Transportation Research Part C 127 (2021) 103088

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Regulating ridesourcing services with product differentiation and congestion externality

Daniel A. Vignon [a], Yafeng Yin [a,*], Jintao Ke [b]

[a] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI, USA
[b] Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Ridesourcing service
Regulation
Pooling
Congestion externality
Commission cap
Congestion toll

ABSTRACT

This paper proposes a model of the ridesourcing market in presence of traffic congestion and with the provision of both solo and pooling services. Our analysis of the first-best solution shows that, under a highly congested scenario, the ridesourcing platform may enjoy non-negative profits. However, when congestion is low, the ridesourcing market must be subsidized due to low marginal costs of operation. This mirrors previous findings in the traditional taxi literature. We also demonstrate that a commission cap on the solo service combined with a congestion toll (however small) on ridesourcing vehicles can induce any desired, sustainable equilibrium under the assumption of homogeneous value of travel time and sufficient supply of homogeneous drivers. Furthermore, numerical experiments suggest that the most important problem that a regulator should address when faced with a monopoly may not be that of congestion but rather that of market power. Indeed, when congestion is high, similar to previous findings in the literature, decisions by the monopolist tend to mirror that of the regulator. This is because customers on the platform must also bear the congestion cost, which hurts the platform's revenues. Additionally, numerical examples reveal that, even when accounting for heterogeneity in the value of travel time, the commission cap is able to achieve the second-best–whether combined with a toll or not. This confirms the effectiveness of the commission cap strategy illustrated in previous literature and provides decision makers with a simple, non-intrusive mechanism for regulating the market.

1. Introduction

In recent years, ridesourcing services such as Uber, Lyft, and Didi Chuxing have grown in popularity and accrued a stable user base. For example, Uber's quarterly earning reports show that between 2018 and 2019 the number of trips served by the platform grew by twenty-eight percent while the number of monthly users grew by twenty-two percent over the same period (Uber Inc., 2020). This growth in popularity can be attributed, among other things, to the convenience and expediency in booking rides and completing trips. This surge in popularity has, however, generated widespread criticism that ridesourcing services increase congestion, particularly in metropolitan cities like New York City and San Francisco.

Evidence that ridesourcing services are a major contributor to increased traffic congestion in cities has been surfacing in recent years. Using the National Household Travel Survey data, Schaller (2018) found that replacing a private vehicle trip with a ridesourcing trip at least doubles the number of miles traveled. Moreover, he also reported that only a little over 20% of trips taken with Uber or Lyft in New York City in February 2018 using their pooling service resulted in actual sharing. This puts into question the claim that ridesourcing

* Corresponding author.
E-mail address: yafeng@umich.edu (Y. Yin).

https://doi.org/10.1016/j.trc.2021.103088
Received 28 November 2020; Received in revised form 20 February 2021; Accepted 8 March 2021
Available online 20 April 2021

services might positively impact congestion by inducing sharing. After accounting for population, employment growth, and roadway modifications, Castiglione et al. (2018) and Erhardt et al. (2019) showed that ridesourcing companies are a major contributor to the drop in travel speed and the increase in vehicle miles travelled in San Francisco between 2010 and 2016. More recently, using exogenous variation provided by Uber and Ola drivers' strike in three major Indian cities, Agarwal et al. (2019) showed that the absence of ridesourcing drivers resulted in a reduction in delay nearly equivalent to half of that observed on major holidays.[1] Ridesourcing vehicles may, in certain instances, alleviate private car use, although available research suggests that a higher proportion of users switch from taxis or higher-occupancy modes, or would forego their trip altogether (Rayle et al., 2016; Clewlow and Mishra, 2017; Hampshire et al., 2017; Schaller, 2018). However, as compared with private car use, in serving their customers, ridesourcing vehicles generate massive vacant or empty trips. These vacant trips create additional vehicular traffic demand. Unless pooling plays a larger part and the fleet is efficiently managed, ridesourcing will likely worsen traffic conditions, at least in the short term (Beojone and Geroliminis, 2021; Wei et al., 2020).

Cities have already started to take steps to mitigate congestion caused by ridesourcing vehicles. For example, in 2019, New York City announced new regulations that impose a cap on new licenses issued to for-hire vehicles, mandate a minimum percent of time ridesourcing vehicles must carry a passenger while operating in Manhattan below 96th Street, and collect a congestion surcharge to trips that begin in, end in or pass through the area. While ridesourcing companies have generally been favorable to the congestion surcharge, some have criticized that the regulations hurt customers–especially low-income customers–and drivers by locking them out of the app at times of low demand (Bellon, 2019; Dobbs, 2019; Honan, 2019). This debate highlights the importance of understanding the welfare implications of these regulations.

