The Efficiency of A Dynamic Decentralized Two-sided Matching Market

Tracy Xiao Liu
Tsinghua University

Zhixi Wan
University of Hong Kong

Chenyu Yang[*]
University of Maryland

Preliminary

Abstract

This paper empirically studies a decentralized dynamic peer-to-peer matching market. We use data from a leading ride-sharing platform in China to estimate a continuous-time dynamic model of search and match between drivers and passengers. We assess the efficiency of the decentralized market by how much centralized algorithms may improve welfare. We find that a centralized algorithm can increase the number of matches by making matches less frequently and matching agents more assortatively.

1 Introduction

Decentralized two-sided matching markets serve millions of people every year across the world with the rise of the sharing economy. Prominent examples include the accommodation platform Airbnb and ride-sharing platforms such as DiDi, Grab and BlaBlaCar. These markets have several features: (1) agents enter and leave the market over time, (2) agents on one side publicize their willingness to be matched and wait for the other side to choose, and (3) agents have heterogeneous preferences for partners. We use data from DiDi to empirically study the strategic incentive in a decentralized market and quantify the gains from improving the market design. We use the number of matches and average match quality to measure market efficiency.

A decentralized dynamic matching market with non-transferable utility may not operate efficiently due to two sets of externality. For exposition purposes, we refer to the side that publishes the willingness to match as "passengers", and the other side as "drivers". Drivers choose passengers. On one hand, a forward-looking driver deciding whether to match with a passenger does not internalize the benefit of the match for the passenger, and potentially could forgo a match that would otherwise be realized in the social optimum. In other words, a driver could wait too long to form a match compared with the social optimum. On the other hand, a driver could wait too little: a driver does not internalize the negative externality of a match on a competing driver, and

[*] Correspondence to: chny.yang@gmail.com

may form a match with a passenger who might be more compatible with a driver that has not yet arrived to the market.

More broadly, we hope that the analysis will shed light on how to increase the efficiency of a market similar to the one we study. We address this question by exploring how a centralized algorithm could improve welfare upon the equilibrium in a decentralized matching market. A centralized algorithm requires information that may not be known to agents in a decentralized market. We discuss what information is valuable and quantify the value of information in different centralized algorithms. We consider algorithms that are easily implementable. These algorithms improve the market efficiency in two ways. First, a centralized algorithm can match more agents by keeping agents in the market longer and creating additional match opportunities. Secondly, a centralized algorithm can improve match qualities by taking into account the externalities.

We use proprietary driver search and ridership data from DiDi Chuxing to study our research questions. DiDi Chuxing is the leading firm in the enormous Chinese ride-sharing market. DiDi offers tiered ride-sharing platforms. According to the *2017 DiDi Factsheet*, the company's platforms served 450 million passengers in 2017, and 25 million trips were completed every day. The main operation of the company, DiDi Express, is similar to Uber or Lyft, where a central dispatch algorithm matches drivers to passengers. Our empirical context is the smaller and decentralized peer-to-peer platform of DiDi. In 2017, 2.23 million rides occurred on this platform during the peak day.

The DiDi peer-to-peer platform provides a unique setting to study a decentralized matching market. The platform features low fares to target long-distance passengers and non-professional drivers. The drivers are mostly commuters who need to reach a specific destination by a certain time. To receive a ride, a passenger sends a request to the platform. The request consists of a pickup and a dropoff location and a departure time. If a request is answered, the answering driver will deliver the passenger according to the conditions specified in the request. The passenger cannot see available drivers or request a particular driver. The only actions available to the passenger are either to wait or cancel the request. To find and answer the most suitable passenger request, a driver specifies a route on the ride-sharing app that sorts all ride requests according to a compatibility index. Most drivers search for passengers once a day. The drivers usually have a different day job, and they use the service to defray the commuting costs. This empirical context allows us to focus on the effect of the driver strategic choice between waiting for a better match and answering a request now. Throughout the rest of the paper, we specifically refer to the additional driver waiting (compared with a myopic driver) induced by the desire to wait for a more compatible passenger as strategic waiting. We conceptualize the implication of strategic waiting in a simple framework in Section 3. In particular, we show that preference heterogeneity plays a key role in determining the equilibrium number of matches.

To address our research questions and exploit the unusually rich data on driver and passenger behavior, we develop a continuous time dynamic model of search and match (Doraszelski and Judd (2012); Arcidiacono, Bayer, Blevins and Ellickson (2016)). Passengers and drivers with preferences for different routes arrive at the market stochastically. A route is defined as a pair of pickup and dropoff locations. Passengers send out their trip requests consistent with their true preferences immediately after arriving. Passengers and drivers leave the market without a match when they reach their maximum waiting time. The driver receives a reservation utility in this case. The lengths of the maximum waiting time are heterogeneous. While in the market, a driver stochastically receives move opportunities to check her phone, and she can choose to wait or answer a request. If a driver answers a request, the driver receives utility as a function of (1) how much detour the driver would have to travel to reach her final destination compared with traveling alone, (2) the length of the passenger route and (3) unobserved driver-passenger match value, and both the driver and the passenger leave the market. When answering a request, a driver compares the utility from picking up the current most compatible passenger with the option value of waiting for a potentially better future match. To evaluate the option value of waiting, the driver is assumed to know the equilibrium distribution of the value of the best match. We additionally allow for heterogeneous driver discount rates. We define a stationary equilibrium based on this driver dynamic optimization. The distribution of the best match is endogenously determined by the entry rates, the lengths of maximum waiting time and the equilibrium driver strategies. In particular, the distribution of the best match depends on how fast passengers leave the market, which is composed of the exogenous exits when they reach their maximum waiting time and the endogenous exits due to drivers answering requests. Drivers effectively play against a stationary distribution of the best match in this equilibrium.

We use a three-step estimation procedure to recover model primitives. In the first step, we estimate the arrival rates directly from the data on passenger and driver arrivals. Next, we use the variations in the set of waiting passengers to identify driver preferences. In particular, the dispersion of the detour lengths and passenger route lengths of the answered requests conditional on different sets of passengers helps to identify the distribution of the heterogeneous tastes for the detours and passenger route lengths. We also observe that a driver often does not answer an acceptable request immediately: about 40% of the answered requests are available for the answering drivers when these drivers enter the market, but these drivers wait on average 9 minutes before answering. We use the variations in how long the answered requests are kept waiting to identify driver time preferences (the driver discount factor or waiting cost). We flexibly incorporate multiple levels of heterogeneity in driver preferences and use a simulated method of moments estimator to estimate the parameters. Finally, we estimate the distribution of passenger maximum waiting time. Because passengers who wait longer are more likely to be answered, the observed average waiting time of the unmatched passengers would be a biased estimator of the mean maximum waiting time. We thus simulate the full search and match game using the estimated driver parameters and estimate the maximum waiting time distribution parameters with a second simulated method of moments estimator.

Using the estimates, we conduct counterfactual simulations to address our research questions. We first simulate the market with myopic drivers. The comparison with the decentralized equilibrium market reveals the implication of the drivers waiting strategically. We find that although waiting increases market thickness and allows drivers to obtain better matches (higher average driver utility), the option value of waiting also causes drivers to forgo more matches and results in a lower number of matches. Next, we assume that the platform has additional information on agent preferences and the maximum waiting time, and we simulate two centralized matching algorithms. The simulations use the centralized greedy and patient algorithms in Akbarpour, Li and Oveis Gharan (2017), adapted to account for the two-sidedness of the empirical application. With agents ex ante identical in terms of how likely they are compatible with other agents, Akbarpour et al. (2017) shows that the patient algorithm can increase market thickness and the number of matches by keeping agents in the market longer. A main motivation for considering these two algorithms is that the pricing on our platform is pegged to distance by regulation. We thus consider non-price mechanisms that may increase market thickness and reduce no-matches and mismatches. In particular, comparing the greedy and patient algorithms shows the interaction between market thickness, the number of matches and the quality of matches. We find that the patient algorithm generates the highest market thickness and achieves the highest match rate at our estimates. The increase is most significant for shorter route passengers (<40KM), but the match rate slightly decreases for the longer route passengers. In addition, drivers are better off on average with the patient algorithm compared with the decentralized equilibrium.

Contributions and Relations to the Literature

First, we demonstrate that a platform can improve efficiency by increasing the waiting time of consumers and increasing market thickness. In doing so, we measure the returns to investment in soliciting or predicting both the consumer preferences for the quality of matches and the ability to wait. Secondly, we study an important market where millions in China rely on this service for their commuting needs and suggest possible ways to improve this service. Our framework may also apply in a wide variety of peer-to-peer platforms similar to ours: many platforms feature one side who "posts and waits" and another side who "chooses or waits": Airbnb's hosts wait for guests to choose their rooms, babysitters wait for customers on Care.com, and drivers wait for passengers on BlaBlaCar, to name a few.[1]

The market design literature shows that having a thick market is important for efficiency,[2] where market thickness is defined as the number of market participants. Some of the recent work (e.g., Arnosti et al. (2015); Baccara et al. (2016); Loertscher et al. (2016); Ashlagi et al. (2016); Akbarpour et al. (2017)) study optimal dynamic matching and thickness.[3] In particular, Akbarpour et al. (2017) shows reducing the matching frequency can increase the number of matches between ex ante identical agents. Our environment and objective are similar, and we generalize the theoretical

[1] For a survey of the recent literature on the peer-to-peer markets, see, for example, Einav et al. (2016) and Farronato and Fradkin (2018).

[2] See the surveyed literature in, for example, **?** and **?**.

[3] There is also a theoretical literature on kidney exchanges (e.g., Roth et al. (2005, 2007); Ünver (2010)) that examines properties of centralized dynamic matching algorithms.

framework for our empirical setting in two ways: we consider ex ante heterogeneous agents who face (stochastic) deadlines to either make a match or leave the market, and we incorporate and estimate waiting costs. Importantly, the agents in our model are not identical in terms of how likely they are compatible with other agents. In our context, a driver might be compatible with passengers 1 and 2, but another driver might be compatible with just passenger 1. Our empirical results suggest the patient algorithm still generates more matches than the greedy algorithm, although we show in theory the patient algorithm may perform worse at some parameterization.

This study is also related to the empirical matching literature.[4] These papers typically examine long-term relationships, and the matches can reasonably be assumed to satisfy a notion of stability (Roth and Sotomayor (1992); Hatfield, Kominers, Nichifor, Ostrovsky and Westkamp (2013)). Choo (2015) studies the gains of marriage in a frictionless dynamic matching market. Fox (2008) studies repeated matching between forward-looking workers and firms. Relationships in our empirical contexts are typically short-term and agents are unlikely to coordinate to swap partners after the initial match (and post match cancellations are typically discouraged). We use this setting to study the efficiency loss due to the frictions that prevent matches from being stable as defined in the cited work.

There has been recent interest in empirically studying frictions in various dynamic matching markets. One empirical context is the New York City taxi market (Lagos (2003); Frechette et al. (2016); Buchholz (2017), among others), where a chief source of inefficiency is the costly process of taxi drivers physically searching for passengers whose locations are unknown to the drivers. Our focus is to consider frictions that endogenously arise from the strategic choices in the matching game where technology has substantially reduced the cost of physical search. Another context is the allocation of deceased donor kidneys. **?** model a patient's choice to accept a deceased donor kidney in a continuous-time dynamic framework. Their objective is to identify a mechanism that increases the number of matches and patient welfare (i.e. increasing match quality). In that context, increasing the market thickness by increasing agents' waiting time would not be a viable strategy due to the particular matching rule (a prioritized market) and the short shelf life of a kidney (in comparison with how long a patient can survive before transplant).

Road Map In the rest of the paper, we first discuss the empirical context and the data. We next present a simple theoretical model of search and match to highlight when the decentralized market may produce a fewer number of matches than a social planner. We then describe our empirical structural model in detail, followed by the estimation and counterfactual experiments.

2 Empirical Context

We use six weekdays of detailed passenger request and driver search data on the DiDi platform between 4:30PM and 5:00PM in a prefecture-level city (population 5 to 10 million) in southern

[4] See the surveyed literature in, for example, Choo and Seitz (2013); Chiappori and Salanié (2016); Fox (2017).

China in the summer of 2018. We ensured that the weather was similar throughout the sample period (sunny or overcast) and there was no major construction project in the chosen city. The first subsection discusses institutional details. The second subsection presents features of the data that support modeling assumptions.

