DESIGNING PROMOTIONS TO SCALE MARKETPLACES

Ashish Kabra[†], Elena Belavina[*], Karan Girotra[†]

[†]INSEAD, [*]Chicago-Booth School of Business

ABSTRACT. Achieving scale is key to the efficacy, survival and eventual domination of peer- to-peer marketplaces. Marketplace operators often run aggressive promotions and incentive schemes to attract new users or increase the usage of existing users. This study quantifies and compares the effect of incentives given to the "buyer" side and "seller" side of the marketplace. Specifically, using data from one of the leading ride-hailing marketplaces, we estimate the effect of passenger incentives and driver incentives on number of trips arranged through the marketplace. Further, the incentives can be designed in different formats, i.e. they could be given for every use (linear incentives) or could be given only upon a certain level of use (threshold incentives). We build a structural model to accurately capture the driver and passenger response to incentives, and the nature of incentives. We take into account the effect of service levels on passenger and driver behavior and their endogenous realization in the system, as well as the cross externalities and economies of scale effects. Driver effort on the platform is unobserved, for which we devise a novel local matching model based imputation method. We find that in short-term (current week) passenger incentives are more effective than similar driver incentives. In long-term (next 3 months), the opposite is true; driver incentives are more effective than passenger incentives. This change of effects is determined by the differential stickiness of passengers and drivers to the platform, as well as differential response to evolving service levels. When structuring driver incentives, it is more effective to use threshold incentives compared to linear incentives. The marketplace exhibits substantial economies of scale. For every doubling of passenger and driver numbers, the number of trips more than doubles, it increases by further 25%.
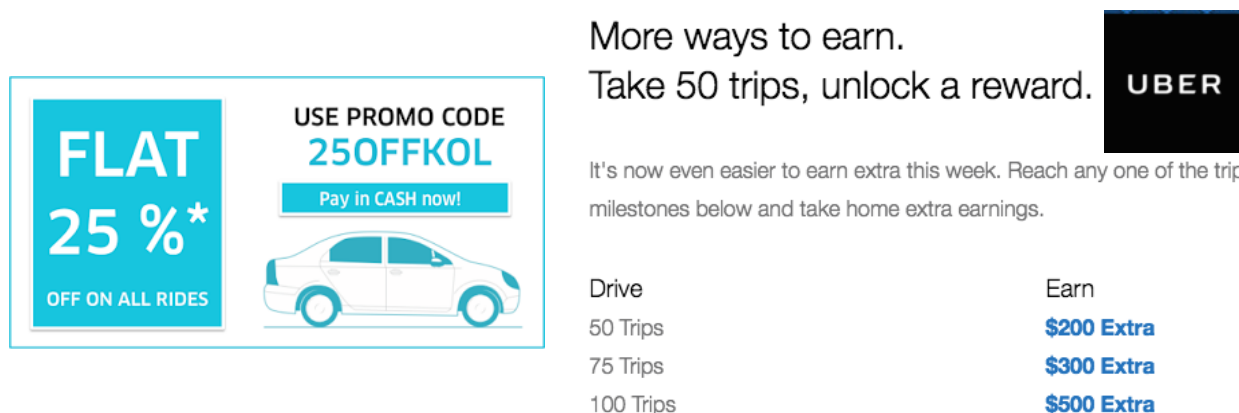
1. INTRODUCTION

On-demand ride-hailing platforms like Uber, and Lyft have radically changed how demand and supply are matched. Uber operates in 77 countries, 527 cities and many other similar services operate in all major cities around the world. These services are digital platform where passengers go to seek drivers and drivers go to seek passengers. These platforms exhibit three key properties which differentiates them from many of the other traditional services (eg: licensed taxi services) or digital marketplaces (eg: ebay). Firstly, the supply side is composed of independent agents. Based on their utility from the marketplace they decide how much to work and when to work. Second,

these marketplaces match users online but the service fulfillment takes place offline. Therefore in-spite of non-remarkable heterogeneities among two users or two drivers characteristics, significant frictions exists in matching them because of the location they are at. Only users who are located nearby each other at a given time, can be matched together. This is related to a third characteristic, that these services are on-demand. Users (passengers) on the platform expect to be matched to a driver in a short-period of time. Passenger requests than cannot be fulfilled within a few minutes are denied by the platform. This makes the notion of service levels key to the platform.

There are two key aspects that these marketplaces exhibit. First, they have significant economies of scale. Let us compare two marketplaces, the first one with ten seeking drivers and ten requesting passengers, while a second one with a hundred drivers and a hundred passengers. Either of them have the same ratio of passengers to drivers. However the platform with a hundred passengers and drivers because of statistical pooling benefits is likely to have passengers and drivers closer to each other and could be matched. Thus, the probability of matching would be higher when the marketplaces has a higher scale. A second key aspect of these marketplaces is that they exhibit cross externalities. Passenger derive a higher utility from using the platform if there is increase in driver use of the platform. Similarly, drivers get a higher utility if there is increase in passenger requests on the platform. The mechanism through which these externalities play out on the platform is service-levels. Passenger service level is the probability of their request being accepted. Driver service level is the frequency of requests he gets on the platform. With increase in use on other side of the platform, the service levels of the focal side improves which increase their use too. Thus when the scale of the marketplace increases, the economies of scale effect and the cross-externalities effect imply that the probability of matches are higher as well as each of the passenger and driver would use the platform much more, leading to a disproportional increase in number of matches or trips handled by the platform. This implies that higher scale in such on-demand marketplaces is quite desirable from passenger, driver and of course the platform perspective.

Indeed, we see that most such marketplaces spend billions of dollars every year on different incentives to passenger and drivers to encourage more of them to join the platform and increase their usage. Uber would spend more than a billion dollars in 2016 on such incentives. Lyft would spend more than 600 million on such incentives. The nature of these incentives for passengers could be a fixed amount or a fraction of trip charge off. For drivers it could be milestone or threshold

TABLE 1. Incentives Example

based payment such as $200 for completing 50 trips in a week. Typical such incentives are shown in Figure 1.

The following statement by one of the senior executives of a leading taxi marketplace in conversation with us sums up the importance efficiently designing these incentives. "The Key to winning in these markets is to make your incentives more efficient". Efficiency of incentives measures the increase in number of trips on the platform for every dollar spent on incentives. There are two ways which can influence the effectiveness of incentives. First is which side of your platform do you give incentives to? If you give incentives to the passenger side of the platform, the utility of passengers from using the platform increases and the marketplaces sees an increase in their participation and usage. That is however not the only effect. Drivers see an increase in the frequency of requests they get and therefore they too participate and use in higher numbers. This increased passenger requests and driver use combined with economies of scale effects leads to an increase in number of system trips. Similarly driver incentives leads to an increase in number of system trips due to the direct effect on drivers, indirect effect on passengers and economies of scale effect. Which of these mechanisms is more efficient relies on understanding the magnitudes of the underlying effects, and is an empirical question. We compare these effects in a short-term horizon, i.e. the effect of system trips in the period in which these incentives are given, and in a long-term horizon, i.e. the effect of these incentives on the evolution of the platform in future periods after the incentives are taken away. For long-term comparison, we use a discounted value of future trips on the platform. The mechanism behind a long-term effect is due to the additional users who are active on the platform

in the future periods even when incentives are discontinued as a result of incentives in current period.