In the literature, several studies have analyzed ridesourcing markets and investigated the welfare implications of potential regulations. Using an aggregate model of the ridesourcing market, Zha et al. (2016) demonstrated that, under the assumption of homogeneous value of time and labor supply, capping the commission that the ridesourcing firm takes on each ride is sufficient to achieve a second-best. Zha et al. (2018a,b) further examined the commission cap in a market with spatial or temporal heterogeneity and demonstrated that the commission cap can, in effect, significantly improve welfare. It appears that the cap may be imposed per trip, by unit distance, or time, and can even vary with respect to location or time of day. The choice of the granularity level will depend on the trade-off between implementation complexity and policy effectiveness. Ke et al. (2020a) compared the pricing equilibrium in two different types of ridesourcing market: a pooling market and a non-pooling market. They show that, at the monopoly, first-best, and second-best equilibria, both markets operate in an efficient regime (as opposed to the wild-goose-chases or WGC regime, first described by Castillo and Weyl (2018)). Additionally, in either equilibrium, the fare in the pooling market is lower than that in the non-pooling market. Their study does not, however, consider the joint operation of both services. Additionally, all the aforementioned studies do not consider traffic congestion. In fact, most of the analyses of regulations accounting for congestion externality in the for-hire vehicle market have focused on the taxi market. Yang et al. (2005) proposed a model of taxi service with a fare structure that implicitly incorporates congestion. Under the proposed framework, they showed that the first-best solution might be sustainable when congestion is high. Yang et al. (2014) extended this work by incorporating bilateral taxi-customer search frictions and showed that taxi demand may decrease with taxi fleet size in the presence of congestion externality. Moreover, under the assumption of constant returns to scale for the matching function, taxi utilization rates decrease along the Pareto frontier from the system optimal solution to the monopolist solution. Albeit insightful, these results should be re-examined against the ridesourcing market due to its distinctive features, such as its two-sided nature, matching technology and workforce flexibility, etc. A few recent studies seek to address the question of congestion and regulation in the ridesourcing market. Li et al. (2019) investigated different regulations in the ridesourcing market: the minimum wage, the congestion fee, and the driver cap. Their analyses show that a minimum wage can actually result in higher welfare for consumers and drivers but lower profits for the firm. In contrast to Li et al. (2019),Zhang and Nie (2019) modelled both the solo and pooling options offered by the ridesourcing platform. This allows them to capture the trade-offs between maximizing vehicle occupancy and mitigating congestion on the one hand, and ensuring that drivers are incentivized to provide service on the other. However, both analyses of congestion regulation fall short by not incorporating the mechanism through which ridesourcing vehicles affect traffic congestion. In contrast, Ke et al. (2020b) incorporated such a mechanism using the concept of macroscopic fundamental diagram (Geroliminis and Daganzo, 2008). They are able to show that pooling can, under certain circumstances and with an appropriate pairing time window, reduce the total travel cost experienced by ridesourcing customers. Their analysis assumes, however, that pooling does not coexist with solo rides on the same platform. Thus, in the present work, we propose a stylized framework that captures the workings of a ridesourcing market, its two different services, i.e., riding alone or sharing with someone else, and their effects on traffic congestion. We then derive optimal solutions to the monopoly and social-optimum problems, and analyze the impact of a commission cap and a toll on system performance and social welfare.

This paper is organized as follows. In Section 2, we present our main assumptions and our model. Section 3 analyzes the monopolist and first-best solutions. Section 4 discusses our proposed mechanism to regulate the market and Section 5 illustrates the results through numerical examples.

2. Model

Consider a ridesourcing market with one platform offering two types of services: *solo* rides (denoted by s) and *pooling* rides (p), which customers choose to use based on their preferences. However, drivers are required to provide both services, and the assignment of customers will be made by the platform via a matching algorithm. In addition, the platform decides the fares F^s and F^p that customers

[1] These results, however, represent short-term effects and do not necessarily reflect longer term patterns resulting from a stoppage in ridesourcing services.

will pay to use each service as well as the amount that drivers will receive for delivering customers. Thus, by setting the fares and drivers' share, the platform essentially determines the demand rates $\widetilde{\lambda}^s$ and $\widetilde{\lambda}^p$ for the solo and pooling services.

2.1. Pairing and matching

In this paper, we consider a stylized matching process that is consistent in spirit with the matching algorithms implemented by some ridesourcing platforms in practice. To avoid confusion, we use *pairing* to refer to the process by which a pooling customer is grouped with other customers with whom they will split their ride. We use *matching* to refer to the process by which both solo and pooling customers are assigned to their drivers.

We consider a pairing process where pooling customers wait for a pairing time window to be matched to other pooling customer. Customers are paired if they depart and arrive within a given pooling radius of each other. Once a target occupancy (the maximum number of customers to be paired) is reached, the pairing process will be terminated and the pooling customers will be subsequently matched with a driver. If the target occupancy is not achieved at the end of the pairing period, currently pooled customers will be assigned to a driver. It implies that if no other customer is found, the waiting customer will ride alone. With this consideration, given the target occupancy, pooling radius and pairing time window exogenously determined by the platform, the average pairing time w^p experienced by a pooling customer is given by:

$$w^p = W^p(\widetilde{\lambda}^p)$$

with $W^{p'} < 0$, suggesting that an increase in pooling demand leads to a decrease in pairing time. Correspondingly, the relationship between average occupancy o and the pooling demand can then be described by another function:

$$o = O(\widetilde{\lambda}^p)$$

where $O' > 0$.

On the other hand, we assume, as in Castillo and Weyl (2018), that the matching time for customers is zero, owing to instantaneous matching and a sufficiently large matching radius. In other words, customers are matched to drivers as they arrive in the matching queue. Pooling customers enter the matching queue either as a unit (with the other customers they will ride with) or alone (if suitable pooling partners were not found before the expiration of the pairing time window).

2.2. Meeting and delivery

Upon being matched to a driver, solo customers experience an expected pickup time w^m and an expected delivery time w^r. In addition to w^m and w^r, pooling customers will also experience an expected detour time Δw that stems from picking up and dropping off their pooling partners. We have:

$$w^r = \frac{d^r}{v}$$
$$w^m = \frac{d^m}{v}$$
$$\Delta w = \frac{\Delta d}{v}$$

where v is an average traffic speed; d^r is the average trip distance for ridesourcing trips, which is assumed to be given; d^m is the average pickup distance and is a decreasing function of the density of idle drivers N^I i.e., $d^m = D^m(N^I)$ and $D^{m'} < 0$; and Δd is the expected detour distance. For a given pooling radius, we assume that the detour distance is a decreasing function of pooling demand as follows:

$$\Delta d = \Delta D(\widetilde{\lambda}^p)$$

with $\Delta D' < 0$. The exact functional form of ΔD will necessarily depend on the operational decisions of the platform as shown, for example, in Daganzo (1978).

Following the network macroscopic fundamental diagram approach (Geroliminis and Daganzo, 2008), it is possible to describe the average traffic speed v using the number of ridesourcing vehicles N and the number of background vehicles N^b in the network:

$$v = V\left(\theta \cdot N + N^b\right)$$

with $V'(\cdot) < 0$. The parameter $\theta \geqslant 1$ reflects higher marginal effect of ridesourcing vehicles on congestion compared to regular background vehicles, because, for example, ridesourcing vehicles often drive slower as they await their next assignment.