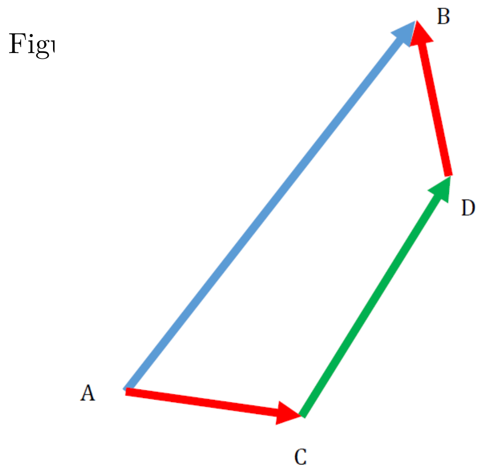
2.1 Institutional Details

On the DiDi platform, a prospective passenger sends a request to the platform to receive a ride. The request consists of the desired pickup location, dropoff location and the pickup time. The passenger does not observe the driver and cannot choose the driver. A prospective driver can view all requests via the platform's app. The first driver that answers a request "wins" the trip. A driver can input her own desired pickup and dropoff locations and sort trip requests by a notion of compatibility. For each request, the driver observes a single percentage compatibility rate calculated by the system, in addition to the detailed specifications of the request. A number of factors go into the calculation of this rate, but the rate is mainly determined by the commonly traveled distance as a proportion of the total driver distance if the driver transports a passenger and heads to the driver final destination: in Fig. 1, suppose a driver i would like to travel $A \to B$ and a passenger j requests $C \to D$. The detour length for i to pick up j: $d_{ij} = AC + CD + DB - AB$, and the fare for the driver is largely determined by $x_j = CD$. The main component of the compatibility index highlighted on the driver end of the app is $\dfrac{x_j}{d_{ij} + AB}$. DiDi's surveys also show that the traffic condition around driver origin and detailed geographical features (e.g. left turns) along the route are also important factors in the driver decision. A driver must reveal its preferred route to the platform to see the list of passengers. We observe the desired origination and destination coordinates of waiting drivers (up to a 0.25 KM^2 rounding). A passenger can cancel the ride at any time before any driver answers the request without penalty. The pricing on the platform is a two-part tariff strictly based on distance.[5] The platform charges passengers \$0.7 for the first 2KM and \$0.14/KM for the rest of the trip. The platform collects 10% of the total charge and the driver earns the rest. In comparison, a taxi costs \$1.3 for the first 3KM and \$0.29/KM for the additional distance.

When a driver answers a ride, a match is formed and there are penalties for the party that cancels the ride. Multiple cancellations can lead to suspension of the account. About 8% of the matches are canceled in our data. According to conversations with DiDi, many of the cancellations appear to be "random", such as the passenger no longer needing the ride, instead of "strategic", such as a driver noticing a better fit.[6] We view increasing the rate at which drivers answer rides

[5] A number of recent papers study pricing on the ride-sharing network. Many of these papers focus on the use of dynamic pricing to increase professional drivers' labor supply and passenger-driver matching on Uber (e.g., Hall et al. (2015, 2017); Castillo et al. (2017) in economics and Banerjee, Riquelme and Johari (2015); Ozkan and Ward (2016); Hu and Zhou (2016); Feng, Kong and Wang (2017) in operation research). Ostrovsky and Schwarz (2018) study how proper road pricing can implement the efficient allocation of passengers to the road capacity in a carpooling context with autonomous cars and non-professional drivers whose waiting cost is low. In this paper, we focus on how to improve the allocative efficiency through non-pricing channels.

[6] The matched driver would have to use a different phone to see the requests in this case.



A driver i would like to travel $A \to B$. A passenger j requests $C \to D$. The detour length for i to pick up j: $d_{ij} = AC + CD + DB - AB$, and the fare for the driver is largely determined by $x_j = CD$.

as a first order issue and leave post-match cancellations for future research.

DiDi's peer-to-peer platform is distinguished by the low fares and the decentralized matching process that cater to commuters (i.e. non-professional drivers). Unlike the taxi drivers (Frechette et al. (2016); Buchholz (2017)), the drivers on the DiDi peer-to-peer platform have a different search problem: they have a larger information set, in the sense that these drivers observe and could answer the request of any nearby (the requested trip origin is within the 10 KM radius area of the searching driver) waiting passengers regardless of the physical distance between the driver and passenger. More importantly, these drivers can act on their preferences by choosing the most compatible passengers (after trading off detour lengths against the value of the fares or the (un)willingness to travel with a stranger). In this context, the focus of our study is on quantifying the extent of the temporal, instead of the spatial, mismatch, and this mismatch is aggravated by the presence of preference heterogeneity.

2.2 Data Summary

Within the half-hour period of our sample, we observe all waiting passengers and drivers across six days. On the passenger side, we observe the geographical coordinates and the desired departure time of the passenger requests. We use requests where passengers ask to leave within one hour, which account for 80% of the total number of requests. We also observe the outcome of the requests: whether the request is answered or canceled by the passenger. On the driver side, we observe the information on drivers who refreshed the passenger list. We thus observe the coordinates of the routes drivers used to search for passengers, the frequency of the search (refresh) and the set of potential passengers in each search. We define the length of the time a driver is in the market as the time between the first and last search. Table 1 reports the summary statistics. About half of the passenger requests (52%) are answered, and 60% of the searching drivers find passengers. We call the percentage of answered requests the "match rate". This rate is low compared to the centralized platforms such as DiDi Express, where almost all passenger requests can be fulfilled, but not uncommon on a decentralized platform: for example, the occupancy rate on Airbnb ranges from 50% to 60% in most major American cities (Andreevska (2016)).

The table also shows that these passengers wait a significant amount of time for a ride on the platform. Conditional on a passenger's request not being answered, the passenger cancels the request at about the 7th minute. Conditional on a request being answered, the request is answered under 5 minutes. In comparison, getting a ride on DiDi Express, Uber or Lyft requires a passenger to wait no more than a few seconds. According to conversations with DiDi, many passengers use the platform as a first choice, and if the requests are not answered by certain time, the passengers can always count on getting a DiDi Express car, a taxi or taking the public transit. We can compute a back-of-the-envelope number for the benefit of waiting: by waiting 5 minutes, with probability 0.5, a passenger going on a 30-KM trip (mean requested trip length) could save

$$\underbrace{\$1.3 + \$0.29 \times (30-3)}_{taxi} - \underbrace{(\$0.7 + \$0.14 \times (30-2))}_{DiDi \text{ peer-to-peer}} = \$4.51,$$

or 49% of the taxi fare. We thus think there is substantial benefit to waiting. Any additional waiting cost likely is psychological, because passengers do not actively choose drivers and waiting does not prevent the passenger from engaging in otherwise productive activities. In our empirical analysis, we estimate driver discount factor on the platform and find the waiting cost of drivers is also quite low under appropriate normalization.

Passenger	Mean	Std
Time in Market (Sec)	439.73	439.16
Time Till Answer (Sec)	279.37	322.07
Percentage Answered	0.52	
Trip length (KM)	30.6	35
# New Requests/10 Sec	4.49	6.59
# Waiting Requests	215.1	16
# Observations	6071	

Driver	Mean	Std
Time in Market (Sec)	449.16	455.86
Time Till Answer (Sec)	514.46	409.81
Percentage Answering	0.58	
Trip length (KM)*	31.43	35.83
# New Drivers/10 Sec	4.02	6.12
# Waiting Drivers	191.9	14
# Observations	5552	

Table 1: Summary Statistics

*: Average Answered Trip Length is 24.5 KM

We also find that, although the sampled half hour is close to what would usually be the evening rush hour, the city traffic conditions seem to be stable from 4:30PM till 6:30PM, right after our sampled period of matching. This observation helps to motivate the assumption of a stationary environment in the model later. We track the travel time on routes between key landmarks (major commercial centers or municipal halls) across the city every 4 minutes. The city can be divided into four regions around four population centers, as shown in Figure 2. The city seat is located in the green area. The red and black areas cover densely populated metro areas and suburbs, and an airport in located in the red area. The magenta region is less densely populated and covers large swaths of suburbs further away from the metro area. To understand the layout of the city, consider

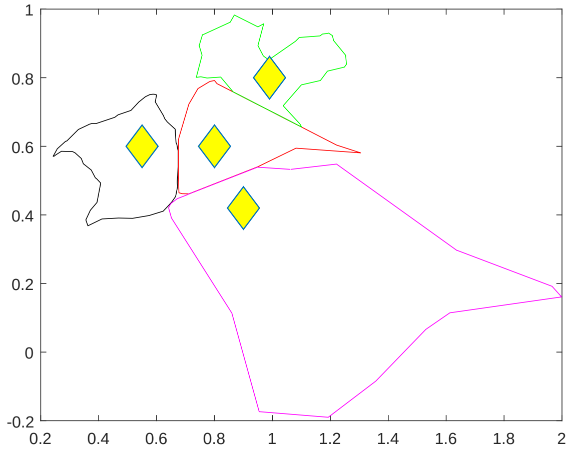


Figure 2: Main Pickup and Drop-off Regions

There are four population centers across the city, and the locations are marked with the yellow diamonds. The city can be divided into the four regions as colored. The city seat is located in the green area. The red and black areas cover densely populated suburbs, and an airport in located in the red area. The magenta region is less densely populated and covers large swaths of suburbs further away from the metro area. We re-scale x and y-axis to avoid disclosing the identity of the city.

an analogy to the Greater Boston area: the green area is the downtown Boston and Cambridge, and the red area is the east and south Boston. Both are political and commercial centers. The black area is the “commuter” region, such as Allston and Newton. The magenta area is the suburb further away, such as Quincy and Braintree. We select 2 landmarks for the black region, 3 for red, and 1 each for the green and magenta region. We first examine the traffic condition across time after aggregating over routes: we plot the average speed in Fig. 3 for traveling at 4:30PM, 4:34PM, ..., 6:30PM across 6 days. The speed slightly declines from 39KM/h to 36KM/h over the 2-hour period. We next examine the traffic condition across routes, after aggregating over time: in Fig. 4, we plot the maximum and minimum travel time across 2 hours against the respective route distance. The average ratio of the maximum to the minimum is 1.12. Considered together, the data suggest that the traffic condition is quite stable for our drivers.

We also find evidence that driver becomes more likely to accept a request as the driver waits longer. We leverage this data pattern to estimate the discount factor in the driver model. We observe that about 75% of the drivers who answer a request do not answer the request during their first search when the answered request is available. If drivers use a cut-off rule to evaluate the fitness of a request, the observation means that the threshold of acceptance decreases over time. In Fig. 5, we plot the histogram of how long an answered request is kept waiting as a proportion of the total wait time of the answering driver, after the driver see the request on the DiDi app. Merely 25% of the drivers who answer requests do so when the drivers see the requests for the first time.

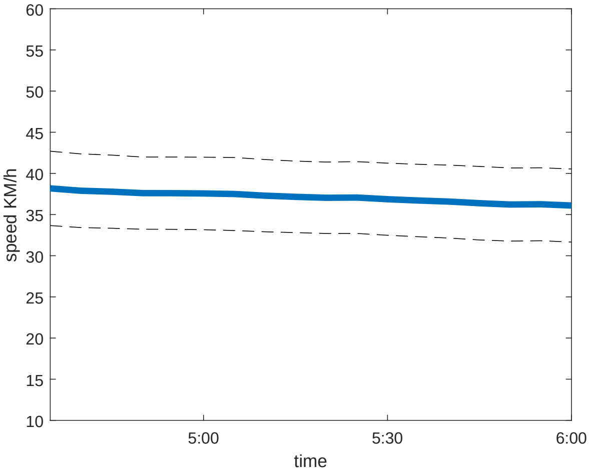


Figure 3: Average Speed, 4:30PM-6:30PM

The average is taken over 6 days for traveling at 4:30PM, 4:34PM, ..., 6:30PM. The dotted lines represent the 95% confidence interval for the average speed across 42 routes.

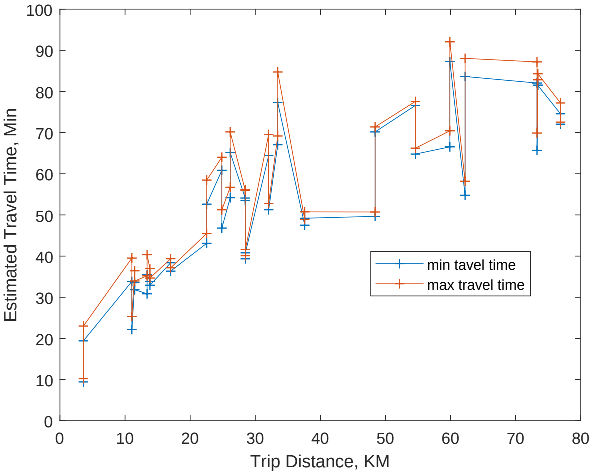


Figure 4: Travel Time Across Routes

The maximum time and minimum travel time on the 42 routes we select between 4:30PM and 6:30PM. The min and max are taken within a route across the travel time measured at 4:30PM, 4:34PM, ..., 6:30PM.

