholz (2015) provides a dynamic equilibrium of taxicabs and studies how taxi regulation leads to inefficiencies as well as how an optimal pricing increases the social welfare.

Among the papers that study peer-to-peer platforms, literature on matching and pricing in the ride-sharing market is also closely relevant to our study. Banerjee, Riquelme, and Johari (2015) investigate the value generated by dynamic pricing in a queueing framework. They illustrate that while dynamic pricing cannot yield higher throughput than static pricing, dynamic pricing is more robust than static pricing to fluctuations in system parameters. Bai, So, Tang, Chen, and Wang (2016) show that the profit of the ride sharing platform would increase when charging a higher price, paying a higher wage to the drivers and setting a higher ratio of wage of price during times of high demand. They also show that, for wait-time sensitive customers, the platform should lower its price to sustain the demand. Bimpikis, Candogan, and Daniela (2016) study a spatial price setting problem in a ride-sharing network. They establish that among all network structures, balanced demand structure leads to the highest platform profit, and that under unbalanced demand pattern, pricing the trip based on the pickup location would increase the platform profit. Ozkan and Ward

(2016) investigate the dynamic matching problem in a ride-sharing setting, in which examine the performance of LP(linear program)-based policy and CD(closest driver) policy are tested. They demonstrate through simulation that CD policy doesn't perform well under circumstances of high demand and supply imbalance, and that LP-based policy should be implemented in such cases. Hu and Zhou (2016) studies a multi-period dynamic matching problem for which the profit from matching is to be maximized. They come up with a sufficient and robustly necessary conditions on matching rewards that make the optimal decision.

Several recent papers have modeled the roles of surge pricing and self-scheduled capacity on peer-to-peer service platforms. Cachon, Daniels, and Lobel (2015) develop a game theoretical model to study the roles of dynamic/surge pricing and dynamic wage offered by a peer-to-peer service platform. They find that surge pricing achieves nearly optimal profit for the platform. Also, service providers and consumers on the platform can always benefit from the combination of dynamic prices and dynamic wages offered by the platform. Gurvich, Lariviere, and Moreno (2015) studies how capacity management when workers self-schedule affects total benefits and service levels in so-called on-demand economy. In contrast to the findings of Cachon, Daniels, and Lobel (2015), they show that no flexibility can provide better service levels, and full flexibility is bad for agents in aggregate in the on-demand economy. Instead of developing theoretical models, our paper conducts empirical studies of price formation and welfare on a peer-to-peer service platform using real data.

We apply the multinomial logit (MNL) model to model both consumer and

driver choices on the Didi platform. The MNL model has been widely used in both theoretical and empirical research (e.g., McFadden 1978, Anderson, De Palma, and Thisse 1992, Berry, Levinsohn, and Pakes 1995). A classic application of the MNL model is on consumer choices in travel and transportation industries (Ben-Akiva and Lerman 1985, Ben-Akiva and Bierlaire 1999). So, modeling consumer choices on the Didi platform, where consumers look for rides for their trips, using the MNL model in our paper is similar to this classic application. We further extend the application of MNL model to model driver choices on the supply-side of the Didi platform. This extension is a natural application of the MNL model because drivers on the Didi platform are all individuals who make individual choices between driving and not driving at particular times.

2.3 Model

Didi is a mobile phone-based transportation platform where consumers who submit a request for a transportation service using the Didi mobile app are matched with available registered drivers. There are several categories of service available on the Didi platform including Taxi, Kuaiche (i.e., Kuaiche in Chinese, which is similar to Uber X), Car Sharing (i.e., Shunfengche in Chinese), Bus, etc. Among these available services, Taxi and Kuaiche are the two largest services, which will be the focus of this study. One unique difference between Didi and other ride sharing platforms such as Uber is that regular Taxi companies (e.g., Yellow Cabs) are allowed to provide their services on Didi's platform.

A Kuaiche driver provides the service using his/her personal car, while a Taxi driver provides the service using a Taxi owned by an independent Taxi company. Kuaiche' pricing is determined by Didi whereas Taxi's pricing is determined by local city governments. When a consumer is looking for a service on Didi's platform, the consumer needs to select which service between the two (either Taxi or Kuaiche) to request for a trip. Then, the request will be received as an order which will be routed and matched to an available registered driver in the service category requested by the consumer. The drivers, however, can register as a driver in only one category of service (i.e., either as a Taxi driver or a Kuaiche driver) in Didi. A registered driver in one category would only be matched to or can only take orders submitted to that category. For example, if a driver registered as a Taxi (Kuaiche) driver, then he/she would only be matched with orders from the Taxi (Kuaiche) category. In addition, Kuaiche drivers are prohibited from accepting customers/rides from street hails, yet Taxi drivers can pick up customers both through Didi app or directly from street hails.

We will develop a supply-demand model to estimate consumer (demand) and driver (supply) choices on Didi's platform. We let A denote the consumer (demand) side and B denote the driver (supply) side of Didi's platform.

2.3.1 Consumers

When a Didi consumer needs a ride for a trip, the consumer would open the Didi mobile app to view the current pricing and availability of the two services. The consumer has three choices: submit a request for a Taxi, submit a request for a Kuaiche,

or choose an outside option such as taking public transportation. Consumers will choose service/option $j \in \{K, T, O\}$ where K denotes *Kuaiche*, T denotes *Taxi* and O denotes the outside option. At any given time t, consumer i's utility function of choosing Kuaiche can be written as:

$$u_{it}^{KA} = \alpha^{KA} N_t^{KB} + \beta^{KA} p_t^{KA} + \gamma^K w_t^K + \delta^{KA} D_t + \xi_t^{KA} + \epsilon_{it}^{KA}, \tag{2.1}$$

where N_t^{KB} is the average number of available Kuaiche drivers at time t; p_t^{KA} is the average total price of Kuaiche trips at time t; w_t^K is the average waiting time for a Kuaiche order to be taken by a Kuaiche driver at time t; D_t is a vector of control variables including day dummies, highest and lowest temperatures, weather and air quality. ξ_t^{KA} and ϵ_{it}^{KA} are random effects of the Kuaiche service faced by consumers where ϵ_{it}^{KA} follows an *i.i.d.* Gumbel distribution.

Consumer i's utility function of choosing Taxi is given in a similar form:

$$u_{it}^{TA} = \alpha^{TA} N_t^{TB} + \beta^{TA} p_t^{TA} + \gamma^T w_t^T + \delta^{TA} D_t + \xi_t^{TA} + \epsilon_{it}^{TA}, \tag{2.2}$$

where N_t^{TB} be the average number of available *Taxi* drivers at time t; p_t^{TA} is the average total price of Taxi trips at time t; w_t^T is the average waiting time for a Taxi order to be taken by a Taxi driver at time t; ξ_t^{TA} and ϵ_{it}^{TA} are random effects of the Taxi service faced by consumers where ϵ_{it}^{TA} follows an *i.i.d.* Gumbel distribution.

We consider a consumer's choice in different time segments. In each individual time segment, the time factor stays constant. Since each consumer differs in their

routes, the total price each of them pays is also different. Therefore, we use the average total prices of all trips at time t, i.e., p_t^{TA} or p_t^{KA}, to account for the price factors in the consumer utility functions.

We assume consumer i's utility of choosing the outside option is given as,

$$u_{it}^{OA} = \mu^{OA} + \epsilon_{it}^{OA}, \tag{2.3}$$

where μ^{OA} is the fixed effect of the outside option, and ϵ_{it}^{OA} is the random effect of the outside option faced by consumer i which follows an $i.i.d.$ Gumbel distribution. Without loss of generality, we normalize the mean utility from choosing the outside option to 0.

Consumers will choose the service/option that generate the highest utility for them. So, consumer i will choose service/option $j \in \{K, T, O\}$ at time t, if $u_{it}^{jA} > u_{it}^{kA}$, for all $k \neq j$, and $k \in \{K, T, O\}$. Based on the utility functions, the market share of service $j \in \{K, T\}$ offered on the platform at time t is given by the Multinomial Logit formula:

$$S_t^{jA} = \frac{exp(\alpha^{jA} N_t^{jB} + \beta^{jA} p_t^{jA} + \gamma^j w_t^j + \delta^{KA} D_t + \xi_t^{jA})}{1 + \sum\limits_{j \in \{K,T\}} exp(\alpha^{jA} N_t^{jB} + \beta^{jA} p_t^{jA} + \gamma^j w_t^j + \delta^{KA} D_t + \xi_t^{jA})}, \tag{2.4}$$

Similarly, the market share for the outside option is given as:

$$S_t^{OA} = \frac{1}{1 + \sum\limits_{j \in \{K,T\}} exp(\alpha^{jA} N_t^{jB} + \beta^{jA} p_t^{jA} + \gamma^j w_t^j + \delta^{KA} D_t + \xi_t^{jA})}. \tag{2.5}$$

Therefore, the system of demand equations on the consumer side of the Didi platform to estimate is as follows:

$$\log(S_t^{jA}) - \log(S_t^{OA}) = \alpha^{jA} N_t^{jB} + \beta^{jA} p_t^{jA} + \gamma^j w_t^j + \delta^{KA} D_t + \xi_t^{jA}, \text{ for } j \in \{K, T\}. \quad (2.6)$$

2.3.2 Drivers

Recall that drivers can only take orders from the service category in which they registered as a driver. Therefore, at any given time t, a registered Didi driver, either a Kuaiche driver or a taxi driver, faces only two choices: driving or not. As a result, the driving decision in a service category are independent of that in the other category so that we can deal with them separately.

Let's consider the decision for a Kuaiche driver first. For Kuaiche driver i, we assume that the utility of driving, i.e., taking Kuaiche orders from the Didi platform, at time t is given as:

$$u_{it}^{KB} = \alpha^{KB} N_t^{KA} + \beta^{KB} p_t^{KB} + \gamma^K v_t^K + \delta^{KB} D_t + \xi_t^{KB} + \epsilon_{it}^{KB}, \quad (2.7)$$

where N_t^{KA} denotes the average number of Kuaiche orders submitted on the Didi platform at time t; p_t^{KB} is the average total payoff a Kuaiche driver receives on one trip at time t; v_t^K is the average cumulative money earned up to time t in a particular day; D_t is the vector of control variables. And ξ_t^{KB} and ϵ_{it}^{KB} are random effects. ϵ_{it}^{KB} follows an *i.i.d.* Gumbel distribution. To deal with the heterogeneity of payoffs each driver may receive, we use the average total price of a Kuaiche trip,

p_t^{KA} , to account for the driver payoff factor, i.e., $p_t^{KB} = p_t^{KA}$.

If a Kuaiche driver decides to not driving, the driver will take the outside option. We use superscript O_K to denote the not driving option for Kuaiche drivers. For Kuaiche driver i, we assume that the utility of not driving at time t is

$$u_{it}^{O_K B} = \mu^{O_K B} + \epsilon_{it}^{O_K B}, \tag{2.8}$$

where $\mu^{O_K B}$ is the fixed effect of not driving for a Kuaiche driver, and $\epsilon_{it}^{O_K B}$ is the random effect which follows an $i.i.d.$ Gumbel distribution at each time t. Without loss of generality, we normalize $\mu^{O_K B}$ to 0.

Based on the utility functions, the proportion of registered Kuaiche drivers that decide to drive, which we call active driver ratio (ADR) of Kuaiche drivers at time t, denoted by s_t^{KB}, is

$$s_t^{KB} = \frac{exp(\alpha^{KB} N_t^{KA} + \beta^{KB} p_t^{KB} + \gamma^K v_t^K + \delta^{KB} D_t + \xi_t^{KB})}{1 + exp(\alpha^{KB} N_t^{KA} + \beta^{KB} p_t^{KB} + \gamma^K v_t^K + \delta^{KB} D_t + \xi_t^{KB})}, \tag{2.9}$$

and the percentage of registered Kuaiche drivers that decide not to drive at time t, denoted by $s_t^{O_K B}$, is

$$s_t^{O_K B} = \frac{1}{1 + exp(\alpha^{KB} N_t^{KA} + \beta^{KB} p_t^{KB} + \gamma^K v_t^K + \delta^{KB} D_t + \xi_t^{KB})}. \tag{2.10}$$

Then, the system of supply equations for the Kuaiche drivers of the Didi

platform to estimate is as follows:

$$\log(s_t^{KB}) - \log(s_t^{O_K B}) = \alpha^{KB} N_t^{KA} + \beta^{KB} p_t^{KB} + \gamma^K v_t^K + \delta^{KB} D_t + \xi_t^{KB}. \quad (2.11)$$

We assume similar utility functions for Taxi drivers. The utility of driving on the Didi platform for Taxi driver i at time t is:

$$u_{it}^{TB} = \alpha^{TB} N_t^{TA} + \beta^{TB} p_t^{TB} + \gamma^T v_t^T + \delta^{TB} D_t + \xi_t^{TB} + \epsilon_{it}^{TB}, \quad (2.12)$$

where μ^{TB} denotes the fixed effect of Taxi service on the platform in the driver's side; N_t^{TA} is the average number of Taxi orders submitted on the Didi platform at time t; p_t^{TB} is the average payoff taxi drivers receive from one trip at time t; v_t^K is the average cumulative money earned up to time t; D_t is the vector of control variables. ξ_t^{TB} and ϵ_{it}^{TB} are random effects. ϵ_{it}^{TB} follows an *i.i.d.* Gumbel distribution.