Assuming that background traffic trips originate at a rate λ^b with an average travel distance of d^{rb}, we have:

$$w^{rb} = \frac{d^{rb}}{v}$$
$$N^b = \lambda^b \cdot w^{rb}$$

where w^{rb} represents the average travel time of background traffic. The second equation holds as per Little's law. We further note that while it is straightforward to consider λ^b to be congestion-dependent, we treat it to be exogenous for simplicity and clarity.

2.3. Demand

Customers face the following costs from taking a trip on the platform:

$$\begin{aligned}\mu^s &= F^s + \beta \cdot (w^m + w^r) - \xi^s \\ \mu^p &= \frac{F^p}{o} + \beta \cdot \left(w^p + w^m + w^r + \Delta w\right)\end{aligned}$$

where F^s and F^p denote the total fare collected by the platform for completing a transaction for the solo and pooling services, respectively; β represents customers' value of time; and $\xi^s > 0$ is a constant, reflecting the fact that, all else equal, riding alone provides higher utility to customers. Additionally, customers also have access to an outside option (such as public transit or driving) with cost μ^0. We further assume that customers' value of travel time β is a variable across the customer population with cumulative density function $G(\cdot)$. Thus, each customer chooses which service to use by comparing the costs μ^s, μ^p and μ^0 and choosing the cheapest option. Thus, we have:

1. if $\mu^s \leqslant \mu^p$ and $\mu^s \leqslant \mu^0$, then customers with $\beta_2 \leqslant \beta \leqslant \beta_1$ choose the solo service. The proportion of these customers is $G(\beta_1) - G(\beta_2)$.
2. if $\mu^p \leqslant \mu^0$ and $\mu^p < \mu^s$, then customers with $\beta \leqslant \beta_3 = \min\left\{\beta_2, \frac{\mu^0 - \frac{F^p}{o}}{(w^p + w^m + w^r + \Delta w)}\right\}$ choose the pooling service, whose proportion is $G(\beta_3)$.

Here:

$$\begin{aligned}\beta_1 &= \frac{\mu^0 - F^s + \xi^s}{w^m + w^r} \\ \beta_2 &= \frac{F^s - \frac{F^p}{o} - \xi^s}{w^p + \Delta w}\end{aligned}$$

If we let λ^0 denote the population size, then demand for each service is given by:

$$\begin{aligned}\widetilde{\lambda}^s &= \lambda^0 \cdot [G(\beta_1) - G(\beta_2)] \\ \widetilde{\lambda}^p &= \lambda^0 \cdot G(\beta_3)\end{aligned}$$

2.4. Supply

We further assume that drivers decide to provide service if their average hourly earnings during the study period ω exceed their opportunity cost. That is, the supply of drivers is given by:

$$N = S(\omega)$$

where $S'(\cdot) > 0$ and $S(\cdot)$ captures the distribution of drivers' opportunity cost. The above relationship between driver supply and driver earnings can be further simplified if one considers the inverse supply function $S^{-1}(\cdot)$:

$$C(N) = S^{-1}(N) \cdot N = \omega \cdot N$$

where $C' > 0$ by construction. $C(\cdot)$ can be understood as the cost for the platform of using N drivers. Now, let e denote the amount that drivers receive per unit service time.[2] Then:

$$\omega = e \cdot \frac{w^r \cdot \left(\widetilde{\lambda}^s + \frac{\widetilde{\lambda}^p}{o}\right)}{N}$$

2.5. Equilibrium

At equilibrium, we consider a steady state in the system where the following conservation equation holds as per Little's law:

[2] Here, note that we assume that drivers do not get paid for the extra detour time induced by pooling. Our model could easily be extended to include such a possibility. In this formulation, we however seek to capture the reality that, for many drivers, they work more and earn almost the same (if not less) when serving pooling trips as opposed to solo trips.

$$N = N^l + \left(\widetilde{\lambda}^s + \frac{\widetilde{\lambda}^p}{o}\right) \cdot \left(w^m + w^r\right) + \frac{\widetilde{\lambda}^p}{o} \cdot \Delta w^d$$

where Δw^d is the additional time that drivers spend on pickup and delivery for a pooling ride relative to a solo ride. As pointed out in Ke et al. (2020a), Δw^d and Δw are correlated, though this correlation, once again, depends on operational decisions from the platform. To simplify our analysis, we also assume that $\Delta w^d = \gamma \cdot \Delta w$.

All of the above considerations yield our model of Eqs. (1a) -(1o), which is a system of 14 equations and 17 unknowns. By specifying exogenous variables F^s, F^p, and e, we can solve the system to evaluate the performance of the ridesourcing market at the steady state.

$$\widetilde{\lambda}^s = \lambda^0 \cdot [G(\beta_1) - G(\beta_2)] \tag{1a}$$

$$\widetilde{\lambda}^p = \lambda^0 \cdot G(\beta_3) \tag{1b}$$

$$\beta_1 = \frac{\mu^0 - F^s + \xi}{w^m + w^r} \tag{1c}$$

$$\beta_2 = \frac{F^s - \frac{F^p}{o} - \xi}{w^p + \Delta w} \tag{1d}$$

$$\beta_3 = \min\left\{\beta_2, \frac{\mu^0 - \frac{F^p}{o}}{(w^p + w^m + w^r + \Delta w)}\right\} \tag{1e}$$