Second influencer of effectiveness of incentives, is that driver incentives could be designed in two different ways which are commonly used in these settings. First is to use a linear structure so that driver gets paid a fixed amount for every trip completed during the incentive period. Second is to use a threshold incentive which awards a relatively higher amount after completing a threshold number of trips. Each of them could imply different efficiency levels depending on their effect on drivers to increase their use and participation and the cost of the incentives.

In this study using detailed data from one of the popular ride-hailing marketplaces, we address the leading questions around offering incentives, which are- who to give incentives to, and how to structure driver incentives. In the process we also identify the economies of scale in this marketplace, which is an interesting and useful measure in itself.

Driver use of the platform is captured in the form of driver effort. The notion of driver effort captures the effort or interest the driver puts in engaging with the system. In our system, this is different than number of hours drivers are online because some of the drivers sign-on and are "online", but still do not accept any or some of the requests. The primary reason is that some drivers don't switch off the app when they are busy serving a passenger they picked up from the street (these are licensed taxis). Our system (unlike Uber) did not penalize taxi drivers in any way for not accepting rides. We need a notion that captures how interested the driver is in using the app. Driver effort serves that purpose—so, in terms of effort a driver who is online 10 hours and accepted 70% of the requests would be equivalent to a more diligent/interested driver who is online 7 hours and accepted all the requests.

We use a structural model to capture the effect of incentives and structure of incentives. Our model has three main components- a passenger side, a driver side and a matching model. Each of the active passengers on the platform is a utility maximizer who determines the number of requests to place as a function of the incentives offered to her and the service level she experiences. The number of active passengers is modeled using another model which determines the number of active passengers in current week as a function of previously active passengers, current passenger incentives and previous period passenger service level. Similarly for drivers, each of the active drivers is a utility maximizer who determines his use level as a function of his incentives and his service level. The number of active drivers responds to previously active drivers, driver incentives

and previous driver service levels. The third main component is a matching model which determines the number of trips in a week as a function of total weekly requests and total weekly driver use. The passenger service level in a week is the probability of a request being accepted. It is modeled as the ratio of total weekly trips and total weekly requests. Since driver use affects the trips in the system, any change in driver use affects passenger service level and therefore affects passenger use of the platform. Similarly, driver service level is the ratio of total trips to total driver use and is affected by passenger requests on the platform.

The primary variation used for identification of incentives and service level effects is weekly variation in incentives offered on the platform and the variation in scale. Service levels are determined endogenously in the system- a higher number of passenger requests would imply a lower passenger service level, similarly higher total driver use would imply a lower driver service level. Therefore the direct estimates of the model would be biased. We use a Generalized Method of Moments based estimation approach for the passenger and driver use models. We use driver incentives as instrumental variables for passenger service levels and passenger incentives as instrumental variables for driver service levels.

A challenge in estimating the above structural model is that the driver effort is not directly observed. The effect of their effort is only observed in terms of number of trips. Separately measuring trips and effort is however crucial for our model to estimate the matching model. We observe that the matching model when looked at a local level, i.e. in a small region and a given time is parameter free. An advantage of the parameter free nature of this local matching model is that given the knowledge of two of three components this model relates, one can impute the value of the third quantity. The model relates the requests, driver effort and trips and we see in our data the local realizations of requests and trips. We thus build a procedure to impute weekly driver effort.

We use individual level passenger adoption, use and leaving data, driver adoption, trips and leaving data from a popular ride-hailing marketplace. To impute the unobserved driver effort we use three additional data sources. We use obtain precise Singapore shape and size data along with latitudes and longitudes of its border using Google maps. To model the locational distribution decisions of drivers in our imputation procedure, we obtain data at frequency of every minute from API's from Singapore Land Transport Authority. This displays the location of all available licensed

taxi drivers in Singapore at the given minute. We also use the time and location of requests and trips from our marketplace data.

We find that there are substantial economies of scale in the platform. If the marketplace has twice as many passenger and drivers, then the number of trips would be more than twice, it would be about 25% higher. This increase comes from increase in passenger and driver service levels by about 15% and an increase in passenger requests and driver effort by 10%. When comparing the short-term effect of similar passenger and driver incentives, we find that it is more efficient to give incentives to the passengers. The average spending on passenger incentives is S$6.77 for every increase in system trip, which is lower than the spending on driver incentives which is S$17.17/trip. In longer-term (discounted number of trips over next 3 months), we find that the opposite of short-term is true- it is more efficient to give incentives to the driver side. The average spending for passenger incentives is 21.96S$/trip, while it is lower, 3.15S$/trip for driver incentives. The objective of managers of these marketplaces have different time-horizons. Short-term gains are sometimes necessary to show value to investors, while long-term gains dictate the health or sustainability of these marketplaces. Our findings cater to each of these needs and interestingly we find different recommendations based on the scenario. The primary drivers of these recommendation is that adoption of passengers is quite responsive to incentives offered. Therefore in short-term we see a higher take-up by passengers and therefore a higher effectiveness. In future periods, passengers are more likely to leave the system when incentives are taken off. In addition, passengers are more responsive to previous passenger service levels in their continuation decisions, while drivers decisions to be active or not is not as responsive to their own service levels. This results in net decrease of passenger service levels when offering passenger incentives, which leads to a relative decrease in the number of active passengers in future. Combined, we see lower long-term benefits of passenger incentives compared to driver incentives. We also compare the effectiveness of structures of driver incentives. We use a linear incentive and two kinds of threshold incentives, one with only one jump and the other with many more jumps. We find that threshold incentives are more effective than linear incentives.

We make three important contributions. We provide first evidence on the effectiveness of different incentives extensively used by most of the on-demand marketplaces that exist today. Our method explicitly accounts for effect of the incentives on the number of active passenger and drivers and usage behavior of passenger and drivers in the system. We carefully model and estimate the effect of the service level which is endogenously determined within the system and affects both passengers

and drivers. In addition we take into account the effect of economies of scale which is substantial in this marketplace. We also devise a novel method to account for the un-observability of effort on one side of the platform. We hope that our analysis and methodology provides a useful template for future research on consumer-behavior in many of the existing and upcoming on-demand models of transportation and delivery related marketplaces.

2. Literature Review

This paper contributes to the fledging research on marketplaces or two-sided markets in operations management, economics and marketing. We also relate to existing research in marketing on the effect of incentives in marketplaces for digital goods. We extend this stream by studying the incentives effect in on-demand marketplaces and bring in the notion of service levels.