Taxi drivers, different from Kuaiche drivers, can also choose to pick up passengers directly from streets while not driving for the Didi platform. For Taxi driver i, we assume that the utility of not driving for the Didi platform, denoted as O_T, for Taxi driver i at time t is:

$$u_{it}^{O_T B} = \mu^{O_T B} + \epsilon_{it}^{O_T B}, \quad (2.13)$$

where $\mu^{O_T B}$ is the fixed effect of not driving for a Didi driver (e.g., including revenue from picking up passengers from streets), and $\epsilon_{it}^{O_T B}$ is the random effect of not driving for Taxi driver i which follows an *i.i.d.* Gumbel distribution at each time t. As we

discussed before, drivers can only drive and take orders from the service category in which they registered as a driver on the Didi platform. Therefore, the driving decisions of the Kuaiche drivers and the Taxi drivers are independent of each other, which allows us to normalize $\mu^{O_T B}$ to 0 without loss of any generality.

Based on the utility functions, the proportion of registered Taxi drivers that decide to drive as a Didi driver, i.e., the active driver ratio (ADR) of Taxi drivers at time t, denoted by s_t^{TB}, is:

$$s_t^{TB} = \frac{exp(\alpha^{TB} N_t^{TA} + \beta^{TB} p_t^{TB} + \gamma^T v_t^T + \delta^{TB} D_t + \xi_t^{TB})}{1 + exp(\alpha^{TB} N_t^{TA} + \beta^{TB} p_t^{TB} + \gamma^T v_t^T + \delta^{TB} D_t + \xi_t^{TB})}, \tag{2.14}$$

and the proportion of drivers that decide not to drive as a Didi driver at time t, denoted by $s_t^{O_T B}$ is

$$s_t^{O_T B} = \frac{1}{1 + exp(\alpha^{TB} N_t^{TA} + \beta^{TB} p_t^{TB} + \gamma^T v_t^T + \delta^{TB} D_t + \xi_t^{TB})}. \tag{2.15}$$

Then, the system of supply equations for the Taxi drivers of the Didi platform to estimate is as follows:

$$\log(s_t^{TB}) - \log(s_t^{O_T B}) = \alpha^{TB} N_t^{TA} + \beta^{TB} p_t^{TB} + \gamma^T v_t^T + \delta^{TB} D_t + \xi_t^{TB}. \tag{2.16}$$

From the system of demand and supply equations for both types of drivers, we can see that consumers' choices and drivers' decisions are closely related to each other and estimations of corresponding demand and supply equations enable us to track the change over time. The effect of price as well as other parameters can also

be estimated and learned from this structure.

2.4 Data and Model Estimations

In this section, we investigate how prices and ADRs of Kuaiche and Taxi affect consumers' choices between the two services on the Didi platform, and how in turn the market shares and prices of the services affect Kuaiche and Taxi drivers' decisions of whether to drive or not.

2.4.1 Data Description and Empirical Strategy

Primarily we use three datasets, two from Didi and one collected from the Internet, to conduct the empirical analysis. The first dataset we obtained from Didi is a transaction-level dataset that contains a random sample of 3000 Kuaiche drivers as well as 1000 Taxi drivers registered on Didi's platform. The dataset spans from December 3rd, 2015 to January 3rd, 2016. In this dataset, for each ride, it includes the longitudes and the latitudes at which each customer is picked up and dropped off by the driver, the accurate-to-the-second time the customer submits a request through the Didi App and the time a driver accepts the request, the time the customer is dropped off, and the total fare of the ride. The summary statistics of the transaction dataset is shown in table 2.1. The second dataset from Didi contains the information of the drivers of both services, including each driver's total online time on the Didi App in each of the days and the brand of the car of each driver. In addition, we collect an independent dataset online on the control variables

during the timespan of the Didi dataset. More specifically, the dataset documents the highest and the lowest temperature, precipitation and air pollution condition of each of the days in the Didi datasets[2].

To estimate the system of demand and supply derived from Equations (2.6), (2.11) and (2.16) in the previous section, we adopt a random-coefficients multinomial logit (MNL) model to capture the impact of customer's (driver's) elasticities of price and the number of drivers (customers) on customer's(drivers) decision. To this end, we first develop the estimation strategy to estimate the average price per order, the consumer's and the driver's choices, and customer's average waiting time from the time the order is placed to the time the order is accepted. We then identify two potential causes of endogeneity in our estimation. One is the price which is a usual endogenous variable in demand (supply) estimation due to the lack of supply (demand) shifter as a control variable. The other is the simultaneity problem in our estimation caused by having the number of drivers and customers on the two sides of the regression. We propose two IVs in the two-stage least squares (2SLS) regression to cope with the endogeneity issue. After that we estimate the customer's (driver's) elasticities of price and the number of drivers (customers) in both Taxi and Kuaiche services. We consider peak time and non-peak time in each day and assume that during peak time and non-peak time, each time segment is identical to the others. That is to say, if we assume that the morning peak time lasts from 6:00 a.m. to 10:00 a.m., and we estimate the variables every three minutes, we will then have

[2] To gauge the level of air pollution, we employ the commonly adopted measure known as atmospheric particulate matter (PM). In particular, we use fine particles with a diameter of 2.5 μm or less (PM 2.5) as the indicator or air quality, where a higher PM 2.5 value suggests a greater level of air pollution.

Tab. 2.1: Summary Statistics

	Total Fare (Yuan)		Trip Time (Min)		Avg. Transaction per Day		
	Mean	Std.Dev	Mean	Std.Dev	Mean	Std.Dev	Sample Size
Didi Kuaiche	18.74	15.50	23.38	299.92	3.27	4.02	309466
Didi Taxi	51.23	35.98	35.79	26.13	1.53	1.10	47526

$32 \times (4 \times 60/3) = 2560$ observations for the variables during morning peak time. The assumption is reasonable especially when we shorten the duration of each time segment, so that the next time segment would be almost identical to the current one. Also, by separately estimating the variables during peak and non-peak hours, we are able to learn how the price elasticities of demand and supply change over time.

2.4.2 Estimations of Auxiliary Variables

We do not observe all auxiliary variables in our datasets like average price per order, market share as well as ADR and thus we need to estimate these variables using the data we have in the datasets and some public data.

2.4.2.1 Estimation of Prices

In order to study the impact of prices on consumer choices and driver choices, we need to estimate the aggregated price effect on the market level. We look at the case for Kuaiche first. We first decompose the individual total price for each Kuaiche trip, which is recorded in our Didi transaction level dataset, according to Kuaiche's pricing structure which includes the base price per kilometer, the base price per

hour, the extra price per kilometer if the total distance is longer than the threshold distance and the extra price per kilometer if the trip happens at night. Note that Didi also apply surge pricing to Kuaiche service. So, for each Kuaiche trip, there might also be a surge multiplier on the base total price. Specifically, according to the pricing structure of Kuaiche service on Didi, the total price customer i pays for her trip at time t can be expressed as follows:

$$p_{it}^{KA} = surge_t^K (p^{Kd} Distance_{it}^K + p^{kh} Time_{it}^K + I_{\{Distance_{it}^K\}} p^{Kd1} (Distance_{it}^K - D_0^K) + I_{\{Time_{it}^K\}} p^{Kd2} Distance_{it}^K), \tag{2.17}$$

where $surge_t^K$ is the surge multiplier for the Kuaiche service at time t; p^{Kd} and p^{Kh} are the base price per kilometer and per hour, respectively; $Distance_{it}^K$[3] and $Time_{it}^K$ are the distance and time duration of the trip; $I_{\{Distance_{it}^K\}}$ is an indicator for whether it exceeds the threshold of distance D_0^K, if so there will be an extra price p^{Kd1} per kilometer for exceeded distance; and similarly $I_{\{Time_{it}^K\}}$ is an indicator for whether the trip happens between 11pm and 5am where there is an extra price per kilometer p^{Kd2} for the whole trip. Notice that all the variables in the above equation are recorded in our Didi transaction level dataset, except the surge multiplier $surge_t^K$. Therefore, using the observations in the dataset, we can estimate

[3] In the formulation, we use the starting and destination latitude and longitude to estimate the total distance of each order. *Distance* (in kilometer) is calculated from the starting and destination longitude and latitude as $Distance = |Dest.Lat - Start.Lat| \times 110.574 + |Dest.Lng - Start.Lng| \times 111.320 \times \cos\left(\frac{Dest.Lng + start.Lng}{2}\right)$.We assume that the distance can be obtained by calculating the summation of latitude and longitude differences due to the fact that the roads in Beijing are mostly perpendicular to each other.

the average surge multiplier $surge_t^K$ at time t. Figure 2.1 demonstrates the pattern

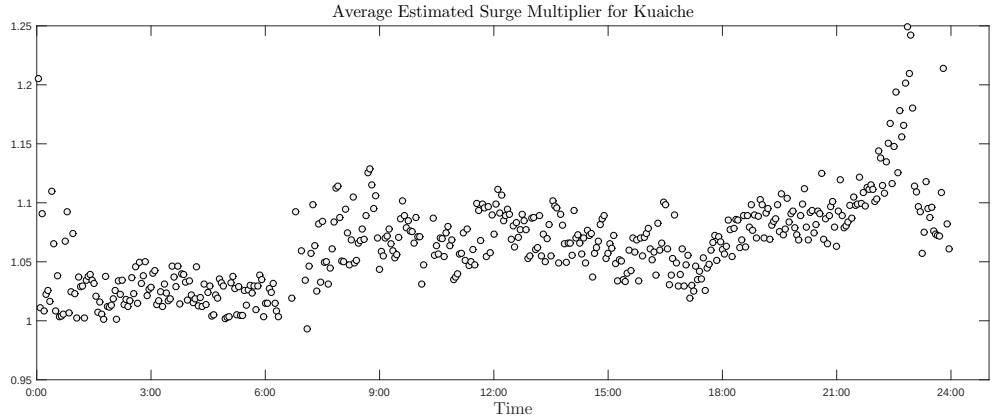

Fig. 2.1: Average Estimated Surge Multiplier for Kuaiche service over Time

of average estimated surge multiplier over different time segments. We can see that on average the surge reaches its first highest level during the morning peak hours, i.e., 7 -9 a.m., and remains very close to 1 at other non-peak hours. The mismatch between insufficient supply and excessive demand reflects itself on surge during the evening peak hours. It shows that the surge starts to climb during the evening peak hours and keeps increasing till late night, especially around 10 p.m..

After obtaining the estimate of average surge rate $surge_t^K$ at time t, we calculate the average distance, $Distance_t$ and trip duration, $Time_t$ of Kuaiche trips at time t using the observations in the dataset. Then, we approximate the average total price of a Kuaiche trip in the market at time t as:

$$p_t^{KA} = surge_t^K(p^{Kd}Distance_t + p^{kh}Time_t + I_{\{Distance_t\}}p^{Kd1}(Distance_t - D_0^K) + I_{\{Time_t\}}p^{Kd2}Distance_t) \tag{2.18}$$

which will be used in the estimation of demand as shown in Equation (2.6). For the

Taxi service, the average total price of a Taxi trip at time t can be estimated in a similar way without the surge multiplier.

2.4.2.2 Estimations of the Market Shares

To estimate the system of the demand and supply as derived by Equations (2.6, 2.11, 2.16), we need the market shares of Kuaiche, Taxi, and the outside options (i.e., S_t^{KA}, S_t^{TA} and S_t^{OA}) as well as the active driver ratios for both Kuaiche drivers and Taxi drivers (i.e., s_t^{KB} and s_t^{TB}), which all are not available in our datasets directly.

We start with the market shares. Let N be the number of time segments in a day, and M be the total number of days in our datasets. The market share of Kuaiche at time segment $n \in [1, N]$ on day $m \in [1, M]$ is defined as:

$$S_{t_{mn}}^{KA} = \frac{N_{t_{mn}}^{KA}}{N_{t_{mn}}^{A}}, \tag{2.19}$$

where $N_{t_{mn}}^{KA}$ is the total number of Kuaiche trips starting at time t_{mn} in the whole city and $N_{t_{mn}}^{A}$ is the total number of trips in the city at time t_{mn} including all means of transportations. However, we do not directly observe both $N_{t_{mn}}^{KA}$ and $N_{t_{mn}}^{A}$ in our datasets, which we need to estimate. We know the overall average number of Kuaiche trips in a day during the period of time of our datasets in the city, $\overline{N}^{KA}$ from Didi, and the average number of Kuaiche trips in our one-month dataset, which can be calculated as $\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} N_{t_{mn}}^{KAS}$. Based on the above two averages, we can calculate the average total Kuaiche trips' population-to-sample ratio at time t_{mn}.