$$o = O(\widetilde{\lambda}^p) \tag{1f}$$

$$w^p = W^p(\widetilde{\lambda}^p) \tag{1g}$$

$$w^m = \frac{D^m(N^l)}{v} \tag{1h}$$

$$w^r = \frac{d^r}{v} \tag{1i}$$

$$\Delta w = \frac{\Delta D(\widetilde{\lambda}^p)}{v} \tag{1j}$$

$$w^{rb} = \frac{d^{rb}}{v} \tag{1k}$$

$$v = V(\theta \cdot N + N^b) \tag{1l}$$

$$N = N^l + \left(\widetilde{\lambda}^s + \frac{\widetilde{\lambda}^p}{o}\right) \cdot \left(w^m + w^r\right) + \frac{\widetilde{\lambda}^p}{o} \cdot \gamma \cdot \Delta w \tag{1m}$$

$$e \cdot \left[w^r \cdot \left(\widetilde{\lambda}^s + \frac{\widetilde{\lambda}^p}{o}\right)\right] = C\left(N\right) \tag{1n}$$

$$N^b = \lambda^b \cdot w^{rb} \tag{1o}$$

3. Scenario analysis

Below we analyze scenarios that determine the choice of exogenous variables F^s, F^p, and e given our ridesourcing market model presented above. In order to analyze the market and derive policy insights, both in this section and Section 4, we will make a few assumptions and specifications to simplify the model. Later in Section 5, we will conduct numerical experiments to examine the effectiveness of the policies derived in the general case considered in the previous section.

3.1. Model simplification

We first assume that customers are homogeneous in their value of time β. With homogeneous value of time, the demand-side equilibrium can be described by the following set of equations:

$$\widetilde{\lambda}^s + \widetilde{\lambda}^p = \Lambda(\mu)$$
$$\mu^s - \mu \geqslant 0 \text{ and } \widetilde{\lambda}^s \geqslant 0$$
$$\mu^p - \mu \geqslant 0 \text{ and } \widetilde{\lambda}^p \geqslant 0$$
$$\widetilde{\lambda}^s \cdot [\mu^s - \mu] = 0$$
$$\widetilde{\lambda}^p \cdot [\mu^p - \mu] = 0$$

where $\Lambda(\cdot)$ is a demand function for ridesourcing services and is such that $\Lambda' < 0$. The above complementarity condition indicates that, at equilibrium, the cost of services with non-zero demand must be equal, and less than that of non-utilized services.

We also assume, as in Korolko et al. (2018), that pooling customers must walk to and from common meeting and drop-off locations. As such, $\Delta d = 0$. This scenario can be likened to the operation of Uber Express Pool, Uber's low-cost service. In this case, the driver compensation becomes the same for solo and pool services and thus it is convenient to consider the choice of $R = e \cdot w^r$, the compensation per ride, rather than that of e. In light of the above, the ridesourcing market model becomes:

$$\mu^s = F^s + \beta \cdot (w^m + w^r) - \xi^s \tag{2a}$$

$$\mu^p = \frac{F^p}{o} + \beta \cdot \left(w^p + w^m + w^r\right) \tag{2b}$$

$$\widetilde{\lambda}^s + \widetilde{\lambda}^p = \Lambda(\mu) \tag{2c}$$

$$\mu^s - \mu \geqslant 0 \tag{2d}$$

$$\mu^p - \mu \geqslant 0 \tag{2e}$$

$$\widetilde{\lambda}^s \cdot [\mu^s - \mu] = 0 \tag{2f}$$

$$\widetilde{\lambda}^p \cdot [\mu^p - \mu] = 0 \tag{2g}$$

$$o = O(\widetilde{\lambda}^p) \tag{2h}$$

$$w^p = W^p(\widetilde{\lambda}^p) \tag{2i}$$

$$w^m = \frac{D^m(N^l)}{v} \tag{2j}$$

$$w^r = \frac{d^r}{v} \tag{2k}$$

$$w^{rb} = \frac{d^{rb}}{v} \tag{2l}$$

$$v = V(\theta \cdot N + N^b) \tag{2m}$$

$$N = N^l + \left(\widetilde{\lambda}^s + \frac{\widetilde{\lambda}^p}{o}\right) \cdot \left(w^m + w^r\right) \tag{2n}$$

$$R \cdot \left(\widetilde{\lambda}^s + \frac{\widetilde{\lambda}^p}{o}\right) = C\left(N\right) \tag{2o}$$

$$N^b = \lambda^b \cdot w^{rb} \tag{2p}$$

In our analyses, we assume that the above modeling system defines continuously differentiable functions between endogenous variables and exogenous variables F^s,F^p, and R, as per the implicit function theorem. To facilitate the presentation of our analysis results, we define the *vehicle* trip rates for solo and pool services as follows:

$$\lambda^s = \widetilde{\lambda}^s$$
$$\lambda^p = \frac{\widetilde{\lambda}^p}{o}$$

To conclude our model presentation, below we present a few results that highlight some useful properties of our model and that will be used in analyzing our different scenarios.

From Eq. (2n), we can obtain the following derivative for the total vehicle trip rate $\lambda^s + \lambda^p$ with respect to the number of vacant

vehicles:

$$\frac{\partial(\lambda^s+\lambda^p)}{\partial N^I} = -\frac{1+\left(\lambda^s+\lambda^p\right)\cdot\frac{D^{m'}}{V}}{w^m+w^r} \tag{3}$$

When $\frac{1+\left(\lambda^s+\lambda^p\right)\cdot\frac{D^{m'}}{V}}{w^m+w^r} < 0$, then $\frac{\partial(\lambda^s+\lambda^p)}{\partial N^I} > 0$, which corresponds to the WGC described in Castillo and Weyl (2018).