Cullen and Farronato [2014] look at the supply responsive and effect on prices for increase in demand on TaskRabbit, an errand outsourcing platform. Li et al. [2015] look at the differences in pricing behavior of professional and non-professional hosts on Airbnb, an online platform for short-term rental of vacant rooms. Ming et al. [2016] measure the inefficiencies due to price and capacity caps for a ride-hailing marketplace. Fradkin [2014] looks at the sources of inefficiencies in matching on Airbnb. Wilbur [2008] measures the externalities between television advertisers and viewers. Buchholz [2015] and Cohen et al. [2016] look at the effect on consumer surplus due to more efficient matching platforms.

Many scholars have used stylized models in the context of ride-hailing platforms and two-sided marketplaces to study pricing issues. Cachon et al. [2015] study the effect of surge pricing on welfare of passengers and drivers. Taylor [2016] studies the effect of time-sensitive passengers and independent driver agents on the prices and wages in ride-hailing marketplaces. Tang et al. [2016] study the optimal prices, wages and their ratios. Banerjee et al. [2015] compare dynamic and static pricing. Gurvich et al. [2015] study the compensation structure to agents of a platform in presence of their strategic scheduling behavior.

Albuquerque et al. [2012] compare the effect of promotional activities and word of mouth effects in an online two-sided market where users can create and purchase content. Singh et al. [2016] compare the uptake of price promotions and philanthropic promotions on ride-hailing platforms.

3. DATA DESCRIPTION

We estimate our model using data from one of the leading ride-hailing marketplace in Singapore. We use detailed data for about 7 months of their operations since their inception. At their peak they had over 10% of the entire ride-hailing marketshare. This marketplace was one of the first major entrants in Singapore. and their major competitors during our data sample period were traditional taxi services. Other major competitors like Uber entered much later, after the end of our data sample.

Operating Model. This marketplace provided a mobile application for passengers who were seeking to get a cab ride instantaneously. Upon placing a request a passenger would be connected to the closest driver who accepts the request. If no driver within a predefined radius accepts the request within a few minutes, then the request is declined. Drivers on the platform are provided with a separate mobile application which allows them to receive passenger requests if they are "online" or signed-on on the app at that time. The first driver to accept a request gets to serve the customer. The drivers on the platform are licensed taxi drivers who use this platform to match with passengers more efficiently. They use the same pricing structure as existing taxi operators which are regulated by the Singaporean Land Transport Authority. These vary slightly across different taxi operators but typically has a booking fee of S$3 when booking online or calling the taxi operator call center, a flag down fee of about S$3 for first 1km or less, and a S$ 0.22 charge for every 400m distance covered thereafter. In addition there are some peak hour surcharges, midnight surcharges and locational surcharge at specific times for trips starting near airport and downtown areas. This marketplace follows the Singapore LTA pricing and charges no additional amount to either its passengers or drivers for access to the platform during our sample period.[1]

Data Statistics. A sample snapshot of our data is presented in Table 2. For every request placed on the platform during the sample period, we observe the unique passenger ID, and the date, time and pick-up location of the request. If the request gets accepted by a driver, we also observe the unique ID of the driver who accepted the request.

Using this data we construct information about use, and active status of each passenger and driver. Our primary unit of time in our analysis is a week. For every week we can construct the number of requests placed by a passenger and the number of trips completed by a driver. The first

---

[1]Later on, much after the end of our sample data period, they introduced a S$ 0.10 charge on drivers for every trip taken on the platform.

Passenger Id	Date	Requested	Driver Id	Accepted	Boarded	Requested Address
1	17/02	0:12:53	1	0:13:24	0:15:55	Outram,
1	17/02	1:13:22				alexandra road lobby
2	17/02	5:31:01	1	5:31:10	5:33:13	Serangoon North Avenue
4	17/02	5:52:20	4	5:52:49	6:08:39	telok Blangah crescent, 17
5	17/02	6:11:48				Jelapang Rd, 505A
1	17/02	6:21:37	6	6:21:44	6:24:57	Prince Edward Road , 8

Table 2. Sample Snapshot of Data

week in which a passenger or driver uses the platform by placing a request or completing a trip is denoted as his or her adoption week. The last week of use of a passenger or driver of the platform is denoted as their week of leaving. They are denoted to be active in between their first and last week of use.

During our sample period, the total number of passenger requests on the platform is over 800,000, while the total number of trips is about 400,000. The unique number of drivers on the platform is over 5,000 while the number of unique passengers is about 80,000.

We use the notion of driver effort to capture driver use of the platform. This captures the effort or interest the driver puts in engaging with the system. In our system, this is different than number of hours drivers are online because some of the drivers sign-on and are "online", but still do not accept any or some of the requests. The primary reason is that some drivers don't switch off the app when they are busy serving a passenger they picked up from the street. We need a notion that captures how interested the driver is in using the app. Driver effort serves that purpose—so, in terms of effort a driver who is online 10 hours and accepted 70% of the requests would be equivalent to a more diligent/interested driver who is online 7 hours and accepted all the requests. Driver effort is thus defined as the number of online hours times the fraction of requests accepted.

Incentives Data. During our sample period, the platform offered several incentives to its passenger and drivers. These incentives were typically offered for a week or several weeks. The nature of incentives to its passenger was in the form of waiving off the S\$3 booking fee that passengers had to pay to the drivers. This was offered twice during the sample period. During the first offering, drivers were not compensated for not receiving booking fee from the passengers. In the second

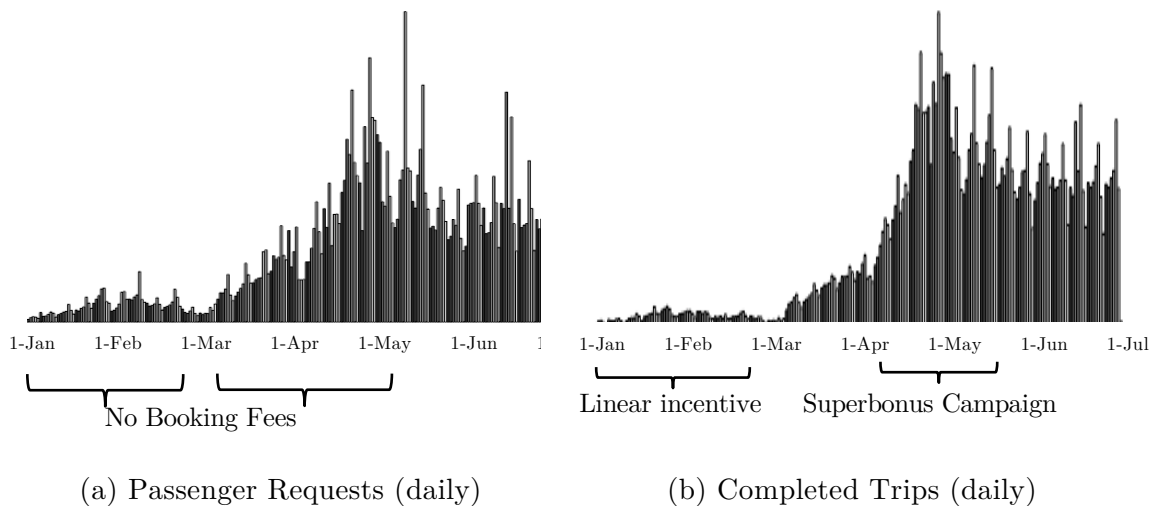(a) Passenger Requests (daily)   (b) Completed Trips (daily)

FIGURE 3.1. Daily Requests and Trips

offering, drivers were compensated for the booking fee waiver to passengers, so that their income from passenger fares remained the same.

