To pool or not to pool: Equilibrium, pricing and regulation

Kenan Zhang[1] and Yu (Marco) Nie [*1]

[1]Department of Civil and Environmental Engineering, Northwestern University
2145 Sheridan Road, Evanston, IL 60208, USA

July 1, 2021

Abstract

We study a transportation network company (TNC) that offers on-demand solo and pooling e-hail services in an aggregate mobility service market, while competing with transit for passengers. The market equilibrium is established based on a spatial driver-passenger matching model that determines the passenger wait time for both solo and pooling rides. We prove, under mild conditions, this system always has an equilibrium solution. Built on the market equilibrium, three variants of pricing problems are analyzed and compared, namely, (i) profit maximization, (ii) profit maximization subject to regulatory constraints, and (iii) social welfare maximization subject to a revenue-neutral constraint. A comprehensive case study is constructed using TNC data collected in the city of Chicago. We found pooling is desirable when demand is high but supply is scarce. However, its benefit diminishes quickly as the average en-route detour time (i.e., the difference between the average duration of solo and pooling trips) increases. Without regulations, a mixed strategy—providing both solo and pooling rides—not only achieves the highest profit and trip production in most scenarios, but also gains higher social welfare. The minimum wage policy can improve social welfare in the short term. However, in the long run, the TNC could react by limiting the size of the driver pool, and consequently, render the policy counterproductive, even pushing social welfare below the unregulated level. Moreover, by maintaining the supply and demand of ride-hail at an artificially high level, the minimum wage policy tends to exacerbate traffic congestion by depressing the use of collective modes (transit and pooling). A congestion tax policy that penalizes solo rides promotes pooling, but may harm social welfare. However, it promises to increase both social welfare and pooling ratio when jointly implemented with the minimum wage policy.

Keywords: e-hail; pooling; pricing; equilibrium; regulation; minimum wage; congestion tax

1 Introduction

The rise of transportation network companies (TNCs), such as Uber, Lyft and Didi Chuxing, has reshaped the ride-hail industry in the past decade. The e-hail service offered by TNCs is praised

*Corresponding author, E-mail: y-nie@northwestern.edu; Phone: 1-847-467-0502.

for its advanced real-time matching, dynamic pricing and flexible labor supply (Azevedo and Weyl, 2016; Cramer and Krueger, 2016; Nie, 2017). Many TNCs also provide on-demand ride-sharing (e.g., UberPool and LyftShared), referred to as *pooling* hereafter. Unlike conventional ride-sharing (e.g., carpooling), pooling occurs in real time and is served by dedicated drivers who are dispatched and paid by TNCs. When requesting a ride, a passenger may select a full-price solo ride or a discounted pooling ride. If she chooses to pool, the platform would search for fellow passengers who may share a portion of her trip. Generally, pooling tends to increase the distance and duration of a trip due to detours. Hence, for passengers, the primary trade-off is between a lower trip fare and a longer travel time. As for TNCs, pooling helps increase the service capacity without hiring more drivers. Also, as drivers are primarily paid according to the time and distance associated with the passenger-delivering portion of the trip[1], referred to as *occupied time and distance* hereafter, TNCs may generate a greater profit by serving more pooling rides. However, the detours due to pooling consume extra vehicle time, which, if not properly controlled, could compromise the overall level of service and compel passengers to choose alternative modes, notably transit. Therefore, the TNCs often have to cut the price of pooling to make it sufficiently attractive, which creates a downward pressure on the profit. On the other hand, pooling promises a smaller fleet size that can be used more efficiently. Hence, it is also an effective response to the outcry of the cities that have been blaming TNCs for, among other things, intensified traffic congestion (Schaller, 2017, 2018; Erhardt et al., 2019).

Central to the complex interdependent relationship among the stakeholders—passengers, drivers, the TNCs and the regulator—are two choices: (i) the passengers' mode choice among solo ride, pooling ride and transit; and (ii) the TNCs' pricing strategies for solo and pooling rides and for the compensation to drivers. The objective of this paper is to model these decision processes, identify various factors that may influence them and quantify their relative effects on the conditions of an aggregate personal mobility service market. Of particular interests are the TNCs' profitability, the market share of the collective modes (pooling and transit) and the social welfare. We shall also examine how regulations might affect these outcomes.

To the above end, we consider an e-hail platform (referred to as *the platform* hereafter) that offers both solo and pooling rides. The two services are jointly priced, along with the payment received by drivers[2]. On the demand side, passengers choose a *mode* from solo ride, pooling ride and transit based on a generalized cost that incorporates trip fare, wait time and trip duration. On the supply side, the platform has access to a pool of qualified drivers who enter the service only if the average earning rate exceeds a *reservation rate* (which may be considered as the opportunity cost for drivers). We note that, even though the platform dominates the e-hail service, it does not monopolize the whole mobility market since we assume the three modes (transit, solo ride and pooling ride) substitute each other. The platform possesses no monopoly power over its supply either, because TNC drivers are allowed to choose flexible work schedules based on their reservation rates and the earning rate of the platform.

The interaction between passengers and drivers is captured by a physical matching process that dictates the passenger wait time. We extend the spatial matching theory developed in Chen

[1]See e.g., Uber (`https://www.uber.com/us/en/drive/services/shared-rides/`) and Lyft (`https://help.lyft.com/hc/en-us/articles/115012926987-Shared-ride-driver-pay`) policies.

[2]Unlike a typical wage, this payment fluctuates with market conditions. We note that drivers are considered as "contractors" rather than "employees", and hence they do not earn a wage.

et al. (2018) and Zhang et al. (2019) to model the case of pooling between two passengers. Specifically, the wait time of each pooling ride is separated into two parts. The first corresponds to the pickup process of the passenger closer to the matched vehicle, while the second is the detour to pickup the other passenger. Accordingly, the total wait time depends on both the density of vacant vehicles and the density of pooling passengers in the market. The market equilibrium is formulated as a nonlinear equation system, on which we build the platform's pricing problem with the objective of (i) profit maximization, (ii) profit maximization subject to a regulatory constraint, and (iii) social welfare maximization subject to a revenue neutral constraint. We investigate two regulatory policies that have already been enacted or are currently under consideration in major US cities, namely, a minimum wage policy and a congestion tax policy. Our case study, built on TNC data collected in Chicago, reveals that the two policies each have pros and cons, and yet combining them promises to simultaneously promote collective modes and improve social welfare.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 presents the physical model that describes the matching process with pooling, from which the expected wait times are derived. A model of the aggregate market equilibrium is presented in Section 4, along with a discussion of its existence. Optimal pricing problems are formulated in Section 5, as well as corresponding solution algorithms. Section 6 reports and discusses the results of the case study. Section 7 summarizes the main findings and comments on the directions for future research.

2 Related work

The ride-hail market, traditionally dominated by taxis, has been studied extensively. Douglas (1972) and Arnott (1996) characterize the aggregate equilibrium between passenger demand and vehicle supply in street-hail and radio-dispatch taxi services, respectively. In both studies, as well as in many following studies, passenger demand is assumed to decrease with trip fare and expected wait time. In addition, the wait time—a primary measure of the level of service (LOS)—is typically modeled as a decreasing function of vacant vehicle density (Douglas, 1972; Beesley and Glaister, 1983; Arnott, 1996; Zha et al., 2018b). This gives the salient feature of the ride-hail market: to achieve a desired LOS, a portion of the supply has to be "wasted" deliberately in the form of vacant vehicle time. In this regard, the radio-dispatch service differs from the street-hail taxis because it wastes supply not only on searching, but also on the *pickup time* spent by a "matched" vehicle to reach a passenger. This difference turns out to be rather consequential for e-hail, which is our focus here and can be viewed as a modern and more sophisticated version of radio-dispatch. Castillo et al. (2018) shows that this extra pickup phase can lead the system to an inefficient state called *Wild Goose Chase* (WGC), where most vehicles are stuck on the way to pick up passengers due to demand surge or suboptimal dispatching polices. The central question for the equilibrium analysis of ride-hail is how passenger wait time is jointly determined by demand, supply and the matching process that links them. Hence, in what follows, we first review previous studies on matching, before turning to pricing and regulations. We end with recent studies on modeling e-hail with pooling.

2.1 Matching

The matching process has been characterized by different approaches in the literature. Some studies assume it is frictionless, i.e., the number of matched trips simply equals to the demand or the supply, whichever is smaller (e.g., Lagos, 2000; Bimpikis et al., 2019). Others introduce an aggregate matching function to describe the pickup rate as a function of two primary inputs: the number of unmatched vehicles and the number of waiting passengers. One widely used function is the Cobb-Douglas function (e.g., Yang et al., 2010b; Yang and Yang, 2011; He and Shen, 2015; Wang et al., 2016; Zha et al., 2016), and the other is the urn-ball matching function (e.g., Shapiro, 2018; Buchholz, 2019), which models the matching process as Bernoulli trials. Another possibility to obtain an aggregate matching function is through simulation (e.g., Frechette et al., 2019). The third line of research attempts to simplify the matching process as a queue (e.g., Banerjee et al., 2015; Afeche et al., 2018; Xu et al., 2019). However, the queuing models often rely on the assumptions that passengers are picked up immediately after being matched with a vacant vehicle. The two exceptions are Besbes et al. (2018a) and Feng et al. (2017). Both studies incorporate a state-dependent pickup time into the service rate of an $M/M/n$ queuing system. However, the impact of passenger competition on the pickup time is described by the total number of passengers in the system, including those that are waiting for pickup and those that are en-route to their destination. Further, empirical evidence has shown that the pickup time in e-hail does not follow an exponential distribution, as usually assumed in the queuing models (Yan et al., 2019; Zhang et al., 2019).

A few recent studies start to pay closer attention to the physical nuances that are unique in the matching problem for ride-hail. Zha et al. (2018b) propose a geometric matching model to estimate the average matching and pickup time, though it still relies on a hypothetical functional form of matching. Yang et al. (2020a) propose a physical matching model based on the notion of the "dominant-zone", within which there is only one waiting passenger, and thus she is always matched to the closest idle vehicle. Chen et al. (2018) and Zhang et al. (2019) develop a spatial matching theory that encapsulates the matching process of both cruising taxi and e-hail as special cases. Their model yields closed-form expected passenger wait time, and is calibrated and validated by empirical data.

2.2 Pricing

The pricing problem has attracted a great deal of interest in ride-hail research, with a significant amount of work devoted to the real-time pricing and price-aware dispatching (e.g., Asghari and Shahabi, 2018; Tong et al., 2018; Xu et al., 2018; Nourinejad and Ramezani, 2019). The focus of the present study, however, is pricing in the context of equilibrium analysis. In this line of research, the most popular topic is *surge pricing*, i.e., dynamically adjusting price in response to demand surge (e.g., Cachon et al., 2017; Besbes et al., 2018b; Hu et al., 2018; Garg and Nazerzadeh, 2019). Castillo et al. (2018) show that surge pricing helps prevent the system from falling into a catastrophic failure caused by WGC. Others have demonstrated that surge pricing is more robust than static pricing policies (Banerjee et al., 2015) and, contrary to the conventional wisdom, could be useful even in areas that are less profitable or have excessive supply as it incentivizes drivers to relocate (Besbes et al., 2018b; Guda and Subramanian, 2018). Although the platform and

drivers in general benefit from surge pricing, passengers are the ones who have to bear the cost. Thus, it has been argued that regulations may be needed to check this practice, especially when a platform has de facto monopoly (Zha et al., 2018a). Yang et al. (2020b) demonstrate that capping the surge price leads to losses in profit and social welfare. They suggest integrating a reward scheme with surge pricing to solve this problem.

Pricing strategies are also designed to achieve a spatial balance between demand and supply, often known as spatial pricing (e.g., Zha et al., 2018b). Recently, Bimpikis et al. (2019) show that both the platform's profit and the consumer surplus can be maximized by an optimal pricing policy under a "balanced" demand pattern. In addition, a few studies examine the impacts of demand sensitivity to wait time and the service capacity on the optimal price of an on-demand service platform (e.g., Bai et al., 2018; Taylor, 2018).

Our focus here is the joint pricing problem for a mixture of solo and pooling rides, which has not been fully examined in the literature. It is worth emphasizing that the objective of "pricing" here is not to split the cost between passengers and drivers to achieve a stable match, as pursued by Furuhata et al. (2014) and Wang et al. (2018), but rather to maximize the platform's profit. While the potential of pooling has been well established in the literature (e.g., Santi et al., 2014; Alonso-Mora et al., 2017), it remains unclear how much of this potential can be achieved by a profit-maximizing platform.

2.3 Regulations

Before the launch of Uber in 2009, the ride-hail industry in most cities around the world had been tightly regulated, in terms of both price and entry. As explained above, conventional taxi firms must supply vacant vehicles to maintain certain LOS but cannot directly price this part of the operation. Since a greater use could help offset this uncovered cost, ride-hail is essentially a decreasing-average-cost industry (Douglas, 1972; Beesley and Glaister, 1983). This in turn implies that the industry should display increasing returns to scale, and thereby is subject to natural monopoly that would produce below the efficient level (Hotelling, 1938; Arnott, 1996). On the other hand, full competition in a ride-hail market is unlikely to maximize social welfare (Douglas, 1972; De Vany, 1975; Cairns and Liston-Heyes, 1996). Both observations suggest regulatory interventions may help, prompting recurrent debates about whether and how to regulate (Frankena and Pautler, 1986; Cairns and Liston-Heyes, 1996; Flores-Guri, 2003; Yang et al., 2005, 2010a). Other than controlling price and entry, some even argue the taxi industry should be subsidized in order to achieve the social optimum (e.g., Arnott, 1996).

While e-hail appears to share a similar cost structure with taxi (Zha et al., 2016), it has a radically different supply model that challenges both operators and regulators. In particular, e-hail drivers are considered "independent contractors" who neither earn a fixed wage nor commit to a fixed work schedule. In other words, the platform does not have a full control over its service capacity. The regulations currently under consideration focus on capping the number of licences and setting a minimum "wage" to protect drivers (Joshi et al., 2019). Gurvich et al. (2019) analyze the service capacity management problem for a service provider relying on a flexible supply model similar to that of e-hail. The authors argue that imposing a minimum wage will force the provider to restrict the number of agents on the platform during certain time periods. This implies that, in response to a minimum wage regulation, e-hail platforms may cap the number

of drivers online when the supply is sufficient, effectively limiting their scheduling flexibility. A similar argument is given by Asadpour et al. (2019). The authors show that, under certain market conditions, the platform cannot satisfy the required wage floor while maintaining a non-negative profit. As a result, it would respond to the regulation by either exiting the market, or reducing drivers' flexibility. However, Parrott and Reich (2018) conclude that TNCs "could easily absorb an increase in driver pay with a minimal fare adjustment and little inconvenience to passengers". Specifically, their simulation results indicate the minimum wage policy proposed by the New York City will only lead to a relatively minor increase in passenger wait time. A recent study by Li et al. (2019) shows that a cap on the number of vehicles benefits the platform but hurts drivers. On the other hand, imposing a minimum wage benefit both drivers and passengers because it pushes the platform to hire more drivers. Yu et al. (2019) analyze the welfare effects of the entry control policy in a market where e-hail and traditional taxi services compete for passengers. They conclude that there exists an optimal capacity cap that can best balance competing objectives of various stake holders. Our reading of literature above suggests no study has examined what partially motivates our study: the impact of regulations on an e-hail platform serving both solo and pooling rides.

2.4 E-hail with pooling

Studies on modeling e-hail with pooling in the context of equilibrium analysis emerged only recently. In Jacob and Roet-Green (2018), a TNC platform offers solo and/or pooling rides to passengers with two different levels of willingness to pay. They derive the optimal operational and pricing strategies under various demand conditions. The analysis only considers two discrete price levels and assumes drivers are paid at a fixed commission rate. Since passenger wait time is not explicitly considered, the model is unable to fully capture the spatial demand-supply interaction in the ride-hail market. Yan et al. (2019) propose and analyze a pool-matching model called *dynamic waiting*. Assuming pooling passengers be picked up simultaneously at the midpoint of their origins, they model and calibrate the pickup time as as a power function of the number of idle vehicles nearby. The model, however, seems to lack an explicit mechanism to link important features in pooling, such as detours and pooling demand, to passenger wait time. Ke et al. (2020) formulate the market equilibria with either solo or pooling rides. The authors show that, under certain conditions, the equilibrium price in a pooling market is lower than that in a market without pooling at the monopoly equilibrium, the social optimum, or a second-best equilibrium. This is because, in a pooling market, a unit decrease in trip fare not only attracts more passengers due to negative price elasticity, but also improves the level of service by reducing the detour, which is negatively correlated with demand. The latter, however, is derived from the assumption that the average detour time is inversely proportional to passenger demand. Another main difference from this study is that, when solving the optimal pricing problem, Ke et al. (2020) treat the vehicle fleet size as a decision variable that is fully controlled by the platform.

3 Matching with pooling

In this study, a *pooling ride* is defined as a trip shared by two passengers. It starts when an idle driver is assigned to pick up both passengers, one at a time, and ends when both passengers arrive at their destinations. A *solo ride* is referred to a regular e-hail trip with only one passenger. A driver cannot be assigned to another ride until he finishes serving his current ride, solo or pooling.

To simplify notations, we name the passengers who choose solo rides as *solo passengers*, and those choosing pooling as *pooling passengers*. If the platform does not offer pooling rides, all passengers must choose solo rides. We name passengers in such a case as *e-hail passengers* to distinguish them from solo passengers who have the pooling option. Accordingly, we use subscripts s, p and e to denote variables associated with solo, pooling and e-hail rides, respectively. Notations used in this study are listed in Table A1 in Appendix A.

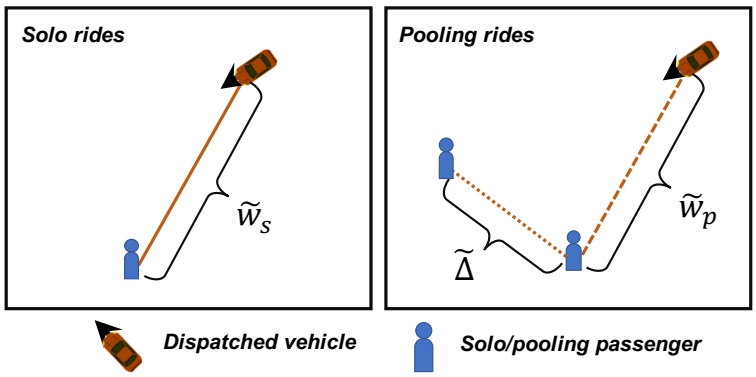


Figure 1: Pickup process of solo and pooling rides.

Figure 1 illustrates how passenger wait times for solo and pooling rides are modeled in this study. The wait time for a solo ride, denoted as $\tilde{w}_s$[3], depends on the distance between the passenger and the vehicle assigned to her (see the left panel in Figure 1).

A pooling ride is matched as follows. After a passenger makes a request, she will be paired with another passenger with relatively close origin and destination. Then, an idle driver is assigned to the pair. This process is similar to what is employed in Jacob and Roet-Green (2018) and Ke et al. (2020) except that no upper bound is imposed on the maximum matching interval in our model. Instead, we rely on the matching mechanism itself to dissuade passengers from choosing pooling, when the matching interval becomes unbearably long. Specifically, the wait time for a pooling ride consists of two parts (see the right panel of Figure 1). The first part, denoted as $\tilde{w}_p$, is the *pickup time* for the passenger closer to the matched vehicle. The second part, denoted as $\tilde{\Delta}$, is the *detour time* consumed in picking up the other passenger, determined by the distance between the two passengers sharing the trip. Since the pooling trip *starts* when both passengers are picked up, $\tilde{\Delta}$ is included in both passengers' wait time, even though the first passenger is already in the vehicle in this period. Note that the demand for pooling is closely related to $\tilde{\Delta}$: the lower is the demand, the longer is $\tilde{\Delta}$.

[3]Random variables used in this study are labelled with a header "~".