From Eqs. (2l), (2m), (2p), the derivative of the total traffic $N^T = N + N^b$ with respect to the fleet size is:

$$\frac{\partial N^T}{\partial N} = 1 - \frac{\theta\cdot d^{rb}\cdot\lambda^b\cdot V'}{V^2+\lambda^b\cdot d^{rb}\cdot V'} \tag{4}$$

Given our assumption of inelastic background traffic demand, it must be that $\frac{\partial N^T}{\partial N} > 0$ so that $\frac{\theta\cdot d^{rb}\cdot\lambda^b\cdot V'}{V^2+\lambda^b\cdot d^{rb}\cdot V'} < 1$. If $V^2+\lambda^b\cdot d^{rb}\cdot V' < 0$, then it follows that $V^2 < (\theta-1)\cdot\lambda^b\cdot d^{rb}\cdot V' < 0$ (given that $\theta\geqslant 1$), which is absurd. Thus, it follows that when background traffic demand is inelastic, we must have $V^2+\lambda^b\cdot d^{rb}\cdot V' > 0$.

3.2. First-best

Here, a social planner seeks to maximize social welfare by solving the following optimization problem:

$$W = \max_{\substack{F^s\geqslant 0,\\ F^p\geqslant 0,\\ R\geqslant 0}} \overbrace{\int_\mu^\infty \Lambda\left(x\right)\cdot dx}^{\text{Consumers' surplus}} + \overbrace{(F^s-R)\cdot\lambda^s + (F^p-R)\cdot\lambda^p}^{\text{Platform profit}} + \overbrace{\int_0^{\frac{R\cdot(\lambda^s+\lambda^p)}{N}} S\left(x\right)\cdot dx}^{\text{Drivers' surplus}} - \overbrace{\gamma^b\cdot\lambda^b\cdot w^{rb}}^{\text{Background cost}} \tag{FB}$$

where γ^b is the value of time for the background traffic.[3] Note that the dependent variables $\lambda^s, \lambda^p, N, w^{rb}$ and μ are functions of the decision variables through the system defined in Eq. (2). Without loss of generality, we do not consider the platform's operation cost.

Assuming that $\lambda^s > 0$ and $\lambda^p > 0$, the first-order optimality conditions (FONC) of the problem yield the following formulae for the price of a solo and a pool ride and for drivers' hourly income:

$$F^s = mc \tag{5a}$$

$$\frac{F^p}{o} = mc\cdot C^o + \beta\cdot W^{p'}\cdot\lambda^p\cdot o \tag{5b}$$

$$\frac{C(N)}{N} = \frac{mc}{w^m+w^r} - \tau^{int} - \tau^b \tag{5c}$$

where:

$$mc = -\beta\cdot\frac{D^{m'}}{V}\cdot\frac{(w^m+w^r)}{1+\left(\lambda^s+\lambda^p\right)\cdot\frac{D^{m'}}{V}}\cdot\left(\lambda^s+\lambda^p\cdot o\right)$$

$$C^o = \frac{1-\lambda^p\cdot O'}{o} = \frac{O'}{\frac{\partial o}{\partial\lambda^p}} < \frac{1}{o}$$

$$\tau^b = -\theta\cdot\gamma^b\cdot\frac{d^{rb}\cdot\lambda^b\cdot V'}{V^2+\lambda^b\cdot d^{rb}\cdot V'} > 0$$

$$\tau^{int} = -\theta\cdot\beta\cdot\frac{V'}{V^2+\lambda^b\cdot d^{rb}\cdot V'}\cdot\frac{(d^m+d^r)}{1+\left(\lambda^s+\lambda^p\right)\cdot\frac{D^{m'}}{V}}\cdot\left(\lambda^s+\lambda^p\cdot o\right) > 0$$

Here, mc represents the marginal cost, to the platform, of providing a ride. Since $F^s > 0$, it follows from Eq. (5a) that, at the social optimum, $mc > 0$, which implies that $1+\left(\lambda^s+\lambda^p\right)\cdot\frac{D^{m'}}{V} > 0$. Thus, at the first-best, the equilibrium of the ridesourcing market lies in

[3] Truly, this is a *quasi first-best* since we take λ^b as given. Under the first-best, the planner would also be able to determine λ^b by considering the demand function for background traffic.

the non-WGC regime. Additionally, we note that Eq. (5b) contains the negative term $\beta \cdot \lambda^p \cdot o \cdot W^{p'}$, which captures the matching externality that each additional pooling customer creates on the platform. It then follows that, since $F^p > 0, C^o > 0$ at the first-best equilibrium. Thus, increasing the pooling *vehicle* trip rate increases occupancy for the pooling service ($\frac{\partial o}{\partial \lambda^p} > 0$). By combining Eqs. (2a), (2b), (5a), (5b), we obtain:

$$mc \cdot \left(1 - C^o\right) = \beta \cdot W^p \cdot \left(1 + \lambda^p \cdot o \cdot \frac{W^{p'}}{W^p}\right) + \xi^s \tag{6}$$

If we assume that the mode-specific constant $\xi^s \approx 0$, then $1 + \lambda^p \cdot o \cdot \frac{W^{p'}}{W^p} > 0$ from which it follows that the solution to the first-best problem lies in the non-elastic portion of the pairing time function for the pooling service. When $\xi^s > 0$, this condition is relaxed and the market outcome may lie in the elastic region of the pairing time function. Here, the disutility of using the pooling service is so large that the resulting equilibrium number of pooled rides is relatively low. It is also evident from Eq. (5) that $F^s - \frac{F^p}{o} > 0$ and $F^s - F^p > 0$. In other words, the fare per customer and the revenue per trip for the single service are higher than their respective counterparts for the pooling service at the first-best.

From Eq. (5c), the average driver hourly income is equalized with the drivers' marginal social benefit. This marginal benefit is composed of three main components:

- the benefit of a marginal driver to the platform $\frac{mc}{w^m + w^r}$, which captures drivers' impact on meeting distance;
- the intra-platform congestion externality τ^{int} that a marginal driver imposes on ridesourcing customers;
- the extra-platform congestion externality τ^b that a marginal driver imposes on the background traffic.

From this latest point, we notice that the externality component is independent of occupancy. This suggests that, irrespective of the number of passengers they carry, ridesourcing vehicles impose the same externality on background traffic.