Using the ratio and the total number of Kuaiche trips within the sample at time segment t_{mn}, denoted as $N_{t_{mn}}^{KAS}$, we are able to estimate the average total number of Kuaiche trips in the whole city at time segment t_{mn} as:

$$\hat{N}_{t_{mn}}^{KA} = N_{t_{mn}}^{KAS} \frac{\overline{N}^{KA}}{\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} N_{t_{mn}}^{KAS}}. \tag{2.20}$$

We then estimate the average total number of trips made in the city on each day, $N_{t_{mn}}^{A}$. We obtained data on the average number of total trips per day in Beijing in 2015, $\bar{N}^{A}$, and the average percentage of trips in a time segment t_n in one day, γ_{t_n} from the report by China Internet Network Information Center[4]. Then to estimate the average total number of trips in each day, we use the variation in the number of daily trips in the sample to account for the real-world daily fluctuations. Denote N_t^{AS} as the total number of Kuaiche and Taxi trips in our data sample in each day, i.e., $N_t^{AS} = N_t^{KA} + N_t^{TA}$. Then we can calculate the average number of total trips at different times as followes:

$$\hat{N}_{t_n}^{A} = \gamma_{t_{mn}} \sum_{n=1}^{N} N_{t_{mn}}^{AS} \frac{\overline{N}^{A}}{\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} N_{t_{mn}}^{AS}}. \tag{2.21}$$

With both $\hat{N}_{t_{mn}}$ and $\hat{N}_{t_{mn}}^{A}$ estimated as above, we have the estimation of the market share of Kuaiche at time t_{mn} as:

$$\hat{S}_t^{KA} = \frac{\hat{N}_{t_{mn}}^{KA}}{\hat{N}_{t_{mn}}^{A}}. \tag{2.22}$$

[4] The data is obtained from the report by China Internet Network Information Center: The Research Report of Ride-Sharing Market Development in 2015

For Taxi, using the average number of taxi trips in a day in the whole city, $\bar{N}^{TA}$ including the taxi rides that were not offered on the Didi platform, and the number of Taxi trips at time t in our dataset, N_t^{TAS}, we have the estimation of the average total number of Taxi trips at time t_n on day m in the city as:

$$\hat{N}_{t_{mn}}^{TA} = N_{t_{mn}}^{TAS} \frac{\bar{N}^{TA}}{\frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} N_{t_{mn}}^{TAS}}, \tag{2.23}$$

and the estimated market share for Taxi on Didi platform at time t is:

$$\hat{S}_t^{TA} = \frac{\hat{N}_t^{TA}}{\hat{N}_t^{A}}. \tag{2.24}$$

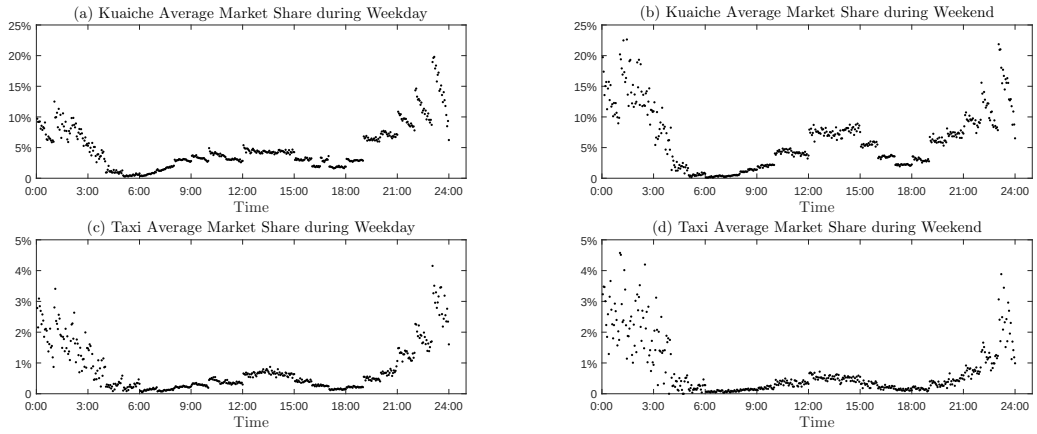

Fig. 2.2: Average market share per three minutes of Kuaiche and Taxi on weekdays and weekends. Panel (a) and (b) illustrates the average market share of Kuaiche during weekdays and weekends; Panel (c) and (d) are the market shares for Taxi during weekdays and weekends.

We know that the total numbers of Kuaiche and Taxi in Beijing were 95,000

and 66,000 in June 2015[5], respectively. So, we can obtain the total number of Kuaiche and Taxi trips according to Equations (2.21) and (2.23). Then, the market shares of Kuaiche and Taxi can be estimated according to Equations (2.22) and (2.24). Figure 2.2 presents the estimated market shares of Kuaiche and Taxi during weekdays and weekends, respectively. As we can see from the figure, the market shares of Kuaiche and Taxi exhibit similar patterns, and Kuaiche appears to have higher market shares than taxi due to the fact that the taxi market shares only capture the customers who hail taxi using the Didi App. From 6:00 a.m. to 6:00 p.m., the market shares of Kuaiche are around 10% at best, and the market shares of Taxi never exceed 3% during the same period. On the other hand, as the public transportations start to shut down after 10:00 p.m., Kuaiche and Taxi become the major means of transportations. We thus observe that the market shares of both Kuaiche and Taxi on Didi platform surge quickly after 9:00 p.m. and maintain at high levels until 3:00 a.m..

2.4.2.3 Estimation of Active Driver Ratios

We next estimate the active driver ratios on the supply side, s_t^{KB} and s_t^{TB} for both Kuaiche and Taxi. For Kuaiche drivers, the estimated percentage of active Kuaiche drivers at time t is

$$\hat{s}_t^{KB} = \frac{N_t^{KBS}}{N^{KBS}}, \tag{2.25}$$

[5] The number is obtained from the report by China Internet Network Information Center: The Research Report of Ride-Sharing Market Development in 2015

where N_t^{KBS} is the total number of active Kuaiche drivers on Didi platform at time t in our datasets, and N^{KBS} is the total number of registered Kuaiche drivers in our datasets. Also notice that the estimated number of active Kuaiche drivers in the whole city at time t is $\hat{N}_t^{KB} = \hat{s}_t^{KB} N^{KB}$, where N^{KB} is the total number of registered Kuaiche drivers in the city which we obtained from Didi.

For Taxi drivers, denote $N_{t_{mn}}^{TBS}$ as the number of active Taxi drivers at time t_{mn} in our datasets, and $N_{t_m}^{TBS}$ as the total number of active Taxi[6] drivers at day m in our datasets. Taxi drivers drive in shifts in a day. Assume that there are L shifts in each day. Then we have the estimated active driver ratio of Taxi driver as:

$$\hat{s}_{t_{mn}}^{TB} = \frac{N_{t_{mn}}^{TBS}}{N_{t_m}^{TBS}/L}. \tag{2.26}$$

Then, the total active taxi at time t in the whole city can be estimated as, $\hat{N}_t^{TB} = \hat{s}_t^{TB} N^{TB}$, where N^{TB} is the total number of all taxi in the city (including the ones who do not register with Didi).

Since drivers are not allowed to switch platforms, the decisions for them to make are whether to drive or not at any given time. When a driver chooses to drive, she will be in one of the two possible status: idle and open to new order or in the middle of an order. The first case would be considered as a unit of available capacity. The latter case, however, is ambiguous because the driver can either refuse to take any further order by logging out of the didi app or still be open to other orders, for example, when the current order is about to end. When the driver is online and

[6] A drivers is considered as active that day if she takes at least one order from Didi platform.

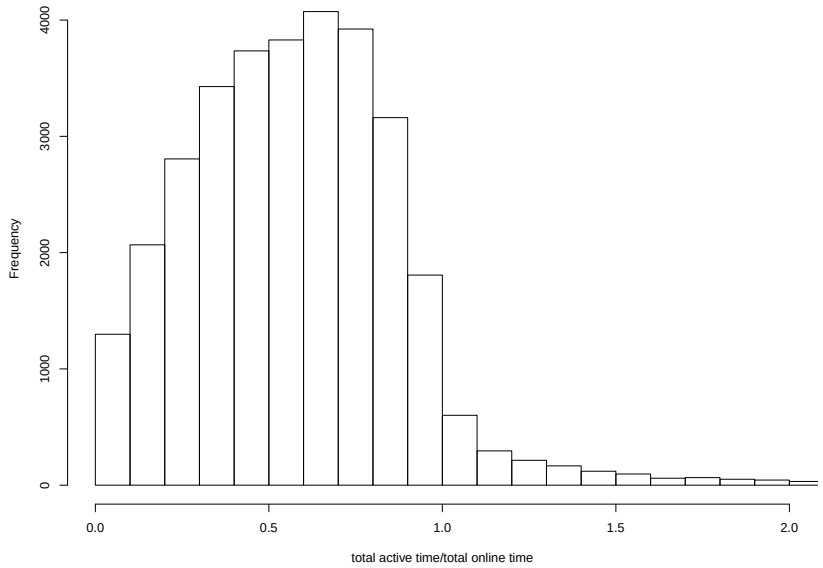

Fig. 2.3: Histogram of total active time of a driver per day over her total online time

open to order at a given time point, she can then be considered as active. Otherwise we will not include her into the total number of active drivers at that time point.

However, since we only have the total online time of each driver on each day, not the detailed data about when the driver is online or offline, we need to identify when drivers are online or offline at any time point to calculate $N_{t_{mn}}^{TBS}$.

First let's take a look at the pattern and data summary of the driver's total online time each day compared to her total work time that day. Figure 2.3 shows the histogram of the total work time of a driver per day over her total online time that day, where work time per order is defined as the finish time minus the strive time, i.e., the total driving time. Most drivers have less total work time than online time, which implies that drivers tend to keep the app open so that they can react in time to future orders. Still we have a number of drivers having less online time than

active time. It is possible that drivers close the app after picking up the customer to save cellular data.

Tab. 2.2: Active time vs Online time

	Didi Kuaiche	Didi Taxi
Avg. active time per day (Hour)	3.51	2.02
Avg. online time per day (Hour)	6.60	7.95

Table 2.2 shows the significant difference between online time and active time. According to the recent news[7], however, the average online time per day for Didi Kuaiche drivers is 4 hours instead of 6.6 hours from our data. Thus we need to calibrate the driver's online time. The average online time of 4 hours together with average active time per day of 3.5 hours of Kuaiche suggest a 30 minutes idle online time by assuming drivers are still online with a consumer in the car. Given the average number of orders per driver per day being around 3, we can approximately assign a fixed time length of 10 minutes to the beginning of each order indicating drivers' online time before any order. Since we have the strive time of each order taken by drivers together with the average online time for each driver, we allocate 10 minutes to the beginning of every order in that day. Any overlap of online time of the current order with the strive time of previous order will be moved to the online time of the previous order. In this way we make sure drivers' status to be online before each order. Notice that be it idle or in the middle of a ride 10 minutes before the next order, the driver always has to log in to Didi's application to pick up the next order. Therefore, the allocation algorithm does not conflict with the driver's

[7] https://kknews.cc/tech/l9j26g.html

status before she picks the next order.

We can then derive the active driver ratio at any given time in each day by aggregating each individual's online status. The active driver ratio of Taxi drivers can be derived in the same way. Figure 2.4 shows the active driver ratios of both

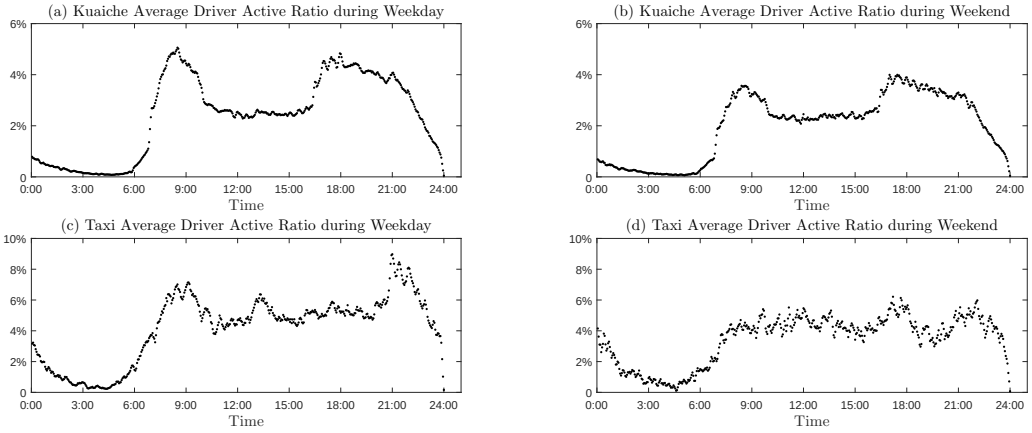

Fig. 2.4: Capacity share per three minutes of Kuaiche and Taxi during weekdays and weekends. Panel (a) illustrates the capacity share of Kuaiche during weekdays; Panel (b) is the capacity share on weekend; Panel (c) and (d) show the capacity share for Taxin during weekdays and weekends respectively.

Kuaiche and Taxi on weekdays and weekends. For Kuaiche, the active driver ratios in weekdays and weekends have similar pattern: the active driver ratios reach the peaks during morning and evening peak hours, and they stay flat at a relatively low level between morning and evening two peak hours. After the evening peak hours, the active driver ratios decrease sharply and keep decreasing from midnight to the morning. Overall, the active driver ratios of Kuaiche are responsive to the raise in demand during peak hours.

In contrast to Kuaiche, Taxi's active driver ratios in both weekdays and week-

ends we estimated are more noisy and do not exhibit significant difference in magnitude between peak hours and non peak hours, especially from 10 a.m. to 4 p.m.. Interestingly, our estimation is consistent with the business model of Taxi whose drivers are full time employed drivers who would be mostly active in a shift independent of many factors such as demand.

2.4.3 Control Variables and Estimation Specification

To estimate the structure specified in Equations (2.6), (2.11) and (2.16), we include a vector of control variables containing dummies for rush hour, daily dummies of weekday/weekend, weather variables of highest and lowest temperatures, precipitation and PM 2.5 to represent the air pollution condition where a higher PM 2.5 means more pollution.

Since the traffic condition and the surge price change dynamically within a day, we add rush hour dummies to control for the effect of the variation in traffic condition and surge rate on both consumers' and drivers' choices. The morning and evening rush hours in Beijing are 7 to 9 a.m. and 5 to 7 p.m.[8] but it is also reported by Beijing transportation department that these rush hours are usually started early and ended late by one hour, which we also observe from Figure 2.4. Thus we define our morning rush hour to be 6:00-10:00 a.m. and evening rush hour to be 4 to 8 p.m.. In particular, during the morning rush hour, we treat the official 7-9 a.m. as one time interval and 6-7 a.m.& 9-10 a.m. as another, as the latter time intervals are typically not as congested as the former one. Apart from the rush hours, we

[8] https://zhidao.baidu.com/question/210261531.html

have three non-rush hour dummies from 11 p.m. to 6 a.m., 10 a.m. to 4 p.m. and 8 to 11 p.m.. The patterns for the first two non-rush periods are similar, yet the 8-11 p.m. slot contains significant variation because the post evening rush hour period usually comprises a decreasing volume of trips as can be seen from Figure 2.4. Thus, we treat each hour in this period as an independent time interval. To include them in one regression, we assign a set of 0/1 dummies to time periods with 11 p.m. to 6 a.m. as the base, followed also by interactions between time periods and other predictors like the number of available drivers, estimated price per trip and waiting time.