Before presenting the matching model, we wish to emphasize that the objective here is to represent the relationship between macroscopic variables that can be meaningfully measured and predicted in a highly idealized macroscopic market, such as average wait time, and the density of passengers/vacant vehicles. As a compromise, most details of the matching process, including matching time (i.e., the duration from the moment when the platform receives a request to the moment when it assigns a vehicle to serve the ride), matching criteria (e.g., distance, shareability), and dynamic pooling strategies, are left out. Because these details vary strongly across time, space and platform, it is difficult to represent them explicitly and satisfactorily by the equilibrium state of our model. In addition, their effects on wait time and detour time, which are the most important attributes of the ride-hail service, appear to be of secondary importance. For example, the matching time is usually at least an order of magnitude smaller than the pickup time (Zha et al., 2018b), whether vehicles are assigned to requests immediately (Castillo et al., 2018) or in a batching process (Yang et al., 2020a). The only exception is when there is a severe supply-demand imbalance and passengers must wait to be matched in a virtual queue. However, such an extreme case need not concern us because (i) the focus here is on a relatively long-term market equilibrium; and (ii) transit would serve as a fallback option. Although we do not model the detailed matching process, we shall implicitly capture its aggregate effect on wait time and detour time by introducing exogenous parameters that can be calibrated from empirical data.

3.1 Expected wait time of e-hail passengers

Consider a passenger who requests a ride from an e-hail platform. Her wait time depends on the density of vacant vehicles Λ and the density of waiting passengers Π in the market (Zhang et al., 2019). Vacant vehicles consists of *unmatched* vehicles, with a density $\Lambda_0 = b_\Lambda \Lambda$, and *matched* vehicles, with a density $\Lambda_1 = (1 - b_\Lambda)\Lambda$. Similarly, waiting passengers can be divided into *unmatched* passengers, with a density $\Pi_0 = b_\Pi \Pi$, and *matched* passengers, with a density $\Pi_1 = (1 - b_\Pi)\Pi$. In this study, an unmatched passenger (vehicle) is waiting to be matched, whereas a matched passenger (vehicle) is waiting to be picked up (en-route to pick up the passenger). To simplify the analysis, the following assumptions are introduced (Chen et al., 2018; Zhang et al., 2019),

Assumption 1 *Vacant vehicles and waiting passengers, matched or unmatched, are all uniformly distributed with their respectively densities. In addition,*

1. *all vehicles are cruising at the same speed v, and*
2. *passengers keep waiting at the same location before pickup.*

Assumption 2 *Through its matching algorithm, the platform can achieve a stable ratio between b_Λ and b_Π, defined as $k := b_\Lambda / b_\Pi$. k is a parameter that measures the matching efficiency. A larger k indicates a higher efficiency.*

Define $\tilde{N}_v(d)$ as the counting process of the number of unmatched vehicles within a distance d from the passenger. With Assumption 1, Chen et al. (2018) prove $\tilde{N}_v(d)$ is an Inhomogeneous Poisson process with intensity function $2\pi d\Lambda_0$[4]. Due to the competition from fellow passengers,

[4]For the convenience of the reader, a proof is provided in Appendix C.

the passenger can only be matched with a fraction of unmatched vehicles. In the case of e-hail, such a fraction may be approximated by $1/\Pi_0$, which dictates unmatched vehicles are evenly distributed among unmatched passengers (Zhang et al., 2019)[5]. Accordingly, the number of *matchable vehicles* forms a subprocess $\tilde{N}_{mv}(d)$ with intensity function $2\pi d\Lambda_0/\Pi_0$. Per Assumption 2, we have

$$\frac{\Lambda_0}{\Pi_0} = \frac{b_\Lambda \Lambda}{b_\Pi \Pi} = \frac{k\Lambda}{\Pi}, \tag{1}$$

and thus the intensity function for $\tilde{N}_{mv}(d)$ can be written as $2\pi dk\Lambda/\Pi$.

Suppose an e-hail passenger is picked up by the closest matchable vehicle at distance $\tilde{D}_e$, then her wait time is given by $\tilde{w}_e = \delta \tilde{D}_e / v$. Here, δ is defined as the ratio between line distance and path distance between two points in the road network. Previous studies (e.g., Boscoe et al., 2012; Yang et al., 2018) suggest δ is primarily determined by the network topology, and its value is relatively stable (ranging between 1.1 and 1.3) . Therefore, for simplicity, we set δ as an exogenous constant in this study. The probability that at least one matchable vehicle is within d from the passenger is given by (see Zhang et al., 2019, for detail)

$$\begin{aligned} F_{\tilde{D}_e}(d) = Pr(\tilde{D}_e \le d) &= 1 - Pr\left(\tilde{N}_{mv}(d) = 0\right) \\ &= 1 - \exp\left(-\int_0^d \frac{2\pi k\Lambda}{\Pi} u \mathrm{d}u\right) = 1 - \exp\left(-\frac{k\Lambda}{\Pi}\pi d^2\right), \end{aligned} \tag{2}$$

and accordingly, the expected wait time is

$$w_e = \frac{\delta}{2v}\sqrt{\frac{\Pi}{k\Lambda}}. \tag{3}$$

3.2 Expected wait time of solo and pooling passengers

To accommodate the pooling service with the model presented above, we first add one more assumption about the distribution of passenger densities.

Assumption 3 *Passengers waiting for solo rides and pooling rides are uniformly distributed with densities Π_s and Π_p. Among those waiting for pooling rides, the unmatched passengers are uniformly distributed with a density $\Pi_0^p = b\Pi_p$.*

Similar to b_Λ and b_Π, b is related to the efficiency of pairing pooling passengers, and it is also treated as an exogenous variable in this study.

Recall the wait time of solo passengers is $\tilde{w}_s$ and that of the pooling passengers consists of two parts, i.e., $\tilde{w}_p$ and $\tilde{\Delta}$. We first discuss the detour time $\tilde{\Delta}$, which only depends on the distance between the two passengers involved.

Denote $\tilde{N}_{mp}(l)$ as the number of *matchable passengers* (i.e., *other* unmatched pooling passengers) within a distance l from a passenger. With Assumptions 1 and 3, and following exactly

[5]Although this assumption is reasonable, in reality the actual fraction may still depend on the matching strategies employed by the platform. In order to capture such a dependence, without modeling the underlying mechanism, one may adjust the fraction with another parameter (e.g. $1/\Pi_0^\epsilon$) that can be calibrated against data. We leave an in-depth analysis of such a possibility to a future study.

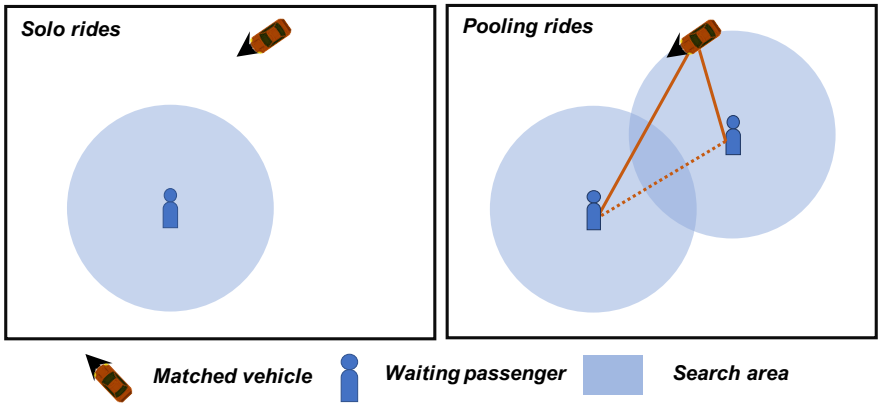


Figure 2: Access to unmatched vehicles through search area for solo passengers (Left) vs. pooling passengers (Right).

the same reasoning used in proving Proposition A2, we can show $\tilde{N}_{mp}(l)$ is an Inhomogeneous Poisson process with intensity function $2\pi l \Pi_0^p = 2\pi l b \Pi_p$. Assume the passenger be matched with the closest matchable passenger at distance $\tilde{L}$. Accordingly, we derive the expectations of $\tilde{L}$ and $\tilde{\Delta}$ as, respectively,

$$L = \frac{1}{2\sqrt{b\Pi_p}}; \quad \Delta = \frac{\delta L}{v} = \frac{\delta}{2v\sqrt{b\Pi_p}}. \tag{4}$$

We next examine the derivation of the expected solo wait time w_s and the expected pickup time for pooling w_p. In both cases, the wait time is determined by the distance to the closest matchable vehicle. For a solo passenger, the probability of finding a matchable vehicle with a distance d from her depends on the circular *search area*; see Figure 2 (left panel). A pair of pooling passengers, however, should be viewed as a team when it comes to competing for unmatched vehicles; see Figure 2 (right panel). Since either member of the team has its own search area, together they have a greater access to unmatched vehicles. As illustrated in Figure 2, the solo passenger in the left panel fails to find a match within her own search area. However, if she chooses to pool, she would succeed because the vehicle falls in her peer's search area. Therefore, pooling passengers enjoy *collective competing power* in the market, which implies $w_p \leq w_s$ in general. Nevertheless, this advantage diminishes with the distance between the pooling passengers. This is because their search areas begin to overlap when they are getting too close to each other. Evidently, if the two passengers happen to wait at the same location, their competing advantage will be wiped out as their search areas completely overlap.

The fact that pooling passengers compete as a team also means the total level of competition for unmatched vehicles will drop. In the extreme case, when everyone is pooling, the total number of vehicles required will be cut in half. Thus, compared to e-hail passengers, the solo passengers in a ride-pooling market tend to have a lower wait time, even if they choose not to pool. Taking the above observations into consideration, the following main result provides a closed-form formula for both w_s and w_p. The latter is obtained through some mild approximations, as explained in detail in Appendix D.

Proposition 1 *With Assumptions 1 and 3, the expected wait time of solo passengers and the expected pickup time of pooling passengers are respectively*

$$w_s = \frac{\delta}{2v}\sqrt{\frac{\Pi_s + \Pi_p/2}{k\Lambda}}, \tag{5}$$

$$w_p \simeq \frac{\delta}{2v}\sqrt{\frac{\Pi_s + \Pi_p/2}{k\Lambda}\frac{m + 4b\Pi_p}{2m + 4b\Pi_p}}, \tag{6}$$

where m is a parameter introduced to approximate the overlapping portions of the pooling passengers' search areas.

Proof. See Appendix D.

In this study, we assume passengers are always pooled together before their trips start. In reality, TNC platforms often allow more flexibility. For example, they may let a pooling passenger leave for her destination without a partner, and attempt to find one en-route. In what follows, we consider the case when each pooling passenger is matched to the closest matchable vehicle, which may be fully or partially vacant. If the passenger rides alone, she could be detoured to pick up another pooling passenger en-route.

Let $\tilde{D}_0$ and $\tilde{D}_h$ be the distance between the passenger and the closest matchable vehicles that are, respectively, fully and partially vacant. Then, the expected pickup time and detour to pick up a second passenger may be specified as

$$w_p = \frac{\delta}{v}\left(E[\tilde{D}_0]Pr(\tilde{D}_0 < \tilde{D}_h) + E[\tilde{D}_h]Pr(\tilde{D}_0 \geq \tilde{D}_h)\right); \tag{7}$$

$$\Delta = \frac{\delta}{v}E[\tilde{D}_h]\left(1 - Pr^{\beta}(\tilde{D}_0 < \tilde{D}_h)\right), \tag{8}$$

where β is the average number of potential pooling passengers along the route.

To interpret Eq. (7), note that the passenger may be picked up by either a fully or partially vacant vehicle, and thus her expected pickup time is the average weighted by the probability of each scenario. As for Eq. (8), the detour trip occurs only if at least one potential pooling passenger cannot find a closer and fully vacant vehicle. Since $\tilde{D}_0$, $\tilde{D}_h$ and β above depend on the density of partially and fully vacant vehicles and the density of pooling passengers, it seems rather difficult to specify the probabilities in Eqs. (7) and (8) in a way that can be empirically calibrated and validated.

Additionally, it is unclear whether explicitly modeling en-route pooling would make a meaningful difference in an idealized macroscopic market. The proposed pooling model centers on two basic trade-offs: pooling helps increase the competing power of pooling passengers, and it becomes more attractive when its market share increases (because higher demand reduces the detour distance). In en-route pooling, these basic trade-offs not only exist, but also are expected to play the same dominating role. Because the origin and destination of all trips are uniformly distributed, fully and partially vacant vehicles would also be evenly distributed relative to passengers waiting for pooling rides. Accordingly, pooling passengers still enjoy a greater access to supply because they can hail both fully and partially vacant vehicles. Also, a higher density of pooling passenger still leads to a shorter detour, even if it occurs en-route. Therefore, the total

trip time of a pooling ride (inclusive of detour and wait time) should not vary much in en-route pooling. The main difference is where detours occur, which need not concern us.

For the above reasons, en-route pooling is not explicitly modeled in this study.

3.3 Discussion

We end this section with a few remarks on the key assumptions that underlie the matching model. Assumptions 2 and 3 imply that the platform can and will dynamically adjust the matching and dispatching algorithms to achieve a desired efficiency. We make these assumptions for two reasons. First, the e-hail matching is such a complex process that itself is being actively researched. Previous studies have shown that decision variables like matching interval, matching radius and maximum allowed detour in pooling are all critical to matching performance (e.g., Yang et al., 2020a; Xu et al., 2019; Ke et al., 2020). In this study, we choose not to explicitly model these details. Instead, we use k and b to represent the overall efficiency obtained by the platform's matching policy, and calibrate them from empirical observations. This enables us to focus on the main effect of the demand-supply relationship on the matching process. Secondly, by setting k and b as exogenous, the platform's pricing strategies—the focus of this study— is isolated from its matching strategies. This simplification allows the former to drive the passenger demand and vehicle supply, while the latter's effect is incorporated through the parameters k and b.

We note that Assumptions 2 and 3 may be violated in some cases. If the platform employs a fixed matching strategy (e.g., a constant matching interval), it may not be able to maintain k and b at a stable level when the market conditions vary. While an increasing number of studies consider dynamic matching policies (e.g., Özkan and Ward, 2020; Qin et al., 2021), using a constant matching interval/radius is common in practice (e.g., Yan et al., 2019). Such a potential violation of the assumptions may introduce estimation errors in passenger wait time and market equilibrium. Assumptions 2 and 3 may also be violated when the market enters a hyper-congested state known as Wild Goose Chase (WGC) (Castillo et al., 2018), which is often accompanied by exceedingly long matching time. The reader is referred to Appendix E for additional discussions about WGC. Nevertheless, as mentioned before, WGC is unlikely to arise in our setting because the demand for e-hail services is elastic, in the sense that transit is always a feasible fallback option. Hence, passengers would begin to leave the e-hail market long before WGC is materialized.

In summary, Assumptions 2 and 3 not only simplify the matching model significantly but also separate the optimization of pricing from matching. However, as these assumptions sometimes deviate from the practice in the industry, they may also become a source of estimation errors. We leave it to a future study to relax these assumptions and to refine the matching model for equilibrium analysis.

4 Market equilibrium

The market equilibrium is dictated by the interaction between demand and supply. On the demand side, passengers choose among solo, pooling and transit based on their generalized cost. On the supply side, drivers decide whether or not to enter the market according to the average earning rate. The matching model presented in Section 3 connects the demand and

the supply, by characterizing the passenger wait time and vehicle occupancy that directly affect passengers' mode choice and drivers' entry choice.

Before we present the equilibrium model, let us first state the main assumptions as follows

Assumption 4 *Transit is a viable mode to all passengers, and is supplied at a constant generalized cost. Also, the transit operator always breaks even, i.e., the fare equals the marginal cost.*

Assumption 5 *Registered e-hail drivers enjoy flexible working schedules. Their decision to enter the service solely depends on the average earning rate offered by the platform relative to that provided by other job opportunities that they can freely pursue.*

Assumption 6 *The amount of vehicular traffic contributed by e-hail services is small, and hence the extra congestion effect is not explicitly considered in the utility of all three modes.*

4.1 Passenger demand

We characterize passengers' choice from a discrete set of modes $\mathcal{M} = \{s, p, t\}$, where s, p, and t refers to solo, pooling and transit, respectively. Let ν be the value of time, and f_i and $\tau_i, i \in \mathcal{M}$ be, respectively, the trip fare and the average duration of mode i. Typically, $\tau_p > \tau_s$ because pooling tends to prolong a trip due to the pair's different destinations . We define the generalized cost for passengers choosing one of the three modes as follows.

$$\text{Solo ride: } u_s = f_s + \nu(w_s + \tau_s), \tag{9a}$$

$$\text{Pooling ride: } u_p = f_p + \nu(w_p + \Delta + \tau_p), \tag{9b}$$

$$\text{Transit: } u_t = f_t + (\nu + \zeta)\tau_t. \tag{9c}$$

According to Assumptions 4 and 6, f_t, τ_s and τ_t are all treated as constants. In theory, τ_p should be endogenous, as a decreasing function of the pooling passenger density. Two recent studies investigate this issue for pooling rides shared by two passengers. By simulating the matching process of on-demand pooling using empirical demand data collected in three large cities, Ke et al. (2021) found that the ratio between the detour distance and the average trip distance is inversely proportional to a function of the batch demand (i.e., the number of requests accumulated in a matching interval). Lobel and Martin (2020) analyze the detour and the value associated with pooling. The latter is measured by the reduced total travel distance. The authors define the detour (value) ratio as the total detour (value) normalized by the total travel distance when both passengers take solo rides. They show that the sum of the two ratios is bound by 0.5, which also implies the detour ratio cannot exceed 0.5. Although these results bring valuable insights, implementing them in our framework is impractical. On the one hand, the model proposed in Ke et al. (2021) cannot be properly calibrated with the data we have. On the other hand, Lobel and Martin (2020) only set an upper bound on the detour ratio. Hence, in this study, we simply assume τ_p to be a constant, which is calibrated from the empirical data (see Appendix B). In Section 6.3, we test our model's sensitivity to τ_p. For the reader who wonders how much a difference an endogenized detour time could make, in Appendix J, we implement a version of our equilibrium model using the detour model of Ke et al. (2021), and perform a sensitivity analysis against key coefficients.

Note that ζ in Eq. (9c) represents additional disutility of transit relative to ride-hail services (associated with privacy, comfort, crowdedness, etc.). Following Schwieterman (2019), we set $\zeta = 0.25\nu$ in this study.

Suppose the total demand is D_0 and the share of each mode is a continuous and differentiable function $q : \mathbb{R}^3_+ \to (0, 1)$ of the general costs of all three modes, i.e.,

$$Q_i = D_0 q(u_i, \mathbf{u}_{-i}),\ i \in \mathcal{M}, -i := \mathcal{M} \setminus \{i\}, \tag{10}$$

where $\mathbf{u}_{-i}$ refers to modes except for i. Without loss of generality, we assume $\partial Q_i / \partial u_i < 0$ and $\partial Q_i / \partial \mathbf{u}_{-i} > 0$. Accordingly, the pooling ratio is given as $r_Q = Q_p / (Q_s + Q_p)$.

4.2 Vehicle supply

Let S_0 be the potential supply, which may be viewed as the total number of drivers registered on the platform. As per Assumption 5, we define $\tilde{e}_0$ as the earning rate of the alternative employment opportunity available to the drivers. Often known as the *reservation rate*, $\tilde{e}_0$ is modeled as a random variable with a cumulative distribution function (CDF) $g(\cdot)$. Drivers enter the e-hail market only if doing so yields an earning rate $e \geq \tilde{e}_0$. This assumption aligns with several studies that empirically observe a positive wage elasticity of supply among ride-hail drivers (e.g., Angrist et al., 2017; Chen and Sheldon, 2017; Sun et al., 2019) Accordingly, the fleet size is derived as

$$N = S_0 Pr(\tilde{e}_0 \leq e) = S_0 g(e). \tag{11}$$

The earning of an e-hail driver is determined by the compensation rate per unit occupied time, denoted as η. In addition, a driver serving pooling rides may also be paid a fixed fee c_p for each additional pickup. Thus, the average earning rate of a driver is computed as

$$e = \frac{1}{N}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_p Q_p\right], \tag{12}$$

where $Q_s\tau_s + \frac{1}{2}Q_p\tau_p$ is the total occupied time and $\frac{1}{2}Q_p$ denotes the number of additional pickups.