Driver incentives were more varied. They were offered with two different structures. First were linear incentives which compensated drivers a constant amount for every additional trip completed on the platform. An example of linear incentive which was offered to drivers for first two months of our data sample is a S\$5 payment for every trip completed on the platform. The other type was threshold incentive which compensated drivers non-linearly with increasing trips. They would offer a relatively large payment only after completing a certain number of trips. An example of this type of incentive is Superbonus campaign which was offered with some variations over weeks for a period of 1.5 months. In this incentive, a S\$100 was given for completing 40 trips in a week, an additional S\$1000 given for completing 100 trips and in addition S\$1.5/trip given for 12 or more trips and S\$2.5/trip given for 24 or more trips. Several other driver incentives of either linear or threshold nature were offered during the sample period.

The sample period thus exhibits sufficient variation in both passenger and driver incentives and in the scale of the platform. The daily volume of requests and trips on the platform is shown in Fig. 3.1.

4. Model

We first present the utility based model for passenger decisions of number of requests in a week and driver decision of amount of effort in a week. We then model the number of active passengers and drivers in a week. We then present the matching model which determines the number of system trips as a function of total passenger requests and total driver effort.

Model of Passenger Requests in a week. An active passenger i has β_{2iw} needs or opportunities in a week w when she can place a request for a trip using the platform. On each of these request opportunities she decides wether or not to use the platform. This is based on a comparison of utilities from placing a request on the platform and of the outside option. An opportunity k has a direct value S_k from completing the trip which is heterogeneous across all opportunities and passengers and follows an exponential distribution with rate ν. The cost of this trip is $p - p_w^r$, where p is the average price of a trip and p_w^r is the incentive amount given to passengers in current week. The probability of the request being accepted or passenger service level is q_w^r. Cost of placing each request is c, and ξ_w^r are the unobservable component in passenger trip value in week w. The utility of passenger i, placing a request in week w is given by u_{iwk}, which is formulated as,

$$u_{iwk} = (S_k - \beta_1 \cdot (p - p_w^r) + \xi_w^r) \cdot q_w^r - c \tag{4.1}$$

The passenger also has access to an outside option which could be a public transport, walking, etc. The utility of that option is u_0^r.

The probability of using the platform on a request opportunity in week w is

$$\begin{aligned}
p_{iw} &= \Pr\left(u_{iwk} \geq u_0^r\right) \\
&= \exp\left(-\nu\left(\frac{c + u_0^r}{q_w^r} + \beta_1 \cdot (p - p_w^r) - \xi_w^r\right)\right)
\end{aligned} \tag{4.2}$$

The number of requests of passenger i in week w is,

$$r_{iw} \sim Bin\left(\beta_{2iw}^r, p_{iw}\right) \tag{4.3}$$

β_{2iw}^r is distributed as $\sim \Gamma\left(\mu_2^r, \sigma_2^r\right)$. We use a gamma distribution here for its long-tail properties which is evident in our data. Each of the β_{2iw}^r opportunities is then requested with probability p_{iw}.

The total number of requests in week w is,

$$tr_w = \sum_{i \in Active\_Passengers_w} r_{iw} \tag{4.4}$$

where $Active\_Passengers_w$ is the set of passengers who are active in week w.

Model of Driver Effort in a week. An active driver j puts in an effort of h_{jw} hours in a week w. The distribution of number of trips completed in h_{jw} hours is given by

$$t_{jw} \sim \Gamma \left(h_{jw} \cdot \mu^m, \sigma^m \right) \cdot q_w^d$$

i.e. t_{jw} is Gamma distributed with shape parameter $h_{jw} \cdot \mu^m$ and scale parameter $\sigma^m \cdot q_w^d$. This is based on trip distribution being $\Gamma \left(\mu^m, \sigma^m \right)$ for unit effort and unit service level. Adding of h_{jw} such pieces results in increase of the shape parameter, while q_w^d service level results in increase of the scale parameter. The choice of gamma distribution is motivated by two of its properties, firstly its long-tail behavior which is reflective of the behavior seen in our data, and secondly the addition of identical gamma distributions being a gamma distribution.

A driver decides his effort level using by maximizing a utility model. The expected utility from effort h_{jw} in week w is given by,

$$u_{jw} \left(h_{jw} \right) = \beta_0 + \underbrace{\beta_1^d \cdot \int_t p_w^d \left(t \right) \cdot f \left(t;\, h_{jw} \cdot \mu^m, \sigma^m \cdot q_w^d \right) \cdot dt}_{expected\_incentive\_revenue} + \underbrace{\beta_2^d \cdot h_{jw} \cdot q_w^d}_{revenue\_fares} + \underbrace{\beta_3^d \cdot h_{jw} + \beta_{4jw}^d \cdot h_{jw}^2}_{quadratic\_cost}$$

$$\tag{4.5}$$

$$+ \quad \underbrace{\xi_w^d \cdot h_{jw}}_{unobserved\_shock}$$

This is composed of an expected incentive revenue component, an expected revenue component from fares from passengers and a quadratic cost of driving. The expected incentive revenue is an integral over the incentive amount from t trips, $p_w^d \left(t \right)$, multiplied with the probability of driver j finishing t trips in week w. The expected revenue from passenger fares scales linearly in effort levels and driver service levels. The marginal cost of driver effort is increasing in his effort level as he has limited time in a day and therefore every additional unit of effort is more costly. Additionally, to increase his effort levels, he might have to forgo other revenue making opportunities (e.g. from

picking up fares from streets). Drivers are heterogeneous in their quadratic cost parameters, i.e. $\beta_{4jw}^d \sim \text{Inv-Gamma}\left(\mu_4^d, \sigma_4^d\right)$. Finally, there is an unobserved component ξ_w^d in driver cost which affects his effort level.

The expected utility from outside option is u_0^d. Driver j chooses his effort level h_{jw} such that

$$u_{jw}\left(h_{jw}\right) \geq u_{jw}\left(h'_{jw}\right) \forall h'_{jw} \tag{4.6}$$

$$u_{jw}\left(h_{jw}\right) \geq u_0^d$$

If no such h_{jw} exists, he chooses 0 effort level and the outside option.