Within each time interval, we treat the average total price, the number of drivers and the market shares in every three minutes as a data point. Doing so not only guarantees enough number of observations for each time interval, peak or non-peak, but also ensures that there are sufficient number of transactions within each three-minute window. We also replicate our estimation results using a five-minute window, and our main results remain the same.

2.4.4 Endogeneity

Two sources of endogeneity normally emerge when estimating the demand system of a two sided-market. The first is the common price endogeneity in demand (supply) estimation[9], which implies that price and the error terms in the estimation Equations (2.6), (2.11) and (2.16) can be correlated. The second is the simultaneity issue.

[9] It's also possible that the price is autocorrelated with lagged demand and supply pattern. However, the Vice President of Research of Didi suggests during our interview that Didi is setting the surge price based on only current supply and demand, we thus rule out this potential endogeneity issue.

When we estimate each service's market share and ADRs simultaneously, we have the number of drivers as a predictor for market share and the number of consumer orders as a predictor for ADRs. Then for instance, more customers requesting Kuaiche service could induce an increase in the number of Kuaiche drivers, yet more Kuaiche drivers could also attract more customers to choose the Kuaiche service. Thus, we allow for the possibility that the number of drivers and customers of a service to be correlated with the error terms in their regressions.

To address the price endogeneity issue, the classic approach is to control for supply (demand) shifter for the demand (supply) estimation for each time interval of the day. Because we have the number of drivers (customers) as control variable in the consumer market share (active driver ratio) estimation, once the supply (demand) shifter is instrumented, the price endogeneity issue would be solved. To address the simultaneity issue, we introduce two instruments, each of which is also an exogenous regressor for the other side in the simultaneous equation system, one for each regression and apply 2SLS. More specifically, we use waiting time w_t^j, $j \in \{K, T\}$ and its interaction with time dummy as instrumental variables (IVs) for the number of available drivers and its interaction with time in market share regression, and the current average total earning v_t^j, $j \in \{K, T\}$ and its interaction with time periods for each driver as IVs in the ADR regression. With demand and supply both instrumented, we don't need a demand/supply shifter for price endogeneity. Since the surge pricing is solely dependent on the current demand and supply, the price endogeneity can be addressed together with the simultaneity issue.

2.4.5 Estimation Results

Table 2.3 shows the results of the first stage least square on the effect of each factor on Kuaiche's market share. Price has a significantly negative effect on consumer's choice of requesting Kuaiche during the night time, i.e., 11 p.m.-6 a.m., which is expected. The same price effect also applies to all other time periods. Notice that price is not significant in the official morning rush hour from 7 to 9 a.m., illustrating that consumers are much less sensitive to price during morning rush hours. The after-dinner periods, from 8-11 p.m., also show the same result that consumers are less sensitive to price. This can be explained from two aspects: there are less transportation options available in the night, and consumers value convenience and comfort more in the night.

The number of available Kuaiche drivers imposes a positive effect on encouraging consumers to requesting Kuaiche service, and the effect is significant during the night time as well as all other time periods. This indicates the importance of the availability of Kuaiche on the Didi platform on incentivize consumers to request the Kuaiche service, because the first information a consumer observe right after opening the Didi App is the number of Kuaiche cars around her location on a map. Another important factor is the waiting time between when a Kuaiche order is placed and when the Kuaiche order is taken by an active Kuaiche driver. The waiting time has a positive but not significant effect on the Kuaiche market share during the night time, because there are limited transportation options available in the night time. So, consumers may care less about the Kuaiche's order waiting time. Similar to the

price effect, waiting time during 10 a.m.-4 p.m. and 4-8 p.m. imposes a significantly negative effect on the Kuaiche market share. However in the morning rush hours, waiting time has a positive effect on the Kuaiche market share, which may suggest that consumers are more patient during these periods. Together with price effect, consumers seem to care more sensitive to the availability of Kuaiche cars than to price and waiting time during the morning rush hours and after-dinner hours.

For the control variables, Kuaiche's market share on Monday is lower than other days and reaches its highest levels on weekends, indicating a higher preference of choosing Kuaiche on weekends from consumers. For weather variables, it is shown that Kuaiche's market share is positively related to the temperature and PM 2.5, but not to rain. The effect of time periods on Kuaiche's market share indicates that the Kuaiche's market share is much lower during morning and evening rush hours than during night hours from 11pm to 6am. This might also be driven by the fact that there are more transportation options available during the day than during the night.

For Kuaiche's active driver ratio, the first-stage least square utilizes the waiting time as an instrument and its interaction with time periods for the endogeneity of the number of Kuaiche consumers and its interaction with time periods. Table 2.4 shows the results on the effect of each factor on Kuaiche's active driver ratio. The number of consumers is always a positive incentive for Kuaiche drivers to drive during all time periods. Since when a Kuaiche driver decides whether or not to drive, she can observe the available Kuaiche orders around her location. The more available orders, the higher the probability is for a driver to take one. The price

Tab. 2.3: Kuaiche Consumers' Market Share

Fitted number of available Kuaiche drivers	0.001^{***} (0.00005)
Waiting time	0.00004 (0.0004)
Price per order	-0.016^{***} (0.001)
Highest Tempareture	0.003 (0.004)
Lowest Tempareture	0.010^{**} (0.004)
Rain	−0.018 (0.029)
PM 2.5	0.001^{***} (0.0001)
Number of drivers*6-7 a.m. & 9am-10am	0.0001^{**} (0.00005)
Number of drivers*7-9 a.m.	0.001^{***} (0.0001)
Number of drivers*10 a.m.-4 p.m.	-0.001^{***} (0.0001)
Number of drivers*4-8 p.m.	0.0001^{**} (0.0001)
Number of drivers*8-9 p.m.	0.0001 (0.0001)
Number of drivers*9-10 p.m.	0.0001 (0.0001)
Number of drivers*10-11 p.m.	0.0003^{***} (0.0001)
Price per order*6-7 a.m.& 9-10 a.m.	0.013^{***} (0.002)
Price per order*7-9 a.m.	0.015^{***} (0.001)
Price per order*10 a.m.-4 p.m.	0.012^{***} (0.002)
Price per order*4-8 p.m.	0.014^{***} (0.002)
Price per order*8-9 p.m.	0.015^{***} (0.002)
Price per order*9-10 p.m.	0.016^{***} (0.002)
Price per order*10-11 p.m.	0.013^{***} (0.003)
Waiting time*6-7 a.m.& 9-10 a.m.	0.006 (0.043)
Waiting time*7-9 a.m.	0.234^{***} (0.072)
Waiting time*10 a.m.-4 p.m.	-0.332^{***} (0.060)
Waiting time*4-8 p.m.	-0.135^{***} (0.026)
Waiting time*8-9 p.m.	1.024^{***} (0.278)
Waiting time*9-10 p.m.	0.273^{*} (0.148)
Waiting time*10-11 p.m.	−0.052 (0.105)
Constant	-2.555^{***} (0.065)
Observations	11,826
Adjusted R^2	0.583
F Statistic	403.661^{***} (df = 41; 11784)
Note:	*p<0.1; **p<0.05; ***p<0.01

effect, however, is negative and significant in 11 p.m.-6 a.m., 6-7 a.m. and 9-10 a.m. and positive during other time periods. This can also be explained by driver's target-earning driven behavior. Part-time Kuaiche drivers have more incentive to drive under a lower price per order during the night time because to reach their target or maximize their earning, they need to take more orders to compensate the lower price for each order. Drivers care less about the price during the morning rush hours since the number of orders is dominating given the higher volume and traffic. In other time periods higher price per order encourages them to drive because these part-time drivers are not as flexible as the night time. The current total earning has a negative and significant effect on the driver's willingness to driver during morning and evening rush hours together with 8-9 pm and stay positive or insignificant during all other non-peak hours. This matches our analysis of Kuaiche drivers' background and flexibility. Rush hours limit driver's choices because although there are more orders during rush hours, each order takes more time to complete given the traffic, and part-time Kuaiche drivers are less likely able to spend much time taking orders during these daytime rush hours. Actually, many Kuaiche drivers will only take the orders with the same directions as their routes to their work places or home during rush hours. The non-rush hours are more flexible for Kuaiche drivers to take more orders.

Interestingly the control variables have the opposite effect on Kuaiche drivers as compared to on consumers. Kuaiche drivers choose to take more orders on Monday than other days except Saturday when they are free to drive all day. Temperatures have a negative effect on Kuaiche drivers meaning that they prefer to drive on

Tab. 2.4: Kuaiche Drivers' Capacity Share

Fitted number of Kuaiche Consumers	0.0004^{***} (0.00004)
Price per order	-0.004^{***} (0.001)
Current total earning	0.002^{***} (0.0002)
Highest Tempareture	-0.004 (0.003)
Lowest Tempareture	-0.024^{***} (0.003)
Rain	0.112^{***} (0.019)
PM 2.5	-0.0005^{***} (0.00004)
Number of consumers*6-7 a.m. & 9-10 a.m.	0.00004 (0.0001)
Number of consumers*7-9 a.m.	0.00000 (0.00005)
Number of consumers*10 a.m.-4 p.m.	-0.0001^{**} (0.00004)
Number of consumers*4-8 p.m.	-0.0001^{*} (0.00004)
Number of consumers*8-9 p.m.	-0.00003 (0.0001)
Number of consumers*9-10 p.m.	-0.00001 (0.0001)
Number of consumers*10-11 p.m.	0.0001 (0.0001)
Price per order*6-7 a.m. & 9-10 a.m.	-0.002^{**} (0.001)
Price per order*7-9 a.m.	0.006^{***} (0.001)
Price per order*10 a.m.-4 p.m.	0.009^{***} (0.001)
Price per order*4-8 p.m.	0.005^{***} (0.001)
Price per order*8-9 p.m.	0.005^{***} (0.001)
Price per order*9-10 p.m.	0.004^{***} (0.001)
Price per order*10-11 p.m.	0.007^{***} (0.002)
Current total earning*6-7 a.m. & 9-10 a.m.	-0.001 (0.002)
Current total earning*7-9 a.m.	-0.007^{***} (0.002)
Current total earning*10 a.m.-4 p.m.	-0.002^{***} (0.0004)
Current total earning*4-8 p.m.	-0.003^{***} (0.0004)
Current total earning*8-9 p.m.	-0.003^{***} (0.001)
Current total earning*9-10 p.m.	-0.002^{***} (0.001)
Current total earning*10-11 p.m.	-0.001^{*} (0.001)
Constant	-5.740^{***} (0.043)
Observations	11,826
Adjusted R^2	0.776
F Statistic	998.136^{***} (df = 41; 11784)
Note:	*p<0.1; **p<0.05; ***p<0.01

cold days when more consumers are willing to choose to take Kuaiche due to bad weather. Also they are more likely to drive on raining days and less-polluted days. Kuaiche drivers prefer driving during the morning and evening rush hour as well as other daytime hours to driving during night time.

Table (2.5) shows the results of the two stage least square on the effect of each factor on Taxi's market share. Taxi's market share is affected by the number of available Taxi drivers positively for all time periods. Since the mechanism of ordering a Taxi and a Kuaiche on Didi platform is the same, consumers can observe the number of available Taxi around, which is proportional to the total number of active Taxi drivers. Price per order for Taxi has a negative but insignificant effect on Taxi's market share during all time periods except the night time. Limited public transportations forces consumers to take either Taxi or Kuaiche during night time even if the night price is higher. During non-peak hours, since the base price of Taxi is fixed and the total fare for a trip is quite predictable, consumers are less sensitive to the price. But the traffic in the morning rush hours makes the travelling time of a trip by Taxi fluctuating. Thus, consumers are more concerned about the price during morning rush hours. The market share is also higher for weekends than weekdays illustrating that consumers prefer ordering on Didi platform to other transportation options on weekends. Different from Kuaiche, Taxi's market share is negatively correlated with PM 2.5 indicating that taking Taxi orders is less preferable on days with worse air pollution. On the driver side, a Taxi driver has a more complicated choice to make: whether to drive as a Didi driver or a traditional Taxi driver, i.e., taking orders from streets without logging into the Didi App. Table (2.6) shows