4.3 Equilibrium

With the demand and supply specified above and the wait and detour times derived in Section 3, the aggregate equilibrium in a unit time period is characterized by the following system of

equations:

$$\text{mode choice:} \quad Q_i = D_0 q(u_i, \mathbf{u}_{-i}),\ i \in \{s, p\}, \tag{13a}$$

$$\text{Fleet size:} \quad N = S_0 g\left(\frac{1}{N}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_p Q_p\right]\right), \tag{13b}$$

$$\text{Flow conservation:} \quad N = V + Q_s\tau_s + \frac{1}{2}Q_p\tau_p, \tag{13c}$$

$$\text{Solo wait time:} \quad w_s = \frac{\delta}{2v}\sqrt{\frac{Q_s w_s + Q_p(w_p + \Delta)/2}{kV}}, \tag{13d}$$

$$\text{Pooling wait time:} \quad w_p = w_s\sqrt{\frac{m + 4bQ_p(w_p + \Delta)}{2m + 4bQ_p(w_p + \Delta)}}, \tag{13e}$$

$$\Delta = \frac{\delta}{2v}\frac{1}{\sqrt{bQ_p(w_p + \Delta)}}. \tag{13f}$$

Eq. (13c) states that the total vehicle operation time (represented by the fleet size N) consists of three parts: (i) the vacant vehicle time (V) that includes both idle and pickup time; (ii) the time spent on delivering solo passengers; and (iii) the time spent on delivering pooling passengers. Since the market equilibrium is established over unit time period, V substitutes Λ in Eqs. (5) and (6). Eqs. (13d)–(13f) are obtained from substituting Π_s and Π_p in Eq. (4)–(6) by $\Pi_s = Q_s w_s$ and $\Pi_p = Q_p(w_p + \Delta)$ as per Little's formula (Little, 1961).

Define $\mathbf{x} = (w_s, w_p, \Delta)$. Then, Q_s, Q_p, N and V can be viewed as functions of $\mathbf{x}$ according to Eqs. (13a)–(13c). Plugging them into Eqs. (13d)–(13f) thus reduces the equilibrium to a fixed-point system $\mathbf{x} = F(\mathbf{x})$. With mild assumptions, we prove the solution existence of such a fixed-point system by invoking Brouwer's theorem (Brouwer, 1911), as summarized in the following proposition.

Proposition 2 *Suppose* $\mathbf{x}$ *is bounded from above by* $\bar{\mathbf{x}} = (\overline{w}_s, \overline{w}_p, \overline{\Delta})^T$. *Then, there exists an* $\mathbf{x}^* = (w_s^*, w_p^*, \Delta^*)^T$ *that satisfies Eq.* (13).

Proof. see Appendix F □

The assumption made in Proposition 2 effectively sets upper bounds on passenger wait times for both solo and pooling rides. At first glance, this seems at odds with Eqs. (13d)–(13f), which allow these wait times to grow infinitely. Nevertheless, if passengers have to wait exceedingly long, the demand would be suppressed below a level of practical interest. In other words, Proposition 2 ensures a solution always exists provided that the demand for solo and pooling rides has a lower bound that can, in theory, be arbitrarily close to zero.

An equilibrium solution to Eq. (13) can be obtained through an iterative fixed-point algorithm. While implementing such an algorithm is straightforward, we note that, as a highly nonlinear system, Eq. (13) may not have a unique equilibrium. In addition, an equilibrium solution may or may not be stable. Appendix H summarizes the efforts of detecting the occurrence of multiple equiliria and unstable solutions in numerical experiments. In a nutshell, we found that both events are too rare to be of a practical concern, at least for the range of parameters tested in our experiments.

5 Optimal pricing

5.1 Profit-maximization pricing without regulation

By choosing a combination of f_s, f_p and η, the platform could control the demand split between solo and pooling rides, thereby the pooling ratio, to maximize its profit. If pooling is not profitable, the platform can simply set its price equal to or higher than solo rides (i.e., $f_s \leq f_p$) to eliminate pooling. On the other hand, the platform may increase the gap between f_s and f_p to encourage pooling during a demand peak. In addition to raising the service capacity, pooling rides could also help boost profits. Drivers are largely paid based on the occupied duration and distance, regardless of how many riders may share the vehicle at any given time. Hence, the platform is poised to reap most extra revenues contributed by pooling rides.

Without loss of generality, we assume the transit fare f_t is fixed, and the platform aims to maximize its gross profit by choosing a price vector $\mathbf{y} = (f_s, f_p, \eta)^T$. The pricing problem is then formulated as

$$\text{(P1)} \quad \max_{\mathbf{y}} \; R = f_s Q_s + f_p Q_p - \eta \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right) - \frac{1}{2} c_p Q_p. \tag{14}$$

Here, the platform's gross profit R equals the total revenue less the expense directly related to the production of trips. Specifically, the first two terms (i.e., $f_s Q_s$ and $f_p Q_p$) are revenues generated from solo and pooling rides, respectively, while $\eta \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right)$ is the payment to drivers based on their occupied time and $\frac{1}{2} c_p Q_p$ is the payment to drivers due to additional pickups in pooling.

To simplify the notation, let $Q_s^{(1)}$, $Q_s^{(2)}$ and $Q_s^{(3)}$ denote the partial derivative of Q_s with respect to f_s, f_p, η, respectively. $Q_p^{(1)}$, $Q_p^{(2)}$ and $Q_p^{(3)}$ are defined similarly. Accordingly, the first-order conditions of P1 are reduced to

$$f_s = \eta \tau_s - \frac{Q_s Q_p^{(2)} - Q_p Q_p^{(1)}}{Q_s^{(1)} Q_p^{(2)} - Q_s^{(2)} Q_p^{(1)}}; \tag{15a}$$

$$f_p = \left(\frac{1}{2} \eta \tau_p + \frac{1}{2} c_p \right) - \frac{Q_p Q_s^{(1)} - Q_s Q_s^{(2)}}{Q_s^{(1)} Q_p^{(2)} - Q_s^{(2)} Q_p^{(1)}}; \tag{15b}$$

$$Q_s \left[1 + \left(\tau_s + \frac{1}{2} \frac{Q_p}{Q_s} \tau_p \right) \frac{Q_s^{(1)}}{Q_s^{(3)}} \right] = 0; \tag{15c}$$

$$Q_p \left[1 + \left(\frac{Q_s}{Q_p} \tau_s + \frac{1}{2} \tau_p \right) \frac{Q_p^{(2)}}{Q_p^{(3)}} \right] = 0. \tag{15d}$$

Eqs. (15a) and (15b) bear similarity with the Lerner formula (Lerner, 1934), where the first term represents the marginal cost of each solo (pooling) ride (i.e., compensation paid to the driver) and the second term is a mark-up that measures the market power of the platform. It also aligns with the optimal trip fare derived in Zha et al. (2016) and Ke et al. (2020). Eqs. (15c) and (15d) imply that, at the optimal solution, either solo (pooling) demand equals 0 or $Q_s^{(1)}/Q_s^{(3)}$ ($Q_p^{(2)}/Q_p^{(3)}$) is dictated by the market shares and average journey times of the two modes.

For comparison, we derive the first-order conditions under single-mode operation, i.e., when only solo or pooling rides are served. In the case of pure solo rides, the single-mode operation

yields

$$f_s = \eta\tau_s - \frac{Q_s}{Q_s^{(1)}}; \tag{16a}$$

$$Q_s\left(1 + \tau_s\frac{Q_s^{(1)}}{Q_s^{(3)}}\right) = 0, \tag{16b}$$

and for pure pooling rides,

$$f_p = \frac{1}{2}\eta\tau_p + \frac{1}{2}c_p - \frac{Q_p}{Q_p^{(2)}}; \tag{17a}$$

$$Q_p\left(1 + \frac{1}{2}\tau_p\frac{Q_p^{(2)}}{Q_p^{(3)}}\right) = 0. \tag{17b}$$

Eqs. (16) and (17) share the same structures as Eq. (15)—the optimal trip fare is the marginal cost plus the platform's mark-up, and the marginal changes of demand with respect to trip fare and compensation rate should be a constant when demand is positive.

Eqs. (16b) and (17b) further imply that, at the optimal solution with positive demand, the marginal change of demand due to an increase in trip fare should be proportional to that due to an increase in compensation rate. Specifically, the rate is $-1/\tau_s$ for solo and $-2/\tau_p$ for pool. In other words, in response to an increased trip fare, the platform must raise the compensation rate to improve LOS so that the service remains attractive to passengers. In the mix-mode scenario, however, the pressure to raise LOS is relieved, i.e., the absolute value of the rate is smaller in Eqs. (15c) and (15d). Because increasing the trip fare of one mode would make the other more attractive, the platform need not increase the compensation rate as much as in the single mode operation in order to hold on to the market share.

We proceed to compare the platform's market power under single and mixed operation modes. Dividing both the numerator and the denominator of the second term in Eq. (15a) by $Q_p^{(2)}$ yields $\frac{Q_s - Q_p Q_p^{(1)}/Q_p^{(2)}}{Q_s^{(1)} - Q_s^{(2)} Q_p^{(1)}/Q_p^{(2)}}$. Here, $Q_p^{(1)}/Q_p^{(2)} < 0$ as f_s and f_p have opposite influences on Q_p, and $Q_s^{(2)} > 0$ because increasing f_p makes solo rides more appealing. Also, $|Q_s^{(1)}| > |Q_s^{(2)}|$ and $|Q_p^{(2)}| > |Q_p^{(1)}|$ because a change in the trip fare of one mode has larger impact on that mode than the other mode. These observations lead to

$$\left|\frac{Q_s - Q_p Q_p^{(1)}/Q_p^{(2)}}{Q_s^{(1)} - Q_s^{(2)} Q_p^{(1)}/Q_p^{(2)}}\right| > \left|\frac{Q_s}{Q_s^{(1)}}\right|,$$

indicating that serving both pooling and solo rides has the potential to boost the platform's market power (hence profit) compared to the case when only solo rides are served. However, the change of total trips served by the platform is unclear, because the impact on pooling rides is unclear.

Since P1 is a non-convex program, we implement a gradient ascent algorithm to search local optimal solutions. In each iteration, given the current solution (corresponding to a market equilibrium), we compute the gradient and update the solution with the following iteration rule:

$$\mathbf{y}^{(i+1)} = \mathbf{y}^{(i)} + \alpha\nabla R(\mathbf{y}^{(i)}), \tag{18}$$

where ∇R is the gradient of the revenue function with respect to $\mathbf{y}$ and α is a predefined learning rate (or step size).

∇R is not easy to evaluate because it involves the differentiation of market equilibrium. Appendix I explains how to obtain the gradient in each iteration. In brief, the method bears some similarities with the sensitivity-analysis-based algorithm for network design problems (e.g. Tobin and Friesz, 1988; Yang, 1995; Patriksson, 2004). That is, given the current solution, it guides the movement towards the next solution according to the sensitivity of the equilibrium solution to the decision variables.

Finally, as per Eqs. (15c) and (15d), P1 could have multiple local optima. Hence, in numerical experiments, we solve P1 with multiple initial solutions and select the one with the highest objective value. The same strategy is applied to other two problems defined in the rest of this section.

5.2 Profit-maximization pricing under regulations

Two regulatory policies have received much attention lately: (i) minimum wage, which requires the platform to ensure an average pay rate[6] no less than an minimum value; and (ii) fleet cap, which caps the number of on-line drivers to serve the platform. Both regulations have been implemented in New York City (Wodinsky, 2019; Hawkins, 2019). These two policies, however, display rather similar effects in our framework, because the fleet size N and the average earning rate e are linked to each other through a one-to-one mapping; see Eq. (11). The difference is, while a fleet cap bounds the fleet size from above, a minimum wage sets a bound from below. Hence, which policy leads to a greater social welfare depends on the difference in the fleet sizes achieved by maximizing profit and maximizing social welfare. In general, profit maximization results in a lower-than-socially-optimum production level (Douglas, 1972). Thus, by pushing up the fleet size, the minimum wage policy seems to have a greater potential for social good. With the above discussions in mind, we focus on the minimum wage policy hereafter.

Mathematically, the effect of regulation can be captured by adding an inequality constraint to the original problem, leading to

$$\text{(P2)} \quad \max_{\mathbf{y}} \; R = f_s Q_s + f_p Q_p - \eta \left(Q_s \tau_s + \frac{1}{2} Q_p \tau_p \right) - \frac{1}{2} c_p Q_p, \tag{19}$$
$$s.t. \quad h(\mathbf{y}) \leq 0, .$$

To solve P2, we introduce a Lagrangian multiplier λ and write the dual problem as

$$\text{(P2')} \quad \min_{\lambda} \max_{\mathbf{y}} \; \mathcal{L}(\lambda, \mathbf{y}) = R(\mathbf{y}) - \lambda h(\mathbf{y}) \tag{20}$$
$$s.t. \; \lambda \geq 0.$$

P2' can thus be solved by a dual gradient ascent algorithm as follows:

$$\mathbf{y}^{(j+1)} = \arg\max \; \mathcal{L}(\lambda^j, \mathbf{y}), \tag{21a}$$
$$\lambda^{(j+1)} = \max \left(0, \; \lambda^{(j)} + \rho h(\mathbf{y}^{(j+1)}) \right), \tag{21b}$$

[6]Note that this is different from the compensation rate η, but rather corresponds to the earning rate e.

where ρ is a constant penalty parameter. In each iteration, the maximization problem (21a) is first solved using the same gradient ascent method used for solving P1, with the current estimate of the multiplier λ. Then, Eq. (21b) is invoked to update λ.

For the minimum wage policy (with a wage floor $\underline{e}$), the constraint $h(\mathbf{y})$ is given by

$$\underline{e} - \frac{1}{N}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_pQ_p\right] \leq 0\,. \tag{22}$$

Accordingly, the Lagrangian of P2 is equivalent to

$$\mathcal{L} = f_sQ_s + f_pQ_p - \left(1 - \frac{\lambda}{N}\right)\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_pQ_p\right], \tag{23}$$

and the first-order conditions are reduced to

$$\begin{aligned} f_s = {} & \left(1 - \frac{\lambda}{N}\right)\eta\tau_s - \frac{Q_sQ_p^{(2)} - Q_pQ_p^{(1)}}{Q_s^{(1)}Q_p^{(2)} - Q_s^{(2)}Q_p^{(1)}} \\ & + \frac{\lambda}{N^2}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_pQ_p\right]\frac{N^{(1)}Q_p^{(2)} - N^{(2)}Q_p^{(1)}}{Q_s^{(1)}Q_p^{(2)} - Q_s^{(2)}Q_p^{(1)}}; \end{aligned} \tag{24a}$$

$$\begin{aligned} f_p = {} & \left(1 - \frac{\lambda}{N}\right)\left(\frac{1}{2}\eta\tau_p + \frac{1}{2}c_p\right) - \frac{Q_pQ_s^{(1)} - Q_sQ_s^{(2)}}{Q_s^{(1)}Q_p^{(2)} - Q_s^{(2)}Q_p^{(1)}} \\ & + \frac{\lambda}{N^2}\left[\eta\left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_pQ_p\right]\frac{N^{(2)}Q_s^{(1)} - N^{(1)}Q_s^{(2)}}{Q_s^{(1)}Q_p^{(2)} - Q_s^{(2)}Q_p^{(1)}}; \end{aligned} \tag{24b}$$

$$Q_s\left[1 + \left(\tau_s + \frac{1}{2}\frac{Q_p}{Q_s}\tau_p\right)\frac{Q_s^{(1)}}{Q_s^{(3)}}\right] = 0; \tag{24c}$$

$$Q_p\left[1 + \left(\frac{Q_s}{Q_p}\tau_s + \frac{1}{2}\tau_p\right)\frac{Q_p^{(2)}}{Q_p^{(3)}}\right] = 0, \tag{24d}$$

where $N^{(1)}$ and $N^{(2)}$ denote the partial derivatives of the total vehicle supply to f_s and f_p, respectively.

On the one hand, Eqs. (24c) and (24d) suggest that the regulation does not affect the optimal conditions regarding the compensation rate. On the other hand, comparing Eq. (24a) with Eq. (15a) reveals how minimum wage affects the optimal ride price: the first term is discounted by a factor $1 - \lambda/N$, and the newly added third term tends to reduce the platform's market power, as $N^{(1)}$ and $N^{(2)}$ are positive in general. Hence, the platform's profit is likely to suffer under the minimum wage policy. However, the actual change in the ride price is not clear, because it also depends on η.

Recently, the City of Chicago proposed to charge a congestion tax on TNC trips starting and/or ending in designated areas during peak times[7]. In order to encourage ride-sharing, the proposed tax is lower for a pooling ride than a solo one. Although such a tax is likely to be fully passed on to passengers, the platform is expected to adjust its pricing strategy and fleet size in order to accommodate potential mode shift. To analyze the impact of such a policy in our

[7] https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html

framework, we assume a constant *extra* tax c_s is charged on each solo ride while pooling rides are not subject to such a tax. Then, we only need to incorporate c_s as a fixed cost into the utility of solo passengers. Accordingly, Eq. (9a) becomes

$$u_s = f_s + \nu(w_s + \tau_s) + c_s. \tag{25}$$

Since the congestion tax does not introduce a new constraint, it should have less impact on the platform's pricing strategies than minimum wage. However, a larger shift toward pooling is expected, because the congestion tax is precisely levied against solo rides.

5.3 Social optimal pricing

We finally consider a second-best social optimal pricing problem that seeks to maximize social welfare. The policy is "second-best" because, while the goal is to maximize social welfare, the platform is not allowed to run a deficit. The corresponding optimization problem is formulated as follows:

$$\begin{aligned} \text{(P3)} \quad \max_{\mathbf{y}} \; & W = D_0 \Delta u + f_s Q_s + f_p Q_p \\ & \qquad - \int_0^N g^{-1}(n/S_0)\mathrm{d}n - c_0 N + c_s Q_s \\ & s.t.\ R \geq 0. \end{aligned} \tag{26}$$

The social welfare W consists of five parts: (i) the surplus of passengers, measured by the total expected general cost saving because of switching from transit to ride-hail services, where Δu denotes the saving of each passenger and will be specified in Section 6.1; (ii) the total platform revenue $f_s Q_s + f_p Q_p$; (iii) the opportunity cost of drivers $\int_0^N g^{-1}(n/S_0)\mathrm{d}n$, where $g^{-1}(\cdot)$ is the inverse function of the CDF of the reservation rate; (iv) the approximate congestion cost caused by the ride-hail fleet, where c_0 is a constant cost caused by the entry of each driver (see Appendix B for the estimation of c_0); and (v) the tax revenue due to the congestion tax $c_s Q_s$, if implemented.

It is worthwhile to clarify several issues before we continue the analysis:

1. Ideally, capturing congestion externality of ride-hail requires an explicit traffic flow model. However, since the number of ride-hail trips is expected to be small relative to other trips (see Assumption 6)[8], it is reasonable to assume each additional vehicle's contribution to congestion (i.e., the marginal cost) would be roughly the same. Moreover, as our modeling framework is highly aggregate, the complexity of introducing a more complicated traffic flow model may not be justified.

2. The congestion tax is a transfer payment within the system and hence should neither increase nor decrease the social welfare. Accordingly, the tax revenue (i.e., item (v)) is included in the social welfare to offset the reduction in the passenger surplus of solo rides.

3. According to Assumption 4, the decisions related to ride-hail (pricing and regulations) do not affect transit operations and the transit system always breaks even. Therefore, neither the supply cost nor the revenue of the transit service is included in the social welfare.

[8]For example, Nie (2017) estimates that the extra traffic brought by the introduction of e-hail in Shenzhen, China is no more than 1% of the traffic already on road.

We close this section by noting that, like P2, P3 can also be solved using the dual gradient ascent algorithm.

6 Numerical experiments

In this section, numerical experiments are conducted to (i) examine passenger wait time of different modes (Section 6.2); (ii) analyze mode choice at the market equilibrium with fixed pricing (Section 6.3) under various market conditions; (iii) assess the performance and welfare implications of profit-maximization pricing (Section 6.4), and (iv) evaluate the effects of different regulations on pooling and social welfare (Section 6.5). In Section 6.1, we first specify the demand and supply models, as well as discuss default parameters used in the experiments.

Table 1: Default values and ranges of the main parameters used in numerical experiments.