The total effort of all drivers in a week is,

$$td_w = \sum_{j \in Active\_Drivers_w} h_{jw} \tag{4.7}$$

where $Active\_Drivers_w$ is the set of drivers who are active in week w.

Model of Trips in a week. The number of matches between passenger requests and available drivers determines the number of trips on the platform. Trips in week w, tt_w are given by the following Cobb–Douglas production function formulation constrained above by total requests tr_w.

$$\ln\left(tt_w\right) = \min\left(\alpha_0 + \alpha_r \cdot \ln\left(tr_w\right) + \alpha_d \cdot \ln\left(td_w\right) + \xi_w^m, \ln\left(tr_w\right)\right) \tag{4.8}$$

The matching parameters α_0, α_r and α_d capture many of the city characteristics such as its size and shape which determine the efficiency of matching. In addition they determine the economies of scale that are present within this system. If α_r and α_d add up-to more than 1, then this platform has positive economies of scaling. ξ_w^m is the unobserved shock in weekly trips.

Model of Service Levels. In counterfactual analysis, we will allow the service levels to be determined within the system. The passenger service levels are given by

$$q_w^r = \frac{tt_w}{tr_w}.$$

Its effect on passenger use reflects the cross-externality from the driver side through the total trips tt_w, and the negative externalities or the congestion effects from higher usage on the passenger side through inverse proportionality to total requests tr_w.

Similarly, the driver service levels are given by

$$q_w^d = \frac{tt_w}{td_w}$$

Model of Active Passengers in a week. *Active Passengers* on the platform in a week are passengers who have joined the platform in current or earlier weeks and have not yet left the platform. We model the number of these active passengers to be determined by three factors. Firstly, because of word of mouth effects, more passengers become aware of the platform and therefore actively use the platform in higher numbers when there are more active passengers on the platform. Also many of the previously active passengers might continue using the system. The net effect is captured in the coefficient δ_1^r below. Secondly, the incentives offered to passengers encourages more of them to actively use the platform. Thirdly, higher the passenger service level in previous week, higher the number of passengers who would be willing to use the platform in current week. Thus the number of *active passengers* ta_w^r is given by,

$$\ln\left(ta_w^r\right) = \delta_0^r + \delta_1^r \cdot \ln\left(ta_{w-1}^r\right) + \delta_2^r \cdot p_w^r + \delta_3^r \cdot \ln\left(q_{w-1}^r\right) + \delta_\epsilon^r \tag{4.9}$$

where ta_{w-1}^r is the number of active passengers in week $w-1$.

Model of Active Drivers in a week. Number of *active drivers* on the platform are modeled in a similar way as number of active passengers. They are determined by three similar factors. The number of drivers who join and actively use the platform in a week w are determined by word of mouth and continuation effect from the existing active drivers, incentives offered to drivers, and the service level experienced by drivers in previous week. The incentive effect is captured by a measure of expected average dollar amount earned per trip completed in week w denote by ep_w^d. Drivers base it on the service level experienced in previous week. Thus the number of active drivers ta_w^d is given by,

$$\ln\left(ta_w^d\right) = \delta_0^d + \delta_1^d \cdot \ln\left(ta_{w-1}^d\right) + \delta_2^d \cdot \left(ep_w^d\right) + \delta_3^d \cdot \ln\left(q_{w-1}^d\right) + \delta_\epsilon^d \tag{4.10}$$

where ta_{w-1}^d is the number of active drivers in week $w-1$.

5. ESTIMATION

We use a two-step Maximum Likelihood Estimation and a Generalized Method of Moments method to estimate the passenger requests and driver effort equations. The trips equation, active passengers and active drivers model are estimated using an OLS method after re-writing them in logged form.

Un-observability of driver effort. A challenge with estimation of the driver effort model is that the driver effort is not directly observed. Only the outcome of their effort is reflected in the number of trips they complete, which however is also a function of the number of requests on the platform, a relationship which is unknown and we seek to estimate. The same un-observability also impacts the estimation of trips model which depends on total driver effort in a week. A possible solution to this un-observability could have been to jointly estimate the driver effort and the trips model. This could be executed by substituting Eq. 4.7 in Eq. 4.8. A challenge with this approach is that there is no fundamental variation in the data the can separately identify the economies of scale parameters (α_r, α_d) from driver service level parameter β_2^d. One way to see this challenge is: an increase in total number of requests could explain an increase in number of trips via two different channels. The first is that β_2^d is 0, i.e. driver effort remains the same even when total requests have increased, and that α_r is high enough so that weekly trips are quite responsive to an increase in total requests. A second way is that β_2^d is high so that drivers increase their effort levels due to increase in passenger requests, while α_r and α_d are moderate. Either of these channels can explain the observed data, and without observing driver effort we cannot tell apart which is the true underlying mechanism.

We devise a novel solution to this problem by observing that the trips matching model at a regional level at a given time is parameter free. This is because the number of trips in a local region (say a square grid of length equal to 15 minute cab ride distance) at a given time is equal to the minimum of the number of requests in that local region and time, and the number of available drivers in this region and time. That is,

$$tt_{l,t} = \min\left(tr_{l,t}, \, td_{l,t}\right)$$

where l denotes a local region, t denotes a time instance, $tt_{l,t}$ denotes number of trips, $tr_{l,t}$ denotes number of requests and $td_{l,t}$ the number of accepting drivers in region l and time t. The parameter

free nature of this equation implies that knowing $tt_{l,t}$ and $tr_{l,t}$ allows one to estimate $\mathbb{E}\left[td_{l,t}\right]$ and therefore $E\left[td_w\right]$. Details of the precise algorithm used are given in Appendix A.

GMM Estimation of Passenger Requests Model. We estimate our Passenger Request Model (Eq's 4.1-4.4) using a Generalized Method of Moments (Hansen [1982]).

For identification, ν is restricted to 1.

Endogeneity and Instruments. Passenger service levels q_w^r are inversely proportional to total number of requests in the system tr_w, and therefore the dependent variable r_{iw} and its predictor q_w^r are jointly determined in our system. This has a potential to bias our estimates. We include instruments based on driver incentives which affect the passenger service level q_w^r but are uncorrelated with shocks ξ_w^r, and are therefore valid instruments. Since driver incentives often have a threshold structure, our instruments are $\left\{p_w^d\left(t\right)|t \in t_{ins}\right\}$, where $p_w^d\left(t\right)$ is the total driver incentive amount for finishing t trips, and $t_{ins} = \{12, 40, 100\}$ which are the typical thresholds used in our data. In addition, the set of instruments includes an intercept term, and p_w^r, which is the incentives offered to passengers.