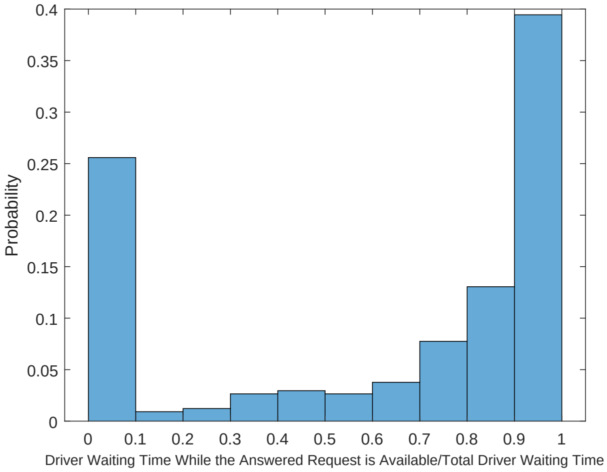


Figure 5: Evidence of Non-stationary Strategy

We show how long an answered request is kept waiting after the answering driver sees the request, as a proportion of the driver wait time.

Conditional on a set of waiting drivers and passengers, the longer time a driver keeps a request waiting suggests that the driver discounts the future less. We use this data feature to identify the discount rate, or the waiting cost, of the drivers.

To aid further analysis, we classify routes by how similar they are. As show in Table 1, there is significant heterogeneity in driver and passenger route preferences, and one may be concerned their arrival rates and unobserved preferences may also be different. To flexibly incorporate and estimate heterogeneity of drivers on different routes in our structural model, we classify drivers and passengers based on the observed route preferences and separately analyze the behaviors of agents in each class. A straightforward way to group the routes would be to classify them by the regions (green, red, black and magenta) of the origin and destination, but given the irregular shape of each region, the routes from the magenta region to the red region could have drastically different distances. **?** suggests using the K-means algorithm to cluster on observables, and we find that the K-means classification based on routes[7] effectively amounts to classifying routes by distance and regions of origins and destinations. We use the K-means algorithm to classify all driver routes into 28 classes so that the smallest class contains more than 0.5% of all observed driver routes.

[7] To use the K-means algorithm, we define the distance ℓ_{ij} between route i going from $a_i = (a_i^x, a_i^y)$ to $b_i = (b_i^x, b_i^y)$ and route j using a city-block distance measure:

$$\ell_{ij} = \left|a_i^x - a_j^x\right| + \left|a_i^y - a_j^y\right| + \left|b_i^x - b_j^x\right| + \left|b_i^y - b_j^y\right|. \tag{1}$$

The approach in **?** clusters on both the covariates and outcome variables to capture latent driver heterogeneity. The outcome variables in our case (waiting time and driver choices conditional on the sets of waiting drivers) are high dimensional. For simplicity, we cluster on just the stated routes of the drivers, which still allows us to capture the heterogeneity of drivers across different types of routes.

42% of the routes are between 0 and 20KM, 42% are between 20 and 40KM, 11% between 40 and 60KM, and 3% between 60 and 80KM. Figure 6 shows the origins and destinations of the top 3 classes of driver routes and their corresponding proportions among all observed driver routes. More than 25% of all driver routes are between and within the red and black regions. We use the algorithm-generated classification rule from the driver classification and classify passenger routes into 28 classes as well. The proportion of each class of passenger as a percentage of all passenger routes is similar to that of the corresponding driver class.

We next discuss the properties of the empirical passenger arrival and exit process.The passenger arrival process can be approximated by a constant hazard model reasonably accurately and we provide support for the assumptions in the model that driver and passenger arrivals are independent, and the arrivals of each class of agents are also independent. Figure 8 plots the model-predicted and actual frequency of the arrival time intervals of the most popular passenger class and driver class. We cannot reject at 90% confidence level the hypothesis the class designations of arriving agents are independent within drivers, within passengers, and between drivers and passengers. Specifically, we test whether the class designation of agent i is correlated with that of agent j, where j arrives immediately after i. We conduct three tests:

1. i is a passenger, j is the passenger arriving after i

2. i is a driver, j is the driver arriving after i

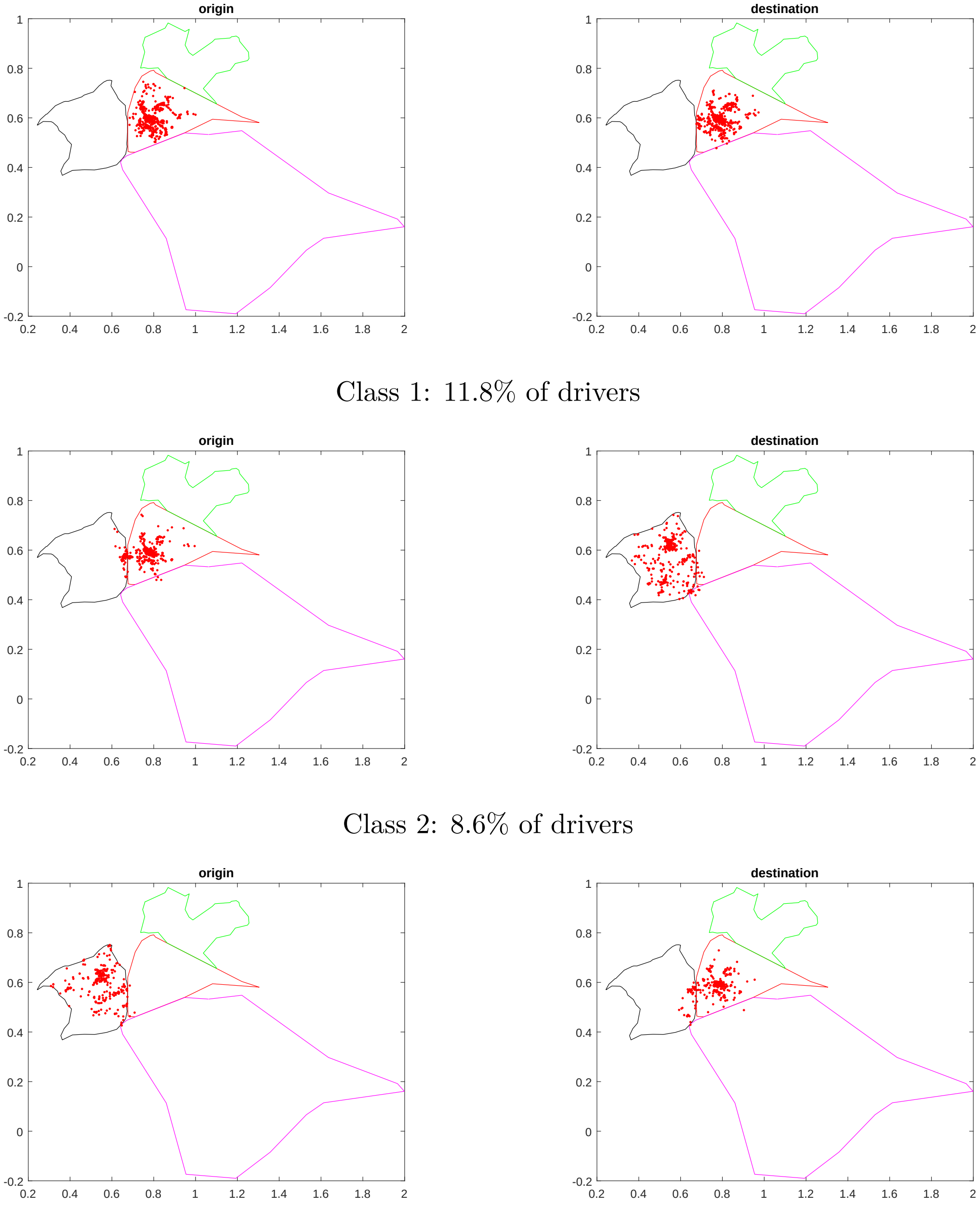
3. i is a passenger, j is the driver arriving after i.

The p-values of the three χ^2 tests are 0.30, 0.37 and 0.94. Furthermore, we find that the numbers of driver and passenger arrivals are weakly correlated across time. Fig. 7 shows the number of arriving passengers and drivers every 10 seconds during the half-hour sample period and averaged across 6 days. The correlation between the number of driver and passenger arrivals within the 10 second intervals is -0.0265.

3 A Simple Model of Driver and Passenger Match

We consider a simple model of match formation between drivers and passengers in an environment similar to our empirical application. We use this model to show the intuition that underlies our more complicated structural model and why the welfare effect of strategic waiting is an empirical question.

Consider a continuous time, infinite horizon matching market of four potential entrants: two drivers (A, B) and two passengers (a, b). Drivers arrive at rate ρ^0 and passengers arrive at κ^0. At any moment, a driver in the market receive opportunities to move (that arrive independently at rate $\gamma > 0$) and choose whether to wait or pick up a passenger in the market. Only a driver can actively pick up a passenger and form a match. Upon the formation of a match with passenger i, the driver j receives utility u_{ij}, and the matched driver and passenger leave the market. An



Class 1: 11.8% of drivers

Class 2: 8.6% of drivers

Class 3: 6.6% of drivers

Figure 6: Pickup (Left) and Drop-off (Right) Locations of the 3 Most Popular Classes of Driver Routes

We re-scale x and y-axis to avoid disclosing the identity of the city.

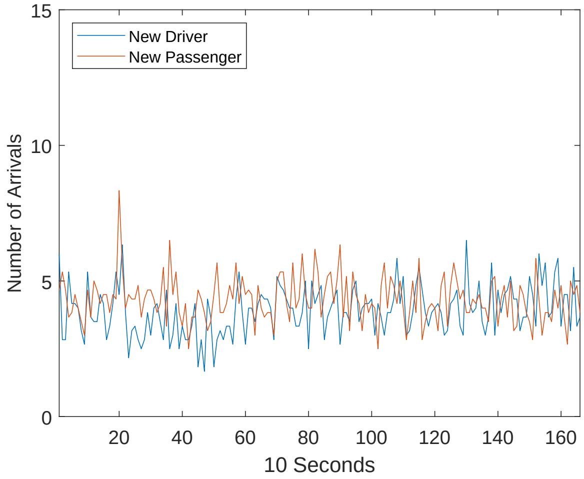


Figure 7: Average Number of Arrivals per 10 Seconds

The time period is 4:00PM to 4:30PM. We average the number of arrivals across 6 days.

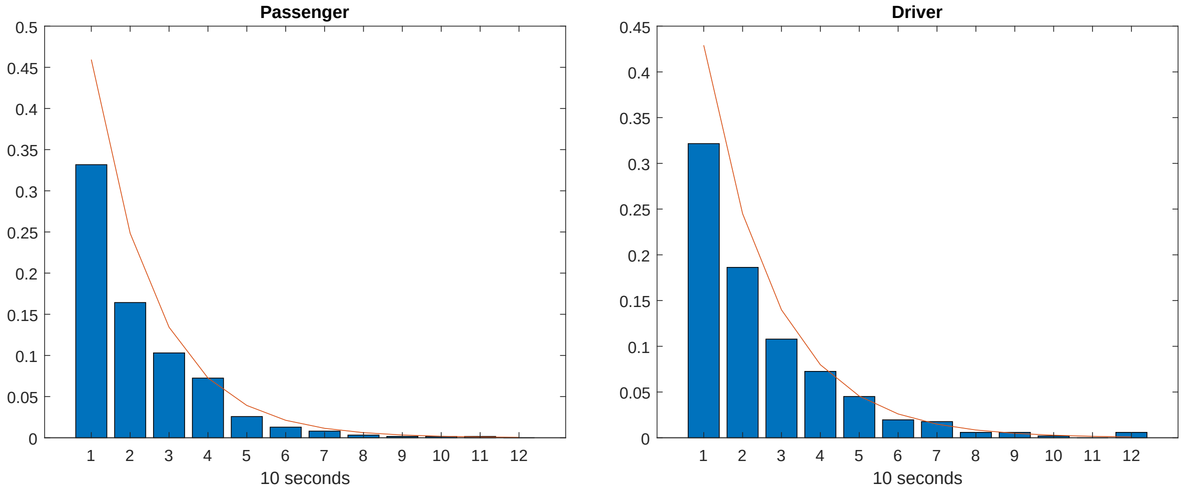


Figure 8: Arrival Frequency of the Most Popular Passenger and Driver Classes

The bar graph represents the empirical frequency of the time between two arrivals within the same class. The solid line represents the predicted probability of arriving between the tth and $t+1$th 10 second interval from the estimated exponential distribution.

unmatched driver or passenger randomly exit the market at rate ρ and κ. If a driver leaves the market without a match, she receives a reservation utility 0. The entry and (unmatched) exit processes are independent across agents and over time.