Parameter		**Unit**	**Default value**	**Variation**
Detour ratio of road network	δ		1.3	
Cruising speed	v	mph	13.6	
Matching efficiency	k	/mi^2	0.16	
Pooling efficiency	b		0.05	
Approximation parameter	m		4	
Average solo trip duration	τ_s	hr	0.28	
Average pooling trip duration	τ_p	hr	0.40	
Average transit trip duration	τ_t	hr	0.53	
Passengers' value of time	ν	\$/hr	27.69	
Relative disutility rate of transit	ζ	\$/hr	6.92	
Mode choice uncertainty	θ		1	
Average reservation rate	e_0	\$/hr	19.84	
Vacant vehicle density	Λ	/mi^2	70	40–100
Waiting passenger density	Π	/mi^2	24	8–40
Fraction of waiting passenger for pooling	r_Π		0.4	0.2–0.6
Total demand	D_0	/mi^2/hr	1200	500–2000
Potential supply	S_0	/mi^2/hr	550	200–800
Solo trip fare	f_s	\$/ride	14	
Pooling trip fare	f_p	\$/ride	10	
Transit trip fare	f_t	\$/ride	2.69	
Compensation rate	η	\$/hr	20	
Additional pickup fare	c_p	\$/ride	0	0–2
Congestion cost per vehicle	c_0	\$/hr	2.9	

6.1 Experiment setting

We characterize passengers' choice based on the random utility theory (Ben-Akiva and Lerman, 1985) and adopt the Multinomial Logit (MNL) model. Therefore, the share of each mode $q(\cdot)$ is

estimated as

$$Q_i = D_0 \frac{\exp(-\theta u_i)}{\sum_{j\in\mathcal{M}} \exp(-\theta u_j)},\ i \in \mathcal{M}, \tag{27}$$

where θ is a non-negative parameter that reflects the degree of uncertainty in mode choice.

Accordingly, the upper bounds in Proposition 2 can be set based on the minimum demand level considered as meaningful by the modeler. More details are provided in Appendix G.

It is well known that in the MNL model the logsum term measures the expected utility of all alternatives (Small and Rosen, 1981; De Jong et al., 2007). Since we treat transit as a benchmark mode, the expected general cost saving of each passenger is given as

$$\Delta u = \frac{1}{\theta} \log \sum_{i\in\mathcal{M}} \exp[\theta(u_t - u_i)],\ i \in \mathcal{M}. \tag{28}$$

For simplicity, we assume the drivers' reservation rate $\tilde{e}_0$ follows a uniform distribution on $[0, 2e_0]$, where e_0 is the average reservation rate. Therefore, Eq. (11) is simplified as

$$N = S_0 \frac{e}{2e_0} = \frac{S_0}{2e_0 N} \left[\eta \left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_p Q_p\right], \tag{29}$$

which yields

$$N = \sqrt{\frac{S_0}{2e_0} \left[\eta \left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_p Q_p\right]}. \tag{30}$$

Accordingly, the drivers' opportunity cost defined in Eq. (26) is reduced to

$$\int_0^N g^{-1}(n/S_0)\mathrm{d}n = \frac{1}{2}\left[\eta \left(Q_s\tau_s + \frac{1}{2}Q_p\tau_p\right) + \frac{1}{2}c_p Q_p\right]. \tag{31}$$

We set up the experiments based on empirical TNC data collected in the City of Chicago; see Appendix B for details. Table 1 reports the default values of key model parameters and the range of parameter values tested in sensitivity analyses. Note that the default prices, i.e., f_s, f_p and η, are only used in the sensitivity analysis presented in Section 6.3.

6.2 Sensitivity analysis on passenger wait times

Eqs. (3) and (5) suggest that offering pooling always lowers the wait time of solo passengers. However, the benefit is less clear for pooling passengers. Therefore, we first conduct a sensitivity analysis on the wait time of e-hail, solo and pooling rides.

Figure 3 examines how waiting passenger density Π, vacant vehicle density Λ and fraction of waiting passenger for pooling $r_\Pi = \Pi_p/\Pi$[9] affect the total passenger wait time and the fraction of pickup detour. Note that the total wait time for both passengers in a pooling ride is $w_p + \Delta$, because the first passenger to be picked up also endures the pickup detour on the way to pick up the other passenger.

[9]Note that r_Π is close to, yet not exactly equal to r_Q because it involves the wait time

As shown in Figure 3(a), while the wait times of solo and e-hail passengers *increase* with the waiting passenger density, that of pooling passengers *decreases*, thanks to shorter pickup detours. In other words, pooling becomes more desirable when the demand level is high. Although a larger waiting passenger density intensifies the competition and increases the total wait time, its impact on pooling passengers is better mitigated by the collective competing power and the reduction of the pickup detours. Nevertheless, when the system becomes overly congested (larger than 30 passengers per square mile in this example), the benefit of pooling diminishes, slightly pushing up wait time. On the other hand, when the vacant vehicle density increases, all passenger wait times drop, as shown in Figure 3(b). Yet, pooling benefits less than the other two due to the existence of pickup detour. According to Figure 3(c), while both solo and pooling passengers benefit substantially from the increase in the fraction of waiting passenger for pooling, pooling passengers' gain is greater thanks to the collective competing power of the pair. Figure 3 also reveals that the fraction of detour time in the pooling passengers' wait time drops quickly as the waiting passenger density and fraction of pooling passenger among them increase.

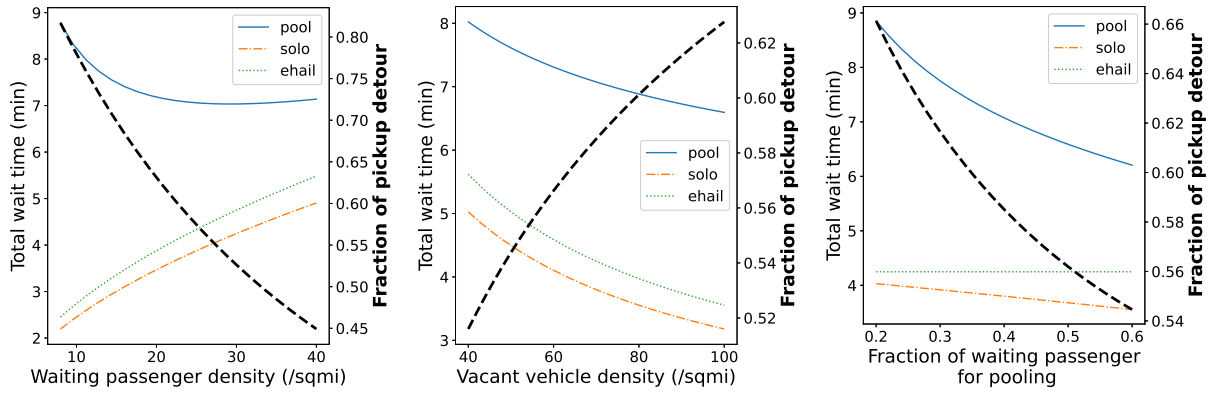


Figure 3: Sensitivity of total wait time and fraction of pickup detour to (a) waiting passenger density, (b) vacant vehicle density and (c) fraction of waiting passenger for pooling.

To summarize, pooling becomes more appealing as the demand level grows, and the rise of fraction of waiting passenger for pooling further reduces the wait time and thereby attracts more demand. However, this seemingly positive feedback leaves out an important caveat—pooling tends to prolong a trip due to the pair's different destinations, on top of the detour time incurred in the pickup phase. This *en-route detour time* also squeezes the supply because a vehicle serving pooling rides will be occupied longer on average. We will show how this effect is captured in an equilibrium model in the next section.

6.3 Mode choice with fixed trip fare

In this section, we examine how mode shares at the market equilibrium vary with the total demand D_0, potential supply S_0, additional pickup fee c_p and en-route detour $\tau_p - \tau_s$. Other parameters, including the platform's pricing strategy, will be fixed as reported in Table 1.

As shown in Figure 4(a), the share of pooling rides first increases and then decreases with the

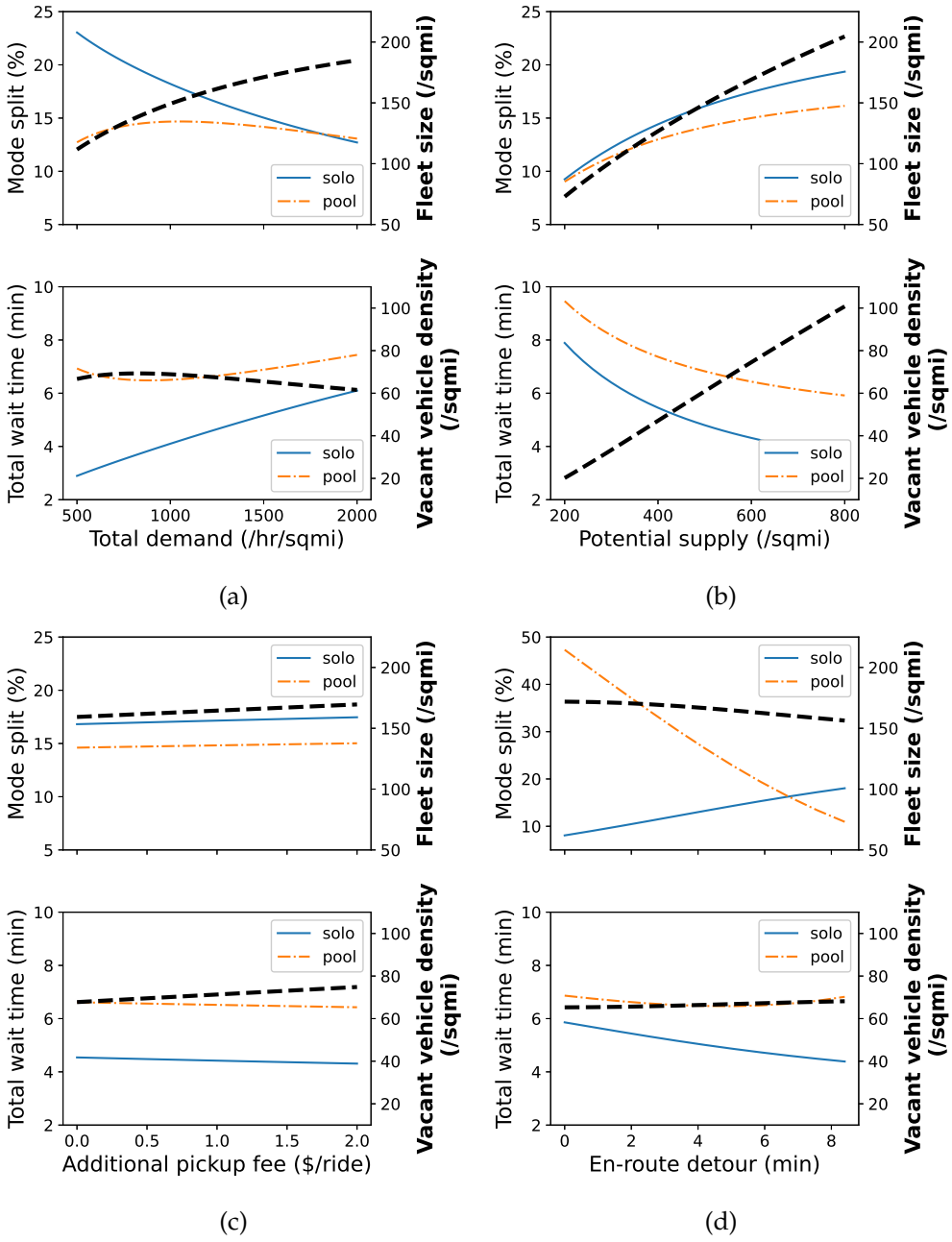


Figure 4: Sensitivity of mode share, vehicle supply and passenger wait time to (a) total demand D_0, (b) potential supply S_0, (c) additional pickup fee c_p and (d) en-route detour $\tau_p - \tau_s$.

total demand. The share of solo rides drops faster yet remains higher than pooling for most parts. This shift is related to the different impacts of the rising demand on the wait time for the two modes. The rising demand intensifies the competition among waiting passengers, thus steadily increasing the solo wait time. On the other hand, higher demand helps reduce the pickup detour for pooling rides, leading to its initial growth of market share. On the supply side, the growth of total demand induces more drivers to enter the market (top plot). However, it only leads to a mild increase of vacant vehicle density at the very beginning (bottom plot). As the total demand further increases, the level of service for both solo and pooling modes deteriorates.

The growth of S_0 reveals a different pattern; see Figure 4(b). The total share of ride-hail increases with the potential supply, while solo rides gain more popularity (top plot). Although the vacant vehicle density increases linearly with the potential supply, passengers enjoy a milder (sub-linear) improvement, owing to competitions on the demand side.

Figure 4(c) reveals that the effect of the additional pickup fee for pooling rides (c_p) is almost negligible. When c_p increases from \$0 to \$2, the share of both pooling and solo modes rises less than 0.5% (top plot). Thus, it cannot serve as an effective incentive to encourage drivers to take more pooling rides. Nor could it bring a meaningful improvement to the level of service (the wait time barely changes, see the bottom plot). Given these observations, the additional pickup fee will not be discussed in following experiments, and will be simply set to zero hereafter.

Figure 4(d) highlights the importance of the en-route detour time. Specifically, the share of pooling drops sharply as the detour rises from 0. As per Eq. (4), the loss of pooling demand results in a longer pickup detour (hence a longer wait time) for pooling passengers, which further discourages pooling. On the other hand, the loss of pooling demand reduces the overall demand level, and as a result, the solo wait time drops despite a shrinking fleet size (top plot). Overall, these findings suggest that *promoting pooling may not be a good strategy for the platform if trip destinations are too scattered to keep the average en-route detour time under control*.

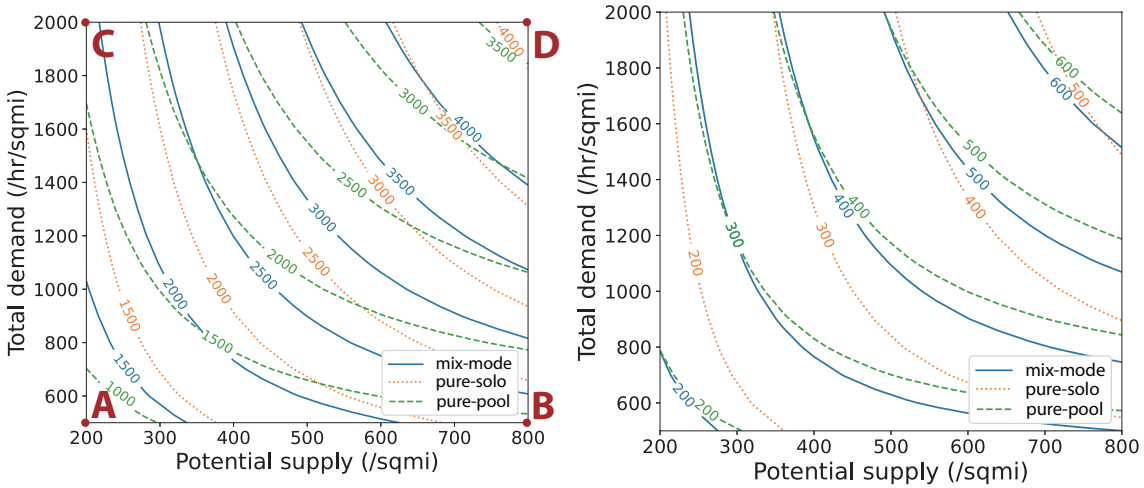
It is worth noting that results in Figure 4 may overestimate the market share of pooling in real practice. In Chicago, from which many of our model's inputs are drawn, the average pooling ratio is less than 15%; see Figure A2 in Appendix B. A few factors might contribute to this discrepancy. First, our model assumes the trip origins and destinations are uniformly distributed in an aggregate market. The heterogeneous distribution of the real demand is likely to produce strong spatiotemporal imbalance that could undermine the matching efficiency for pooling. The high en-route detour time in Chicago (around 7 min, close to the upper bound in Figure 4(d)) may reflect such inefficiency. Second, our model excludes some negative features of pooling (e.g., the loss of privacy and comfort) in favor of simplicity and tractability. Ignoring these factors might underestimate the general cost of pooling.

6.4 Performance of optimal pricing

In this section, we examine the performance of the profit-maximization pricing problem P1 under various operational strategies and market conditions (i.e., combinations of total demand D_0 and potential supply S_0). The platform may use one of the following three operational strategies: (i) offering both solo and pooling rides, (ii) only offering solo rides, and (iii) only offering pooling rides. These strategies are referred to as "mix-mode", "pure-solo" and "pure-pool", respectively.

Figures 5 presents contours of the platform's profit and total trip output (i.e., $Q_s + Q_p$) gained

by P1 under the three operational strategies. It can be read from Figures 5(a) that mix-mode always achieves the highest profit, followed by pure-solo strategy. This finding is predicted by the analytical solutions, as a platform serving both solo and pooling rides enjoys a higher market power (see Eqs. (15a) and (15b)). As expected, pure-pool is the least profitable because the platform has to keep the price sufficiently low to sustain a reasonable level of service for pooling (otherwise it would lose much of the market share to transit). On the other hand, mix-mode and pure-pool produce much more trips than pure-solo, and the difference increases as the market expands; see Figures 5(b). When the potential supply becomes scarce, pure-pool gradually achieves a leading position in terms of trip production. Overall, mix-mode achieves a more favorable balance between trip production and profitability than the other two strategies.



(a) Platform's profit

(b) Trip production

Figure 5: System performance under different operational strategies and market conditions.

We further pick four corner cases (labeled in Figure 5(a)), namely, low-demand-and-low-supply (A: "low-low"), low-demand-and-high-supply, (B: "low-high"), high-demand-and-low-supply (C: "high-low"), and high-demand-and-high-supply (D: "high-high"), and examine more details of the system performance. Besides the three operational strategies, we also solve the second-best pricing problem (P3) under mix-mode, denoted as "SO", for comparison.

Figure 6(a) compares the social welfare obtained by the three operational strategies with the social optimal (SO) state. The different patterns indicate the passenger surplus (logsum term in Eq. (26)), the platform profit and the driver surplus (income less opportunity and congestion cost). We plot the ratio of these three components in each market scenario with respect to the total welfare at SO (normalized as one). We first note that mix-mode consistently yields the highest social welfare, around 80% of the SO level. Pure-pool is the worst in three out of the four corner cases. It slightly outperforms pure-solo only when demand is high but supply is low (Case C: high-low), which is expected because pooling is most helpful in such a case. Combined with the results from Figure 5, we conclude that *serving a mixture of solo and pooling rides benefit both the platform and the society.* On the other hand, only serving pooling rides does not necessarily yield a higher social welfare than the regular ride-hail service, because the passengers suffer from a degraded level of service and the drivers earn a lower wage.

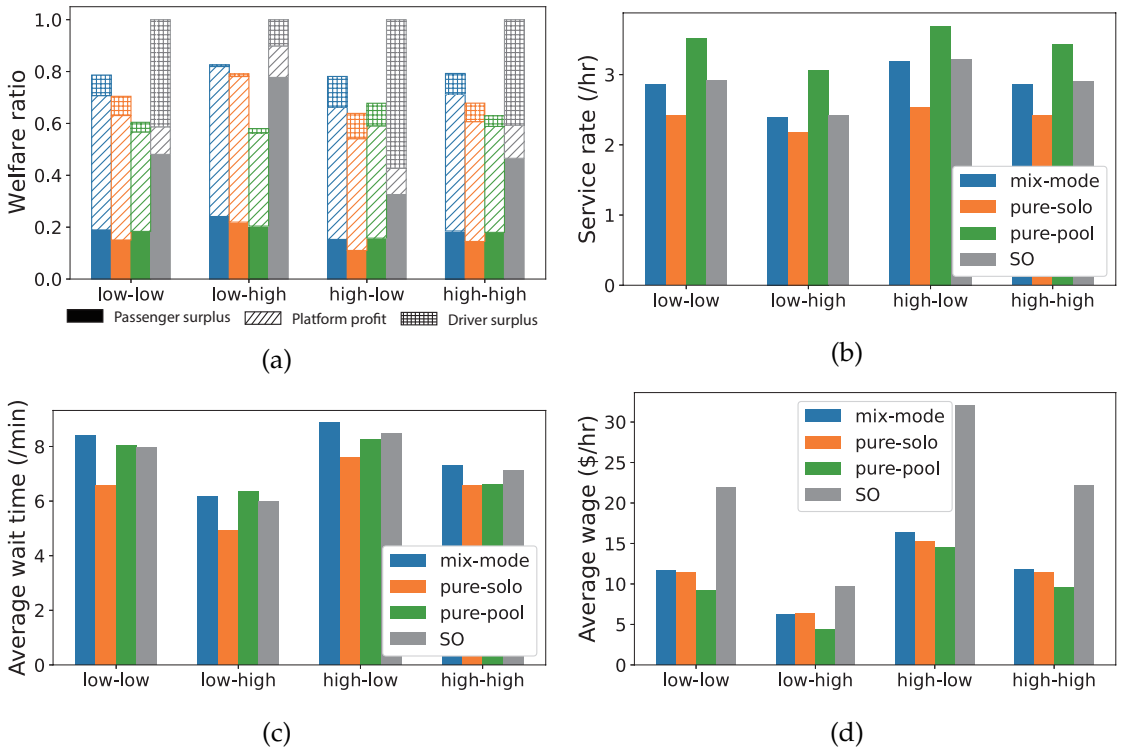


Figure 6: System performance under different operational strategies and representative market conditions.