The set of moment conditions are:

based on the intercept term

$$\mathbb{E}_w\left[\xi_w^d\right] = 0;$$

based on passenger incentives instrumenting for themselves

$$\mathbb{E}_w\left[p_w^r \cdot \xi_w^d\right] = 0;$$

based on driver incentives instrumenting for passenger service levels

$$\mathbb{E}_w\left[p_w^d\left(t\right) \cdot \xi_w^d\right] = 0 \forall t \in t_{ins};$$

where $t_{ins} = \{12, 40, 100\}$ are typical threshold values as explained in the previous paragraph.

In addition, for efficient identification of distribution parameters, we include micro-moment conditions similar in principle to that of Petrin [2001], so that the predicted probability of number of requests being between certain r_1 and r_2 is in expectation same as observed probability of requests being in the range r_1 and r_2 in a week w.

$$\Pr{}_{w}\left(r_1 \leq r_{iw} \leq r_2\right) = \Pr{}_{w}\left(r_1 \leq \hat{r}_{iw} \leq r_2\right)$$

We choose the value of $\{r_1, r_2\}$ as $\{0, 0\}, \{1, 5\}, \{6, 10\}, \{11, \infty\}$.

Finally, we include balance conditions in the style of Berry et al. [1995] to estimate the unobserved terms ξ_w^d by equating the total requests in a week to observed requests in that week,

$$tr_w\left(\vec{\beta}, \mu_2^r, \sigma_2^r, \xi_w^r\right) = \hat{tr}_w.$$

GMM Estimation of Driver Effort Model. Similarly, we estimate our Driver Effort Model (Eq's 4.5-4.7) using a Generalized Method of Moments.

Similar to Passenger Requests Model, the driver service levels q_w^d are endogenously determined in the system. We use instruments based on passenger incentives to get unbiased estimates, i.e., $p_w^r \in Z_w^d$. Other instruments we use are an intercept, and driver incentive instruments $\left\{p_w^d\left(t\right) | t \in t_{ins}\right\}$, where $t_{ins} = \{12, 40, 100\}$.

Our set of moment conditions are:

based on intercept term,

$$\mathbb{E}_w\left[\xi_w^d\right] = 0;$$

driver incentives instrumenting for themselves,

$$\mathbb{E}_w\left[p_w^d\left(t\right) \cdot \xi_w^d\right] = 0 \forall t \in t_{ins};$$

passenger incentives instrumenting for endogenous driver service levels

$$\mathbb{E}_w\left[p_w^r \cdot \xi_w^d\right] = 0;$$

micro-moment conditions

$$\Pr{}_{w}\left(d_1 \leq t_{jw} \leq d_2\right) = \Pr{}_{w}\left(d_1 \leq \hat{t}_{jw} \leq d_2\right)$$

where we choose the value of $\{d_1, d_2\}$ as $\{0, 0\}, \{1, 10\}, \{11, 50\}, \{51, 90\}, \{91, 110\}, \{111, \infty\}$.

and balance conditions

$$tt_w\left(\vec{\beta}_.^d, \mu_4^d, \sigma_4^d; \xi_w^d\right) = \hat{tt}_w.$$

For identification σ_4^d is restricted to 1. Parameter σ^m which is used to model the variance of mapping from driver effort to trips is calibrated using the local matching model. We use trips as dependent variables instead of imputed driver hours, as at individual driver level the imputed hours would have some measurement error. We used the imputed total weekly driver effort data in estimating trips model which serves as a bridge in this model for writing the likelihood of trips as a function of the parameter values in the model.

Endogeneity and Instruments. Similar to Passenger Requests Model, the driver service levels q_w^d are endogenously determined in the system. We use instruments based on passenger incentives to get unbiased estimates, i.e., $p_w^r \in Z_w^d$. Other instruments we use are an intercept, and driver incentive instruments $\left\{ p_w^d\left(t\right) | t \in t_{ins} \right\}$, where $t_{ins} = \{12, 40, 100\}$.

Estimation of Trips Model and Active users' Models. To estimate the trips model in Eq. 4.8, we estimate it using the OLS method. The upper constraint of total requests on trips only plays a role in counterfactual analysis.

$$\ln\left(tt_w\right) = \alpha_0 + \alpha_r \cdot \ln\left(tr_w\right) + \alpha_d \cdot \ln\left(td_w\right) + \xi_w^m$$

The active passengers and drivers models (Eq's 4.9,4.10) are estimated using a 2SLS method to account for endogeneity of service levels. We use similar instruments as those used in usage models; passenger service levels are instrumented using driver incentives and driver service levels are instrumented using passenger service levels.

$$\ln\left(ta_w^r\right) = \delta_0^r + \delta_1^r \cdot \ln\left(ta_{w-1}^r\right) + \delta_2^r \cdot p_w^r + \delta_3^r \cdot \ln\left(q_{w-1}^r\right) + \delta_\epsilon^r$$

$$\ln\left(ta_w^d\right) = \delta_0^d + \delta_1^d \cdot \ln\left(ta_{w-1}^d\right) + \delta_2^d \cdot ep_w^d + \delta_3^d \cdot \ln\left(q_{w-1}^d\right) + \delta_\epsilon^d$$

6. Results

Results and Marginal Effects. The estimates for our system of equations are presented in Table 3. We report the marginal effects in Table 4. The trips model estimates indicate that there are substantial economies of scale. The sum of α^r and α^d coefficients is 1.172. This implies for a 10% increase in total number of requests in a week and total driver effort, the number of trips increases by more than 10%, about 18% more. In next subsection we further explore the effect of economies of scale.

Passenger Requests Model

β_1	$c + u_0^r$	μ_2^r	Pseudo-R²
0.240	0.272	1.645	0.931
(0.024)***	(0.046)***	(0.161)***	

Driver E ort Model

$\beta_0^d - u_0^d$	β_1^d	β_2^d	β_3^d	μ_4^d	Pseudo-R²
-5.025	0.802	5.892	2.537	0.362	0.834
(1.125)***	(0.058)***	(0.036)***	(0.062)***	(0.012)***	

Trips Model

α_0	α_r	α_d	Adj-R²
-2.275	0.572	0.600	0.997
(0.295)***	(0.063)***	(0.049)***	

Active Passenger Model

δ_0^r	δ_1^r	δ_2^r	δ_3^r	Adj-R²
0.926	0.850	0.053	1.228	0.888
(0.171)***	(0.021)***	(0.011)***	(0.200)***	

Active Driver Model

δ_0^d	δ_1^d	δ_2^d	δ_3^d	Adj-R²
0.545	0.943	0.012	-0.047	0.577
(0.082)***	(0.014)***	(0.004)**	(0.070)	

TABLE 3. Estimation Results

Economies of Scale Effect. We consider a counterfactual scenario of the effect on number of trips when the number of passengers and drivers on a platform doubles. This affects the passenger and driver service levels on the platform because of economies of scale effect. This further changes the usage levels of passengers and drivers. Service levels too change with changing usage levels, which we take into account.