While the passengers are not "strategic" in the sense that they do not actively choose partners or how much they wait, the model captures a key feature of a matching market: one driver's action limits the choice of another driver by changing the availability of the matching partners. The simplification allows us to characterize the key economic trade-offs. This barebone model also represents the main features of our empirical application.

To solve for driver strategies, we assume that drivers have full knowledge of the past history once they enter the market. An agent (a driver or a passenger) has three possible states: the agent has not entered the market, the agent is in the market or the agent has exited (either by a match or by a random exit). The state space of a driver problem consists of the Cartesian product of other agents' states and thus has $3^3 = 27$ elements. Use S_{jt} to denote the state of a driver at t. Given an opportunity to move, the driver observes the set of available passengers R_t and the utility of matching with the most compatible passenger $s_{jt}\left(R_t\right) = \max_{i \in R_t} u_{ij}$. Driver j decides between matching with $i^{\star} = \arg\max_{i \in R_t} u_{ij}$ or waiting. Without loss of generality, we consider A's problem. The value of waiting $V\left(S_{At}\right) \geq 0$ depends on S_{At}. A forward-looking A compares $\max_{i \in R_t} u_{iA}$ against $V\left(S_t\right)$ and a myopic A compares $\max_{i \in R_t} u_{iA}$ against the reservation value 0, and given preference ordering, a forward-looking agent will pass over some match opportunities to wait for a better partner. In Appendix A, we present the parameterization of two scenarios where strategic waiting has different implications for the number of matches in equilibrium and solve for a pure strategy Bayesian equilibrium explicitly. Below, we discuss the intuition of the two cases.

1. Preference ordering:

$$A : a \succ b \succ \text{unmatched}$$

$$B : b \succ \text{unmatched} \succ a$$

   When ρ and κ are sufficiently small, forward-looking A would always wait for a and B always for b. In this case, both drivers will wait for their most compatible partners and all drivers and passengers will be matched. In contrast, if b and A arrive at the market first, a myopic A would match with b, leaving the late arrivals (B, a) unmatched. In this case, strategic waiting increases the expected number of matches. In fact, the strategic waiting incentives allow the decentralized market equilibrium to achieve the social optimum.

2. Preference ordering:

$$A : a \succ b \succ \text{unmatched}$$

$$B : a \succ \text{unmatched} \succ b$$

   When ρ and κ are small, and A sufficiently prefers a to b, A would wait for a when the market contains (A, b). If a arrives before B, or when a, A and B are in the market but A gets to move before B, A will match with a, which leads to B and b unmatched. In contrast, when the market consists of only (A, b), a myopic A would match b and both would leave the market, and B can always match with a when they enter the market. If agents appear in different orders (and the event of exits before all agents have entered is close to 0), the same number of agents will be matched with either forward-looking or myopic agents. In this case, strategic waiting decreases the expected number of matches. The expected number of matches is lower than the social optimum regardless of whether the agents are myopic or forward-looking.

4 Empirical Model

We use an infinite horizon continuous time dynamic model to study the matching between passengers and drivers. Each agent (driver or passenger) has a preferred route: agent i wants to travel from location a_i to location b_i. Drivers and passengers arrive to the market at rates ρ^0 and κ^0. Use F_I and F_J to denote the distributions of the drivers and passengers' preferred routes. The preferred route is drawn independently from the respective distribution and persistent throughout her stay in the market. The maximum waiting time of drivers and passengers is distributed exponentially with means $\frac{1}{\rho}$ and $\frac{1}{\kappa}$. Drivers additionally have heterogeneous preferences over the compatibility of passenger routes and unobserved match values. In estimation, we also allow drivers to have heterogeneous ρ, but for the simplicity of the presentation, we for now assume that ρ is the same across drivers. We specify the driver preferences for passengers below.

Drivers are forward-looking. At rate γ, a driver checks her phone and searches for a compatible passenger. In data, a driver checks the phone every 36 seconds on average. Upon checking at time t, a driver sees the available requests R_t and chooses between answering a request or waiting. If a driver chooses to answer a request j, the driver and the chosen passenger leave the market and the driver receives utility

$$u_{ijt} = \underline{u}_i + \underbrace{u_{i0} - \alpha_i d_{ij} + \beta_i x_j + \xi_{ij}}_{\delta_{ij}} + \varepsilon_{it}, \tag{2}$$

$\underline{u}_i$ is the value of the trip if the i drives alone. u_{i0} is an intercept, reflecting the base value of picking up a passenger. d_{ij} is the detour length, and x_j is the length of the trip, as defined in Section 2.1. Because most trips are much longer than 2KM (the maximum range of the base fare), the price schedule is linear for these trips. Therefore x_j reflects both the preferences for the fare and the distance trip to travel with a stranger. ξ_{ij} is the unobserved driver-passenger synergy, which captures time-persistent but idiosyncratic match values, such as whether there is a difficult left turn on the route to pick up the passenger. ε_{it} is a time-varying unobservables that reflect other factors that affect driver decisions, such as local traffic conditions. d_{ij} and x_j are defined as in Section 2.1. This formulation can incorporate rich driver heterogeneity: we allow $\underline{u}_i$, u_{i0} to be specific to the observed driver class, and we additionally allow for random coefficients (α_i, β_i) that are heterogeneous across drivers and whose distribution is specific to a driver class.[8]

[8] This formulation does not take into account the match quality based on the stated departure time of the passengers

We next specify the driver problem. We assume that driver i knows her maximum waiting time T_i.[9] As a driver waits, the option value of waiting changes continuously over time in how long they have waited. The opportunity to check the phone arrives at rate γ. Given an opportunity to check the phone, the driver chooses between waiting and one passenger among R_t passengers available. The driver and passenger leave the market if the driver chooses a passenger, and otherwise will wait till T_i. We assume that ε_{it} is i.i.d across each incidence of checking with distribution $f_\varepsilon\left(\varepsilon_{it}\right)$. This assumption is motivated by (1) the survey response that drivers are particularly concerned with the local traffic around the driver origin, (2) the arrival processes are stable over time, and (3) that traffic conditions also appear to be stable over a much longer horizon than the typical waiting time of a driver (Section 2.2). We also assume that the driver does not know the future values of ε at t. R_t evolves over time because passengers exit when they reach the maximum waiting time or drivers answer their requests. The driver has expectations about the evolution of R_t but does not have perfect knowledge about it.

We now define the state variables and write down the driver dynamic optimization problem. Use τ to denote the time the driver has been in the market. At $\tau = T_i$, the driver checks the phone one last time and the option value of waiting is traveling alone; at $\tau < T_i$ the value of waiting may be higher because a more compatible passenger may arrive between τ and T_i. Therefore the driver decision depends on τ. The decision also depends on the future R_t. With a slight abuse of notation, we use R_{it} to denote the set of passenger routes and their unobserved match values ξ_{ij}. We assume that R_{it} evolves according to a Markov process: R_{it} changes at rate $\lambda_i\left(R_{it}\right)$, and conditional on changing, the new $\tilde{R}_{it}$ follows the distribution $g_{R,i}\left(\tilde{R}_{it} \mid R_{it}\right)$.[10] Both the change rate and the distribution are subscripted with i because the unobserved match values ξ differ across drivers. A driver does not observe how many other drivers are waiting, these rivals' preferences or maximum waiting time, but their decision rules, along with the maximum waiting time of the existing passenger and the arrival processes of drivers and passengers, change R_{it}. The state vector at t after having waited τ is (τ, R_{it}).

Use $V_i\left(\tau, R_{it}\right)$ to denote i's value of waiting after having waited τ. Then in an infinitesimal amount of time $\varDelta$, the Bellman equation is

and drivers. According to the conversations with DiDi, the departure time is a secondary consideration compared with the match quality of routes in that drivers are often willing to move up departure to accomodate a passenger. We find evidence for this anecdote in our data: among 91.5% of the matched driver-passenger pairs, the driver departure time is later than passenger departure time, and the coefficient of variation of this time difference is 2.8. In comparision, the coefficient of variation for the detour lengths d_{ij} is 0.9.

[9] This assumption is consistent with data. Although the data on the stated time of departure appear to have round number biases (many claim to depart at 4:45, 4:50, ...), we find that among the drivers who did not find a passenger, the last searches of the majority (65%) of these drivers are within 1 minute of the stated departure time, and the last searches of 84% of these drivers are no later than the stated departure time.

[10] Appendix B shows that λ_i and $g_{R,i}$ are well-defined.

$$
\begin{aligned}
V_i\left(\tau, R_{it}\right) = \frac{1}{1+\theta_i \Delta} \Bigg[& \underbrace{\Delta \lambda_{R,i}\left(R_{it}\right) E\left(V_i\left(\tau+\Delta, \tilde{R}_{it}\right) \middle| \, R_{it}\right)}_{R_{it}\text{ changes}} \\
& + \underbrace{\Delta \gamma \int_{\varepsilon_{it}} \max\left\{\max_{j \in R_{it}}\left[\underline{u}_i+\delta_{ij}+\varepsilon_{it}\right], V_i\left(\tau+\Delta, R_{it}\right)\right\} f_{\varepsilon}\left(\varepsilon_{it}\right) d\varepsilon_{it}}_{i\text{ receives a move opportunity}} \\
& + \underbrace{\left(1-\Delta \lambda_{R,i}\left(R_{it}\right)-\Delta \gamma\right) V_i\left(\tau+\Delta, R_{it}\right)}_{\text{nothing happens}} \Bigg], \qquad (3)
\end{aligned}
$$

where E denotes the conditional expectation over R_{it}. In the interpretation of Doraszelski and Judd (2012), Δ is sufficiently small such that the probability of R_{it} changing and i moving simultaneously occurs with negligible probability compared with the first order terms in the above equation. If i does pick a passenger, then i receives a onetime payoff $\max_{j \in R_{it}}\left[\underline{u}_i+\delta_{ij}+\varepsilon_{it}\right]$ and leaves. Otherwise τ increases. θ_i is the discount factor while drivers search on the DiDi app. Alternatively, θ_i can be interpreted as a waiting cost, where a higher θ_i reflects higher search cost because the discounted present value of waiting at the beginning of the waiting is lower. Assuming the smoothness of the value function, we can write the Bellman equation in a more compact form:

$$
\begin{aligned}
V_i\left(\tau, R_{it}\right) = & \frac{1}{\theta_i+\gamma+\lambda_{R,i}\left(R_{it}\right)} \\
& \times \left[\gamma \int_{\varepsilon_{it}} \max\left\{\max_{j \in R_{it}}\left[\underline{u}_i+\delta_{ij}+\varepsilon_{it}\right], V_i\left(\tau, R_{it}\right)\right\} f_{\varepsilon}\left(\varepsilon_{it}\right) d\varepsilon_{it}\right. \\
& \left. + \lambda_{R,i}\left(R_{it}\right) E\left(V_i\left(\tau, \tilde{R}_{it}\right) \mid R_{it}\right) + \frac{\partial}{\partial \tau} V_i\left(\tau, R_{it}\right)\right], \qquad (4)
\end{aligned}
$$

with the boundary condition that $V_i\left(T_i, R_{it}\right) = \int_{\varepsilon_{it}} \max\left\{\max_{j \in R_{it}} \underline{u}_i+\delta_{ij}+\varepsilon_{it}, \underline{u}_i\right\} f_{\varepsilon}\left(\varepsilon_{it}\right) d\varepsilon_{it}$ when i departs at t after having waited T_i.

To bring the model to data, we need to solve the Bellman equation. The challenge is that R_t is of very high dimension. Here we employ an approach motivated by the design of the platform and related to the logit inclusive value method in **?**. Note that the identity of the most compatible passenger that solves $\max_{j \in R_{it}} \underline{u}_i+\delta_{ij}+\varepsilon_{it}$ does not change when the driver checks again, if R_t remains the same, because we interpret the time-varying shock ε_{it} as reflecting the local traffic conditions, and ε_{it} is not specific to a passenger. As a result, just like the logit inclusive value, $\max_{j \in R_t} \delta_{ij}$ is a sufficient statistic for the dynamic optimization problem.[11] Tracking just $\max_{j \in R_t} \delta_{ij}$ instead of R_{it} greatly simplifies the computation, because the state space of the simpler problem is two-dimensional, with the state consisting of τ and $s_{it} = \max_{j \in R_{it}} \delta_{ij}$. In practice, the decisions to pick up passengers occur in real time, and while drivers can view hundreds of requests in their

[11] **?** extensively discussed the trade-offs of this approach. In this context, the agent who tracks the first order statistics forfeits some information when forming the expectation of the future R_t. **?** shows that conditioning on both the first and second order statistics does not substantially change the result.

search in theory, most would just view the one with the highest compatibility index. We view our simplification as a reasonable approximation of how drivers actually make decisions.