We can also rewrite Eq. (5) as follows:

$$F^s = \left[\frac{C(N)}{N} + \tau^{int} + \tau^b\right] \cdot \left(w^m + w^r\right) \tag{7a}$$

$$\frac{F^p}{o} = \left[\frac{C(N)}{N} + \tau^{int} + \tau^b\right] \cdot \left(w^m + w^r\right) \cdot C^o + \beta \cdot W^{p'} \cdot \lambda^p \cdot o \tag{7b}$$

Thus, at the first-best, the fare only covers the marginal social cost of vehicles that are in pickup or delivery modes. In particular, it does not cover the marginal cost of drivers when they are idle. Now, from Arnott (1996) and Yang et al. (2005), we know that unless congestion is high, the taxi market must be subsidized at the first-best. Additionally, from Zha et al. (2016), without considering congestion, the ridesourcing market must be subsidized when the production function of rides is increasing returns to scale. We examine here whether, in the presence of pooling and congestion, the ridesourcing market is sustainable at the first-best. Consider the first-best profits π^f:

$$\pi^f = \overbrace{-\frac{mc}{w^m + w^r} \cdot N^l + \left(\beta \cdot W^{p'} \cdot \lambda^p \cdot o^2 - mc \cdot O^l\right) \cdot \lambda^p}^{\text{Operating loss} \leqslant 0} + \left(\tau^b + \tau^{int}\right) \cdot N \tag{8}$$

From examining Eq. (8), we notice two main components. On one hand, there is a congestion-independent component that is negative and captures the industry's losses. Those include the cost of idle drivers–which is not covered by the fare at the first-best (Arnott, 1996; Yang et al., 2005)–and a pooling-service-related cost. This latter cost emerges from the positive matching externality that the pooling service enjoys and the fact that pooling reduces the marginal impact of riders on the platform but not the marginal cost of drivers relative to the single service. On the other hand, there is a congestion related component, $\left(\tau^b + \tau^{int}\right) \cdot N$, which is positive. When congestion externality is high, this term may exceed the operating loss, which would result in positive profits for the platform. In other words, similar to the taxi market investigated by Yang et al. (2005), the first-best is sustainable for the ridesourcing platform when congestion externality is high.

3.3. Monopoly

In this section, we derive the monopoly equilibrium, its properties, and the distortions that arise relative to the social optimum. In this setting, the ridesourcing platform determines exogenous variables F^s, F^p, and R to maximize its profits. Thus its revenue-maximizing decision can be obtained by solving the following optimization problem:

$$\pi = \max_{\substack{F^s \geq 0,\\ F^p \geq 0,\\ R \geq 0}} \left(F^s - R \right) \cdot \lambda^s + \left(F^p - R \right) \cdot \lambda^p \tag{M}$$

The FONC of the optimization problem are given below:

$$F^s = -\frac{\lambda^s + \lambda^p \cdot o}{\Lambda'} + mc \tag{9a}$$

$$\frac{F^p}{o} = -\frac{\lambda^s + \lambda^p \cdot o}{\Lambda'} + mc \cdot C^o + \beta \cdot \lambda^p \cdot o \cdot W^{p'} \tag{9b}$$

$$C'\left(N\right) = \frac{mc}{w^m + w^r} - \tau^{int} \tag{9c}$$

Compared to first-best pricing, F^s and F^p include a markup term $-\frac{\lambda^s + \lambda^p \cdot o}{\Lambda'} > 0$ under monopoly pricing. Thus, fares under the monopolist are higher than under the first-best. Moreover, Eq. (9c) indicates that, at optimality, the platform equalizes the cost of the marginal driver, $C'(N)$, to its marginal benefit. Thus, while the regulator is concerned with the average cost of the fleet, the platform is only concerned with the cost of the marginal driver. Additionally, by comparing the right-hand sides of Eqs. (5c) and (9c), it appears that the monopolist only internalizes part of the congestion externality that arises from running the platform. Indeed, under monopoly pricing, customers only bear the congestion cost they impose on each other but do not bear the cost of congestion on the background traffic. Taken together, it follows that, for a convex driver cost function $C(\cdot)$, assuming equal marginal driver and marginal externality costs, the number of drivers under the monopolist is higher than the first-best number of drivers.[4] Since demand is lower under the monopolist, it follows that utilization rates are also lower under the monopolist, a finding similar to that of Ke et al. (2020a) when considering pooling and solo services separately.

Now, the question arises as to whether it may be optimal for the monopolist to operate in the WGC-regime ($mc < 0$) when accounting for congestion. Indeed, unlike in the first-best case, it is not straightforward to rule out a WGC solution at optimality for the monopolist. It is, however, easy to show that for any solution with $mc < 0$, we can construct another solution with identical N,λ^s, and λ^p but higher fares F^s and $\frac{F^p}{o}$. Thus, though they may exist, solutions with $mc < 0$ are dominated, even in the presence of congestion. This is consistent with the findings of Zha et al. (2018b) in a spatial market without congestion.

Lastly, while $F^s > \frac{F^p}{o}$ and $F^s > F^p$ in the first-best, it is not necessary that $F^s > F^p$ under the monopoly. Indeed:

$$F^s - F^p = -\frac{\lambda^s + \lambda^p \cdot o}{\Lambda'} \cdot \left(1 - o\right) + mc \cdot \left(1 - C^o \cdot o\right) - \beta \cdot \lambda^p \cdot o^2 \cdot W^{p'} \tag{10}$$

Since $o \geqslant 1$, the markup per ride that the platform collects is higher for the pooling service. When occupancy is high and the markup per passenger is significantly higher than the difference in marginal costs between the solo and the pooling service, the monopolist may earn more revenue from the pooling service than from the solo service. In practice, however, such a situation does not seem to occur: numerous reports suggest that the prices required to make the pooling service profitable tend to induce a demand lower than that necessary to achieve high occupancy (Spotwood, 2017; An, 2020). Thus, for the rest of our analysis, we will assume that $F^s > F^p$ at the monopoly equilibrium.