We find that the expected number of trips in the expanded system increases to more than twice of the earlier system, by 25.71%. This is combined effect of increased use and increased service levels. The passenger service level increases by 14.95%, driver service level by 14.11%, total passenger requests by 9.37% and total driver effort by 10.18%. Figure 6.1 shows these effects.

Passenger Model

	Increase in Number Active Passengers	Increase in Number of Requests by an Active Passenger
10% increase in Passenger Service Level	5.33%	6.02%
3$/trip Passenger Incentive	17.39%	106.70%
10% increase in lagged Number of Active Passengers	8.43%	

Driver Model

	Increase in Number Active Drivers	Increase in Effort by an Active Driver
10% increase in Driver Service Level	0.00%	8.01%
3$/trip Driver Incentive	3.60%	32.72%
10% increase in lagged Number of Active Drivers	9.40%	

Trips Model

	Increase in Number of Trips
10% increase in Total Passenger Requests	5.58%
10% increase in Total Driver Effort	5.89%

TABLE 4. Marginal Effects

Short-term and Long-term effect of incentives. We now compare the effect of passenger and driver incentives on trips in short-term, i.e. the week in which an incentive is given, and in long-term, which is the discounted measure of the increase in number of trips in future weeks because of an incentive given in present week. To compute the short-term effect of passenger incentive in a week, we allow our model to determine the number of active passengers and drivers and their use levels, which results in a change in number of trips. This takes into account effect on new passengers who join the system, and the change in number of requests each passenger makes on the platform. This affects the passenger service level and therefore their requests. Cross-externality results in a change in driver effort levels which again affects the passenger requests because of a
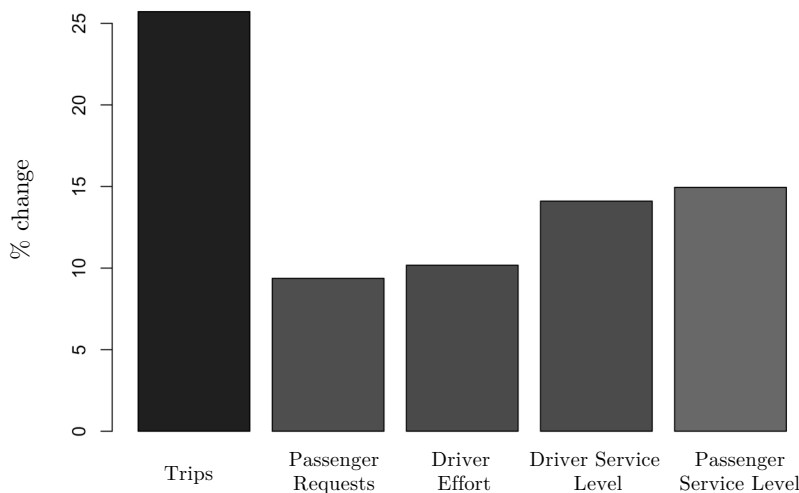
FIGURE 6.1. Counterfactual: Economies of Scale Effect

change in passenger service level. The net effect of passenger incentive on trips takes into account all of these channels. We similarly calculate the short-term effect of driver incentives.

To compute the long-term effect of a passenger incentive, we offer this incentive in a week and discontinue it for future weeks. We then allow our model to predict the evolution of number of active passengers and drivers, changes in service levels of passengers and drivers, and changes in their requests and effort levels. The discounted increase in number of trips over time gives us a measure of long-term effect of passenger incentive. Similarly we calculate the long-term effect of driver incentive.

We use a time period of 3 months (12 weeks) and a discount rate of 2.45% per week to calculate the long-term effect. When offering a 3S$/trip incentive to passengers, we observe a net expected increase of 79.54% in system trips. When offering a 3S$/trip driver incentive, we observe a 21.17% increase in system trips. These incentives affect the platform in long-term through an indirect manner. Even though passengers and drivers derive no directly utility from past incentives, the number of active passengers or drivers in future weeks increases due to past incentives which serves as a channel for the long-term effect. We observe a 0.75% long-term increase in system trips from offering a 3S$/trip passenger incentive and a 3.55% long-term increase in trips from offering a 3S$/trip driver incentive.

To calculate the effectiveness, we calculate the money spent on these incentives in addition to above numbers. The expected number of increase in trips for each dollar spent on incentives gives us the measure of effectiveness of different incentives. Table 5 lists the change in number of trips,

	Short-Term Effect		Long-Term Effect	
	Passenger Incentive 3$/trip	Driver Incentive 3$/trip	Passenger Incentive 3$/trip	Driver Incentive 3$/trip
Cost per Trip Increase	6.77 $/trip	17.17 $/trip	21.96 $/trip	3.15 $/trip
Trips Increase	79.54%	21.17%	0.75%	3.55%

Table 5. Effectiveness of Incentives

and the effectiveness of incentives. We find that in short-term, the passenger incentives are more effective than driver incentives. In long-term however, the effects are reversed, driver incentives are more effective than passenger incentives.

A primary reason of change in the short-term and long-term prescription is that although passengers are more responsive in terms of being active on the platform and higher usage levels in the week when the incentive is offered, they are less sticky on the platform in the longer-term (compare δ_1^r and δ_1^d). In addition passengers are quite responsive to their service levels while drivers are not very responsive to increasing passenger requests. Thus when the passenger incentive is offered, there is an increase in passenger requests and driver efforts, but the increased driver effort is not enough, resulting in decreased passenger service levels. This leads to a further decrease in sustaining the effect of a passenger incentive on longer-term. Drivers on the other hand behave differently. Their active status is less responsive to their service levels, while more passengers actively use the platform and each of them places more requests in response to increased driver effort. This results in a more sustained effect of a driver incentive.

Effect of Incentive Structure. We next compare the effect of different structures of a driver incentive on their efficiency. The two most popular forms of incentives given to drivers in ride-hailing marketplaces are linear and threshold incentives. In a linear incentive, a driver gets a constant dollar amount for each trip he completes on the platform in a week (e.g. S$3/trip). In a threshold incentive, the driver pay is non-linear in trips. An example is that a driver gets paid S$200 upon completing 40 trips in a week and S$500 upon completing 100 trips in a week. In Table 6, we compare the effect of a linear and two different threshold incentives. The linear incentive amount is 3S$/trip. The first threshold incentive is a pay of 30 dollars upon completing 10 trips and 3$/trip for each additional trip after 10 trips in a week. The second threshold incentive is a pay
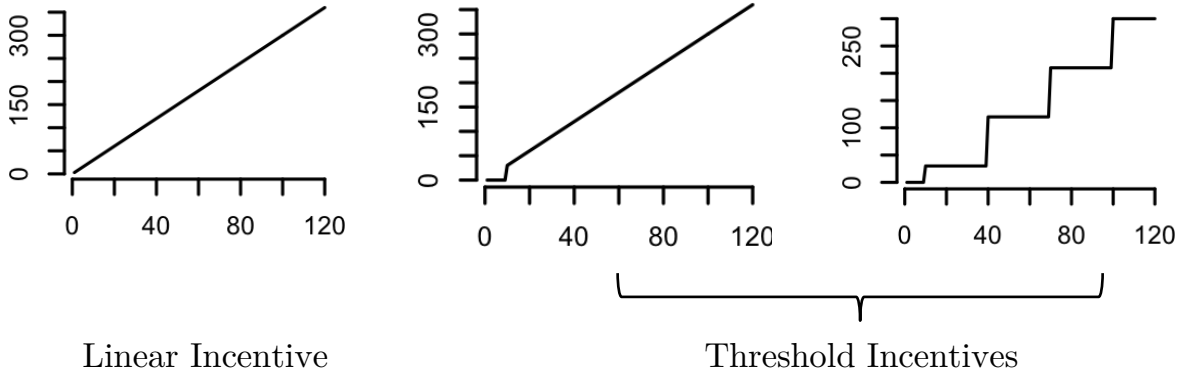
Linear Incentive   Threshold Incentives