With the simplifying assumption, we re-write the Bellman equation as

$$\begin{aligned} V_i\left(\tau, s_{it}\right) = & \frac{1}{\theta_i + \gamma + \lambda_{s,i}\left(s_{it}\right)} \\ & \times \left[\gamma \int_{\varepsilon_{it}} \max\left\{\underline{u}_i + s_{it} + \varepsilon_{it}, V_i\left(\tau, s_{it}\right)\right\} f_{\varepsilon}\left(\varepsilon_{it}\right) d\varepsilon_{it} \right. \\ & \left. + \lambda_{s,i}\left(s_{it}\right) E\left(V_i\left(\tau, \tilde{s}_{it}\right) \mid s_{it}\right) + \frac{\partial}{\partial \tau} V_i\left(\tau, s_{it}\right)\right], \end{aligned} \tag{5}$$

where E is conditional expectation of s_{it} , and we use $\tilde{s}_{it}$ to represent the new state conditional on a change, which occurs at rate $\lambda_{s,i}\left(s_{it}\right)$ and the conditional distribution is $g_{s,i}\left(\tilde{s}_{it} \mid s_{it}\right)$.

We next define the equilibrium concept. The driver strategy is the accept/reject decision σ_i that maps to $\{0, 1, \ldots, \#R_{it}\}$, where 0 means waiting and $\#R_{it}$ is the number of waiting passengers:

$$\sigma_i\left(\tau, R_{it}, \varepsilon_{it}\right) = \left\{\underline{u}_i + \max_{j \in R_{it}} \delta_{ij} + \varepsilon_{it} > V_i\left(\tau, \max_{j \in R_{it}} \delta_{ij}\right)\right\} \cdot \arg\max_{j \in R_{it}} \delta_{ij}.$$

The entry and exit processes of passengers and drivers, combined with σ, govern the evolution of R_{it}. R_{it} in turn implies the distribution of s_{it} for each driver i. In the equilibrium, s_{it} need be stationary and consistent with the driver beliefs. Use $\mathcal{D}$ to denote the set of driver types (routes and preferences). We focus on driver decisions within a relatively short time window (half an hour) compared with the length of the day, and we work with a stationary equilibrium concept.

Definition 1. A stationary equilibrium[12] consists of $\{V_i, \sigma_i, \lambda_{R,i}, g_{R,i}, \lambda_{s,i}, g_{s,i}\}$ such that

- The Bellman equation (5) is satisfied for all $i \in \mathcal{D}$;
- $\forall i \in \mathcal{D}, R_{it}$ has a stationary distribution $g_{R,i}$ consistent with $\{\sigma_i\}_{i \in \mathcal{D}}$, the entry process $\{F_I, F_J, \rho^0, \kappa^0\}$ and the maximum waiting implied by ρ and κ .
- The drivers have rational beliefs consistent with the evolution of s_{it} implied by R_{it}.

We do not separately specify a passenger problem for two reasons. First, passengers do not see drivers or choose drivers, and if a passenger is picked up, the passenger is delivered exactly as specified in the request, and passengers need not have preferences over the extent of the "mismatches"

[12] The taxi literature (e.g. Buchholz (2017); Frechette et al. (2016)) typically focuses on drivers who operate throughout a day and considers a non-stationary equilibrium to capture the time-varying demand conditions. The drivers in our context are "one-shot" labor suppliers, who are chiefly concerned with the demand and travel conditions within the next 40 minutes (for over 80% of the drivers). Demand and travel conditions even later are not relevant, because these commuters, who have their own travel needs to reach their destination by a certain time, will have left the market. Therefore we empirically consider a short time window and focus on the potential drivers who have travel needs during that time window. Empirical evidence in Section 2.2 shows that both the demand and the travel conditions are stable during and beyond our sampling time window. Considered together, we use a stationary equilibrium concept to study the matching in our context.

This equilibrium concept is related to the oblivious equilibrium (Weintraub, Benkard and Van Roy (2008)), the mean field equilibrium (Iyer et al. (2014)) and the equilibrium concepts used in Krusell and Smith (1998), Backus and Lewis (2016), Bodoh-Creed et al. (2017), Buchholz (2017) and others.

as drivers do. Secondly, passengers commonly see DiDi's peer-to-peer platform as the first choice, before resorting to DiDi Express (the Uber version of DiDi) or taxi, which can be hired with little extra delay, or the public transit, which runs on a fixed schedule. Therefore modeling passengers as either waiting till its maximum time or being picked up before the waiting time runs out is a reasonable approximation.[13] The framework laid out here can easily be augmented to incorporate a passenger's waiting problem or choice problem if a relevant empirical context demands.

5 Identification and Estimation

In this section, we discuss the identification and estimation of these parameters:

1. (F_I, F_J): the distribution of the driver and passenger routes upon entry.
2. (ρ^0, ρ_i, γ) : the rates at which (1) a driver arrives to the market, (2) the inverse of the mean maximum driver waiting time (alternatively, the rate at which a driver exits without a match) and (3) the rate at which a driver checks the phone.
3. (κ^0, κ): the rates at which (1) a passenger arrives to the market and (2) the inverse of the mean maximum passenger waiting time.
4. the driver preference parameters: $(\underline{u}_i, u_{i0}, \alpha_i, \beta_i, \xi_{ij}, \varepsilon_{it}, \theta_i)$.

We use the observed driver and passenger route distribution as F_I and F_J and the observed driver and passenger arrival rates for ρ^0 and κ^0. The rate of search γ is identified from the frequency of drivers refreshing the passenger list. One may be concerned that γ may be heterogeneous across drivers. We find limited evidence for heterogeneity in γ. On average, drivers check their phones every 36 seconds, and the drivers who eventually answered a request checked every 33 seconds on average. Drivers who are in the market for less than 449 seconds (observed mean waiting time) check their phones every 32 seconds. For simplicity, we assume that γ is the same for all drivers. The next section discusses the identification of ρ_i, κ and driver preference parameters.

5.1 Identification of ρ_i, κ and Driver Preference Parameters

First, we make the following assumptions that are weaker than the parametric assumptions introduced in Section 4:

Assumption 1.

1. *The distribution of the time between consecutive arrivals of passengers has full support on* R^+.

[13] From a welfare analysis point of view, one should still estimate the passenger problem to recover the cost of waiting to determine whether a longer waiting time substantially decreases passenger welfare (although indirect evidence in Section 2.1 suggests not). In our empirical context, because most passengers will participate in a matching market until the end of their maximum waiting time, the ideal identifying variations would be the values of alternative service providers. In the US, one such variation would be the dynamically set Uber fares. However, DiDi Express at the time had switched to fixed prices due to local regulations imposed in 2017.

2. *The distribution of the time between consecutive arrivals of drivers has full support on R^{+}.*
3. *The support of the distribution of the maximum waiting time of drivers and passengers includes $(0, N]$, where N is a finite positive number.*

The distribution of the maximum waiting time is nonparametrically identified under these assumptions: for any T, with positive probability, we would observe in data a period T where there are no waiting passengers or drivers. The proportion of exiting drivers or passengers identifies the respective distribution of the maximum waiting time.

Restricting to our flexible parameterization, we assume that the driver and passenger exit rates are heterogeneous across observed attributes. Specifically, we assume that ρ is different across each driver class. We could do the same for the passenger exit rate κ, but for practical and computational reasons to be discussed in Section 5.2, we assume that the distribution of maximum waiting time is the same for all passengers (and each individual passenger's waiting time is a realization from this distribution).

To identify driver preferences for passengers, we first normalize the payoff to traveling alone, $\underline{u}_i$ to 0. Note that θ_i and $\underline{u}_i$ cannot be separately identified: a driver that prefers to wait may either do not discount the future much (θ_i close to 0) or have a high $\underline{u}_i$. We argue that the difference between V and (the deflated) $\underline{u}_i$ is still identified with this normalization. This point can be most clearly seen using the discrete version of the Bellman equation (3). At $\tau = T_i$, like a standard choice model, $V_i(T_i, R_{it}) - \underline{u}_i$ is identified. At $\tau = T_i - \Delta$ for a small Δ, the Bellman equation can be re-written as

$$\begin{aligned}
&V_i\left(T_i-\Delta, R_{it}\right)-\frac{\underline{u}_i}{1+\Delta\theta_i}\\
=&\frac{1}{(1+\Delta\theta_i)}\left[\underbrace{\Delta\lambda_{R,i}\left(R_{it}\right)E\left(V\left(T_i,\tilde{R}_{it}\right)-\underline{u}_i\middle|\, R_{it}\right)}_{R_{it}\text{ changes}}+\underbrace{\Delta\gamma\int_{\varepsilon_{it}}\max\left\{\max_{j\in R_{it}}\delta_{ij}+\varepsilon_{it}, V_i\left(T_i,R_{it}\right)-\underline{u}_i\right\}f_{\varepsilon}\left(\varepsilon_{it}\right)d\varepsilon_{it}}_{i\text{ receives a move opportunity}}\right.\\
&\left.+\underbrace{\left(1-\Delta\lambda_{R,i}\left(R_{it}\right)-\Delta\gamma\right)\left(V_i\left(T_i,R_{it}\right)-\underline{u}_i\right)}_{\text{nothing happens}}\right],
\end{aligned}$$

where $V_i\left(T_i-\Delta, R_{it}\right)-\dfrac{\underline{u}_i}{1+\Delta\theta_i}$ is the difference between V_i and the deflated value of traveling alone.

With this normalization, we rely on the variations in the changes in waiting passengers that vary across a driver's search instance and across drivers. The identifying restriction is similar to the assumption common in demand estimation, where the variations in choice sets identify consumer tastes. In particular, the variability in α_i and β_i maps into the distribution of the detour lengths and passenger trip lengths of successful matches conditional on a set of waiting passengers. The validity of this identifying restriction relies on the assumption that the variations in the choice sets are orthogonal to an individual driver's unobserved heterogeneity. We assume the following:

Assumption 2.

1. *Passenger arrivals are independent.*
2. *Driver arrivals are independent.*
3. *Driver and passenger arrivals are independent.*
4. *ξ_{ij}'s are independent across i and j.*

Combined with Assumption 1, R_{it} is independent of an arriving driver's route, α_i and β_i, and any combination of passenger routes is in the support of R_{it}. We directly test the first three assumptions in Section 2.2. The last assumption warrants some discussion. ξ_{ij} is the unobservable that captures persistent driver-passenger match value, such as whether there is a left turn on the driver route to pick up the passenger. This part is reasonably i.i.d given the coarseness of our data (the precision of the location data is 0.25KM^2), but there may also be an unobserved component $\tilde{\xi}_j$ in ξ_{ij} that is common to passengers, such as the passenger's reputation score, which is not observed by the researcher but observed by drivers. The common knowledge unobservable would cause a selection problem, because the passengers with a higher $\tilde{\xi}_j$ will be answered more quickly and less likely in the data. We show that the selection bias is likely quite small. In Appendix C, we take the estimates and simulate passengers as having a vertical attribute ξ_j that has the same impact on all drivers. In other words, the utility for i to pick up j in this robustness check is

$$u_{ijt} = \underline{u}_i + u_{i0} - \alpha_i d_{ij} + \beta_i x_j + \xi_j + \varepsilon_{it}.$$

We then simulate the outcomes of the decentralized market as well as the counterfactuals. The difference with the results from the model with i.i.d ξ_{ij} is quite small.

We assume that $\varepsilon_{it} \sim N(0,1)$, and $\xi_{ij} \sim N\left(0, \sigma^2_{\xi,i}\right)$. Like ρ_i, $\sigma_{\xi,i}$ differs across driver classes. The effect of ξ_{ij} is similar to the logit shocks in a random coefficient logit demand model: the number of waiting requests (choices) increases the probability of a driver answering a request. Therefore with a large $\sigma_{\xi,i}$, the conditional mean number of waiting requests for drivers who answer requests should be greater than the unconditional mean.

The last parameter to be identified is θ_i, the time preference or the waiting cost of drivers. The variations in how long an answered request is kept waiting by the answering driver provides the identification. At the extreme, if drivers are myopic ($\theta_i = \infty$), no such requests would be kept waiting; if the drivers prefer to answer the requests at the end of their wait ($\theta_i = 0$ or even $\theta_i < 0$), few drivers would answer requests during their first search.

5.2 Estimation

We use a three-step estimation approach based on the method of simulated moments. In the first step, we estimate the arrival rates and the distribution of driver and passenger routes directly from data. The arrival rates of the drivers and passengers are 0.4335 and 0.4823, which means that on average a driver or a passenger enters every 2 seconds. In the second step, we simulate the outcomes of drivers, and match moments from the simulated outcomes with data. We separately estimate the preference parameters for drivers in each class. We assume that within a class, α_i and β_i are normally distributed, and we estimate their means and standard deviations for each class. Therefore drivers in each class could still have different preferences, although the preferences are drawn from the same distribution specific to a class. Across classes the random coefficients have different distributions. The moments we use follow closely the discussion of the identification of the parameters in the previous section.