Under profit-maximizing pricing, the platform's profit dominates the total social welfare; see Figure 6(a). In contrast, the SO pricing clearly prioritizes the surplus of passengers and drivers. Interestingly, SO did not wipe out the platform's profit completely (even it is allowed to do so). This suggests that the only constraint in P3 (that the profit should not be negative) is inactive at the optimum. In other words, the solution obtained here is indeed a true SO, rather than a second-best optimum. Incidentally, this finding also implies that no subsidy is needed to sustain an SO solution, which seems at odds with previous findings about taxi markets (Douglas, 1972; Arnott, 1996, e.g.,). Upon further inspection, we note that the discrepancy arises from the externality term (congestion cost $c_0 N$) included in the objective function. When that term is removed, a negative profit would indeed show up and the profitability constraint would be activated. Effectively, the congestion externality discourages over-supply in the ride-hail market.

Figure 6(b) examines the service rate, defined as the ratio between trip production and fleet size, i.e., $(Q_s + Q_p)/N$. As expected, pure-pool achieves the highest service rate while pure-solo has the lowest. The service rate under mix-mode is almost the same as that at SO, which suggests a profit-maximizing platform serving both solo and pooling rides could operate at a socially optimal service rate.

Figures 6(c) and (d) present, respectively, the average passenger wait time for ride-hail services (weighted by solo and pooling demand) and the average driver wage. In all cases, pure-solo offers the shortest wait time because it serves fewer passengers with a relatively large fleet of drivers. Also, the wait time is the highest in the case of high-demand-and-low-supply (high-low)

and the lowest in the case of low-demand-and-high-supply (low-high). Except for the low-high case, mix-mode always leads to the longest average passenger wait time but the highest driver wage. Yet, the difference from the second place is minor (less than half minus in wait time and \$1/hr in earning rate). Pure-pool is the worst strategy for driver income in all cases. At SO, however, the drivers are treated much better than all operational strategies under profit-maximizing pricing. Their hourly income more than doubled in some case (e.g., high-low). This is clearly linked to higher driver surplus of SO solution shown in Figure 6(a).

In summary, the mix-mode strategy seems the ideal choice for a profit-maximizing platform. Compared to the other two strategies, it maximizes both profit and social welfare, and brings greater benefits to passengers and drivers. Interestingly, although a profit-maximizing platform would not operate at the socially optimal scale, it tends to achieves a service rate (the number of trips served per vehicle) close to the SO level.

6.5 Impact of regulations

6.5.1 Minimum wage

To assess the impact of the minimum wage policy, we first solve P3 to obtain the "socially optimal" earning rates. These rates are then imported in P2 to derive a profit-maximizing platform's pricing strategy with an SO minimum wage constraint. A moment of reflection suggests that the policy would encourage more drivers to enter the service, which, in turn, attracts more passengers and boosts trip production. Indeed the entire system could move closer to SO. However, such a policy could be potentially detrimental to profitability, as the platform is now obligated to guarantee a minimum earning rate to anyone entering the market. Note that the price is the only "legal" tool available to the platform to manage the fleet size in the short term, and consequently, it has little recourse to reduce the fleet size below the lower bound now dictated by the government-mandated minimum wage. In the long run, however, the platform can reduce its driver pool S_0 to manage this downward pressure on profits. In fact, after New York City enacted the minimum wage policy, both Uber and Lyft have stopped hiring drivers (Edelstein, 2019b). Over the time, such a hiring freeze, along with other tactics, could reduce S_0. To examine this effect, we assume that the platform seeks to achieve a profit-maximizing S_0 for the minimum wage imposed by the regulator. More specifically, by solving P_2 over a range of possible values for S_0, we identify the S_0 that yields the highest profit and the scenario is then used to represent the long-term impact of the minimum wage policy.

Below, we compare the system performance under the four representative market conditions. For each condition, four scenarios are examined: profit maximization without minimum wage ("MO"), profit maximization with minimum wage and a fixed potential supply ("short"), profit maximization with minimum wage and an "optimized" potential supply ("long"), and SO ("SO"). Note that in Scenario "long", the value of S_0 differs from those used the other three scenarios due to the platform's presumed reaction. In all scenarios, the platform is assumed to adopt the mix-mode strategy.

Figure 7(a) plots the normalized welfare under each market condition. As expected, the minimum wage policy significantly improves social welfare in the short term, especially when the potential supply is relatively small. This improvement can be attributed to prioritizing driver

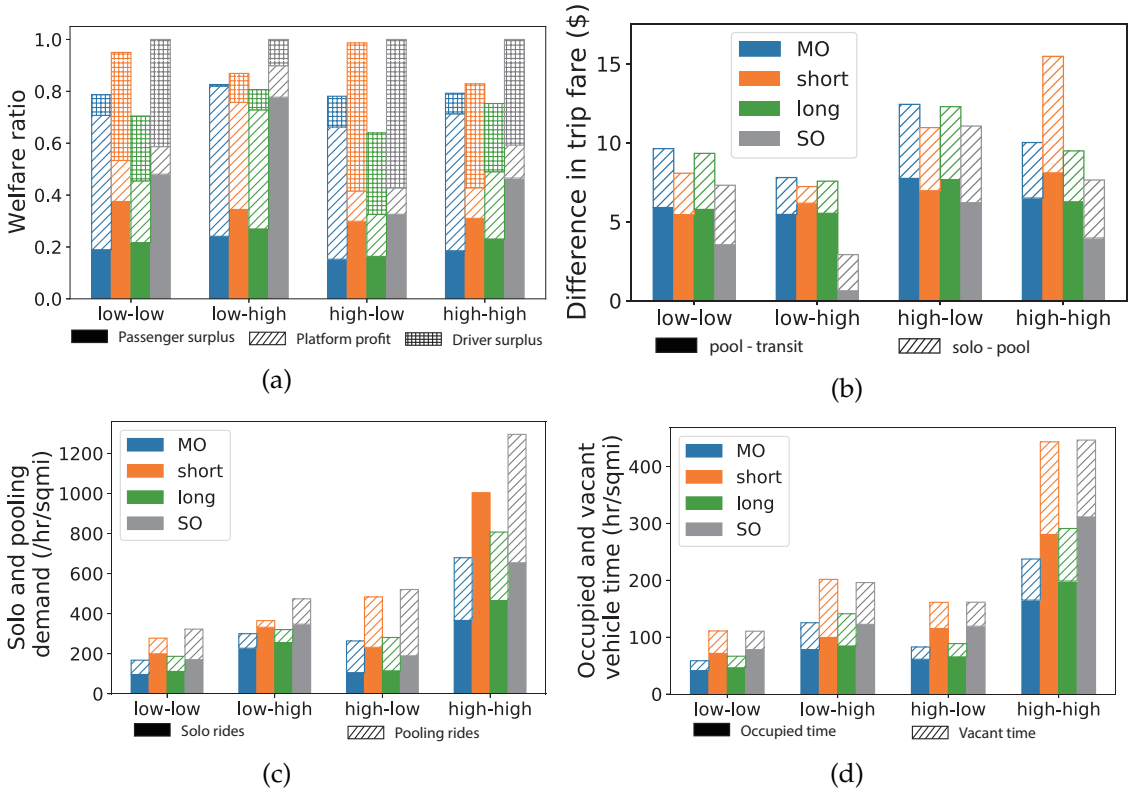


Figure 7: System performance under the minimum wage policy.

and passenger surplus at the expense of the platform's profit, which is related to the reduced market power predicted in Eqs. (24a) and (24b). Interestingly, the welfare distribution among the three stakeholders in Scenario "short" closely resembles that at SO. However, the result looks rather different in Scenario "long", where the platform is allowed to manipulate S_0 to its own advantage. By doing so, it manages to take back some of the lost profits, but unfortunately inflicts greater damages on driver and passenger surplus. *Under all conditions, the minimum wage policy ends up lowering the welfare in the long term, and the effect is more damaging when the supply pool is small to begin with.*

Figure 7(b) visualizes the platform's optimal pricing strategies. The bar chart in the plot represents the difference in trip fare between pooling and transit (solid filled) and between solo and pooling (pattern filled). In the short run, the platform tends to lower the trip fare of solo rides to attract passengers from transit. In this way, the platform could exhaust the additional service capacity induced by the minimum wage. An exception is the case of "high-high", where the platform raises both solo and pooling fares, but more for pooling than solo rides. As a result, no passenger would choose pooling at all; see Figure 7(c). Here, the rationale is to shift all pooling demand to solo rides, which not only keeps every driver busy but also generates higher revenue. These findings imply that *in a dense market, pooling may be completely eliminated by the minimum wage policy*. In the long-term, as the platform takes back more control on the supply side, it is able to keep the price close to the pre-regulation level for both pooling and solo rides.

Figures 7(c) and (d) show SO induces the highest demand and supply, followed by Scenario

"short". This result again confirms, in the short run, imposing a socially optimal minimum wage will force the platform to scale up and discourage pooling by adjusting its pricing strategy. In addition, Scenario "short" significantly increases both occupied and vacant vehicle time. In fact, under all market conditions tested, the vacant vehicle time induced by the minimum wage policy in the short run is even higher than that at SO. In the long run, the ride-hail market is scaled back to the unregulated level. However, the pooling ratio does not fully recover. Instead, it remains modestly lower than what is achieved in both unregulated scenario and SO. Although the long-term adjustment made by the platform will largely eliminate the supply surge achieved by the minimum wage policy, the supply in Scenario "long" remains above the unregulated level.

To summarize, although regulating the minimum wage does protect drivers from being unfairly exploited, it could create a host of problems. For one thing, the policy will surely draw opposition from the platform (e.g., Edelstein, 2019a). More importantly, by maintaining the supply and demand of ride-hail at an artificially high level, it could depress the use of collective modes (transit and pooling), thus exacerbating traffic congestion. In the long run, the platform might limit the potential supply in order to recover the lost profits. As a result, *the regulation in the name of social justice might even hurt the social welfare*.

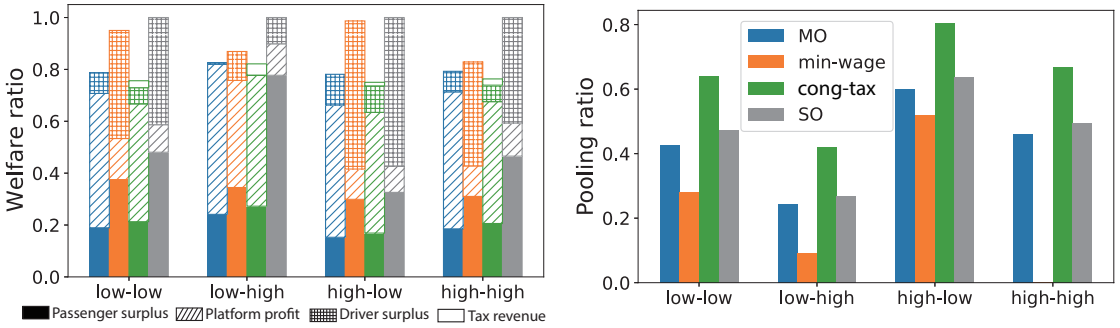
6.5.2 Congestion tax

We set the congestion tax $c_s = \$1$, which is in par with the actual policy implemented in Chicago[10]. Figure 8 compares the social welfare and pooling ratio in four scenarios: profit maximization without regulation ("MO"); profit maximization with minimum wage ("min-wage")[11]; profit maximization with congestion tax ("cong-tax"); and SO. The results show the congestion tax actually hurts the social welfare, though it slightly improves the passenger surplus; see Figure 8(a). Interestingly, the seemingly small congestion tax diverts a large number of passengers from solo to pooling rides. As shown in Figure 8(b), with the congestion tax, the pooling ratio rises by more than 20% under all market conditions. In fact, a close look reveals that this effect is so dramatic that it significantly reduces the wait time for pooling rides, which is eventually translated to a greater passenger surplus.

Therefore, while the minimum wage policy improves the social welfare (at least in the short run) but undermines ride-sharing, the congestion tax has exactly the opposite effect. One wonders, naturally, whether or not jointly implementing these two policies would lead to a win-win solution. The results reported in Figure 9 offers a preliminary but promising answer to the question. The joint policy achieves a higher social welfare than each individual policy in all but one case. The exception is the case of "high-low", where the minimum wage policy delivers a slightly higher social welfare. Several factors contribute to the rise of social welfare. First, the higher earning rate attracts more drivers to enter the market. The improvement in LOS of ride-hail services thus retains the demand for solo rides, which translates into a significantly higher tax revenue compared to the congestion tax itself. Second, the increase in vehicle supply, along

[10]The current policy charges congestion tax on all TNC trips, yet differently, based on the trip origin and destination as well as the time period. The price difference between solo and pooling rides is \$1.75 with downtown zone and \$0.6 otherwise. More details see `https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html`

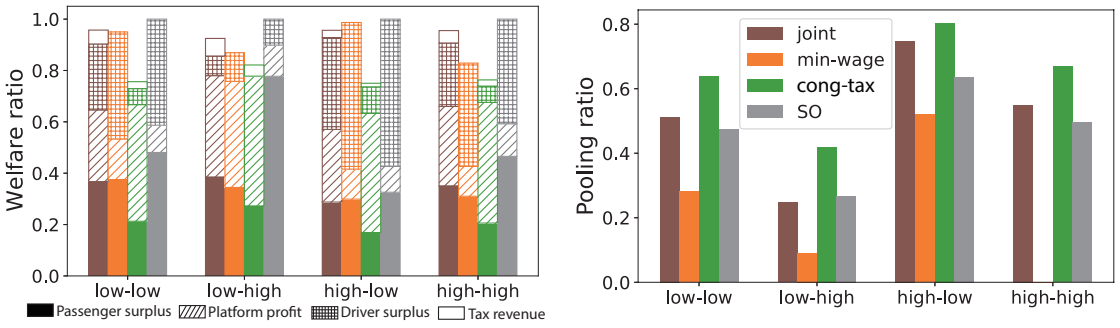
[11]This is equivalent to Scenario "short" in the previous section.



(a) (b)

Figure 8: System performance under congestion tax.

with the higher solo trip fare due to the congestion tax, makes pooling rides more appealing to passengers. Although the pooling ratio under the joint policy falls behind that only with the congestion tax policy, it is consistently higher than the pooling ratio at SO. It is reasonable to expect that the joint policies considered here is not "optimal" for a combined objective of minimizing social welfare, maximizing toll revenue and promoting ride-sharing. We leave the problem of finding such an optimal policy to a future study.



(a) (b)

Figure 9: System performance under joint regulation of minimum wage and congestion tax.

7 Conclusions

To pool or not to pool? This is the primary question for both passengers and the e-hail platform in our idealized aggregate market of personal mobility service. On the one hand, passengers choose a mode among pooling, solo and transit to maximize their own utility, in response to prices set by the platform. On the other hand, the platform determines an ideal pooling ratio and achieves that through pricing, while anticipating the movement of the market equilibrium.

To answer this question, we formulated the market equilibrium as a nonlinear equation system based on a spatial driver-passenger matching model that captures the operational characteristics of e-hail with pooling. We showed that, under mild conditions, this system always has an equilibrium that can be found using a simple iterative fixed-point algorithm. We then examined

and compared the platform's pricing decisions under three scenarios: profit maximization subject to market equilibrium constraint, profit maximization subject to both market equilibrium and regulatory constraints, and social welfare maximization subject to both market equilibrium and revenue neutral constraints. Main findings from numerical experiments are summarized below.

- As expected, pooling is desirable when demand is high but supply is scarce. However, its benefit diminishes quickly as the average en-route detour time (i.e., the difference between the average duration of solo and pooling trips) increases. Therefore, keeping this value under control is the key to the success of pooling.
- Without regulations, a mixed strategy, i.e., providing both solo and pooling rides, is the best choice for a profit-maximizing platform. Besides profit, it also achieves the highest social welfare compared to alternative strategies. Importantly, maintaining the system optimal output does not require subsidies if traffic congestion externality is considered in social welfare.
- The minimum wage policy can improve social welfare in the short term. However, in the long run, the platform might limit supply in an effort to recover the lost profits. As a result, the policy could end up undermining social welfare, and the damage is greater when the potential supply is small. Moreover, by maintaining the supply and demand of ride-hail at an artificially high level, it could depress the use of collective modes (transit and pooling), thus exacerbating traffic congestion.
- The congestion tax policy encourages pooling but hurts social welfare. Combining it with the minimum wage policy, however, achieves a desired balance between the two seemingly conflicting objectives in the short term.

In this study, we assume a single e-hail platform monopolizes the ride-hail market, even though it does not have full control on either side of the market. In reality, however, it is common to have multiple platforms competing with each other, as well as with conventional taxis. Hence, extending the analysis to accommodate such competitions is our immediate next step. A future study can also relax the assumptions made to simplify the matching process. Such an extension may endogenize the "matching parameters" (k and b) by linking them to such variables as matching interval/radius and maximum allowed detour. Accordingly, the platform may consider jointly optimizing the matching and the pricing decisions.

As argued in Castillo et al. (2018), surge pricing can protect the system from the "catastrophic consequence" of WGC. The underlying logic is that certain amount of demand must be "priced out" so that the system can return to an efficient state of operation. Could pooling solve WGC without leaving a portion of demand unserved? This is also an intriguing question worthy of consideration in future studies. Another possible direction is to extend the aggregate equilibrium model to a network equilibrium model. We expect that the spatiotemporal demand pattern in a network would become a crucial factor that influences the pooling ratio in each local market. Moreover, the platform may control the supply across local markets by encouraging or discouraging pooling. Hence, the spatial pricing problem would become more complex when pooling is involved.

Acknowledgment

The research was supported by the US National Science Foundation under the award number CMMI 1922665. We wish to thank the anonymous reviewers and the Associate Editor, Professor Robin Lindsey, for their constructive comments that were extremely helpful. The remaining errors are those of the authors' alone.

References

P. Afeche, Z. Liu, and C. Maglaras. Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. *Columbia Business School Research Paper*, (18-19): 18–19, 2018.

J. Alonso-Mora, S. Samaranayake, A. Wallar, E. Frazzoli, and D. Rus. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3):462–467, 2017.

J. D. Angrist, S. Caldwell, and J. V. Hall. Uber vs. taxi: A driver's eye view. Technical report, National Bureau of Economic Research, 2017.

R. Arnott. Taxi travel should be subsidized. *Journal of Urban Economics*, 40(3):316–333, 1996.

A. Asadpour, I. Lobel, and G. van Ryzin. Minimum earnings regulation and the stability of marketplaces. Available at SSRN 3502607 (Accessed: 2020-10-25), 2019.

M. Asghari and C. Shahabi. Adapt-pricing: a dynamic and predictive technique for pricing to maximize revenue in ridesharing platforms. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 189–198. ACM, 2018.

E. M. Azevedo and E. G. Weyl. Matching markets in the digital age. *Science*, 352(6289):1056–1057, 2016.

J. Bai, K. C. So, C. S. Tang, X. Chen, and H. Wang. Coordinating supply and demand on an on-demand service platform with impatient customers. *Manufacturing & Service Operations Management*, 2018.