4. Policy discussion

Before we dive into the analysis of potential policies, it is important to characterize an efficient policy. First, such a policy must target the two sides of the market, i.e., it must address the monopolist's market power on the demand and supply sides. Second, it must address congestion externality by ensuring that drivers (and by extension the passengers) bear the social cost of the congestion they impose on the background traffic. Lastly, the policy should be easy to implement.

We seek for an optimal policy with which the monopoly problem admits the first-best solution. Suppose that the fares $\widehat{F}^s$ and $\widehat{F}^p$ and the per-ride driver revenue $\widehat{R}$ solve the above first-best problem. An obvious policy would be to regulate the fares and the earnings to be at this first-best level. Such regulations, however, may be unpopular as they would restrict the operational freedom and flexibility of the platform. Since the modeling system of the ridesourcing market, Eqs. (2a)–(2c), (2l)–(2p), enjoys three degrees of freedom (14

[4] This can be seen by noting that:

- C and C' are both increasing in N and that the right hand side of Eq. (5c) is lower than that of Eq. (9c);
- $C'(N) \geqslant \frac{C(N)}{N}$ for convex functions.

equations and 17 unknowns), a policy that can ensure that any three variables stay at their first-best level will be optimal, if the reduced system admits a unique solution. However, this implies that regulating fares, driver per-ride earnings, or the fleet size alone cannot induce the first-best; it is necessary to explore a combination of regulatory instruments.

4.1. New York City's regulatory scheme

We first briefly analyze NYC's approach to regulating the for-hire vehicle market to determine whether it meets the aforementioned criteria. We preface our discussion by noting that NYC's ridesourcing market operates with multiple companies while our setting only considers a single firm. Thus, we do not comment on the effectiveness of NYC's policy as it pertains to NYC's current market, but rather on the effectiveness of that policy when applied to a market similar to the one in our setting. In 2019, New York City announced new regulations that impose a cap on new licenses issued for for-hire vehicles, mandate a minimum percent of time ridesourcing vehicles must carry a passenger while operating in Manhattan below 96th Street, and collect a congestion surcharge on trips that begin in, end in or pass through the area. In December 2018, NYC additionally implemented an effective minimum wage requirement of 15 $\frac{\$}{\text{hr}}$ for ridesourcing drivers. We investigate here whether such policies are effective in our setting, where, under these regulations, the problem for the monopolist would be as follows:

$$\begin{aligned}
\max_{\substack{\mathbf{F}\geqslant 0,\\ \omega\geqslant 0,\\ N\geqslant 0}} \quad & \left(F^s - R - \widehat{\tau}^s\right)\cdot\lambda^s + \left(F^p - R - \widehat{\tau}^p\right)\cdot\lambda^p \\
\text{s.t.} \quad & (\lambda^s+\lambda^p)\cdot w^r \geqslant \widehat{\rho}\cdot N \quad \text{(Occupied time constraint)} \\
& N \leqslant \widehat{N} \quad \left(\text{Fleet size constraint}\right) \\
& C(N) \leqslant \widehat{\omega}\cdot N \quad \text{(Minimum wage constraint)}
\end{aligned} \tag{11}$$

where $\widehat{\tau}^s$ and $\widehat{\tau}^p$ are the congestion surcharges on single and pooled trips respectively; $\widehat{\rho}$ is the first-best utilization rate for the ridesourcing fleet; $\widehat{N}$ is the first-best vehicle fleet cap; and $\widehat{\omega}$ is the wage under the first-best.

First, we note that the fleet size constraint and the minimum wage constraint cannot simultaneously be binding at equilibrium. Indeed, the fleet size constraint is only necessary and effective when the congestion externality imposed on the background traffic exceeds the monopolist's market power, so that the monopoly wage is higher than the socially efficient wage. On the other hand, the minimum wage constraint is only effective and necessary when congestion is low and the firm's market power leads to wages lower than socially efficient. We analyze below the two cases.

Suppose that the fleet size constraint is binding so that $N = \widehat{N}$ and $C(\widehat{N}) = \widehat{\omega}\cdot\widehat{N}$. Then, $w^r = \frac{d^r}{V(\widehat{N}+\widetilde{N}^s)}$, i.e., the monopoly and first-best travel times are equal and the monopolist solves the following problem:

$$\begin{aligned}
\max_{\mathbf{F}\geqslant 0} \quad & (F^s - R - \widehat{\tau}^s)\cdot\lambda^s + (F^p - R - \widehat{\tau}^p)\cdot\lambda^p \\
\text{s.t.} \quad & \lambda^s+\lambda^p \geqslant \widehat{\lambda}
\end{aligned} \tag{12}$$

where $\widehat{\lambda} = \widehat{\rho}\cdot\frac{\widehat{N}}{\widehat{w}^r} = \widehat{\lambda}^s + \widehat{\lambda}^p$. Essentially, the occupied time constraint becomes a minimum on the number of trips served by the platform. Thus, it essentially acts as a way to maximize the demand served by the monopolist. However, the fleet size cap renders the congestion management effect of the congestion surcharge unnecessary. Moreover, the occupied time constraint might be redundant. Indeed, following Xu et al. (2017), when congestion is high, the objectives of the platform and of the planner tend to be aligned, so that the platform inherently seeks to maximize its fleet utilization rate.

Now consider the case in which congestion is low and the monopolist's driver supply is lower than that targeted by the regulator. Then, we consider two situations of interest. In the first, $\widehat{N}$ is such that $C(\widehat{N}) = \widehat{\omega}\cdot\widehat{N}$. Then, the monopolist solves the following problem:

$$\begin{aligned}
\max_{\mathbf{F}\geqslant 0, N\geqslant 0} \quad & (F^s - R - \widehat{\tau}^s)\cdot\lambda^s + (F^p - R - \widehat{\tau}^p)\cdot\lambda^p \\
\text{s.t.} \quad & \left(\lambda^s+\lambda^p\right)\cdot w^r \geqslant \widehat{\rho}\cdot\widehat{N} \quad \left(\text{Occupied time constraint}\right)
\end{aligned} \tag{13}$$

This is identical to Eq. (12). Here again, the surcharge is unnecessary (since congestion is not a problem). Moreover, it likely hampers the effectiveness of the minimum trip requirement, since higher prices might discourage consumers from using the service. In the second case, the platform's optimal choice of N is such that $C(N) \leqslant \widehat{\omega}\cdot\widehat{N}$: it is more advantageous for the platform to hire fewer drivers. In such situation, not only would the number of drivers be suboptimal, but the rationing mechanism used by the platform could further decrease welfare.