1. $E\left(\text{driver time in market}\right)$.
2. $E\left(1\left[i \text{ answers a request}\right]\right)$.
3. Conditional on i answering j's request, the 0.25, 0.5 and 0.75 quantiles of d_{ij}.
4. Conditional on i answering j's request, the 0.25, 0.5 and 0.75 quantiles of x_j.
5. $Cov\left(d_{ij}, x_j\right)$, conditional on i answering j's request.
6. The 0.25, 0.5 and 0.75 quantiles of the number requests at the last search of the drivers.
7. The 0.25, 0.5 and 0.75 quantiles of the number requests at the last search of the drivers conditional on answering requests.
8. Conditional on i answering j, the 0.25, 0.5 and 0.75 quantiles of $\frac{\tilde{T}_{ij}}{\tilde{T}_i}$, where $\tilde{T}_{ij}$ is the length of the time when both i and j are present, and $\tilde{T}_i$ is the length of the time i is in the market.

There are a total of 8 parameters per driver class for 28 classes. We estimate the parameters of each class separately: we simulate the outcomes of drivers in a class, and match the simulated moments corresponding with the data moments specialized to the drivers of this class.[14] There could potentially be multiple equilibria, and we focus on the equilibrium in the data: we assume that the data are generated by one equilibrium, and we use the estimated evolution of s_{it} when solving for each driver i's dynamic problem. Specifically, for a given vector of parameters, we first estimate the evolution of s_{it} from the observed time-series variations in the set of waiting passengers, and we use the estimated distribution to calculate the driver strategy.

The estimates of driver primitives are presented in Tables 2 and 3. We sort the drivers by the popularity of the routes among drivers (the proportion of each class of routes as a percentage of all observed driver routes). The first column is the distance range of the driver preferred routes, and the last column is the proportion of the observed drivers that fall into the driver class. The standard errors are from a 100-reptition bootstrap on the days, preserving the correlations across time within a day. On average, drivers trade off 1 KM of detour with the 1.7 KM of passenger trip length (higher fare), although there is substantial heterogeneity across drivers. In particular,

[14] The weighting matrix is the inverse of the variance of each moment, estimated by bootstrap across days.

Table 2: Driver Parameter Estimates,

Class	Distance		u_{i0}	$\bar{\alpha}$	$\bar{\beta}$	σ_α	σ_β	ρ	θ	σ_ξ	Proportion
	Range (KM)										
1	0	20	-2.3647	26.4605	11.8391	1.4984	14.7112	0.0009	-0.0001	0.2218	0.118
		SE	0.7012	6.092	3.9287	1.573	2.6871	0.0001	0.0001	0.3986	
2	20	40	-1.8962	20.0991	10.1218	2.6503	7.0209	0.0008	0	0.281	0.0859
		SE	0.4296	6.1167	3.5801	2.4994	3.1909	0.0001	0.0001	0.413	
3	20	40	-2.6149	14.4541	10.5891	0.0483	6.5277	0.0009	0.0001	0.0928	0.0655
		SE	0.6046	4.6652	7.9281	3.6738	4.5779	0.0001	0.0002	0.1616	
4	0	20	-3.0039	23.3622	20.9678	3.7857	0.8453	0.0008	0.0003	0.6834	0.0608
		SE	0.7877	5.5933	4.6495	3.5053	6.9344	0.0001	0.0001	0.3354	
5	0	20	-1.2163	42.1694	28.2416	1.9269	2.0691	0.0009	0.0001	1.0816	0.0499
		SE	0.5279	4.1114	4.0625	2.1014	3.1472	0	0.0001	0.3308	
6	20	40	-2.8962	32.1694	13.1548	5.6934	6.2384	0.001	0	0.7656	0.0491
		SE	0.8166	5.9346	7.4205	3.4715	4.183	0.0001	0	0.2305	
7	20	40	-1.8649	22.1694	10.5891	5.6936	6.1623	0.001	0.0001	0.1591	0.0484
		SE	0.6701	3.6645	2.7928	1.7716	4.8866	0.0001	0.0001	0.3745	
8	0	20	-0.9329	14.0303	10.4307	0.6263	0.8852	0.0009	-0.0001	0.671	0.048
		SE	0.383	4.1097	3.1079	2.7952	3.3087	0.0001	0.0001	0.391	
9	0	20	-0.36	22.996	18.4441	0.0001	5.7414	0.0006	0	0.3337	0.0432
		SE	0.753	4.6728	3.1863	1.2972	2.7215	0.0001	0.0001	0.4124	
10	0	20	-1.5748	22.1694	16.1748	0.6251	9.5197	0.001	0.0001	0.0938	0.042
		SE	0.9061	7.3882	4.8939	2.3729	4.9218	0.0001	0.0001	0.3083	
11	20	40	-2.8725	24.5859	10.59	0.0392	7.2776	0.0009	0.0002	0.1593	0.0377
		SE	0.5461	4.9143	6.1819	2.1896	2.4977	0	0.0001	0.2142	
12	20	40	-2.2538	17.7814	9.1807	2.7472	5.8803	0.0011	0.0001	1.0896	0.0354
		SE	0.6689	3.268	5.1603	2.4912	2.8043	0.0001	0.0001	0.4013	
13	40	60	-2.8649	25.961	10.5085	0.633	14.9371	0.0009	0.0001	0.0341	0.0337
		SE	1.1934	5.2752	6.1439	1.423	3.7419	0.0001	0.0001	0.2264	

the standard deviation of β_i is much larger than that of α_i across most driver classes, which shows that although most drivers eschew passenger requests that are incompatible with their own, they differ significantly in how willing they are to travel with a stranger per KM, even though the fare is higher from a longer passenger trip. Therefore the β estimates should not be simply interpreted as preferences for trip fares.

Finally, we simulate the decentralized equilibrium and estimate the distribution of the passenger maximum waiting time. In principle, we can also allow the mean maximum waiting time $\frac{1}{\kappa}$ to be heterogeneous, specific to each passenger class. The driver parameter estimates suggest that this is not a key dimension of heterogeneity: the estimated mean maximum waiting time $(\frac{1}{\rho})$ for more than 80% of drivers is between 15 and 20 minutes. We therefore assume that passengers have the same κ. This assumption also substantially reduces the computational burden by turning the estimation into a one-dimensional numerical optimization problem.[15] To simulate the full game,

[15] Evaluating a MSM objective function once in the third step is even more costly compared with the second step. We need to simulate a sufficiently long period of the steady state of the game; simulating the full game requires

Table 3: Driver Parameter Estimates, Continued

Class	Distance		u_{i0}	$\bar{\alpha}$	$\bar{\beta}$	σ_α	σ_β	ρ	θ	σ_ξ	Proportion
	Range										
14	40	60	-3.1149	32.1694	10.4328	1.3966	12.1673	0.0012	0	0.4053	0.0318
		SE	0.8516	5.5497	4.2726	2.1876	2.6324	0.0001	0.0001	0.3079	
15	20	40	-2.8024	32.3257	10.5891	3.1342	19.9371	0.0009	0.0001	0.9646	0.0301
		SE	0.4058	9.8224	6.7777	2.3677	9.1498	0.0001	0.0001	0.3564	
16	20	40	-2.8637	35.9606	11.175	9.7485	13.7652	0.001	0.0001	0.0625	0.0242
		SE	0.6397	5.6644	4.403	2.2968	2.5282	0.0001	0.0001	0.1795	
17	0	20	-0.8704	24.4034	13.1338	6.6009	7.0522	0.0012	0.0002	1.6783	0.0238
		SE	0.5508	5.0413	3.9075	2.1527	4.2137	0.0001	0.0001	0.2069	
18	20	40	-2.7555	33.4194	11.2141	1.4972	12.0953	0.001	0.0001	1.9296	0.0218
		SE	0.9096	5.3721	5.0045	1.8145	5.6983	0.0002	0.0001	0.2314	
19	20	40	-2.0039	23.3343	13.3588	0.6346	13.6872	0.0009	0.0001	0.1594	0.0208
		SE	0.6107	8.1797	4.7494	2.3584	7.9927	0.0002	0.0001	0.304	
20	20	40	-2.066	24.5366	14.166	5.6933	1.5298	0.0011	0.0002	0.4175	0.0196
		SE	0.7702	6.7699	8.1163	2.5049	5.7312	0.0001	0.0002	0.3186	
21	0	20	0.2402	18.7189	10.4307	1.4972	0.8453	0.0009	0.0001	0.5617	0.018
		SE	0.5956	6.3933	4.3684	3.1119	6.8891	0.0002	0.0001	0.2654	
22	0	20	-1.8647	24.6686	20.5413	2.3385	3.3464	0.001	0.0002	0.1592	0.0172
		SE	0.6401	3.5566	4.5503	1.6347	4.4353	0.0001	0.0002	0.2975	
23	60	80	-2.0123	23.6458	-0.6752	0.6342	20.1788	0.001	0	1.0721	0.0163
		SE	0.4861	5.2628	4.6012	1.6641	3.2149	0.0001	0.0002	0.4094	
24	60	80	-1.8647	23.3359	-3.1788	0.6354	14.4236	0.0008	0.0001	0.1592	0.0151
		SE	0.7576	6.3485	5.5501	1.8897	4.7863	0.0001	0.0001	0.4597	
25	60	80	-2.6264	30.9606	4.9773	0.6262	17.6788	0.0012	0	0.1592	0.013
		SE	0.9646	3.561	3.8845	1.1579	2.5103	0.0001	0.0001	0.1792	
26	40	60	-1.5745	15.2969	-1.926	0.0017	11.8896	0.0009	-0	0.1813	0.0122
		SE	0.7191	5.2905	4.0266	2.1168	2.784	0.0001	0.0001	0.3672	
27	40	60	-2.4068	25.6493	5.0335	0.4844	13.8374	0.001	0.0003	0.1592	0.0096
		SE	1.0259	4.6593	7.1171	2.9575	4.1201	0.0001	0.0001	0.256	
28	60	80	-3.1179	32.1664	0.5912	0.1584	13.9996	0.0011	0	0.9337	0.0087
		SE	0.9618	6.9078	5.8643	2.3984	5.1141	0.0001	0.0001	0.3502	

we use the estimated driver strategies. We match three moments:

1. The percentage of matched passengers.
2. Mean driver time in market.
3. Mean passenger time in market.

The estimated κ is 0.00055,[16] which means that the maximum waiting time of passengers is on average 30 minutes.

6 Counterfactual

We conduct four simulations to answer our research questions. We first use the estimates to simulate the "factual" market evolution. Next, we assume that driver maximum waiting time is distributed as in the first simulation, but they make static optimal decisions (myopic drivers): when given the opportunity to search for passengers, driver i answers the request $j \in R_t$ if

$$\max_{j \in R_{it}} \delta_{ij} + \varepsilon_{it} > 0.$$

In contrast, a dynamically optimizing i answers the request j if

$$\max_{j \in R_{it}} \delta_{ij} + \varepsilon_{it} > V_i.$$

The comparison between the first two simulations reveals the effect of strategic waiting.

We next examine two counterfactual centralized algorithms. The platform is assumed to know the preference of the drivers. We first simulate the "greedy" algorithm in Akbarpour et al. (2017) adapted to the two-sided market and adjusted for sorting. The platform matches a new driver or a passenger subject to the driver's incentive constraint (the driver utility must be non-negative) immediately after the agent shows up , and the platform chooses the matching partner to maximize the driver utility. If an agent is not matched, she stays in the market until she leaves without a match or gets matched with another new agent. We next simulate the "patient" algorithm. In this case, the platform is assumed to additionally know the maximum waiting time T_i. The platform matches an agent immediately before she leaves the market to an agent that maximizes the driver utility subject to the driver incentive constraint.

The main idea of the patient algorithm is that given a set of agents, the platform can increase their time in the market and therefore the chances of an agent meeting a compatible partner. We want to point out that there are other mechanisms to increase market thickness and improve match

drawing many unique drivers from the distribution of F_I (route distribution), α_i and β_i, and we need to solve for a dynamic driver problem for each new draw. Estimating a scalar parameter reduces the number of times required to evaluate the objective functions.

[16] The bootstrapped standard error, which takes into account the standard errors of the estimates of the previous two steps, is 0.0001.

efficiency. In markets where pricing is less restricted (e.g., peer-to-peer platforms in the US; airline markets (?)) than ours (the non-transferable utility framework is more similar to matches between students and schools, medical graduates and medical schools, organ donors and receipients and other contexts with a rigid or no price structure), the platform can also set prices dependent on an agent's waiting time to encourage or discourage waiting. Setting these prices correctly also requires the platform to know agent preferences and how long they can wait. Our results thus offer a measure of returns to investment in predicting these preferences.