S. Banerjee, C. Riquelme, and R. Johari. Pricing in ride-share platforms: A queueing-theoretic approach. Available at: SSRN 2568258 (Accessed: 2018-11-12), 2015.

A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, 18(1):5595–5637, 2017.

M. E. Beesley and S. Glaister. Information for regulating: the case of taxis. *The Economic Journal*, 93(371):594–615, 1983.

M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press, 1985.

O. Besbes, F. Castro, and I. Lobel. Spatial capacity planning. 2018a. Available at SSRN 3292651 (Accessed: 2019-10-23).

O. Besbes, F. Castro, and I. Lobel. Surge pricing and its spatial supply response. Available at SSRN 3124571 (Accessed: 2019-8-4), 2018b.

K. Bimpikis, O. Candogan, and D. Saban. Spatial pricing in ride-sharing networks. *Operations Research*, 2019.

F. P. Boscoe, K. A. Henry, and M. S. Zdeb. A nationwide comparison of driving distance versus straight-line distance to hospitals. *The Professional Geographer*, 64(2):188–196, 2012.

L. E. J. Brouwer. Über abbildung von mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115, 1911.

N. Buchholz. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. Available at `https://scholar.princeton.edu/sites/default/files/nbuchholz/files/taxi_draft.pdf` (Accessed: 2020-03-04), 2019.

G. P. Cachon, K. M. Daniels, and R. Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384, 2017.

R. D. Cairns and C. Liston-Heyes. Competition and regulation in the taxi industry. *Journal of Public Economics*, 59(1):1–15, 1996.

J. Castiglione, T. Chang, D. Cooper, J. Hobson, W. Logan, E. Young, B. Charlton, C. Wilson, A. Mislove, L. Chen, et al. TNCs today: a profile of San Francisco transportation network company activity. Technical report, San Francisco County Transportation Authority, 2016.

J. Castillo, D. T. Knoepfle, and E. G. Weyl. Surge pricing solves the wild goose chase. Available at SSRN 2890666 (Accessed: 2018-5-3), 2018.

H. Chen, K. Zhang, X. Liu, and Y. M. Nie. A physical model of street ride-hail. Available at SSRN 3318557 (Accessed: 2019-1-18), 2018.

M. K. Chen and M. Sheldon. Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. *Working Paper, University of California, Los Angeles*, 2017.

G. Cookson and B. Pishue. Global traffic scorecard. Technical report, INRIX, 2017.

J. Cramer and A. B. Krueger. Disruptive change in the taxi business: The case of uber. *American Economic Review*, 106(5):177–82, 2016.

G. De Jong, A. Daly, M. Pieters, and T. Van der Hoorn. The logsum as an evaluation measure: Review of the literature and new results. *Transportation Research Part A: Policy and Practice*, 41 (9):874–889, 2007.

A. S. De Vany. Capacity utilization under alternative regulatory restraints: an analysis of taxi markets. *The Journal of Political Economy*, pages 83–94, 1975.

G. W. Douglas. Price regulation and optimal service standards: The taxicab industry. *Journal of Transport Economics and Policy*, pages 116–127, 1972.

S. Edelstein. Lyft will sue New York City to block driver minimum wage law, 2019a. Available at https://www.thedrive.com/news/26247/lyft-will-sue-new-york-city-to-block-driver-minimum-wage-law (Accessed: 2019-02-03).

S. Edelstein. Uber and Lyft have officially stopped hiring drivers in NYC, 2019b. Available at https://www.thedrive.com/news/27756/uber-and-lyft-have-officially-stopped-hiring-drivers-in-nyc-will-resume-in-2020 (Accessed: 2019-05-02).

G. D. Erhardt, S. Roy, D. Cooper, B. Sana, M. Chen, and J. Castiglione. Do transportation network companies decrease or increase congestion? *Science advances*, 5(5):eaau2670, 2019.

G. Feng, G. Kong, and Z. Wang. We are on the way: Analysis of on-demand ride-hailing systems. Available at SSRN 2960991 (2019-5-23), 2017.

D. Flores-Guri. An economic analysis of regulated taxicab markets. *Review of industrial organization*, 23(3):255–266, 2003.

M. W. Frankena and P. A. Pautler. Taxicab regulation: an economic analysis. *Research in Law and Economics*, 9:129–165, 1986.

G. R. Frechette, A. Lizzeri, and T. Salz. Frictions in a competitive, regulated market: Evidence from taxis. *American Economic Review*, 109(8):2954–92, 2019.

M. Furuhata, K. Daniel, S. Koenig, F. Ordonez, M. Dessouky, M.-E. Brunet, L. Cohen, and X. Wang. Online cost-sharing mechanism design for demand-responsive transport systems. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):692–707, 2014.

N. Garg and H. Nazerzadeh. Driver surge pricing. Available at arXiv:1905.07544 (Accessed: 2019-05-20), 2019.

Y. Guda and U. Subramanian. Your uber is arriving : Managing on-demand workers through surge pricing , forecast communication and worker incentives. 2018.

I. Gurvich, M. Lariviere, and A. Moreno. Operations in the on-demand economy: Staffing services with self-scheduling capacity. In *Sharing Economy*, pages 249–278. Springer, 2019.

A. J. Hawkins. NYC's new driver wage law means the days of cheap Uber rides are over, 2019. Available at https://www.theverge.com/2019/2/1/18206737/nyc-driver-wage-law-uber-lyft-via-juno (Accessed: 2019-09-20).

F. He and Z.-J. M. Shen. Modeling taxi services with smartphone-based e-hailing applications. *Transportation Research Part C: Emerging Technologies*, 58:93–106, 2015.

H. Hotelling. The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica: Journal of the Econometric Society*, pages 242–269, 1938.

B. Hu, M. Hu, and H. Zhu. Surge pricing and two-sided temporal responses in ride-hailing. Available at SSRN 3278023 (Accessed: 2019-8-4), 2018.

J. Jacob and R. Roet-Green. Ride solo or pool : The impact of sharing on optimal pricing of ride-sharing services. Available at SSRN 3008136 (Accessed: 2020-08-20), 2018.

M. Joshi, N. Cowan, O. Limone, K. McGuinness, and R. Rao. E-hail regulation in global cities. Technical report, Rudin Center for Transportation Policy and Management, NYU Wagner School of Public Service, 2019.

J. Ke, H. Yang, X. Li, H. Wang, and J. Ye. Pricing and equilibrium in on-demand ride-pooling markets. *Transportation Research Part B: Methodological*, 139:411–431, 2020.

J. Ke, Z. Zheng, H. Yang, and J. Ye. Data-driven analysis on matching probability, routing distance and detour distance in ride-pooling services. *Transportation Research Part C: Emerging Technologies*, 124:102922, 2021.

S. Krantz and H. Parks. *The implicit function theorem: history, theory and applications*. Springer Science and Business Media, 2012.

R. Lagos. An alternative approach to search frictions. *Journal of Political Economy*, 108(5):851–873, 2000.

A. P. Lerner. The concept of monopoly and the measurement of monopoly power. *The Review of Economic Studies*, 1(3):157–175, 1934.

S. Li, H. Tavafoghi, K. Poolla, and P. Varaiya. Regulating tncs: Should uber and lyft set their own rules? *Transportation Research Part B: Methodological*, 129, 2019.

J. D. Little. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961.

I. Lobel and S. Martin. Detours in shared rides. Available at SSRN 3711072 (Accessed: 2020-11-11), 2020.

Y. M. Nie. How can the taxi industry survive the tide of ridesourcing? evidence from shenzhen, china. *Transportation Research Part C: Emerging Technologies*, 79:242–256, 2017.

M. Nourinejad and M. Ramezani. Ride-sourcing modeling and pricing in non-equilibrium two-sided markets. *Transportation Research Part B: Methodological*, 2019.

E. Özkan and A. R. Ward. Dynamic matching for real-time ride sharing. *Stochastic Systems*, 2020.

J. A. Parrott and M. Reich. An earning standard for new york city app-based drivers: Economic analysis and policy assessment, 2018.

M. Patriksson. Sensitivity analysis of traffic equilibria. *Transportation Science*, 38(3):258–281, 2004.

G. Qin, Q. Luo, Y. Yin, J. Sun, and J. Ye. Optimizing matching time intervals for ride-hailing services using reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 129:103239, 2021.

P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, and C. Ratti. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, 111 (37):13290–13294, 2014.

B. Schaller. Unsustainable? the growth of app-based ride services and traffic, travel and the future of new york city, 2017.

B. Schaller. The new automobility: Lyft, Uber and the future of American cities, 2018.

J. P. Schwieterman. Uber economics: evaluating the monetary and travel time trade-offs of transportation network companies and transit service in chicago, illinois. *Transportation Research Record*, 2673(4):295–304, 2019.

M. H. Shapiro. Density of Demand and the Benefit of Uber. Availabel at `http://www.shapiromh.com.`, 2018.

K. A. Small and H. S. Rosen. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society*, pages 105–130, 1981.

S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press, 2018.

H. Sun, H. Wang, and Z. Wan. Model and analysis of labor supply for ride-sharing platforms in the presence of sample self-selection and endogeneity. *Transportation Research Part B: Methodological*, 125:76–93, 2019.

T. A. Taylor. On-demand service platforms. *Manufacturing & Service Operations Management*, 20 (4):704–720, 2018.

R. L. Tobin and T. L. Friesz. Sensitivity analysis for equilibrium network flow. *Transportation Science*, 22(4):242–250, 1988.

Y. Tong, L. Wang, Z. Zhou, L. Chen, B. Du, and J. Ye. Dynamic pricing in spatial crowdsourcing: A matching-based approach. In *Proceedings of the 2018 International Conference on Management of Data*, pages 773–788. ACM, 2018.

X. Wang, F. He, H. Yang, and H. O. Gao. Pricing strategies for a taxi-hailing platform. *Transportation Research Part E: Logistics and Transportation Review*, 93:212–231, 2016.

X. Wang, H. Yang, and D. Zhu. Driver-rider cost-sharing strategies and equilibria in a ridesharing program. *Transportation Science*, 52(4):868–881, 2018.

S. Wodinsky. In major defeat for Uber and Lyft, New York City votes to limit ride-hailing cars, 2019. Available at `https://www.theverge.com/2018/8/8/17661374/uber-lyft-nyc-cap-vote-city-council-new-york-taxi` (Accessed: 2019-09-20).

Z. Xu, Z. Li, Q. Guan, D. Zhang, Q. Li, J. Nan, C. Liu, W. Bian, and J. Ye. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 905–913. ACM, 2018.

Z. Xu, Y. Yin, and J. Ye. On the supply function of ride-hailing systems. *Transportation Research Part C: Emerging Technologies*, 00, 2019.

C. Yan, H. Zhu, N. Korolko, and D. Woodard. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)*, 2019.

H. Yang. Heuristic algorithms for the bilevel origin-destination matrix estimation problem. *Transportation Research Part B: Methodological*, 29(4):231–242, 1995.

H. Yang and T. Yang. Equilibrium properties of taxi markets with search frictions. *Transportation Research Part B: Methodological*, 45(4):696–713, 2011.

H. Yang, M. Ye, W. H. Tang, and S. C. Wong. Regulating taxi services in the presence of congestion externality. *Transportation Research Part A: Policy and Practice*, 39(1):17–40, 2005.

H. Yang, C. Fung, K. Wong, and S. Wong. Nonlinear pricing of taxi services. *Transportation Research Part A: Policy and Practice*, 44(5):337–348, 2010a.

H. Yang, C. W. Leung, S. Wong, and M. G. Bell. Equilibria of bilateral taxi–customer searching and meeting on networks. *Transportation Research Part B: Methodological*, 44(8):1067–1083, 2010b.

H. Yang, J. Ke, and J. Ye. A universal distribution law of network detour ratios. *Transportation Research Part C: Emerging Technologies*, 96:22–37, 2018.

H. Yang, X. Qin, J. Ke, and J. Ye. Optimizing matching time interval and matching radius in on-demand ride-sourcing markets. *Transportation Research Part B: Methodological*, 131:84–105, 2020a.

H. Yang, C. Shao, H. Wang, and J. Ye. Integrated reward scheme and surge pricing in a ride-sourcing market. *Transportation Research Part B: Methodological*, 134:126–142, 2020b.

J. J. Yu, C. S. Tang, Z.-J. Max Shen, and X. M. Chen. A balancing act of regulating on-demand ride services. *Management Science*, 2019.

L. Zha, Y. Yin, and H. Yang. Economic analysis of ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 71:249–266, 2016.

L. Zha, Y. Yin, and Y. Du. Surge pricing and labor supply in the ride-sourcing market. *Transportation Research Part B: Methodological*, 117(PB):708–722, 2018a.

L. Zha, Y. Yin, and Z. Xu. Geometric matching and spatial pricing in ride-sourcing markets. *Transportation Research Part C: Emerging Technologies*, 92:58–75, 2018b.

K. Zhang, H. Chen, S. Yao, L. Xu, J. Ge, X. Liu, and Y. M. Nie. An efficiency paradox of uberization. Availabel at: SSRN 3462912 (Accessed: 2019-10-15), 2019.

A Notations

Table A1: List of notations

Variable	Description	Unit
w_e (w_s)	e-hail (solo) passenger wait time	hr
w_p	first part of pooling passenger wait time (matching time plus pickup time of the first passenger)	hr
Δ	second part of pooling passenger wait time (pickup time of the second passenger)	hr
Λ (Λ_0)	vacant (unmatched) vehicle density	/sqmi
Π (Π_0)	waiting (unmatched) passenger density	/sqmi
Π_s (Π_p)	solo (pooling) waiting passenger density	/sqmi
$\tilde{N}_v(d)$ ($\tilde{N}_{mv}(d)$)	number of unmatched (matchable) vehicle within a distance d from a passenger	
$\tilde{N}_p(l)$ ($\tilde{N}_{mp}(l)$)	number of pooling (matchable) passenger within a distance l from a passenger	
v	cruising speed of vacant vehicles	mph
k	coefficient of matching efficiency	/sqmi
b	coefficient of pooling efficiency	
m	approximation parameter	/sqmi
δ	detour ratio of road network	
$\tilde{D}_e$ ($\tilde{D}_p$)	the distance between the e-hail (pooling) passenger and the closest matchable vehicle (passenger)	mi
D_0	total demand rate	/hr/sqmi
Q_s (Q_p)	solo (pooling) demand rate	/hr/sqmi
r_Π	fraction of waiting passenger for pooling	
r_Q	pooling ratio	
f_s (f_p, f_t)	trip fare of solo rides (pooling rides, transit)	\$
τ_s (τ_p, τ_t)	travel time of solo rides (pooling rides, transit)	hr
u_s (u_p, u_t)	general cost of solo rides (pooling rides, transit)	\$
Δu	average saving of each passenger due to switching from transit to ride-hail service	\$
ν	value of time	\$/hr
θ	Mode choice uncertainty	/\$
ζ	disutility factor of transit trips	\$/hr
S_0	potential supply	/sqmi
N	fleet size (number of drivers in operation)	/sqmi
V	vacant vehicle time	hr/sqmi
$\tilde{e}_0$ (e_0)	random (average) reservation rate	\$/hr
e	driver's earning rate	\$/hr
η	compensation rate (payment per unit occupied time)	\$/hr
c_s	congestion tax on each solo ride	\$
c_p	additional pickup fare in each pooling ride	\$
c_0	congestion cost of each ride-hail vehicle	\$

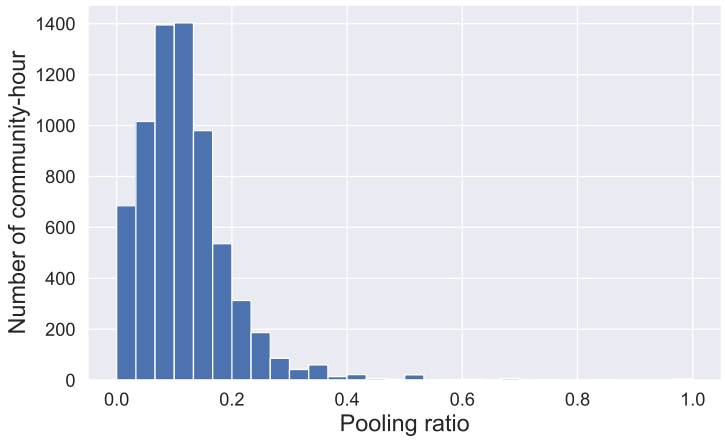


Figure A2: Histogram of pooling ratio in the study area and period.

B Data

Most of the parameters used in this study are estimated using a Chicago TNC dataset[12]. The data contain 7.66 million trips recorded in September 2019 with both pickup and dropoff locations in the City of Chicago. The total number of registered drivers is around 48K. We select ten of 77 communities in Chicago as the study area, which have the most pickups and dropoffs and collectively cover over 70% trips in the city. In addition, we only consider trips starting between 6 AM - 9 PM from Monday to Thursday, during which the demand patterns are relatively stable.

Parameters such as pickup rates, average travel time and distance can be directly obtained from the trip data. The cruising speed is approximated using the average speed of occupied trips. The detour ratio is computed as the average ratio between the recorded trip distance and the length of the straight line connecting the trip's origin and destination (located based on the geometric center of the corresponding census tracts). Drivers' hourly compensation rate is estimated based on hourly trip fare and the platform's commission rate (assumed to be 20%). The transit fare is obtained from Schwieterman (2019), while the transit trip duration is taken as the average of travel time estimates by Google Map API between OD pairs in the study area (weighed by trip numbers).

The data contains two pieces of information about pooling. The first indicates whether the passenger chooses to pool, and the other stores the number of passengers sharing the trip. If the pooling indicator is "true" whereas the number of passengers is one, then it means the passenger had failed to find a partner to share the trip. Using these information, we specify the average travel time and trip fare of solo and pooling rides. We note that this method might underestimate these parameters because we may only observe pooling rides that are more attractive to passengers (with shorter travel time and lower trip fare). Given this consideration, the travel time and trip fare of pooling rides in Table 1 are selected to be slightly higher than the actual estimates (0.26 hr and \$8/ride). The fraction of passengers waiting for pooling, r_{Π}, is estimated based on the observed pooling ratio r_Q (see Figure A2).

The congestion cost per vehicle is approximated based on recent studies on TNC. Erhardt et al. (2019) estimate the launch of TNC services in San Francisco has caused about 26,000 extra

[12] Available at `https://data.cityofchicago.org`.

vehicle delayed hours (VDH) per day in 2016 compared to 2010. Castiglione et al. (2016) report that over 6500 TNC vehicles operate in San Francisco during peak hour on a typical weekday in 2016. If we assume the average number of TNC vehicles in operation is 5000 over the day, then roughly about 0.22 hr of VDH per hour can be attributed to each vehicle. According to Cookson and Pishue (2017), an average American driver lost 99 hours to traffic congestion, translated to a monetary value of $1377. Thus, we estimate the congestion cost per TNC vehicle is $2.9/hr.

The value of time is estimated based on the value of business trips reported in US Bureau of Labor Statistics[13], adjusted to 2019 US dollar value. The total demand D_0 is approximated by summing up the ride-hail demand and transit ridership reported by CTA[14], while the total supply is based on the total number of registered drivers, adjusted according to the size of the study area and period.

The matching efficiency k and pooling efficiency b are not estimated from the Chicago TNC data. The value of k is taken from Zhang et al. (2019), who calibrate the e-hail matching model using TNC data collected in Shenzhen, China. By definition, b is the ratio between unmatched and total waiting pooling passenger densities. Hence, it is set to be the ratio between the average matching time (taken as 15 s) and the average total wait time (take as 5 min).

To determine the parameter m in Eq. (A8), we tested a range of values between 1 and 6. We found $\hat{A}(d,l)$ tends to overestimate (underestimate) $A(d,l)$ when $m = 6$ (1). Also, a value between 2 and 4 delivers similar approximation quality. Importantly, within this range, the performance of the equilibrium model seems insensitive to the choice of m. Based on the above tests, $m = 4$ is finally selected in numerical experiments.