Tab. 2.5: Taxi Consumers' Market Share

Fitted number of available Taxi drivers	0.0002^{***} (0.00003)
Price per order	0.0003^{**} (0.0002)
Waiting time	-0.0005 (0.0004)
Highest Tempareture	-0.006 (0.004)
Lowest Tempareture	-0.008^{*} (0.004)
Rain	0.053^{*} (0.031)
PM 2.5	-0.0005^{***} (0.0001)
Number of drivers*6-7 a.m. & 9-10 a.m.	0.0003^{***} (0.00004)
Number of drivers*7-9 a.m.	0.001^{***} (0.0001)
Number of drivers*10 a.m.-4 p.m.	0.0003^{***} (0.0001)
Number of drivers*4-8 p.m.	0.001^{***} (0.00005)
Number of drivers*8-9 p.m.	0.001^{***} (0.0001)
Number of drivers*9-10 p.m.	0.001^{***} (0.0001)
Number of drivers*10-11 p.m.	0.001^{***} (0.0001)
Price per order*6-7 a.m. & 9-10 a.m.	-0.0002 (0.0003)
Price per order*7-9 a.m.	-0.001^{***} (0.0003)
Price per order*10 a.m.-4 p.m.	-0.0002 (0.0002)
Price per order*4-8 p.m.	-0.001^{**} (0.0003)
Price per order*8-9 p.m.	-0.001^{**} (0.0003)
Price per order*9-10 p.m.	-0.0005 (0.0004)
Price per order*10-11 p.m.	-0.001^{**} (0.001)
Waiting time*6-7 a.m. & 9-10 a.m.	0.082^{*} (0.044)
Waiting time*7-9 a.m.	-0.289^{***} (0.083)
Waiting time*10 a.m.-4 p.m.	0.050 (0.089)
Waiting time*4-8 p.m.	-0.133^{***} (0.029)
Waiting time*8-9 p.m.	-0.582^{**} (0.255)
Waiting time*9-10 p.m.	-0.866^{***} (0.226)
Waiting time*10-11 p.m.	-0.593^{***} (0.163)
Constant	-4.002^{***} (0.066)
Observations	11,826
Adjusted R^2	0.683
F Statistic	621.014^{***} (df = 41; 11784)
Note:	*p<0.1; **p<0.05; ***p<0.01

that the number of consumers incentivize them to drive all the time expect the night time from 11 p.m. to 6 a.m.. The reason could be that orders during night time are much less than earlier daytime hours, and this forces drivers to continuously search for orders on Didi platform. It is also possible that drivers are less willing to driver at night time. Price, however, has no significant effect on Taxi drivers. Unlike Kuaiche orders, Taxi has no surge pricing and drivers care less about the price factor. The current total earning imposes a significantly positive effect on Taxi driver's willingness to take orders on Didi platform during night time. If they receive higher payoff by working as a Didi driver, there is a strong incentive for them to continue especially at night time because there are less consumers hailing on streets. However during the morning rush hours and after-dinner hours drivers with a lower total earning have more incentive to take orders on Didi platform. This can be partially explained by driver's target setting behavior. Also the active driver ratio is higher during day time than night time. The morning and evening rush hours have lower active driver ratios than the non-rush hours 10 a.m.-4 p.m. and 8-11p.m.. This is due to the unique working pattern of Taxi drivers who can take orders directly on the streets, which is very likely to be higher during rush hours. However during the non-rush hours, Taxi drivers work similarly to Didi drivers to try to reduce empty rates.

Tab. 2.6: Taxi Drivers' Capacity Share

Fitted number of Taxi Consumers	-0.001^{***} (0.0003)
Price per order	0.0002^{*} (0.0001)
Current total earning	0.001^{***} (0.0001)
Highest Tempareture	0.001 (0.003)
Lowest Tempareture	0.006^{*} (0.003)
Rain	-0.048^{**} (0.020)
PM 2.5	0.00003 (0.00004)
Number of consumers*6-7 a.m. & 9-10 a.m.	0.004^{***} (0.0004)
Number of consumers*7-9 a.m.	0.005^{***} (0.0003)
Number of consumers*10 a.m.-4 p.m.	0.003^{***} (0.0003)
Number of consumers*4-8 p.m.	0.003^{***} (0.0003)
Number of consumers*8-9 p.m.	0.006^{***} (0.0004)
Number of consumers*9-10 p.m.	0.005^{***} (0.0004)
Number of consumers*10-11 p.m.	0.005^{***} (0.0004)
Price per order*6-7 a.m. & 9-10 a.m.	-0.0004^{*} (0.0002)
Price per order*7-9 a.m.	-0.0002 (0.0002)
Price per order*10 a.m.-4 p.m.	-0.0002 (0.0002)
Price per order*4-8 p.m.	-0.0003^{*} (0.0002)
Price per order*8-9 p.m.	-0.0004^{*} (0.0002)
Price per order*9-10 p.m.	-0.0002 (0.0002)
Price per order*10-11 p.m.	-0.001^{*} (0.0003)
Current total earning*6-7 a.m. & 9-10 a.m.	-0.001^{*} (0.0003)
Current total earning*7-9 a.m.	-0.002^{***} (0.0002)
Current total earning*10 a.m.-4 p.m.	-0.001^{***} (0.0001)
Current total earning*4-8 p.m.	-0.001^{***} (0.0001)
Current total earning*8-9 p.m.	-0.002^{***} (0.0001)
Current total earning*9-10 p.m.	-0.002^{***} (0.0002)
Current total earning*10-11 p.m.	-0.002^{***} (0.0001)
Constant	-4.158^{***} (0.056)
Observations	11,826
Adjusted R^2	0.481
F Statistic	268.645^{***} (df = 41; 11784)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.5 Counterfactual Analysis

We then utilize our estimation results on the demand and supply dynamics to investigate the impact of potential government regulations on pricing and capacity through counterfactual analysis. Governments around the globe are often concerned about the lack of regulations in ride-sharing industry. Based on concerns that existing taxi companies may get severely affected by the low base price and the large number of Kuaiche drivers of Didi, and that customers may suffer from the surge price, in China there are proposed regulations commonly focusing on enforcing constraints on pricing and limiting the number of Kuaiche drivers in a city. To evaluate the impact of government regulations on ride-sharing platforms, we estimate both consumer and driver welfare through log-sum method, and evaluate the difference between the current levels and scenarios that assume the implementation of price caps and capacity limitations on Kuaiche service through counterfactual analysis.

2.5.0.1 Welfare Estimation

Each consumer has the same choice set including Kuaiche, Taxi and the outside options, i.e., other means of transportation. A consumer's utilities of choosing Kuaiche, Taxi and other transportations are u_{it}^{KA}, u_{it}^{TA} and u_{it}^{OA} defined in Equations (2.1), (2.2) and (2.3), respectively. The consumer's welfare, ω_{it}^{A}, is the expected value of the maximum utility among the three utilities of the three options, which is defined as

$$\omega_{it}^{A} = \mathbb{E}(\max\{u_{it}^{KA}, u_{it}^{TA}, u_{it}^{OA}\}). \tag{2.27}$$

Given the utilities u_{it}^{KA}, u_{it}^{TA} and u_{it}^{OA} and the properties of Gumbel distribution, ω_{it}^{A} can be written in the following closed-form as

$$\omega_{it}^{A} = \log(e^{u_{it}^{KA}} + e^{u_{it}^{TA}} + e^{u_{it}^{OA}}). \tag{2.28}$$

With the individual welfare defined in this way, we can now derive the total consumer's welfare, i.e., $\sum_t \hat{N}_t^A \omega_t^A$, by implementing the number of total consumers at time t.

For drivers, both Kuaiche and Taxi drivers face two options: drive or not drive on the Didi platform. Given the utilities u_{it}^{KB}, $u_{it}^{O_K B}$, u_{it}^{TB}, and $u_{it}^{O_T B}$, the welfare of the Kuaiche driver and Taxi driver can be written as

$$\omega_t^{KB} = \log(e^{u_{it}^{KB}} + e^{u_{it}^{O_K B}}), \tag{2.29}$$

and

$$\omega_t^{TB} = \log(e^{u_{it}^{TB}} + e^{u_{it}^{O_T B}}), \tag{2.30}$$

respectively. With the total numbers of Kuaiche driver and Taxi driver, we can write the total welfare functions for the drivers as: $\sum_t \hat{N}^{KB} \omega_t^{KB}$ and $\sum_t \hat{N}^{TB} \omega_t^{TB}$, respectively.

We can then estimate the total welfare of all consumers and drivers in each time segment. To evaluate the welfare on an individual level, we take the weighted average for both parties across different time periods. Thus the average welfare for

a consumer at time segment m among all days is,

$$\frac{\sum_{n=1}^{N} \hat{N}_{t_{mn}}^{A} \omega_{t_{mn}}^{A}}{\sum_{n=1}^{N} \hat{N}_{t_{mn}}^{A}}. \tag{2.31}$$

Since we have a fixed group of Kuaiche and Taxi drivers, i.e., $\hat{N}^{KB}$ and $\hat{N}^{TB}$, which doesn't change over time, we have the average individual Kuaiche and Taxi driver welfare in each time segment across all days expressed as $\sum_{n=1}^{N} \omega_{t_{mn}}^{KB}/N$ and $\sum_{n=1}^{N} \omega_{t_{mn}}^{TB}/N$. Notice that in the following analysis we will take the weighted average not only across all days, but also across other combinations of days or periods, dependant on our research interests. Then, we can take a deeper look at consumer and driver welfare together with the effect of various factors on the welfare. However, since we have simultaneous equations of demand and supply, both price and capacity regulation generate a new set of solutions to our equation system and thus render a changed welfare for each time point.

2.5.0.2 Price Regulation

The pricing practices of ride-sharing platforms have caused some concerns by the regulators and the general public. Many ride-sharing platforms such as Didi offer heavy subsidies to consumers which make their prices artificially lower than the prices of traditional Taxis on average. Taxi companies around the world have complained about unfair competition due to ride-sharing platforms' artificially low prices. In the following counterfactual analysis, we examine the impact of a price

regulation which enforces the average price of a ride-sharing service, Didi Kuaiche in our study, to be no lower than the average price of Taxi.

We have the estimated the average prices per trip, p_t^{KA} and p_t^{TA} for Kuaiche and Taxi on Didi platform at any time t and Figure 2.5.0.2 shows the comparison of the prices. As can be seen from the figure, expect the hour from 12am to 1am, Taxi's average price per trip is higher than Kuaiche's during all other hours in a day.

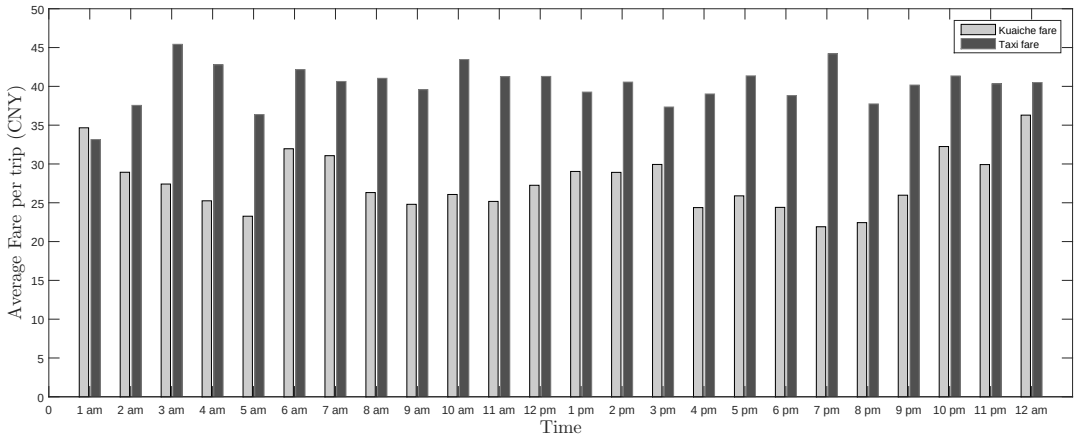

Fig. 2.5: Comparison of estimated price per trip for Kuaiche and Taxi over time

With the price regulation that Kuaiche's average price cannot be lower than Taxi's, we re-calculate the Kuaiche prices at all time segments and make sure that the new Kuaiche prices are as least as high as Taxi's in the same time segments. With effects of other parameters as well as the coefficients remaining the same, we solve the nonlinear equation system for both consumers (2.6) and drivers (2.11 and 2.16) to study the dynamics of consumer and driver responses to the price regulation over different time periods. The results are presented in Table 2.7.

Tab. 2.7: Counterfactual Analysis: The Impact of Price Regulation

Day	Change of Numbers				Change of Welfare		
	K Consumers	K Drivers	T Consumers	T drivers	Consumers	K Driver	T Driver
Weekday	−56.82%	−26.73%	−30.12%	−22.16%	**-43.72%**	**-11.20%**	**-15.89%**
Weekend	−42.02%	−19.33%	−23.61%	−14.58%	**-35.39%**	**8.01%**	**-10.16%**
Avg.	−54.14%	−24.20%	−25.78%	−17.51%	**-39.84%**	**-5.66%**	**-13.29%**

As shown in Table 2.7, when the price of Kuaiche is regulated to be no lower than that of Taxi, the customer welfare will decrease significantly in both weekdays and weekends with an average reduction of 39.84%. The decrease is caused by the equilibrium of inter-dependent factors. First, the the increase of Kuaiche price as a result of the regulation reduces consumer's welfare directly. Second, higher prices lead to a drop in the number of Kuaiche consumers, which in turn attracts less Kuaiche drivers to be active to drive. Although some Kuaiche consumers may benefit from shorter waiting times due to the faster decrease of the number of Kuaiche consumers (demand) than the number of active Kuaiche drivers (supply), price regulation dramatically decrease consumer welfare as a whole.

With increased Kuaiche price, Kuaiche driver's welfare gains 8.01% on weekends, while losses 11.20% on weekdays. On average there is a 5.66% decrease in Kuaiche driver welfare. Notice that the driver welfare changes in opposite direction during weekdays and weekends. The reason could be that the total volume of Kuaiche consumers is higher during weekends, even though the increased price would reduce the total number of consumers taking Kuaiche, there are still high enough volume so that Kuaiche drivers can still earn more given the higher price.

The Taxi driver's welfare, however, is negatively affected by the price regulation on Kuaiche with a average loss of 13.29%. Although the Kuaiche price is regulated to be no less than Taxi price, which would discourage some consumers from taking Kuaiche, this group of consumers is less likely to switch to Taxi because they are price sensitive and Taxi has the same price as Kuaiche in this case. As a matter of fact, the increase of Kuaiche price would discourage a significant amount of consumers from taking Kuaiche but its impact on Kuaiche driver's willingness to drive is not as significant. Consequently, the excess capacity or supply of Kuaiche drivers relative to reduced number of Kuaiche consumers, which likely results in less waiting time, will actually attract some Taxi consumers to switch to Kuaiche, and thus squeeze Taxi's market share on Didi platform. Therefore, our counterfactual analysis shows that price regulation of Kuaiche would not benefit Taxi as expected by many regulators and general public.