A related issue is whether "personalized" matching rule benefits the platform at the expense of the consumers. Recent empirical studies (??) have examined the effects of personalized pricing when firms are able to predict consumer preferences. We show that the majority, but not all, of drivers and passengers are likely to benefit if the platform uses the knowledge of individual preferences to implement the patient algorithm, where high market thickness mitigates the inter-temporal mismatch.

Akbarpour et al. (2017) shows that the patient algorithm matches more homogenous agents in a one-sided market than the greedy algorithm. This result may not hold for two sided markets with heterogeneous agents who enter and leave the market at different rates. We discuss the intuition by considering the greedy and patient algorithms using the second example in Section 3. Recall that in this example, two drivers A and B and two passengers a and b stochastically enter the market. A prefers a to b to being unmatched, and B prefers a to being unmatched to b. We also assume that b will be matched with A in the greedy algorithm if both A and B are present and also in the patient algorithm if b exits first. The exit rate is sufficiently small that the probability of no-match exit is close to 0. In the greedy algorithm, both agents will be matched only if

1. (B, a) or (A, b) show up before other agents; or
2. (a, b) show up and then B shows up.

Therefore all agents are matched with probability $\frac{2}{\binom{4}{2}} + \frac{1}{\binom{4}{2}}\frac{1}{2} = \frac{5}{12}$. With probability close to $1 - \frac{5}{12}$ only one match will be formed. In the patient algorithm, with probability close to 1 matches only occur when all four agents are in the market. Assume that the driver exit rates are equal to the passenger exit rates, then all agents are matched only if B moves first, which occurs with probability $\frac{1}{4} < \frac{5}{12}$. Therefore the expected number of matches is lower under the patient algorithm. The key insight from this calculation is that when the choice of some agent, in this case passenger a or driver A, has a large negative externality on other agents, the greedy algorithm can match more agents than the patient algorithm.

We report the results in Table 4. We report the summary statistics that largely mirror those in Table 1. Because we do not specify the utility of the passengers, we separately report the match rates (percentage of requests answered) as a proxy for passenger participation and utility. The welfare measure for the drivers is in the unit of detour lengths (KM) (we divide u_{ijt} discounted to the time of arrival by α_i before averaging). While the level of such a welfare measure is not interpretable, the difference is interpreted as the length of the detour saved.[17] We therefore normalize the welfare of the factual simulation to 0 and just report the difference (expected detours saved). A higher number in the row of driver utility indicates higher welfare (more detours saved). We calculate the unconditional average trip length as a measure of platform revenue by treating the unmatched trips' length as 0.

The first column of Table 4 uses the model estimates and simulates the steady state market equilibrium. The results fit the data well for the match rate, waiting time and market thickness. In the last row, the corresponding conditional trip length (conditional on being answered) is $11.7/0.52 = 22.5$KM, and the data average is 24.5 KM.

We summarize the key observations below. All percentage differences are statistically significant at the 95% level.

1. The comparison between the "Factual" and "Myopic" results show that removing strategic waiting can increase the match rate by 10%, but the driver utility decreases. In other words, strategic waiting benefits the drivers at the expense of more unmatched passengers. Strategic waiting increases equilibrium numbers of waiting drivers and passengers.

2. The comparison between "Greedy" and "Factual" shows that the greedy algorithm can slightly increase the match rate, but drivers are on average worse off. To put the welfare measure in perspective, note that conditional on being matched, the average driver detour length is 6.14KM.

3. The comparison between "Greedy" and "Patient" shows that additional market thickness increases both the match rate (13%) and driver utility. The improvement quantifies the gains from knowing the maximum waiting time of the drivers and passengers.

4. Both the drivers and passengers wait longer with the patient algorithm. Welfare for drivers still improves because of the low driver waiting cost.

5. The unconditional average trip lengths (the passenger trip length of the drivers who end up traveling alone is 0) increase with the greedy and patient algorithms, indicating higher revenues for the platform.

The counterfactual results also illustrate the role of market thickness when drivers have heterogeneous preferences across matches. The ratio of the number of waiting driver to passengers is well over 70%. Had drivers been indifferent across matches, the match rate (for the passengers) should also be above 70% in each case, but the highest match rate is 60% across simulations. Interestingly, the "Patient" column has the lowest driver-passenger ratio, or fewest drivers per passenger, but a higher percentage of drivers and passengers are matched as a result of the thicker market. The result would be reversed if both sides have homogeneous preferences (as would be the case if the drivers only care about picking up a passenger but not the destination).

[17] The level of the welfare measure is also not identified, but the difference across scenarios is.

	Factual	Myopic	Greedy	Patient
Match Rate	0.52	0.57	0.53	0.6
SE	0.02	0.01	0.02	0.02
Driver Utility (KM)	0*	-1.57	-1.57	0.16
SE	0.31	0.14	0.14	0.24
# Waiting Drivers	198.03	178.87	188.74	260.8
SE	12.08	11.24	11.17	15.11
# Waiting Passengers	220.63	200.05	211.02	329.57
SE	16.26	17.03	17.22	15.11
Driver Time in Market (Sec)	435.01	396.46	410.08	562.55
SE	10.75	11.14	12.35	10.88
Passenger Time in Market (Sec)	423.23	386.81	392.42	598.95
SE	13.2	13.99	14.2	17.93
Unconditional Average Trip Length (KM)	11.7	12.57	12.03	12.84
SE	0.38	0.35	0.37	0.45

Table 4: Counterfactual Results

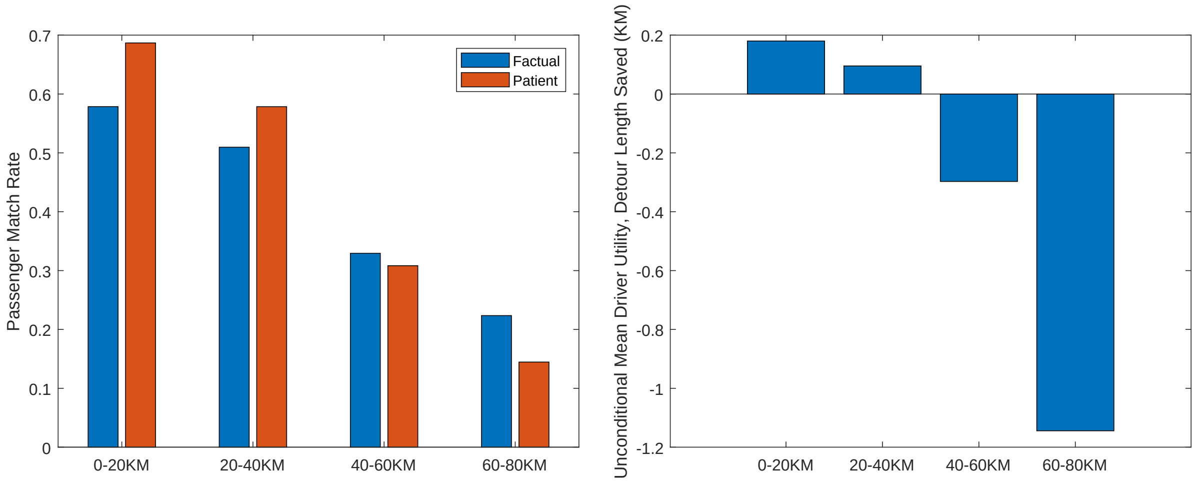
The average driver utility is an unconditional expectation, including those not matched ($u = 0$), deflated to the arrival time of the driver. The standard errors are from a 100-reptition bootstrap on the days, preserving the correlations across time within a day and taking into account the standard errors of the estimates.
*: normalized to be 0. The welfare measures of other scenarios are interpreted in the units of detours saved.

As a benchmark, we also consider an ex post optimal outcome where agents that appear at different time points can be matched. A motivation is a one-day-ahead market where there are sufficiently many agents who know their commuting needs one day ahead, and a centralized algorithm makes the match after all agents submit their preferences. Because the passenger arrival rate is greater than the driver arrival rate, the upper bound of the match rate is $\frac{\rho^0}{\kappa^0} = 90\%$.

We also note that both the outcomes of the greedy and the patient algorithm could be a lower bound on the match rates and driver utility if they are implemented. Match opportunities are triggered by arrivals or departures, and overall drivers "search" less often and receive fewer ε shocks with the two centralized algorithms than the "Factual" and "Myopic" scenarios. Potentially the platform can improve product design or provide dynamically adjusted incentives (such as discounts for temporary traffic congestion) to overcome the lack of positive shocks and achieve even better outcomes with the centralized algorithms.

We then turn to the distributional impact of the patient algorithm. We break down the changes in the match rates and driver utility by passenger and driver distance category (0-20KM, 20-40KM, 40-60KM and 60-80KM) in Fig. 9. Shorter distance passengers and drivers (<40KM), which together account for 84% of the participants, are better off (higher match rates for passengers, positive detour lengths saved for drivers) with the patient algorithm, but longer distance ones are worse off. To put the welfare measures in perspective, the average detour lengths of the four distance groups in data are 3.09KM, 4.55KM, 7.60KM and 8.70KM.

A key assumption for our counterfactual results is that driver and passenger participation (ρ^0



Match Rates

Driver Utility*

Figure 9: Distributional Impact of the Patient Algorithm

*: The average detour lengths of the answered trips in the four distance groups in data are 3.09KM, 4.55KM, 7.60KM and 8.70KM.

and κ^0) and the distribution of the maximum waiting time (ρ and κ) are "structural": i.e. they stay the same if the centralized matching algorithm is implemented. One would expect these rates to be dependent on the expected outcomes of the participation: if passengers expect the match rate to be lower, the entry rates might fall and precipitate an even lower match rate. Similarly, if drivers expect the average utility to be lower, the driver arrival rate might decrease. The counterfactual simulations suggest that the match rate increases, and we should expect the entry rate to be at least as high as before.

7 Conclusion

We present an empirical framework to analyze the efficiency of a decentralized dynamic matching market. Our model flexibly takes into account many dimensions of driver heterogeneity and two types of driver strategies. We find that (1) drivers' strategic waiting increases the market thickness, increases driver welfare but decreases the number of matches relative to myopic drivers, and (2) the patient algorithm, which decreases match frequency and increases market thickness, can substantially increase the number of matches. Our model additionally indicates that the patient algorithm also improves the average driver welfare. We think of the counterfactual analysis as measuring the returns to the platform's or a social planner's investment on inferring agent preferences. If the platform can accurately solicit or infer agent preferences, we show that there exist centralized algorithms that increase the number of matches and the platform revenue.

References

Akbarpour, Mohammad, Shengwu Li, and Shayan Oveis Gharan, "Thickness and information in dynamic matching markets," 2017.

Andreevska, Daniela, "What Kind of Airbnb Occupancy Rate Can You Expect?," *MASHVISOR*, 2016.

Arcidiacono, Peter, Patrick Bayer, Jason R Blevins, and Paul B Ellickson, "Estimation of dynamic discrete choice models in continuous time with an application to retail competition," *The Review of Economic Studies*, 2016, *83* (3), 889–931.

Arnosti, Nick, Ramesh Johari, and Yash Kanoria, "Managing congestion in matching markets," 2015.

Ashlagi, Itai, Maximillien Burq, Patrick Jaillet, and Vahideh Manshadi, "On matching and thickness in heterogeneous dynamic markets," *arXiv preprint arXiv:1606.03626*, 2016.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv, "Optimal dynamic matching," 2016.

Backus, Matthew and Gregory Lewis, "Dynamic demand estimation in auction markets," Technical Report, National Bureau of Economic Research 2016.

Banerjee, Siddhartha, Carlos Riquelme, and Ramesh Johari, "Pricing in ride-share platforms: A queueing-theoretic approach," 2015.

Bodoh-Creed, Aaron, Joern Boehnke, and Brent Richard Hickman, "How Efficient are Decentralized Auction Platforms?," 2017.

Buchholz, Nicholas, "Spatial equilibrium, search frictions and efficient regulation in the taxi industry," Technical Report 2017.

Castillo, Juan Camilo, Dan Knoepfle, and Glen Weyl, "Surge pricing solves the wild goose chase," in "Proceedings of the 2017 ACM Conference on Economics and Computation" ACM 2017, pp. 241–242.

Chiappori, Pierre-André and Bernard Salanié, "The econometrics of matching models," *Journal of Economic Literature*, 2016, *54* (3), 832–861.

Choo, Eugene, "Dynamic marriage matching: An empirical framework," *Econometrica*, 2015, *83* (4), 1373–1423.

_ **and Shannon Seitz**, "The Collective Marriage Matching Model: Identification, Estimation, and Testing," in "Structural Econometric Models," Emerald Group Publishing Limited, 2013, pp. 291–336.