C Number of vacant vehicles as Poisson Process

Proposition A2 *(Chen et al. (2018) Proposition 1) Under Assumption 1, the counting process $\tilde{N}_v(d)$ is an Inhomogeneous Poisson processes with intensity functions $\eta_v(d) = 2\pi d\Lambda_0$.*

Proof: Due to Assumption 1.1, $\tilde{N}_v(d) = 0$ and the increments of $\tilde{N}_v(d)$ are independent. Consider a ring area defined by d and $d + \Delta d$, and equally cut it into n small pieces with area Δs. Then, the number of vacant vehicle in the ring area follows binomial distribution where each piece contains one vacant vehicle with probability $\Lambda_0 \Delta s$. As n approaches to infinity, such a binomial distribution can be approximated by a Poisson distribution with rate

$$np = \frac{\pi(d+\Delta d)^2 - \pi d^2}{\Delta s}\Lambda_0 \Delta s = \pi\Lambda_0(2d + \Delta d)\Delta d. \tag{A1}$$

Hence,

$$\begin{aligned} & Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) = 1) = \pi\Lambda_0(2d+\Delta d)\Delta d \exp[-\pi\Lambda_0(2d+\Delta d)\Delta d] \\ \Rightarrow \lim_{\Delta d \to 0} & \frac{Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) = 1)}{\Delta d} = 2\pi d\Lambda_0, \end{aligned} \tag{A2}$$

$$\begin{aligned} & Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) > 1) = 1 - \exp[-\pi\Lambda_0(2d+\Delta d)\Delta d] - 2\pi d\Lambda_0 \\ \Rightarrow \lim_{\Delta d \to 0} & \frac{Pr(\tilde{N}_v(d+\Delta d) - \tilde{N}_v(d) > 1)}{\Delta d} \approx \lim_{\Delta d \to 0} \frac{1 - [1 - \pi\Lambda_0(2d+\Delta d)\Delta d]}{\Delta d} - 2\pi d\Lambda_0 = 0. \end{aligned} \tag{A3}$$

[13] Available at `https://www.transportation.gov/office-policy/transportation-policy/guidance-value-time`

[14] Available at `https://www.transitchicago.com/ridership`

Therefore, $\tilde{N}_v(d)$ is an Inhomogeneous Poisson Process with intensity function $\eta(d) = 2\pi d\Lambda_0$. □

Following the same reasoning, it can be proved that the counting process $\tilde{N}_{mp}(l)$ defined in Section 3.2 is an Inhomogeneous Poisson process with intensity function $2\pi lb\Pi_p$.

D Proof of Proposition 1

Because the pooling passengers are viewed as a single unit competing for unmatched vehicles, we first define the *effective density of waiting passenger* as

$$\Pi' \approx \Pi_s + \Pi_p/2. \tag{A4}$$

Hence, w_s is derived the same as w_e except that Π is replaced with Π'. From Eq.(3), we have

$$w_s = \frac{\delta}{2v}\sqrt{\frac{\Pi'}{k\Lambda}} = \frac{\delta}{2v}\sqrt{\frac{\Pi_s + \Pi_p/2}{k\Lambda}}. \tag{A5}$$

The derivation of w_p, however, is more complicated. Let us define $\tilde{D}_p$ as the minimum distance from *either* passenger to the closest matchable vehicle. For any given distance d, a search area is defined for each passenger as the area enclosed by a circle of radius d and centered at the passenger waiting location. We further define *effective search area* as the union of the two search areas. Accordingly, $Pr(\tilde{D}_p \leq d|l)$ gives the probability that at least one matchable vehicle appears inside the effective search area with parameter d conditional on the distance between the pooling pair l.

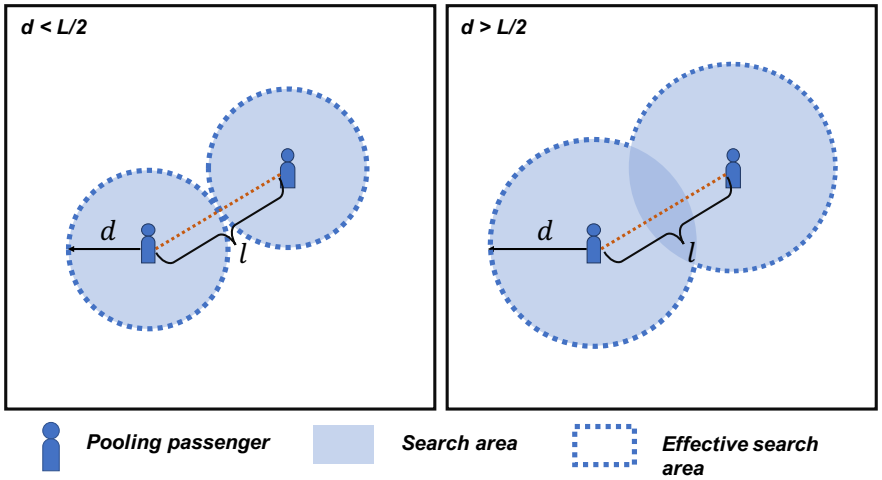


Figure A3: Illustration of search area and effective search area.

Let $A(d,l)$ denote the area of the effective search area. As illustrated in Figure A3, when $d \leq l/2$, the two passengers' search areas do not overlap. Hence, $A(d,l)$ simply equals $2\pi d^2$. When $d > l/2$, $A(d,l)$ is the total search area less the intersection. To summarize,

$$A(d,l) = \begin{cases} 2\pi d^2 & , d \leq l/2 \\ 2\pi d^2 - 2\cos^{-1}\left(\frac{l}{2d}\right)d^2 + dl\sqrt{1-\left(\frac{l}{2d}\right)^2} & , d > l/2 \end{cases}. \tag{A6}$$

Therefore, $Pr(\tilde{D}_p \leq d|l)$ can be written as

$$Pr(\tilde{D}_p \leq d|l) = 1 - \exp\left(-\frac{\Lambda}{k\Pi'} A(d,l)\right). \tag{A7}$$

Compared to Eq. (2), pooling passengers double their competing power when $d \leq l/2$, though this advantage diminishes as l decreases. In an extreme case, when $l = 0$, the pooling passengers are expected to have the same wait time as solo passengers.

The fact that $A(d,l)$ is not smooth makes it difficult to evaluate the expectation of $\tilde{D}_p$. Hence, we propose to approximate it with a smooth function as follows:

$$\hat{A}(d,l) = \left(2 - \frac{1}{1+ml^2}\right)\pi d^2. \tag{A8}$$

It is easy to verify that $\hat{A}(d,l) \to \pi d^2$ as $l = 0$ and $\hat{A}(d,l) \to 2\pi d^2$ as $l \to \infty$. Thus, $\hat{A}(d,l)$ well captures the lower and upper bounds of $A(d,l)$ and the parameter m may be adjusted to achieve good approximation.

Using $\hat{A}(d,l)$, the conditional expectation of $\tilde{D}_p|l$ is derived as

$$E[\tilde{D}_p|l] = \frac{1}{2}\sqrt{\frac{\Pi'}{k\Lambda}}\left(2 - \frac{1}{1+ml^2}\right)^{-1/2}. \tag{A9}$$

Recall that l is a realization of random variable $\tilde{L}$. Thus, the expectation of $\tilde{D}_p$ is given by

$$D_p = E[E[\tilde{D}_p|\tilde{L}]] = \int_0^\infty E[\tilde{D}_p|l]\mathrm{d}F_{\tilde{L}}(l), \tag{A10}$$

where $F_{\tilde{L}}(l)$ is CDF of $\tilde{L}$.

The above integral cannot be derived analytically due to the functional form of Eq. (A9). Instead, we introduce the following approximation:

$$D_p = E[E[\tilde{D}_p|\tilde{L}] \approx E[\tilde{D}_p|E[\tilde{L}]] = \frac{1}{2}\sqrt{\frac{\Pi'}{k\Lambda}}\left(2 - \frac{1}{1+mL^2}\right)^{-1/2}. \tag{A11}$$

Plugging Eq. (4) into Eq. (A11) thus yields

$$w_p = \frac{\delta}{v}D_p = \frac{\delta}{2v}\sqrt{\frac{\Pi'}{k\Lambda}\frac{m + 4b\Pi_p}{2m + 4b\Pi_p}}. \tag{A12}$$

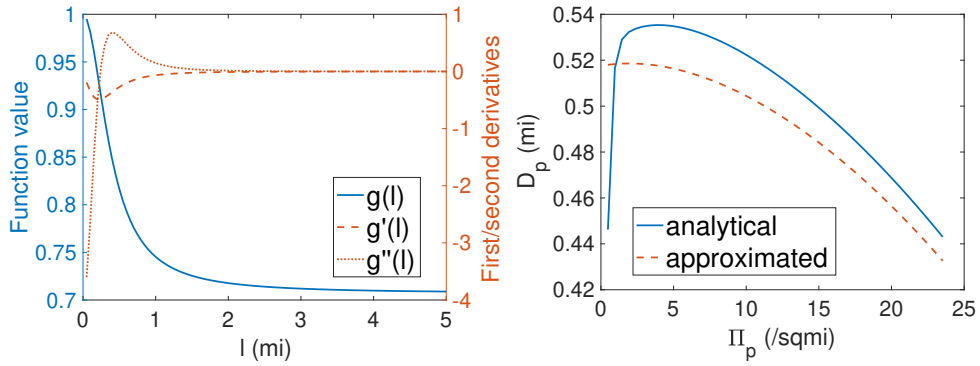
The approximation made in Eq. (A11) warrants some discussions. Introduce a new function

$$g(l) = \left(2 - \frac{1}{1+ml^2}\right)^{-1/2}. \tag{A13}$$

Then, the approximation made in Eq. (A11) is equivalent to $E[g(l)] \approx g(E[l]) = g(L)$. The approximation quality depends on the functional form of $g(l)$. As per Jensen's inequality, if $g(l)$ is a linear function, the approximation is subject to no error. Otherwise, $E[g(l)] \geq g(L)$ holds when $g(l)$ is convex, and $E[g(l)] \geq g(L)$ if $g(l)$ is concave.

Figure A4(a) plots the function value, first and second derivatives of $g(l)$ with $m = 4$. It can be seen that $g(l)$ is convex when $l > 0.25$ and quickly converges to $1/\sqrt{2}$. Hence, Eq. (A11) is likely

to underestimate D_p (hence w_p). This finding is confirmed in Figure A4(b), which compares the approximated value and the analytical result given by Eq. (A10) (the analytical result is computed by numerical integration). Specifically, we fix the total waiting passenger density Π and vacant vehicle density Λ at the default values in Table 1, and vary the waiting passengers density for pooling Π_p. All other parameters are set according to Table 1. As shown in Figure A4(b), the approximated value slightly underestimates D_p except when Π_p is close to 0. For most values of Π_p, the error is within 5% and it decreases as Π_p increases. Therefore, Eq. (A11) does offer a reasonable approximation for D_p.



(a) Functional properties of $g(l)$ (b) Analytical and approximated D_p

Figure A4: Analysis of approximation quality

E Constant matching efficiency k prevents Wild Goose Chase (WGC)

Castillo et al. (2018) defines WGC as a state when the system throughput Q decreases with passenger wait time w. From Eq. (3), we have

$$w_e = \frac{\delta}{2v}\sqrt{\frac{\Pi}{k\Lambda}} \quad \Rightarrow \quad w_e^2 = \frac{\delta}{4v^2}\frac{Qw_e}{kV} \quad \Rightarrow \quad V = \frac{\delta}{4v^2}\frac{Q}{kw_e}.$$

Plugging into the flow conservation $N = V + Q\tau$ yields

$$N = \frac{\delta}{4v^2}\frac{Q}{kw_e} + Q\tau \quad \Rightarrow \quad Q = \frac{N}{\tau + \frac{\delta}{4v^2 k w_e}}.$$

When N and other parameters, including k, are fixed, Q is monotonically increasing with w_e and thus WGC would never emerge. This violation is due to the assumption that k remains constant over time. In reality, it is expected that the matching process becomes inefficient under extreme demand-supply imbalance. In other words, k is more likely to be a piece-wise function of w_e. That is, when w_e is below certain threshold, k is a constant; and as w_e exceeds the threshold, k would decrease with w_e, i.e., $k'(w_e) < 0$. Accordingly,

$$\frac{\partial Q}{\partial w_e} = \frac{N}{\left(\tau + \frac{\delta}{4v^2 k(w_e) w_e}\right)^2} \frac{\delta}{(4v^2 k(w_e) w_e)^2} \left(k(w_e) + k'(w_e) w_e\right),$$

and the system enters WGC when $w_e > -k(w_e)/k'(w_e) \rightarrow \frac{\partial Q}{\partial w_e} < 0$.

F Proof of Proposition 2

Brouwer's fixed-point theorem (Brouwer, 1911) states that: if a continuous function $f : \Omega \subset R^n \to \Omega$ maps a compact and convex set Ω to itself, then there exists $\mathbf{x}^* \in \Omega$ such that $\mathbf{x}^* = F(\mathbf{x}^*)$.

We first prove Ω is compact and convex. By definition, $\mathbf{x} = (w_s, w_p, \Delta)^T \in \Omega \subset \mathbb{R}^3_+$. From the assumption made in the proposition, w_s, w_p and Δ must all be bounded from the above, otherwise the demand for solo and/or pooling rides will disappear all together. We now show that these three variables must also have lower bounds. First, passenger wait times reach lower bounds $\underline{w}_s = \frac{\delta}{2v\sqrt{S_0}}$ and $\underline{w}_p = \frac{\delta}{2v\sqrt{2S_0}}$, when the fleet size approaches its upper bound S_0 and the demand D_0 approaches zero. Second, as Π_p is bounded from above by $D_0(\overline{w}_p + \overline{\Delta})$, the lower bound of Δ is given by $\underline{\Delta} = \frac{\delta}{2v}[bD_0(\overline{w}_p + \overline{\Delta})]^{-1/2}$. Consequently, Ω can be defined as the cubic space $[\underline{w}_s, \overline{w}_s] \times [\underline{w}_p, \overline{w}_p] \times [\underline{\Delta}, \overline{\Delta}]$, which is compact and convex.

We proceed to show the self-map $F(\cdot)$, i.e., Eqs. (13d)–(13f), is continuous. Recall that Eqs. (13d)–(13f) are continuous functions of Q_s, Q_p and V, along with w_s, w_p and Δ. From (13c), we know V is a continuous function of Q_s, Q_p and N. Therefore, to show F is continuous, we only need to prove Q_s, Q_p and N are continuous functions of w_s, w_p and Δ. The continuity of previous two are directly shown from Eqs. (13a) and (9). The last one is more complicated as it involves the implicit function Eq. (13b). We prove this result in Lemma 1.

Lemma 1 *The fleet size N defined in Eq.* (13b) *can be represented as a continuous function of $\mathbf{x} = (w_s, w_p, \Delta)^T$.*

Proof: We apply the implicit function theorem (Krantz and Parks, 2012) to prove the result.

Consider a continuously differentiable function $L : \mathbb{R}^{n+m} \to \mathbb{R}^m$ and a point $(\mathbf{x}_0, \mathbf{y}_0), \mathbf{x}_0 \in \mathbb{R}^n, \mathbf{y}_0 \in \mathbb{R}^m$ such that $L(\mathbf{x}_0, \mathbf{y}_0) = 0$. The theorem states that, if the Jacobian matrix

$$J_{L,\mathbf{y}}(\mathbf{x}_0, \mathbf{y}_0) = \left[\frac{\partial L_i}{\partial y_j}(\mathbf{x}_0, \mathbf{y}_0)\right],\ i = 1, \ldots, m,\ j = 1, \ldots, m \tag{A14}$$

is invertible, then there is a neighborhood of $\mathbf{x}_0$, denoted as $U \subset \mathbb{R}^n$, and a unique continuously differentiable function $g : U \to \mathbb{R}^m$ such that $\mathbf{y} = g(\mathbf{x}), \forall \mathbf{x} \in U$.

To apply the above result, let us rewrite Eq. (13b) as

$$L(\mathbf{x}, N) = S_0 G\left(\frac{1}{N}\left[\eta\left(Q_s(\mathbf{x})\tau_s + \frac{1}{2}Q_p(\mathbf{x})\tau_p\right) + \frac{c_p}{2}Q_p(\mathbf{x})\right]\right) - N = 0. \tag{A15}$$

Therefore, for any point that satisfies Eq. (A15), N is a continuous function of $\mathbf{x}$ in a neighborhood of that point provided $\frac{\partial L}{\partial N}$ is invertible, or equivalently,

$$\frac{\partial L}{\partial N} = -S_0 G' \frac{1}{N^2}\left[\eta\left(Q_s(\mathbf{x})\tau_s + \frac{1}{2}Q_p(\mathbf{x})\tau_p\right) + \frac{c_p}{2}Q_p(\mathbf{x})\right] - 1 \neq 0. \tag{A16}$$

To see why (A16) must hold, note that G' is the probability density function of the drivers' reservation rate. Thus, G', as well as all other variables, must be nonnegative. Accordingly, $\partial L/\partial N \leq -1$ and hence it cannot be zero. This completes the proof. □

Therefore, both conditions stated in Brouwer's fixed-point theorem are satisfied. We hence conclude that the existence of a solution to Eq. (13) is guaranteed.

G Determination of upper bounds for Δ, w_s and w_p

The MNL model implies that the mode share decreases exponentially with its general cost but never reaches zero. However, in reality no one would choose solo or pooling if the corresponding wait time (w_s and $w_p + \Delta$) are too long. Let ε be the minimum demand considered to be meaningful for analysis. We next show the upper bound for Δ, denoted as $\overline{\Delta}$, can be derived as a function of ε. Note that

$$Q_p = D_0 \frac{e^{-\theta u_p}}{\sum_i e^{-\theta u_i}} \leq D_0 \frac{e^{-\theta u_p}}{2e^{-\theta \underline{u}} + e^{-\theta u_p}}, \tag{A17}$$

where $\underline{u} = \min\{u_s, u_t\} = \min\{f_s + \nu(\underline{w}_s + \tau_s), f_t + \nu\tau_t\}$ (recall $\underline{w}_s = \frac{\delta}{2v\sqrt{S_0}}$; see the proof of Proposition 2). Setting

$$D_0 \frac{e^{-\theta u_p}}{2e^{-\theta \underline{u}} + e^{-\theta u_p}} \leq \varepsilon$$

yields

$$\begin{aligned} \left(\frac{D_0}{\varepsilon} - 1\right) e^{-\theta u_p} \leq 2e^{-\theta \underline{u}} \quad &\Rightarrow \quad u_p \geq \underline{u} + \frac{1}{\theta} \log \frac{D_0/\varepsilon - 1}{2} \\ &\Rightarrow \quad \Delta \geq \frac{1}{\nu}(\underline{u} - f_p) + \frac{1}{\theta\nu} \log \frac{D_0/\varepsilon - 1}{2} - w_p. \end{aligned} \tag{A18a}$$

In other words, whenever Eq. (A18a) is satisfied, the demand for pooling would reduce to no more than ε. Since $w_p \geq 0$, we may set

$$\overline{\Delta} \equiv \frac{1}{\nu}(\underline{u} - f_p) + \frac{1}{\theta\nu} \log \frac{D_0/\varepsilon - 1}{2}. \tag{A19}$$

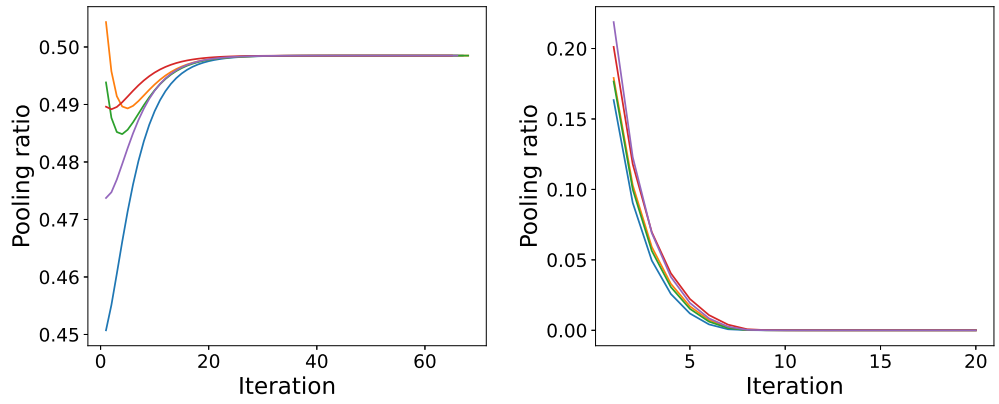
It is clear that the upper bound established this way is likely to be loose. However, it suffices for our purpose to show a finite upper bound does exist. $\overline{w}_s$ and $\overline{w}_p$ can be obtained in a similar fashion, and the details are omitted for brevity.