The decrease in the number of customers together with the increase??? in the number of Kuaiche drivers on the street seems to balance the supply with the demand. However, most mismatches occurs during peak hours, when the price regulation on Kuaiche price to be no lower than Taxi will not help much because the average Kuaiche price charged during peak hour is usually higher than Taxi's due to surge pricing. During non-peak hour, however, the price regulation on Kuaiche will lead to inefficient matching of supply and demand of Kuaiche as customers have less incentive to use Kuaiche due to the higher price, yet more Kuaiche drivers are active on the street lured by the higher price. In our counterfactual analysis, with the price regulation on Kuaiche, we observe a decrease in the number of customers who

choose Kuaiche as their means of transportation. On the supply side, less drivers are expected to show up on the street. This is especially true when the demand far out numbers supply due to inclement weather or other exogenous shocks for which Didi has to set price high enough to encourage more drivers to balance the demand. The insight behind the results of this analysis is that price regulation by itself will lead to inefficiency when supply matches demand with the price coordinated by the market instead of by the regulator.

2.5.0.3 Capacity Regulation

Another major complain against ride-sharing platforms around the world is that they attract too many drivers on the street. Some cities including Beijing in our study have passed regulations to limit the maximum number of registered drivers a ride-sharing platform can have. The government has more concerns about the legitimacy of the ride sharing platform and the potential risks of its fast expansion. Currently, Kuaiche drivers in Beijing need a formal license to legally provide riding service on Didi which results from local government's concern that the ride sharing market is over-saturated. However, among many other cities, the capacity regulation is as strict as Beijing and we want to study the effect of such changes. In this section, we present a counterfactual analysis of the potential impact of a capacity regulation on the ride-sharing platform based on our empirical model and data.

In our counterfactual analysis, we consider a regulation limiting the number of registered Kuaiche drivers in Beijing from the current 95,000 to the number of Taxi cabs, 66,000. Then, we re-solve the demand and supply Equation systems 2.6,

Tab. 2.8: Counterfactual Analysis: The Impact of Capacity Regulation

Day	Change of Numbers				Change of Welfare		
	K Consumers	K Drivers	T Consumers	T drivers	Consumers	K Driver	T Driver
Weekday	−26.17%	−50.01%	−31.30%	−22.43%	**-22.72%**	**-30.49%**	**-14.87%**
Weekend	−16.92%	−33.93%	−19.85%	−15.68%	**-15.62%**	**-17.18%**	**-6.55%**
Avg.	−21.33%	−42.37%	−27.31%	−18.04%	**-18.07%**	**-23.40%**	**-10.93%**

2.11 and 2.16 with the new Kuaiche capacity. The changes in consumer choices (demand) and driver choices of both Kuaiche and Taxi (supply) as well as welfares are listed in Table 2.8.

By limiting the number of Kuaiche drivers from 95,000 to 66,000, we observe on average 18.07% and 23.40% decrease in consumer welfare and and Kuaiche driver welfare, respectively. The shortage of Kuaiche capacity will reduce consumer's willingness to choose Kuaiche which in turn will hurt Kuaiche driver's incentive to drive. Thus, this is a chain-reaction such that the decrease of Kuaiche capacity reduces the number of Kuaiche orders which in turn hurt Kuaiche driver's welfare. Surprisingly, Taxi drivers also experience a loss of welfare as a result of the capacity regulation against Kuaiche. The logic is similar to what happened under the price regulation: the decrease of capacity would in general discourage consumers from using Kuaiche, but affect Kuaiche driver's willingness to drive less dramatically, which could actually improves the demand-supply balance of Kuaiche. This possible improved demand-supply balance of Kuaiche could actually attract some consumers to switch from Taxi to Kuaiche. Also consumers who do not choose Kuaiche due to its lower capacity may choose to hail taxi directly on the streets, which is one of the outside

options. Taxi drivers may choose outside options such as taking orders on the streets than the Didi platform, resulting in a loss of welfare. Thus, the capacity regulation on Kuaiche would lead to welfare losses to all parties in the market, consumers, Kuaiche drivers and Taxi drivers.

In the last a few months many cities in China have established and implemented policies towards Didi and limiting the number of Kuaiche drivers is a main focus. Despite the potential issues brought by the platform, its contributions in mitigating the mismatch between the supply and demand as well as utilizing the idle resource is nonnegligible.

2.5.0.4 Surge Pricing

An important difference between Kuaiche and Taxi is surge pricing. Unlike the fixed prices of Taxi, the Didi platform could increase the price of Kuaiche to respond to short term demand surge on Kuaiche in an area. It is commonly believed that surge pricing can help to relieve the mismatch between supply and demand during high demand periods. This is because higher prices by surging pricing reduce demand, while increase supply of Kuaiche by attracting more drivers to drive. But, how surging pricing affect the welfares of consumers, drivers and the market as a whole is not clear. In this section, we try to provide some insight on these questions by conducting a counterfactual analysis on what could happen if Didi does not apply surge pricing on Kuaiche.

Because we have estimated the surge multiplier $surge_t^K$, then we can force the surged prices back to the normal prices by setting the surge multiplier to 1. We

Tab. 2.9: Counterfactual Analysis: The Impact of Surge Pricing

Day	Hour	Change of Numbers				Change of Welfare		
		K Consumers	K Drivers	T Consumers	T drivers	Consumers	K Driver	T Driver
Weekday	Peak	−25.21%	−19.65%	−15.69%	−18.70%	**-24.83%**	**-19.95%**	**-18.70%**
	Non-Peak	4.96%	3.45%	−5.90%	−5.52%	**5.94%**	**4.43%**	**-4.89%**
Weekend	Peak	−23.31%	−31.89%	16.31%	10.91%	**-25.07%**	**-27.88%**	**12.56%**
	Non-Peak	14.03%	11.81%	−12.71%	−10.74%	**12.13%**	**10.19%**	**-13.82%**
Avg.		−20.65%	−22.93%	−4.99%	−5.51%	**-21.80%**	**-22.02%**	**-6.17%**

then can re-estimate the model to analyze the impact on the demand and supply for both Kuaiche and Taxi, as well as the change of consumer and driver welfare on an individual level.

Table 2.9 summarizes the counterfactual result of no surge pricing. We can see that surge pricing plays an important role in mitigating mismatch between supply and demand such that without surge pricing, all parties including consumers and drivers of both Kuaiche and Taxi would suffer a loss of welfare by 21.80%, 22.02%, and 7.17%, respectively. However, surge pricing does not necessarily make consumers or drivers better off. The effects of surge pricing on consumers and drivers can go either way depending on the exact time period in a day.

As can be seen from the table, during both morning and evening peak hours on weekdays, surge pricing benefits both consumers and drivers in such a way that if there is no surge pricing then every party including consumers and drivers of both Kuaiche and Taxi would experience a loss in their welfare of 24.87%, 19.95%, and 18.70%, respectively. Note that surge pricing frequently happens during peak hours. If we eliminate surge pricing during peak hours, the number of Kuaiche drivers would

decrease since their expected utility of driving decreases due to lower prices. As a result, the mismatch between high demand during peak hours and low supply of Kuaiche drivers will dramatically discourage consumers from choosing Kuaiche. As we can see that the decrease in the number of Kuaiche consumers is more significant than the decrease in the number of Kuaiche drivers so that the mismatch between Kuaiche demand and supply would not become much worse. This together with the lower prices could actually attract a group of Taxi consumers to use Kuaiche instead, and causes a welfare loss to Taxi drivers.

In contrast, the Taxi driver's welfare during peak hours on weekends shows a diverse pattern. Although the welfare of consumers and Kuaiche drivers decrease by 25.07% and 27.88% during the same period, the welfare of Taxi drivers increases 12.56% if surge pricing is not applied. We can see from the table that the numbers of Kuaiche consumers and drivers during peak hours on weekends decrease by 23.31% and 31.89%, respectively without surge pricing. However, the number of Kuaiche drivers decrease more than the number of consumers, which could push some consumers to take Taxi instead. The reason is that consumers with high valuations would find the service level of Kuaiche much lower than before when there is surge pricing and Taxi service will then become attractive As a result consumers are more willing to order a Taxi from the platform and thus Taxi drivers can generate more welfare from it.

During non-peak hours, the impacts of surge pricing are consistent across weekdays and weekends. Without surge pricing, more consumers would choose Kuaiche, which in turn would attract more Kuaiche drivers to drive. In contrast,

without surge pricing, Taxi would become a less attractive option for consumers. The changes in welfare on weekends is more significant than the ones in weekdays because the non-peak hours during weekends have a larger chance of experiencing a demand surge which would trigger surge pricing.

The role of surge pricing depends highly on pattern of the mismatch between demand and supply during various periods of time. We find that surge pricing is able to relieve the mismatch and improve consumers and drivers welfare significantly during peak hours, while in other time in a day, eliminating surge price helps to boost the demand and stimulate drivers' interest of participating.

2.6 Discussion and Conclusion

In this paper, we first build a discrete choice model for two services of Kuaiche and Taxi to analyze the joint effect among the number of open orders around (demand), the number of available driver (supply) and price as well as a range of other factors like weather and different peak hours that affect driver's decision to work and customer's platform choice. And we derive the simultaneous equation system for consumer's market share and driver's capacity share such that we can capture the interdependence between demand and supply.

Utilizing the data including order details and driver information for both Kuaiche and Taxi obtained from Didi, we calibrate customer's and driver's market share and capacity share respectively as well as the price per order for each service over time. We address the simultaneity between demand and supply by

employing the exclusive predictor in each other's function. From the regression results, it's seen that customer's behavior is significantly affected by the capacity of Kuaiche and the price. The higher the Kuaiche's capacity share(or the shorter the distance, or the lower the fare), the higher proportion of the customers will choose Kuaiche platform and vice versa. And the price elasticity depends significantly on different time periods in a day. For example consumers are not sensitive to prices during the morning peak hour from 7 to 9 a.m. while during the non-peak hour from 10 a.m. to 4 p.m. prices impose a significantly negative effect on Kuaiche's market share. In the meantime, the Kuaiche driver's decision of whether to work is significantly affected by the number of customers around as well as the price. A higher number of nearby consumers always encourage drivers to drive. The price, during most time periods, is positively correlated with driver's willingness to drive. However, the night time from 11 p.m. to 6 a.m. and morning non-peak hours of 6-7 a.m. and 9-10 a.m. encourage drivers to work with a lower fare. It is possible that they are target-driven and thus need to drive more hours to reach the goal.

It's shown in the counterfactual analysis that price and capacity regulation alone cannot improve upon the current pricing scheme in which capacity is self-scheduled and incentivized by the price. The problem with price regulation is that the price cap will tear the supply and demand apart in opposite directions, yet it cannot effectively separately affect one side of the market. Also, the capacity cap will harm the social welfare in the sense that it will further lower the supply and lead to a higher percentage of mismatch between supply and demand. Notice that both price and capacity regulations can harm Taxi driver's welfare in a way that

the mismatch between Kuaiche's demand and supply is able to attract consumers from Taxi service because of the excess capacity of Kuaiche. We also discuss the role of surge pricing by analyzing the welfare change without it. It can be shown from the result that overall surge pricing mitigates the mismatch between demand and supply and enables the improvement of both consumer's and driver's welfare. It is also worth to notice that despite of the advantage of surge pricing in a general level it makes consumers and Kuaiche drivers better off by not implementing surge pricing during non-peak hours.

To sum, the government regulations that are most often noted turn out to perform poorly in enhancing the social welfare and balance the demand with the supply. Surge pricing has a significant effect on the enhancement of efficient match between consumers and drivers. A thorough study is needed to explore and design mechanisms that would improve upon the effectiveness generated by the current surge pricing scheme and self-scheduling capacity.

Appendix for

Consumer Equilibrium, Pricing, and Efficiency in Group Buying: Theory and Evidence

2.7 Proofs of Propositions

2.7.0.0.1 Proof of Lemma 1:

Consider a customer with utility $u > p_0 - d$. For such customer $max\{u - p_0, -d\} = u - p_0$, and provided that this customer signs up, he will always stay in after the event, even if neither of the two deals happen. Hence, such a customer arriving at time $t \in [0, 1]$ will sign up if and only if

$$V_k(u,t) = (u-p_1)\pi^1_{k+1}(t)+(u-p_2)\pi^2_{k+1}(t)+(u-p_0)(1-\pi^1_{k+1}(t)-\pi^2_{k+1}(t)) \geq 0\,, \quad (2.32)$$

which holds if and only if

$$u \geq \bar{u}^1_{k,t} \triangleq p_0 - \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t)). \quad (2.33)$$

In a similar manner, a consumer with reservation value $p_1 - d < u \leq p_0 - d$, will drop out after the event only when neither of the two deals materializes, and signs up at time t if and only if

$$\begin{aligned} V_k(u,t) &= (u-p_1)\pi^1_{k+1}(t) + (u-p_2)\pi^2_{k+1}(t) + (-d)(1-\pi^1_{k+1}(t)-\pi^2_{k+1}(t)) \geq 0 \\ &\Leftrightarrow \quad u \geq \bar{u}^2_{k,t} \triangleq p_0 - d + \frac{d - \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))}{\pi^1_{k+1}(t) + \pi^2_{k+1}(t)}; \end{aligned} \quad (2.34)$$

and finally, a customer with utility $u \leq p_1 - d$, will drop out if the second deal does not materialize and signs up if and only if

$$\begin{aligned} V_k(u,t) &= (-d)\pi^1_{k+1}(t) + (u-p_2)\pi^2_{k+1}(t) + (-d)(1-\pi^1_{k+1}(t)-\pi^2_{k+1}(t)) \geq 0 \\ &\Leftrightarrow \quad u \geq \bar{u}^3_{k,t} \triangleq p_0 - 2\delta - d\left(1 - \frac{1}{\pi^2_{k+1}(t)}\right). \end{aligned} \quad (2.35)$$

Note that by (2.33)-(2.35), $\bar{u}^1_{k,t} > p_0 - d$ if and only if $d > \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))$; $p_1 - d < \bar{u}^2_{k,t} \leq p_0 - d$, if and only if $d\pi^2_{k+1}(t) < d \leq \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))$; and $\bar{u}^3_{k,t} \leq p_1 - d$ if and only if $d \leq \delta\pi^2_{k+1}(t)$. Further, $\bar{u}^i_{k,t} \geq p_2$ for $i \in \{1, 2, 3\}$.