FIGURE 6.2. Incentive Structures

	Linear Incentive	Threshold Incentive - I	Threshold Incentive - II
Cost per Increase Trip ($/trip)	17.170	14.039	13.092

TABLE 6. Effect of Incentive Structures

of S\$30 upon completing 10 trips, a pay of S\$120 upon 40 trips, pay of S\$210 upon 70 trips, and S\$300 upon 100 or more trips in a week (Fig. 6.2). We see that when looking at the effectiveness of incentives, threshold incentives are much more effective (by about 31% in this case).

7. DISCUSSION

This paper provides first empirical estimates of passenger and driver response to incentives offered to them. We build and estimate a structural model for the number of active passenger and drivers, and their amount of use, as well as the matching between two sides of this marketplace. The methodology and framework developed here could be used to address similar such question in a variety of different marketplaces. Further, the imputation method devised here based on the local matching model could be applicable in many different spatial on-demand marketplaces where seller interest or effort is unobserved.

Competition between several marketplaces offering competing incentives could likely affect our analysis. Our current setup operates with a competitive model, with the competitor being tradi-tional taxi services. In future work, we plan to work in a setting with more involved competitive dynamics between different marketplaces and study these effects. Further, a larger study similar to ours comparing many cities could provide insight on how some of the demographic and geographic factors shape user preferences and effects.

REFERENCES

P. Albuquerque, P. Pavlidis, U. Chatow, K.-Y. Chen, and Z. Jamal. Evaluating promotional activities in an online two-sided market of user-generated content. *Marketing Science*, 31(3): 406–432, 2012.

S. Banerjee, R. Johari, and C. Riquelme. Pricing in ride-sharing platforms: A queueing-theoretic approach. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 639–639. ACM, 2015.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica*, pages 841–890, 1995.

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

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

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

Z. Cullen and C. Farronato. Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms. 2014.

A. Fradkin. Search frictions and the design of online marketplaces. *Working Paper*, 2014.

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

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

J. Li, A. Moreno, and D. J. Zhang. Agent behavior in the sharing economy: Evidence from airbnb. *Available at SSRN 2708279*, 2015.

L. Ming, T. Tunca, Y. Xu, and W. Zhu. An empirical analysis of price formation, utilization, and value generation in ride sharing services. *Work in Progress*, 2016.

A. Petrin. Quantifying the benefits of new products: The case of the minivan. Technical report, National Bureau of Economic Research, 2001.

J. Singh, N. Teng, and S. Netessine. Philanthropic campaigns and customer behavior: Field experiments in an online taxi booking company. 2016.

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

T. Taylor. On-demand service platforms. *Available at SSRN 2722308*, 2016.

K. C. Wilbur. A two-sided, empirical model of television advertising and viewing markets. *Marketing science*, 27(3):356–378, 2008.

APPENDIX A. IMPUTATION OF DRIVER EFFORT USING LOCAL-MATCHING MODEL

Here we detail the algorithm used by us in imputing the unobserved driver effort.

The city of Singapore is divided in 38 square grids. The grid length is equal to 15 minute driving distance for average Singapore driving speed of about 25km/hr. A day is divided into time windows (tw) of 4 hours (0600-1000, 1000-1400, 1400-1800, 1800-2200, 2200-0200, 0200-0600), which are further divided into time intervals (ti) of 15 minutes each. A driver day shift is 0600-1800 and night shift is 1800-0600. Shifts are indexed by s. Driver j determines the amount of effort (accepting hours) $h_{j,s,y}$ for each day y and shift s. He then allocates these $h_{j,s,y}$ hours among the 4-hours time windows within his shift in proportion to the average passenger demand on that day of the week in our sample. We denote the hours he allocates to a time-window tw on a day y by $h_{tw,y}$. In each of the 15 minute interval on day y and time-window tw, he is online with probability $\frac{h_{tw,y}}{4}$ at grid-location l which is drawn from a distribution $D(tw, dow)$, where dow is the day of the week. We determine driver location distribution $D(tw, dow)$ using data from Singapore government of locations of available taxi-drivers made available at the frequency of 1-minute.

At the start of each time-interval ti, a location l generates $r_{l,ti}$ requests which is observed in our data. For a realization of driver location and online status, $nd_{l,ti}$ is the number of drivers at location l in time-interval ti. The $r_{l,ti}$ requests are randomly assigned to these $nd_{l,ti}$ drivers. Every driver completes one trip in this time-interval if $r_{l,ti} = nd_{l,ti}$. Otherwise there are either unfulfilled requests or vacant drivers. This simulation determines the expected number of trips a driver completes in a shift s on day y. This expectation is a function of the driver effort $h_{j,s,y}$ as well as the effort of other driver $h_{-j,s,y}$. The effort of other drivers is relevant in the simulation as there would be realizations where some of the drivers go empty due to not enough requests in a location and time-interval (congestion effects). We denote expected trips of driver j in shift s and day y by $t_{jsy}(h_{jsy}, h_{-jsy})$. The observed number of trips of driver j in s, y is $\hat{t}_{j,s,y}$. The effort

levels $h_{.,s,y}$ are determined such that for all drivers, the expected trips on a shift-day matches the observed number of trips on a shift-day, that is

$$t_{j,s,y}\left(h^*_{j,s,y}, h^*_{-j,s,y}\right) = \hat{t}_{j,s,y} \ \forall j$$

The driver effort for a week w is then simply given by summing over all shifts and days by

$$h_{jw} = \sum_{s \in \{1,2\}} \sum_{w(y)=w} h_{j,s,y}$$

and total weekly driver effort is

$$td_w = \sum_j h_{jw}$$

This is used in the trips model to estimate α_0, α_r and α_d parameters which are then used in the driver effort model to write the likelihood function.