Doraszelski, Ulrich and Kenneth L Judd, "Avoiding the curse of dimensionality in dynamic stochastic games," *Quantitative Economics*, 2012, *3* (1), 53–93.

Einav, Liran, Chiara Farronato, and Jonathan Levin, "Peer-to-peer markets," *Annual Review of Economics*, 2016, *8*, 615–635.

Farronato, Chiara and Andrey Fradkin, "The welfare effects of peer entry in the accommodation market: The case of airbnb," Technical Report, National Bureau of Economic Research 2018.

Feng, Guiyun, Guangwen Kong, and Zizhuo Wang, "We are on the Way: Analysis of On-Demand Ride-Hailing Systems," 2017.

Fox, Jeremy T, "An Empirical, Repeated Matching Game Applied to Market Thickness and Switching," 2008.

_ , "Specifying a Structural Matching Game of Trading Networks with Transferable Utility," *American Economic Review*, 2017, *107* (5), 256–260.

Frechette, Guillaume R, Alessandro Lizzeri, and Tobias Salz, "Frictions in a Competitive, Regulated Market Evidence from Taxis," 2016.

Hall, Jonathan, Cory Kendrick, and Chris Nosko, "The effects of Uber′s surge pricing: A case study," 2015.

Hall, Jonathan V, John J Horton, and Daniel T Knoepfle, "Labor Market Equilibration: Evidence from Uber," 2017.

Hatfield, John William, Scott Duke Kominers, Alexandru Nichifor, Michael Ostrovsky, and Alexander Westkamp, "Stability and competitive equilibrium in trading networks," *Journal of Political Economy*, 2013, *121* (5), 966–1005.

Hu, Ming and Yun Zhou, "Dynamic type matching," 2016.

Iyer, Krishnamurthy, Ramesh Johari, and Mukund Sundararajan, "Mean field equilibria of dynamic auctions with learning," *Management Science*, 2014, *60* (12), 2949–2970.

Krusell, Per and Anthony A Smith Jr, "Income and wealth heterogeneity in the macroeconomy," *Journal of political Economy*, 1998, *106* (5), 867–896.

Lagos, Ricardo, "An analysis of the market for taxicab rides in New York City," *International Economic Review*, 2003, *44* (2), 423–434.

Loertscher, Simon, Ellen V Muir, and Peter G Taylor, "Optimal Market Thickness and Clearing," 2016.

Ostrovsky, Michael and Michael Schwarz, “Carpooling and the Economics of Self-Driving Cars,” 2018.

Ozkan, Erhun and Amy R Ward, “Dynamic matching for real-time ridesharing,” 2016.

Roth, Alvin E and Marilda Sotomayor, “Two-sided matching,” *Handbook of game theory with economic applications*, 1992, *1*, 485–541.

_ , Tayfun Sönmez, and M Utku Ünver, “Pairwise kidney exchange,” *Journal of Economic theory*, 2005, *125* (2), 151–188.

_ , _ , and _ , “Efficient kidney exchange: Coincidence of wants in markets with compatibility-based preferences,” *American Economic Review*, 2007, *97* (3), 828–851.

Ünver, M Utku, “Dynamic kidney exchange,” *The Review of Economic Studies*, 2010, *77* (1), 372–414.

Weintraub, Gabriel Y, C Lanier Benkard, and Benjamin Van Roy, “Markov perfect industry dynamics with many firms,” *Econometrica*, 2008, *76* (6), 1375–1411.

A Numerical Solution for Section 3

We begin by writing down the driver problem. Consider driver A. Use $\{o, m, e\}$ to denote whether an agent has not entered the market, is in the market or has exited the market. For example,

$$S_{At} = \left\{ \begin{array}{cc} a & o \\ b & m \\ B & e \end{array} \right\}$$

represents the state a has not entered the market, b is in the market, and B has left the market. The value function for A for an infinitely small period Δ thus is

$$\begin{aligned} V_A\left(S_{At}\right) = \underbrace{\Delta\kappa V\left(S_{At}^1\right)}_{b \text{ leaves}} + \underbrace{\Delta\kappa^0 V\left(S_{At}^2\right)}_{a \text{ enters}} + \underbrace{\Delta\gamma \max\left\{u_{bA}, V\left(S_{At}\right)\right\}}_{A \text{ moves}} \\ + \underbrace{\Delta\rho \cdot 0}_{A \text{ exits}} + \underbrace{\left(1 - \Delta\kappa - \Delta\kappa^0 - \Delta\gamma - \Delta\rho\right) V_i\left(S_{At}\right)}_{\text{nothing happens}}, \end{aligned}$$

where

$$S_{At}^1 = \left\{ \begin{array}{cc} a & o \\ b & e \\ B & e \end{array} \right\}, S_{At}^2 = \left\{ \begin{array}{cc} a & m \\ b & m \\ B & e \end{array} \right\}$$

which simplifies to

$$V_A\left(S_{At}\right)=\frac{\kappa V\left(S_{At}^1\right)+\kappa V\left(S_{At}^2\right)+\gamma \max\left\{u_{bA},V\left(S_{At}\right)\right\}}{\kappa+\kappa^0+\gamma+\rho}.$$

The value function is more complicated when B is in the market. Consider

$$S_{At}=\left\{\begin{array}{cc} a & o \\ b & m \\ B & m \end{array}\right\}.$$

We focus on pure strategy equilibrium, and assume that B's strategy is deterministic. For the purpose of the presentation, we assume that B will pick up passenger b if presented the opportunity. Thus the value function can be written as

$$V_A\left(S_{At}\right)=\frac{\left(\gamma+\kappa\right)V\left(S_{At}^1\right)+\kappa V\left(S_{At}^2\right)+\gamma \max\left\{u_{bA},V\left(S_{At}\right)\right\}}{\kappa+\kappa^0+2\gamma+\rho}.$$

We can similarly write down the value functions for other states.

For the two scenarios discussed in Section 3, we use the following parameterization and solve the Bellman equation for the value function and the strategy function. Use $\emptyset$ to denote the case of being unmatched. In the first case, we assume that

$$A: u_{aA}=2>u_{bA}=1>u_{\emptyset A}=0$$

$$B: u_{bB}=2>u_{\emptyset B}=0>u_{aB}=-1$$

We also assume that $\kappa^0=\rho^0=\gamma=1$ and $\kappa=\rho=0.1$. In a pure strategy Bayesian equilibrium where B always waits for b, the minimum V_A when a has not exited the market and b is in the market is 1.5565$>u_{bA}$ for state (omo), which means that A will always wait for a if a has not entered. In the second case, if we assume that

$$A: u_{aA}=2>u_{bA}=1>u_{\emptyset A}=0$$

$$B: u_{aB}=2>u_{\emptyset B}=0>u_{bB}=-1.$$

In the pure strategy Bayesian equilibrium where B always waits for a, the minimum V_A when a has not exited the market and b is in the market is 0.7325$<u_{bA}$ for state (omm), which means that A will answer b and not wait for a if both b and B are in the market. If A more strongly prefers a, the result is reversed: if we assume that

$$A: u_{aA}=5>u_{bA}=1>u_{\emptyset A}=0$$

$$B: u_{aB}=2>u_{\emptyset B}=0>u_{bB}=-1.$$

the minimum V_A when a has not exited the market and b is in the market is 1.2585$>$ u_{bA} for state (omm), meaning that A will always wait for a despite the presence of b and regardless of the presence of B.

B Construction of λ_i and $g_{R,i}$

In this section, we fully specify the driver decision problem, where R_{it}'s evolution is governed by the entry and exit of passengers. The exits are the results of either passengers waiting until the maximum waiting time or being picked up by drivers. We show that λ_i and $g_{R,i}$ can be constructed from the κ^0, κ and rival driver strategies. A key assumption is that the entry rates and no match exit rates (distribution of maximum waiting time) are constant. Use p_j to denote the probability of the arriving passenger being j, which encodes both the passenger route type and unobserved match value ξ. Denote the support of passenger types as $\mathcal{R}$. Denote the equilibrium probability of a driver $\imath$ picking passenger j as $p_{\imath j}\left(R_{it}\right)$. The full Bellman equation is

$$\begin{aligned}
V_i\left(\tau, R_{it}\right) = \frac{1}{1+\theta_i \mathit{\Delta}} \Bigg[& \underbrace{\mathit{\Delta}\kappa^0 \sum_{j\in\mathcal{R}} p_j V\left(\tau+\mathit{\Delta}, R_{it}\cup j\right)}_{\text{new arrival}} \\
& + \underbrace{\mathit{\Delta}\kappa \sum_{j\in R_{it}} V\left(\tau+\mathit{\Delta}, R_{it}\backslash j\right)}_{\text{exits where a passenger reaches its } T_j} \\
& + \underbrace{\mathit{\Delta}\gamma E_{D|R}\left[\sum_{j\in R_{it}}\sum_{\imath\in D_{it}} p_{\imath j}\left(R_{it}\right) V\left(\tau+\mathit{\Delta}, R_{it}\backslash j\right)|R_{it}\right]}_{\text{Another driver picks up } j} \\
& + \underbrace{\mathit{\Delta}\gamma \int_{\varepsilon_{it}} \max\left\{\max_{j\in R_{it}} \underline{u}_i + \delta_{ij} + \varepsilon_{it}, V_i\left(\tau+\mathit{\Delta}, R_{it}\right)\right\} f_\varepsilon\left(\varepsilon_{it}\right) d\varepsilon_{it}}_{i \text{ receives a move opportunity}} \\
& + \underbrace{\left(1-\mathit{\Delta}\left(\kappa^0\sum_{j\in\mathcal{R}} p_j + \#R_{it}\cdot\kappa + \gamma E_{D|R}\left[\sum_{j\in R_{it}}\sum_{\imath\in D_{it}} p_{\imath j}\left(R_{it}\right)|R_{it}\right]\right) - \mathit{\Delta}\gamma\right) V_i\left(\tau+\mathit{\Delta}, R_{it}\right)}_{\text{nothing happens}} \Bigg]
\end{aligned}$$

where $E_{D|R}\left[\cdot\,|R_{it}\right]$ is the conditional expectation with respect to the set of waiting drivers in the stationary equilibrium. The above can be re-arranged into Eq. (3), where

$$\lambda_i\left(R_{it}\right) = \kappa^0 \sum_{j\in\mathcal{R}} p_j + \#R_{it}\cdot\kappa + \gamma E_{D|R}\left[\sum_{j\in R_{it}}\sum_{\imath\in D_{it}} p_{\imath j}\left(R_{it}\right)|R_{it}\right],$$

	Factual	Myopic	Greedy	Patient
Match Rate	0.51	0.55	0.52	0.58
SE	0.01	0.01	0.01	0.02
Driver Utility (KM)	0*	-1.49	-1.42	0.07
SE	0.33	0.13	0.15	0.18
# Waiting Drivers	204.47	185.09	192.48	263.64
SE	12.15	12.38	11.56	15.05
# Waiting Passengers	227.42	205.87	215.08	331.24
SE	16.74	15.96	16.31	14.86
Driver Time in Market (Sec)	445.03	407.43	421.52	568.74
SE	12.29	10.07	10.5	14.45
Passenger Time in Market (Sec)	426.02	392.28	394.48	597.79
SE	13.81	11.15	13.85	14.27
Unconditional Average Trip Length (KM)	11.55	12.48	11.97	12.61
SE	0.38	0.32	0.39	0.4

Table 5: Counterfactual Results

The average driver utility is an unconditional expectation, including those not matched ($u = 0$), deflated to the arrival time of the driver. The standard errors are from a 100-reptition bootstrap on the days, preserving the correlations across time within a day and taking into account the standard errors of the estimates.
*: normalized to be 0. The welfare measures of other scenarios are interpreted in the units of detours saved.

the transitional probability $R_{it} \to R_{it} \backslash j$ is

$$g\left(R_{it} \backslash j \,|R_{it}\right) = \frac{1}{\lambda_i\left(R_{it}\right)} \left(\kappa + \gamma E_{D|R}\left[\sum_{\imath \in D_{it}} p_{\imath j}\left(R_{it}\right) |R_{it}\right]\right),$$

and the transitional probability $R_{it} \to R_{it} \cup j$ is

$$g\left(R_{it} \cup j \,|R_{it}\right) = \frac{\kappa^0 p_j}{\lambda_i\left(R_{it}\right)}.$$

C Passenger Unobserved Heterogeneity

We use the estimates of σ_ξ and simulate the outcomes of alternative model where ξ is a vertical attributes: ξ_j is i.i.d across j, and the value of i picking up j is

$$u_{ijt} = \underline{u}_i + u_{i0} - \alpha_i d_{ij} + \beta_i x_j + \xi_j + \varepsilon_{it}.$$

The results presented in Table 5 are quite similar to the ones in the main text.