Consider a customer with utility u, arriving at time t with k existing sign ups at the time. First suppose $d \leq \delta\pi^2_{k+1}(t)$. Then, since $d \leq \delta\pi^2_{k+1}(t) \leq \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))$, we have $\bar{u}^1_{k,t} \leq p_0 - d$, and $\bar{u}^2_{k,t} \leq p_1 - d$. Therefore, in this case, any customer with reservation value $u \geq \bar{u}^3_{k,t}$ will sign up. Next, consider the case $d\pi^2_{k+1}(t) < d \leq \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))$. Then $\bar{u}^1_{k,t} \leq p_0 - d$, and $\bar{u}^3_{k,t} > p_1 - d$, which means that, all customers with $u \geq p_0 - d$ will sign up and no customer with $u < p_1 - d$ will sign up, and a customer will sign up if an only if $u \geq \bar{u}^2_{k,t}$. Finally, if $d > \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))$, we have $\bar{u}^2_{k,t} \geq p_0 - d$, and $\bar{u}^3_{k,t} \geq p_1 - d$, i.e., no consumer with reservation value $u < p_0 - d$ will sign up. Hence, when $d > \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))$, with k existing arrivals at time t, an arriving customer will sign up if and only if $u \geq \overline{u}^1_{k,t}$.

By (2.33)-(2.35), it then follows that with $k \geq 0$ existing sign ups at time $t \in [0, T]$ an arriving customer with utility u, will sign up if and only if $u \geq \bar{u}_{k,t} \geq p_2$, where

$$\bar{u}_{k,t} = \begin{cases} p_0 - 2\delta - d\left(1 - \dfrac{1}{\pi^2_{k+1}(t)}\right) & \text{if } 0 \leq d \leq \delta\pi^2_{k+1}(t), \\ p_0 - d + \dfrac{d - \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t))}{\pi^1_{k+1}(t) + \pi^2_{k+1}(t)} & \text{if } \delta\pi^2_{k+1}(t) < d \leq \delta\pi^1_{k+1}(t) + 2\delta\pi^2_{k+1}(t), \\ p_0 - \delta(\pi^1_{k+1}(t) + 2\pi^2_{k+1}(t)) & \text{if } d > \delta\pi^1_{k+1}(t) + 2\delta\pi^2_{k+1}(t). \end{cases} \tag{2.36}$$

This completes the proof. ■

Proof of Proposition 1: We start with the derivation of the recursive functional form of $\pi^2_k(t)$. By (1.4), we have

$$\begin{aligned} \pi^2_k(t) &= \int_0^{T-t} (H_k(t+x)\pi^2_{k+1}(t+x) + (1 - H_k(t+x))\pi^2_k(t+x))\lambda e^{-\lambda x} dx \\ &= \int_t^T H_k(u)\pi^2_{k+1}(u)\lambda e^{-\lambda(u-t)} du + \int_t^T (1 - H_k(u))\pi^2_k(u)\lambda e^{-\lambda(u-t)} du \end{aligned} \tag{2.37}$$

Let $g_k(t) = \int_t^T H_k(u)\pi_{k+1}^2(u)\lambda e^{-\lambda(u-t)}du$. Then,

$$\pi_k^2(t) = g_k(t) + \int_t^T (1 - H_k(u))\pi_k^2(u)\lambda e^{-\lambda(u-t)}du = g_k(t) + e^{\lambda t}\int_T^t (H_k(u) - 1)\pi_k^2(u)\lambda e^{-\lambda u}du. \tag{2.38}$$

Taking the derivative with respect to t and by (2.38), we have

$$\begin{aligned}
\frac{\partial \pi_k^2(t)}{\partial t} &= \frac{\partial g_k(t)}{\partial t} + \lambda e^{\lambda t}\int_T^t (H_k(u) - 1)\pi_k^2(u)\lambda e^{-\lambda u}du + e^{\lambda t}(H_k(t) - 1)\pi_k^2(t)\lambda e^{-\lambda t} \\
&= \frac{\partial g_k(t)}{\partial t} + \lambda(\pi_k^2(t) - g_k(t)) + \lambda(H_k(t) - 1)\pi_k^2(t) \\
&= \frac{\partial g_k(t)}{\partial t} - \lambda g_k(t) + \lambda H_k(t)\pi_k^2(t).
\end{aligned} \tag{2.39}$$

Now,

$$\begin{aligned}
\frac{\partial g_k(t)}{\partial t} &= -\lambda e^{\lambda t}\int_T^t H_k(u)\pi_{k+1}^2(u)\lambda e^{-\lambda u}du - e^{\lambda t}H_k(t)\pi_{k+1}^2(t)\lambda e^{-\lambda t} \\
&= \lambda g_k(t) - \lambda H_k(t)\pi_{k+1}^2(t),
\end{aligned} \tag{2.40}$$

which, by substituting into (2.39) yields the recursive differential equation

$$\frac{\partial \pi_k^2(t)}{\partial t} = \lambda H_k(t)\pi_k^2(t) - \lambda H_k(t)\pi_{k+1}^2(t). \tag{2.41}$$

By (2.41), we have

$$e^{-\int_0^u \lambda H_k(v)dv}\left(\frac{\partial \pi_k^2(u)}{\partial u} - \lambda H_k(u)\pi_k^2(u)\right) = -e^{-\int_0^u \lambda H_k(v)dv}\lambda H_k(u)\pi_{k+1}^2(u), \tag{2.42}$$

which implies

$$e^{-\int_0^u \lambda H_k(v)dv}\pi_k^2(u)\Big|_t^T = -\int_t^T e^{-\int_0^u \lambda H_k(v)dv}\lambda H_k(u)\pi_{k+1}^2(u)du. \tag{2.43}$$

By (2.43) and the boundary condition $\pi_k^2(T) = 0$ for all $k \leq M_2 - 1$, it follows that

$$-e^{-\lambda\int_0^t H_k(v)dv}\pi_k^2(t) = -\lambda\int_t^T e^{-\lambda\int_0^u H_k(v)dv}H_k(u)\pi_{k+1}^2(u)du, \tag{2.44}$$

from which we have

$$\pi_k^2(t) = \lambda \int_t^T e^{-\lambda \int_t^u H_k(v)dv} \lambda H_k(u) \pi_{k+1}^2(u) du, \tag{2.45}$$

as stated in (1.6). The corresponding equation for $\pi_k^1(t)$ for $0 \leq k \leq M_1 - 1$, can also be derived similarly.

Now, since $H_k(t) = Pr\{V_k(u,t) \geq 0\}$, and by Lemma 1, a customer with utility u arriving at time t with k existing arrivals will sign up if and only if $u \geq \bar{u}_{k,t}$, where $\bar{u}_{k,t}$ is as defined in (1.3), it follows that $H_k(t) = 1 - F(\bar{u}_{k,t})$. By the boundary conditions, $\pi_k^1(t) = 0$, $\pi_k^2(t) = 1$, and by (1.3), $H_k(t) = 1 - F(p_2)$ for all $t \in [0,T]$, $k \geq M_2$, and hence we have a full unique characterization of (π_k^1, π_k^2, H_k) for $k \geq M_2$. Notice that by (1.3), for all $k \geq 0$, H_k is uniquely determined by π_{k+1}^1 and π_{k+1}^2. Hence, for any k, such that $1 \leq k \leq M_2 - 1$, if we have a full unique characterization of (π_k^1, π_k^2, H_k), we have a full unique characterization of H_{k-1}. Further, given a full unique characterization of $(\pi_k^1, \pi_k^2, H_{k-1})$, by utilizing (1.6) we can uniquely obtain π_{k-1}^2. Further, for all $t \in [0,T]$, for $M_1 \leq k \leq M_2 - 1$, by the boundary conditions, $\pi_k^1(t) = 1 - \pi_k^2(t)$, and for $1 \leq k < M_1$, π_{k-1}^1 can be solved again utilizing (1.6). Therefore, we can obtain π_{k-1}^1, and π_{k-1}^2, which implies that we have a full unique characterization of $(\pi_{k-1}^1, \pi_{k-1}^2, H_{k-1})$. Hence, by backward induction, we have a full unique characterization of (π_k^1, π_k^2, H_k) for all $k \geq 0$. ■

Derivation for the Maximum Likelihood Estimate for Single Pricing: For a product traditionally single-priced at p_s and with the Poission consumer arrival rate λ and consumer utility c.d.f F, the sign up (purchase) process is Poisson with rate $\lambda(1-F(p_s))$, and hence the inter-arrival times for the sign ups are exponentially distributed with mean $1/(\lambda(1 - F(p_s)))$. Hence, if there are N arrivals over a time

period of T, the log-likelihood function for the Maximum Likelihood Estimation is

$$\begin{aligned}
\mathcal{L}(\lambda, F; t_1, \ldots, t_{M_2-1}) &= \log\left(\prod_{j=1}^{N} \lambda(1-F(p_s))e^{-\lambda(1-F(p_s))(t_j-t_{j-1})}\right) \\
&= N\log(\lambda(1-F(p_s))) - \lambda(1-F(p_s))\sum_{j=1}^{N}(t_j - t_{j-1}) \\
&= N\log(\lambda(1-F(p_s))) - \lambda(1-F(p_s))T. \qquad (2.46)
\end{aligned}$$

Defining $z = \lambda(1-F(p_s))$, (2.46) is strictly concave in z with the first order optimality condition $N/T = z$, which implies that the Maximum Likelihood Estimate for λ is $N/((1-F(p_s))T)$. ■

2.8 Category Based Estimation Results

1-Refrigerators							
Arrival Rate (λ)	**No. of Events**	**Min**	**Average**	**Max**	**Standard Deviation**		
	44	22.33	195.65	360.73	89.71		
		Mean			**Standard Deviation**		
Utility Distribution	**No. of Events**	**Min**	**Average**	**Max**	**Min**	**Average**	**Max**
Beta	30	1062.29	3618.66	6399.40	166.64	333.96	480.23
Log-Normal	7	1602.56	3273.54	6160.33	255.27	373.82	473.14
Normal	7	1679.70	2927.02	5183.90	219.69	404.80	521.79
	Pseudo R^2:	Min	Average	Max			
		0.35	0.61	0.87			

2-Air Conditioners							
Arrival Rate (λ)	**No. of Events**	**Min**	**Average**	**Max**	**Standard Deviation**		
	39	39.24	202.64	390.83	96.22		
		Mean			**Standard Deviation**		
Utility Distribution	**No. of Events**	**Min**	**Average**	**Max**	**Min**	**Average**	**Max**
Beta	24	1125.90	3837.67	6365.67	175.91	373.13	481.69
Log-Normal	8	1974.48	3832.79	7231.87	128.91	369.47	514.36
Normal	7	1670.62	3648.18	5541.02	458.05	534.91	578.30
	Pseudo R^2:	Min	Average	Max			
		0.39	0.62	0.87			

3-Television Sets							
Arrival Rate (λ)	**No. of Events**	**Min**	**Average**	**Max**	**Standard Deviation**		
	63	32.25	211.99	408.13	101.02		
		Mean			**Standard Deviation**		
Utility Distribution	**No. of Events**	**Min**	**Average**	**Max**	**Min**	**Average**	**Max**
Beta	37	1055.15	3580.55	6286.38	192.95	369.14	483.99
Log-Normal	12	892.35	4265.38	7084.22	143.71	373.70	511.46
Normal	14	1311.90	3902.11	5835.86	201.49	370.11	577.89
	Pseudo R^2:	Min	Average	Max			
		0.36	0.66	0.86			

4-Water Heaters

Arrival Rate (λ)	**No. of Events**	**Min**	**Average**	**Max**	**Standard Deviation**		
	32	70.71	200.43	376.98	91.40		
		Mean			**Standard Deviation**		
Utility Distribution	**No. of Events**	**Min**	**Average**	**Max**	**Min**	**Average**	**Max**
Beta	23	1002.42	3931.13	6305.64	169.57	348.17	477.84
Log-Normal	3	2665.80	3047.57	3323.54	150.78	308.97	544.22
Normal	6	2093.49	3934.29	5390.35	224.87	400.49	548.57

Pseudo R^2:	Min	Average	Max
	0.39	0.64	0.87

5-Gas Stoves

Arrival Rate (λ)	**No. of Events**	**Min**	**Average**	**Max**	**Standard Deviation**		
	27	29.15	194.44	376.16	101.62		
		Mean			**Standard Deviation**		
Utility Distribution	**No. of Events**	**Min**	**Average**	**Max**	**Min**	**Average**	**Max**
Beta	17	1198.45	3194.38	5788.39	201.85	411.15	485.03
Log-Normal	5	1495.04	3022.77	6429.49	224.37	323.60	453.86
Normal	5	1397.42	2740.75	3955.14	250.91	409.56	538.77