H Equilibrium stability

Because Eq. (13) is a highly nonlinear equation system, theoretically it could have more than one solution. In addition, a solution may or may not be *stable*. An equilibrium is said to be stable if the system always returns to it after a small perturbation. We are particularly interested in stable solutions.

Recall that the solution to Eq. (13) can be represented as a fixed point, i.e., $\mathbf{x}^* = F(\mathbf{x}^*)$, where $\mathbf{x}^* = (w_s^*, w_p^*, \Delta^*)^T$. The stability theory (e.g., Strogatz, 2018) states that a solution $\mathbf{x}^*$ is stable if and only if all eigenvalues of the Jacobian matrix of $F(\cdot)$ at $\mathbf{x}^*$, denoted by $J_{\mathbf{x}^*}$, have absolute values less than 1. Using this result, we screen each equilibrium solution obtained from fixed-point iterations and only keep those that pass the stability test. Since $F(\cdot)$ is explicitly expressed, i.e., Eq. (13), we could evaluate $J_{\mathbf{x}^*}$ using automatic differentiation (Baydin et al., 2017).

Figure A5 reports the convergence performance of the iterative fixed-point algorithm in two examples, both using the default parameters given in Table 1 except for the pricing strategies (f_s, f_p and η). The algorithm is terminated when the gap drops below a predefined threshold, set at 10^{-8} in this study. Each convergence curve (gap vs. number of iteration) in the plots



(a) $f_s = 17, f_p = 12, \eta = 20$ (b) $f_s = 19, f_p = 17, \eta = 21$

Figure A5: Convergence performance of the iterative fixed-point algorithms. (a) All initial solutions lead to the same stable equilibrium; (b) All initial solutions lead to the same unstable equilibrium.

represents a different initial solution. These results are representative of our overall experience with the fixed-point algorithm, which is that it generally converges quite fast and the choice of the initial solutions tends to have a negligible impact on convergence. In this particular experiment, the pricing strategy on the left leads to a stable equilibrium and the algorithm is able to locate it regardless of where it starts. However, an unstable equilibrium emerges when the pricing strategy on the right is employed. This happens because the general cost of a pooling ride is close to that of a solo ride when there is no demand for pooling. As a result, a small perturbation in the inputs could make pooling more attractive, forcing the solution to significantly deviate from the pre-perturbation one. Still, the algorithm converges faithfully to this unstable equilibrium, starting from all tested initial points.

In light of the above observations, whenever the market equilibrium is sought, we run the fixed-point algorithm multiple times with randomly generated initial solutions. If all runs converge to the neighborhood of the same fixed point, we take their average as the final equilibrium solution. If there is more than one equilibrium solution, we test their stability and only keep the stable ones. Curiously, in all numerical experiments conducted in this study, we did not come across a single case where multi equilibrium were found.

I Derivation of ∇R

Gradient ∇R is evaluated as

$$\frac{\partial R}{\partial f_s} = Q_s + (f_s - \eta\tau_s)\frac{\partial Q_s}{\partial f_s} + \left(f_p - \frac{1}{2}\eta\tau_p - \frac{1}{2}c_p\right)\frac{\partial Q_p}{\partial f_s}; \tag{A20a}$$

$$\frac{\partial R}{\partial f_p} = Q_p + (f_s - \eta\tau_s)\frac{\partial Q_s}{\partial f_p} + \left(f_p - \frac{1}{2}\eta\tau_p - \frac{1}{2}c_p\right)\frac{\partial Q_p}{\partial f_p}; \tag{A20b}$$

$$\frac{\partial R}{\partial \eta} = (f_s - \eta\tau_s)\frac{\partial Q_s}{\partial \eta} + \left(f_p - \frac{1}{2}\eta\tau_p - \frac{1}{2}c_p\right)\frac{\partial Q_p}{\partial \eta} - Q_s\tau_s - \frac{1}{2}Q_p\tau_p = 0. \tag{A20c}$$

In what follows, we explain how to compute $\partial R/\partial f_s$ in each iteration. $\partial R/\partial f_p$ and $\partial R/\partial \eta$ can be computed similarly. The two components to be evaluated in Eq. (A20a) are $\partial Q_s/\partial f_s$ and $\partial Q_p/\partial f_s$. Take $\partial Q_s/\partial f_s$ as an example. We expand it as

$$\frac{\partial Q_s}{\partial f_s} = D_0 \left[\frac{\partial q}{\partial f_s} + \frac{\partial q}{\partial w_s}\frac{\partial w_s}{\partial f_s} + \frac{\partial q}{\partial w_p}\frac{\partial w_p}{\partial f_s} + \frac{\partial q}{\partial \Delta}\frac{\partial \Delta}{\partial f_s} \right]. \tag{A21}$$

Here, the partial derivatives of the function q with respect to f_s, w_s, w_p and Δ can be evaluated numerically using automatic differentiation (Baydin et al., 2017).

To obtain the implicit partial derivatives $\partial w_s/\partial f_s$, $\partial w_p/\partial f_s$ and $\partial \Delta/\partial f_s$, we first rewrite Eq. (13) as follows:

$$w_s = \frac{\delta}{2v\sqrt{k}}\sqrt{\frac{\Pi'}{V}}, \tag{A22a}$$

$$w_p = \frac{\delta}{2v\sqrt{k}}\sqrt{\frac{\Pi'}{V}\frac{m+4b\Pi_p}{2m+4b\Pi_p}}, \tag{A22b}$$

$$\Delta = \frac{\delta}{2v\sqrt{b}}\frac{1}{\sqrt{\Pi_p}}, \tag{A22c}$$

where Π' is defined in (A4). Also, Π', V and Π_p can be viewed as functions of f_s, w_s, w_p, Δ as per Eqs. (13a)-(13c).

Taking logarithm and then differentiating with respect of f_s on both sides of Eq. (A22) yields

$$\frac{1}{w_s}\frac{\partial w_s}{\partial f_s} = \frac{1}{2\Pi'}\frac{\partial \Pi'}{\partial f_s} - \frac{1}{2V}\frac{\partial V}{\partial f_s}, \tag{A23a}$$

$$\frac{1}{w_p}\frac{\partial w_p}{\partial f_s} = \frac{1}{2\Pi'}\frac{\partial \Pi'}{\partial f_s} - \frac{1}{2V}\frac{\partial V}{\partial f_s} + \frac{1}{2}\left(\frac{4b}{m+4b\Pi_p} - \frac{4b}{2m+4b\Pi_p}\right)\frac{\partial \Pi_p}{\partial f_s}, \tag{A23b}$$

$$\frac{1}{\Delta}\frac{\partial \Delta}{\partial f_s} = -\frac{1}{2\Pi_p}\frac{\partial \Pi_p}{\partial f_s}. \tag{A23c}$$

$\partial \Pi'/\partial f_s$ can be evaluated similarly as $\partial Q_s/\partial f_s$. Recall that $\Pi' = w_s Q_s + (w_p + \Delta)Q_p$. We may represent Π' as a function $\Pi' = \pi(w_s, w_p, \Delta, Q_s, Q_p)$. Accordingly,

$$\frac{\partial \Pi'}{\partial f_s} = \frac{\partial \pi}{\partial w_s}\frac{\partial w_s}{\partial f_s} + \frac{\partial \pi}{\partial w_p}\frac{\partial w_p}{\partial f_s} + \frac{\partial \pi}{\partial \Delta}\frac{\partial \Delta}{\partial f_s} + \frac{\partial \pi}{\partial Q_s}\frac{\partial Q_s}{\partial f_s} + \frac{\partial \pi}{\partial Q_p}\frac{\partial Q_p}{\partial f_s}. \tag{A24}$$

Again, the partial derivatives of π can be computed by automatic differentiation.

In other words, $\partial \Pi'/\partial f_s$ can be expressed as a linear function of $\partial w_s/\partial f_s$, $\partial w_p/\partial f_s$ and $\partial \Delta/\partial f_s$. $\partial V/\partial f_s$ and $\partial \Pi_p/\partial f_s$ in Eq. (A23) can be derived in the same way. Consequently, Eq. (A23) turns into a linear equation system with respect to $\partial w_s/\partial f_s$, $\partial w_p/\partial f_s$ and $\partial \Delta/\partial f_s$.

Plugging the solution of Eq. (A23) into Eq. (A21), we can obtain $\partial Q_s/\partial f_s$. The computation of $\partial Q_p/\partial f_s$ is similar and omitted here for brevity.

J Sensitivity of market equilibrium to en-route detour

As discussed in Section 4.1, the en-route detour $\tau_p - \tau_s$ is expected to decrease with pooling demand. Yet, in all numerical experiments presented in the main text, we have assumed the

detour to be constant for simplicity. In this appendix, we adopt the results of Ke et al. (2021) and Lobel and Martin (2020) to test the sensitivity of our findings to the endogeneous en-route detour.

Ke et al. (2021) empirically observe the passenger detour distance Δl follows

$$\frac{\Delta l}{\bar{l}} = \frac{1}{\alpha N + \beta}, \tag{A25}$$

where $\bar{l}$ is the average trip distance, N is the number of requests accumulated in a matching interval, and α, β are coefficients[15].

Since vehicles travel at a constant speed v, $\Delta l / l = (\tau_p - \tau_s)/\tau_s$. In addition, the batch demand N in Eq. (A25) can be replaced with $b\Pi_p$, which is the unmatched pooling passenger density. Accordingly, we adjust the average trip duration of pooling rides as follows:

$$\tau_p = \tau_s \left(1 + \min\left(0.5, \frac{1}{\alpha b \Pi_p + \beta}\right)\right). \tag{A26}$$

Note that the upper bound 0.5 on $1/(\alpha b \Pi_p + \beta)$ follows from the result proved in Lobel and Martin (2020).

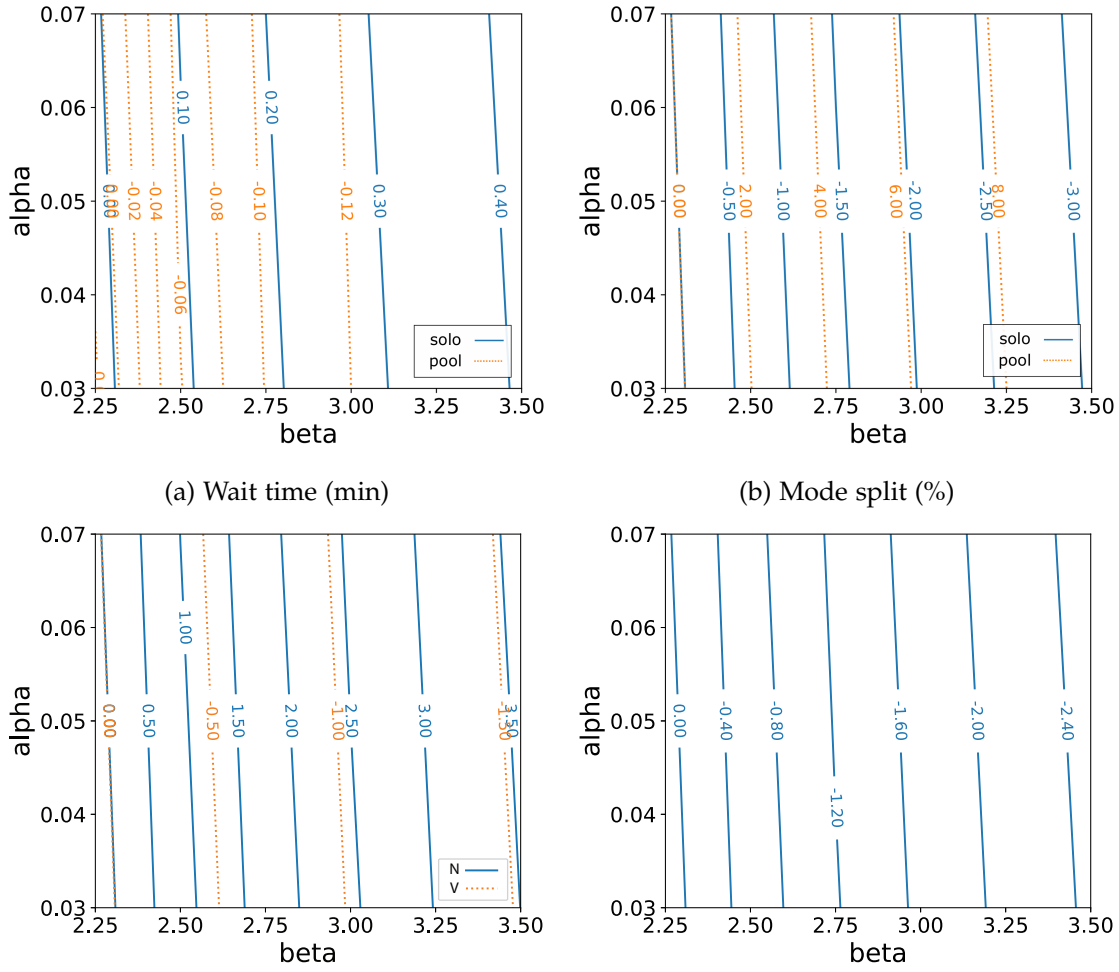
Ke et al. (2021) calibrated the coefficients α, β for New York City and two other cities in China with various matching radius. We take the range of coefficient values associated with New York City and calculate the market equilibrium with default total demand ($D_0 = 1200$/hr/sqmi) and potential supply ($S_0 = 550$/sqmi).

Figure A6 shows how the key outputs of the equilibrium model vary with α and β. First, the market equilibrium is clearly insensitive to α, mainly because the first term $\alpha b \Pi_p$ is much smaller compared to β. In other words, the pooling demand has a minor impact on the en-route detour, as long as the pooling demand is not too large. The differences in passenger wait times and average trip duration of pooling rides are small (less 10%), and those in fleet size and vacant vehicle density are almost negligible (less than 1%). The mode split of pooling rides is the most sensitive to β. As shown in Figure A6(b), the share of pooling increases in the adjusted model by more than 8%, when β increases from 2.25 to 3.5. As a larger β corresponds to a smaller detour ratio, this result is expected.

We continue to examine the sensitivity of market equilibrium to different supply-demand levels using the adjusted model. Since the results are demonstrated insensitive to α, we fix it as 0.05 and consider two extreme cases of β, i.e., $\beta = 2.25$ and $\beta = 3.5$. The results, along with the market equilibria with constant en-route detour, are illustrated in Figures A7 and A8. As expected, when $\beta = 2.25$, the adjusted model barely causes any meaningful changes in the equilibrium solution. The most prominent sensitivity is again shown in the mode split when β is large. Specifically, the difference in pooling share reaches about 10% when the total demand is low or the potential supply is high. The difference in solo share, on the other hand, is quite stable across different supply and demand levels, at around 3-5%. The differences in passenger wait time is even smaller, less than half of an minute in most scenarios.

The above findings imply that by assuming a constant detour, our model may underestimate the attractiveness of pooling, especially when β is large. However, the values of α and β used in the sensitivity analyses, calibrated for New York City, may not fit our case study (based in

[15]Here, we follow the same notations in Ke et al. (2021), hence there are a few conflicts with the notations in our main text



(a) Wait time (min)

(b) Mode split (%)

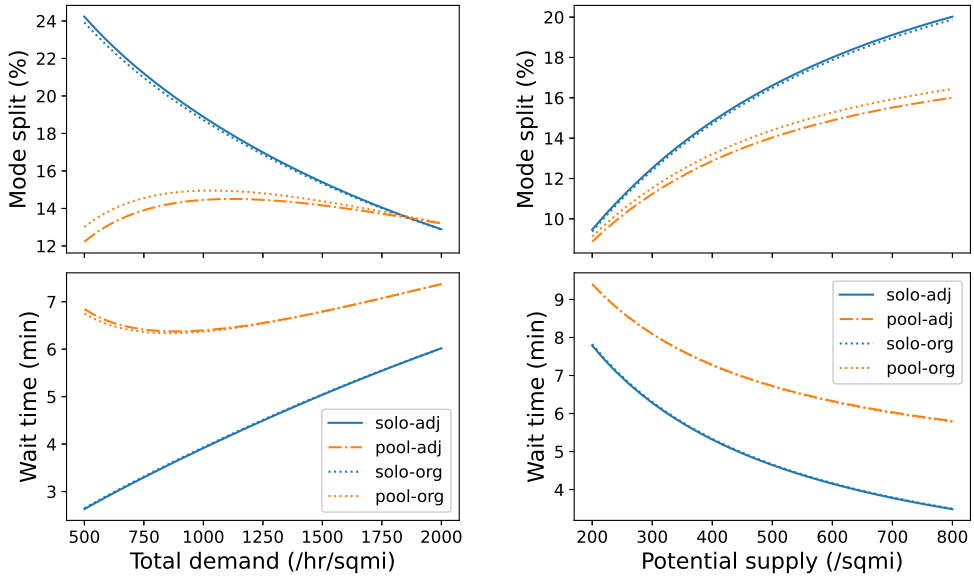
(c) Fleet size and vacant vehicle density (/sqmi)

(d) Pooing trip duration (min)

Figure A6: Differences in equilibrium outputs between the original model (with constant τ_p) and the adjusted model (with τ_p being specified by Eq.(A26)).

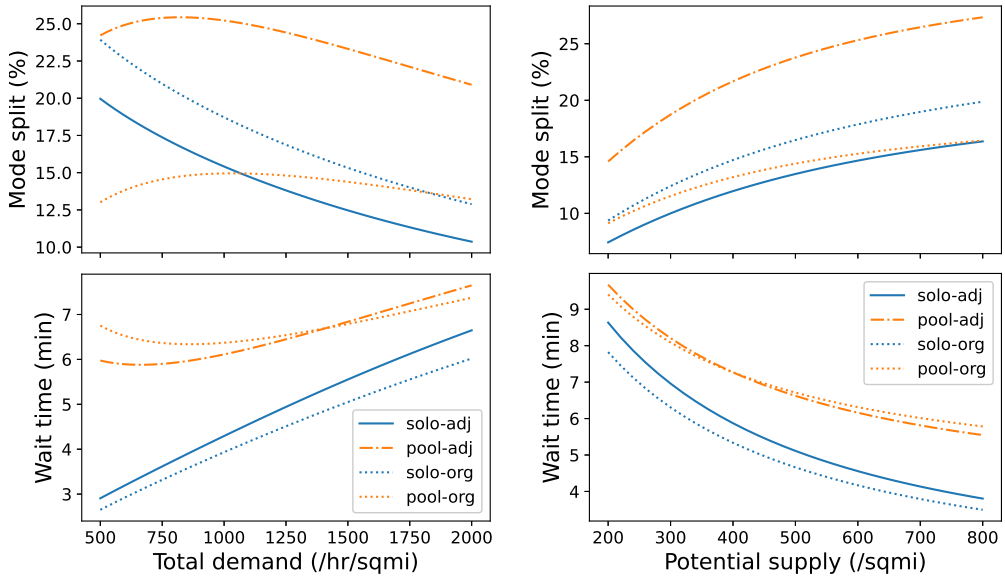
Chicago) well. Hence, without further empirical investigation, we cannot properly qualify how much our simplified model may deviate from reality. Also, the results suggest that the en-route detour is stable when the pooling demand is low. Therefore, assuming it as a constant seems reasonable, given that the pooling ratio in Chicago is mostly below 20% (see Figure A2).

Another difficulty that comes with the endogenous en-route detour is that it would turn the pooling demand into an implicit function, as Q_p appears on both sides of Eq. (13a). This would further complicate the derivation of analytical results.



(a) (b)

Figure A7: Sensitivity of market equilibrium to (a) total demand D_0 and (b) potential supply S_0 with $\alpha = 0.05, \beta = 2.25$.



(a) (b)

Figure A8: Sensitivity of market equilibrium to (a) total demand D_0 and (b) potential supply S_0 with $\alpha = 0.05, \beta = 3.5$.