Pseudo R^2:	Min	Average	Max
	0.37	0.70	0.89

6-Washing Machines

Arrival Rate (λ)	**No. of Events**	**Min**	**Average**	**Max**	**Standard Deviation**		
	61	39.13	195.67	411.31	98.23		
		Mean			**Standard Deviation**		
Utility Distribution	**No. of Events**	**Min**	**Average**	**Max**	**Min**	**Average**	**Max**
Beta	38	1206.72	3509.82	6511.37	165.16	379.74	475.27
Log-Normal	14	906.54	2657.82	6018.87	145.14	303.36	528.70
Normal	9	1463.15	3804.47	5520.93	213.72	418.22	587.58

Pseudo R^2:	Min	Average	Max
	0.36	0.62	0.85

Tab. 2.1: Regressions for Consumer Utility Distributions based on Product Category

	Mean			**Standard Deviation**		
Refrigerators	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
(Intercept)	4.9965***	0.5947	<0.0001	4.1107***	0.3832	<0.0001
Capacity	0.5134***	0.1007	<0.0001	0.2138***	0.0649	0.0022
	F-ratio: 26.00, R^2: 0.40			F-ratio: 10.85, R^2: 0.21		
Air Conditioners	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
(Intercept)	3.8223***	1.4121	<0.0001	2.0587*	1.0952	0.0699
Capacity	1.4121***	0.2569	<0.0001	1.0654***	0.2781	0.0006
Energy Level	-0.2452**	0.0955	0.0155	-0.0709	0.1034	0.4979
	F-ratio: 46.37, R^2: 0.74			F-ratio: 15.65, R^2: 0.45		
Television Sets	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
(Intercept)	-3.3524***	0.4399	<0.0001	-2.2321***	0.6830	0.0019
Size	3.0247***	0.1156	<0.0001	2.0404***	0.1795	<0.0001
Resolution	0.1274***	0.0369	0.0011	0.0822	0.0572	0.1571
	F-ratio: 342.20, R^2: 0.92			F-ratio: 64.63, R^2: 0.71		
Water Heaters	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
(Intercept)	5.0525***	0.5810	<0.0001	3.5480***	0.5598	<0.0001
Capacity	0.5513***	0.1275	0.0002	0.3507***	0.1229	0.0086
Power	0.3524***	0.0821	0.0002	0.2255	0.0791	0.0086
	F-ratio: 15.76, R^2: 0.52			F-ratio: 6.91, R^2: 0.30		
Gas Stoves	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
(Intercept)	3.6399***	0.6508	<0.0001	1.6335**	0.7663	0.0456
Power	0.8903***	0.1464	<0.0001	0.8300***	0.1724	0.0001
Panel	0.1192*	0.0579	0.0527	0.0857	0.0682	0.2231
	F-ratio: 23.39, R^2: 0.67			F-ratio: 13.80, R^2: 0.54		
Washing Machines	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
(Intercept)	6.5756***	1.0361	<0.0001	4.5147**	0.2425	<0.0001
Capacity	1.0361***	0.0870	<0.0001	0.7174***	0.1276	<0.0001
Energy level	-0.1085***	0.0299	0.0007	-0.0432	0.0439	0.3300
	F-ratio: 77.10, R^2: 0.74			F-ratio: 16.24, R^2: 0.36		

*:$p < 0.1$, **:$p < 0.05$, ***:$p < 0.01$

2.10 Category Based Price Trend Regressions

Tab. 2.1: Trend regressions for Normalized Deal discounts

$\boldsymbol{p_0/p_s}$	Cluster I_A			Cluster II_A		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
α	1.0400***	0.0066	<0.0001	1.1100***	0.0135	<0.0001
β	0.0002***	0.0001	0.0048	-0.0002***	0.0001	0.0001
	No. of observations: 131, R^2: 0.06			No. of observations: 135, R^2: 0.13		
	F-ratio: 8.23			F-ratio: 19.70		
$\boldsymbol{p_1/p_s}$	Cluster III_A			Cluster IV_A		
	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
α	0.9700***	0.0080	<0.0001	0.9100***	0.0165	<0.0001
β	-0.0002***	0.0001	0.0025	0.0001**	0.0001	0.0147
	No. of observations: 131, R^2: 0.07			No. of observations: 135, R^2: 0.04		
	F-ratio: 9.52			F-ratio: 6.11		
Category $\boldsymbol{\delta/p_s}$	Left Clusters ($k = 1$)			Right Clusters ($k = 2$)		
	Estimate	Std. Error	p-Value	Estimate	Std. Error	p-Value
α_1	0.1438***	0.0250	<0.0001	0.0783	0.0668	0.2433
α_2	0.0233	0.0353	0.4593	0.2173***	0.0611	0.0005
α_3	0.0447**	0.0225	0.0496	0.1930***	0.0407	<0.0001
α_4	0.0856**	0.0289	0.0111	0.2189***	0.0522	<0.0001
α_5	0.0329	0.0300	0.2747	0.3671***	0.1023	0.0005
α_6	0.0376**	0.0227	0.0495	0.1292**	0.0619	0.0474
β_1	-0.0001	0.0002	0.5332	0.0001	0.0002	0.5842
β_2	0.0008***	0.0003	0.0036	-0.0004*	0.0002	0.0934
β_3	0.0005***	0.0002	0.0073	-0.0004***	0.0001	0.0089
β_4	0.0001	0.0002	0.7210	-0.0005***	0.0002	0.0051
β_5	0.0006**	0.0002	0.0148	-0.0009***	0.0003	0.0069
β_6	0.0006***	0.0002	0.0044	-0.0001	0.0002	0.2616
	No. of observations: 131, R^2: 0.25			No. of observations: 135, R^2: 0.23		
	F-ratio: 3.60, p-value: 0.0002			F-ratio: 3.42, p-value: 0.0004		

*:$p < 0.1$, **:$p < 0.05$, ***:$p < 0.01$

References

Aggarwal, C. C. and C. K. Reddy (Eds.) (2013). *Data Clustering: Algorithms and Applications*. Boca Raton, FL: Chapman & Hall/CRC.

Anand, K. S. and R. Aron (2003). Group buying on the web: A comparison of price-discovery mechanisms. *Management Science 49*(11), 1546–1562.

Anderson, S. P., A. De Palma, and J. F. Thisse (1992). *Discrete choice theory of product differentiation*. MIT press.

Argentesi, E. and L. Filistrucchi (2007). Estimating market power in a two-sided market: The case of newspapers. *Journal of Applied Econometrics 22*(7), 1247–1266.

Bai, J., K. C. So, C. S. Tang, X. M. Chen, and H. Wang (2016). Coordinating supply and demand on an on-demand platform: Price, wage, and payout ratio. *Working Paper*.

Banerjee, S., C. Riquelme, and R. Johari (2015). Pricing in ride-share platforms: A queueing-theoretic approach. *Working Paper*.

Ben-Akiva, M. and M. Bierlaire (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, pp. 5–33. Springer.

Ben-Akiva, M. E. and S. R. Lerman (1985). *Discrete choice analysis: theory and application to travel demand*, Volume 9. MIT press.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Bimpikis, K., O. Candogan, and S. Daniela (2016). Spatial pricing in ride-sharing

networks. *Working Paper*.

Bouman, C. A., M. Shapiro, G. Cook, C. B. Atkins, and H. Cheng (1997). Cluster: An unsupervised algorithm for modeling gaussian mixtures.

Buchholz, N. (2015). Spatial equilibrium, search frictions and efficient regulation in the taxi industry. Technical report, Working paper.

Cachon, G. P., K. M. Daniels, and R. Lobel (2015). The role of surge pricing on a service platform with self-scheduling capacity. *Available at SSRN*.

Cameron, A. C. and P. K. Trivedi (2009). *Microeconometrics using stata*, Volume 5. Stata press College Station, TX.

Cao, Z., K.-L. Hui, and H. Xu (2015). When discounts hurt sales: The case of daily-deal markets. Working paper, Hong Kong University of Science and Technology.

Chen, J., X. Chen, and X. Song (2002). Bidder's strategy under group-buying auction on the internet. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 32*(6), 680–690.

Chen, J., X. Chen, and X. Song (2007). Comparison of the group-buying auction and the fixed pricing mechanism. *Decision Support Systems 43*(2), 445–459.

Chen, R. R. and P. Roma (2011). Group buying of competing retailers. *Production and Operations Management 20*(2), 181–197.

Chen, Y. and T. Zhang (2014). Interpersonal bundling. *Management Science 61*(6), 1456–1471.

Chen, Y.-F. and H.-F. Lu (2015). We-commerce: Exploring factors influencing online group-buying intention in taiwan from a conformity perspective. *Asian Journal of Social Psychology 18*(1), 62–75.

Chien-Wei, W. and C. Hsien-Hung (2016). Price discrimination through group buying. *Hitotsubashi Journal of Economics 57*(1), 27–52.

Cohen, P., R. Hahn, J. Hall, S. Levitt, and R. Metcalfe (2016). Using big data to estimate consumer surplus: The case of uber. Technical report, National

Bureau of Economic Research.

Colin Cameron, A. and P. K. Trivedi (1998). Regression analysis of count data. *ECONOMETRIC SOCIETY MONOGRAPHS 30*.

Coxe, S., S. G. West, and L. S. Aiken (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment 91*(2), 121–136.

Edelman, B., S. Jaffe, and S. D. Kominers (2016). To groupon or not to groupon: The profitability of deep discounts. *Marketing Letters 27*(1), 39–53.

Fraiberger, S. P. and A. Sundararajan (2015). Peer-to-peer rental markets in the sharing economy. *NYU Stern School of Business Research Paper*.

Gurvich, I., M. Lariviere, and A. Moreno (2015). Operations in the on-demand economy: Staffing services with self-scheduling capacity. *Available at SSRN 2336514*.

Gwee, M. Y. and K. T. Chang (2013). Developing e-loyalty amongst impulsive buyers via social influence on group buying websites. In *PACIS*, pp. 141.

Hu, M., M. Shi, and J. Wu (2013). Simultaneous versus sequential group-buying mechanisms. *Management Sci. Forthcoming*.

Hu, M. and Y. Zhou (2016). Dynamic type matching. *Working Paper*.

Hu, M. M. and R. S. Winer (2016). The "tipping point" feature of social coupons: An empirical investigation. *International Journal of Research in Marketing*.

Jing, X. and J. Xie (2011). Group buying: a new mechanism for selling through social interactions. *Management Science 57*(8), 1354–1372.

Kauffman, R. J. and B. Wang (2001). New buyers' arrival under dynamic pricing market microstructure: The case of group-buying discounts on the internet. In *System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on*, pp. 10–pp. IEEE.

Li, J., A. Moreno, and D. J. Zhang (2016). Pros vs joes: Agent pricing behavior in the sharing economy. *Working Paper*.

Liang, X., L. Ma, L. Xie, and H. Yan (2014). The informational aspect of the group-buying mechanism. *European Journal of Operational Research 234*(1), 331–340.

Marinesi, S., K. Girotra, and S. Netessine (2016). The operational advantages of threshold discounting offers. *Working paper, INSEAD*.

McFadden, D. (1978). Modeling the choice of residential location. *Transportation Research Record* (673).

Nelder, J. A. and R. J. Baker (1972). Generalized linear models. *Encyclopedia of statistical sciences*.

Ozkan, E. and A. R. Ward (2016). Dynamic matching for real-time ridesharing. *Working Paper*.

Pujari, A. K. (2001). *Data Mining Techniques*. India: Universities press.

Rochet, J.-C. and J. Tirole (2003). Platform competition in two-sided markets. *Journal of the European Economic Association 1*(4), 990–1029.

Rochet, J.-C. and J. Tirole (2006). Two-sided markets: a progress report. *The RAND journal of economics 37*(3), 645–667.

Schmalensee, R. and R. Willig (Eds.) (1989). *Handbook of Industrial Organization*, Volume 2. Oxford, UK: Elsevier.

Song, M. (2013). Estimating platform market power in two-sided markets with an application to magazine advertising.

Subramanian, U. (2012). A theory of social coupons. *Available at SSRN 2103979*.

Talluri, K. T. and G. J. Van Ryzin (2006). *The Theory and Practice of Revenue Management*, Volume 68. New York, NY: Springer Science & Business Media.

Vassilvitskii, S. (2007). *K-Means: Algorithms, Analyses, Experiments*. Stanford, CA: Stanford University.

Weyl, E. G. (2009). The price theory of two-sided markets. *Available at SSRN 1324317*.

Wilson, R. B. (1993). *Nonlinear Pricing*. New York, NY: Oxford University Press.

Wu, C., Y. S. Liang, and X. Chen (2015). Is daily deal a good deal for merchants? an empirical analysis of economic value in the daily deal market (june 30, 2015). *Available at SSRN: http://ssrn.com/abstract=2625901*.

Wu, J., M. Shi, and M. Hu (2014). Threshold effects in online group buying. *Management Science 61*(9), 2025–2040.

Zervas, G., D. Proserpio, and J. Byers (2016). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Boston U. School of Management Research Paper* (2013-16).

Zhang, J. J. and W.-H. S. Tsai (2015). United we shop! chinese consumers' online group buying. *Journal of International Consumer Marketing 27*(1), 54–68.

Zhang, Z. and C. Gu (2015). Effects of consumer social interaction on trust in online group-buying contexts: An empirical study in china. *Journal of Electronic Commerce Research 16*(1), 1.

Zhou, G., K. Xu, and S. S. Liao (2013). Do starting and ending effects in fixed-price group-buying differ? *Electronic Commerce Research and Applications 12*(2), 78–89.