Our brief analysis shows that, while NYC's regulatory scheme might be effective in mitigating the congestive effect of ride-sourcing

vehicles, it is unnecessarily burdensome and might not always improve welfare in our setting. In the following section, we propose another solution that not only remedies that issue but is also more parsimonious.

4.2. Commission cap regulation and congestion toll

Zha et al. (2016) showed that, when customers are homogeneous in their value of time, regulating the amount of commission that the platform receives can achieve a second-best. Consider such regulation applied to our current framework. First, we note that, given a commission cap on the solo service corresponding to the first-best commission $\widehat{P}^s = \widehat{F}^s - \widehat{R}$, it must be that the firm's choice of pooling commission is such that $P^p \leqslant \widehat{P}^s$.[5] Thus, there exists a natural, non-binding cap on the pooling service commission. It might therefore be possible to regulate the market with a single commission cap. Second, we note that Eqs. (5c), differ by τ^b. *A priori*, a policy that increases the cost of drivers by τ^b should be enough to address the congestion externality. Thus, for the regulated monopoly, Eq. (2o) becomes:

$$\begin{aligned} R = & \frac{C(N)}{\lambda^s + \lambda^p} + \tau^b \cdot \left(w^I + w^m + w^r\right) \\ = & \left[S^{-1}(N) + \tau^b\right] \cdot \left(w^I + w^m + w^r\right) \end{aligned}$$

with $w^I = \frac{N^I}{\lambda^s + \lambda^p}$. We consider then the following regulated problem for the monopolist:

$$\begin{aligned} \pi = \max_{\substack{F^s \geqslant 0, \\ F^p \geqslant 0, \\ R \geqslant 0}} \quad & (F^s - R) \cdot \lambda^s + (F^p - R) \cdot \lambda^p \\ \text{s.t.} \quad & F^s - R \leqslant \widehat{P}^s \quad \left(\text{Single service cap}\right) \end{aligned} \tag{M-CAPT}$$

The FONC of (M-CAPT) satisfy:

$$\begin{aligned} F^s = & -\left[\lambda^s - \nu_1 + o \cdot \lambda^p\right] \cdot \frac{1}{\Lambda'} + \left[C'\left(N\right) + \tau^b + \tau^{int}\right] \cdot \left(w^m + w^r\right) + \left[S^{-1}(N) + \tau^b\right] \cdot w^I \cdot \frac{\nu_1}{\lambda^s + \lambda^p} - \\ & \tau^{int} \cdot \left(w^m + w^r\right) \cdot \frac{\nu_1}{\lambda^s + \lambda^p \cdot o} \end{aligned} \tag{14a}$$

$$\begin{aligned} \frac{F^p}{o} = & -\frac{\lambda^s - \nu_1 + \lambda^p \cdot o}{\Lambda'} + C^o \cdot \left[\left[C'\left(N\right) + \tau^b + \tau^{int}\right] \cdot \left(w^m + w^r\right) + \left[S^{-1}(N) + \tau^b\right] \cdot w^I \cdot \frac{\nu_1}{\lambda^s + \lambda^p} \right. \\ & \left. - \tau^{int} \cdot \left(w^m + w^r\right) \cdot \frac{\nu_1}{\lambda^s + \lambda^p \cdot o}\right] + \beta \cdot \lambda^p \cdot o \cdot W^{p'} \end{aligned} \tag{14b}$$

where $\nu_1 \geqslant 0$ is the Lagrangian multiplier associated with the commission cap. Now consider a constructed, constrained social welfare maximization problem below:

$$\begin{aligned} W = \max_{\substack{F^s \geqslant 0, \\ F^p \geqslant 0, \\ R \geqslant 0}} \quad & \int_{\mu}^{\infty} \Lambda\left(x\right) \cdot dx + \left(F^s - R\right) \cdot \lambda^s + \left(F^p - R\right) \cdot \lambda^p - \gamma^b \cdot \lambda^b \cdot w^{rb} \\ \text{s.t.} \quad & F^s - R \geqslant \widehat{P}^s \end{aligned} \tag{SB}$$

It is straightforward to see that the first-best solution solves this constrained social welfare maximization problem. Now, the FONC of SB yield:

$$\begin{aligned} F^s = & -\frac{\eta_1}{\Lambda'} + \left[\frac{C(N)}{N} + \tau^b + \tau^{int}\right] \cdot \left(w^m + w^r\right) - S^{-1}\left(N\right) \cdot w^I \cdot \frac{\eta_1}{\lambda^s + \lambda^p} + \\ & \tau^{int} \cdot \left(w^m + w^r\right) \cdot \frac{\eta_1}{\lambda^s + \lambda^p \cdot o} \end{aligned} \tag{15a}$$

$$\begin{aligned} \frac{F^p}{o} = & -\frac{\eta_1}{\Lambda'} + C^o \cdot \left[\left[\frac{C(N)}{N} + \tau^b + \tau^{int}\right] \cdot \left(w^m + w^r\right) - S^{-1}\left(N\right) \cdot w^I \cdot \frac{\eta_1}{\lambda^s + \lambda^p} + \right. \\ & \left. \tau^{int} \cdot \frac{\eta_1}{\lambda^s + \lambda^p \cdot o} \cdot \left(w^m + w^r\right)\right] + \beta \cdot \lambda^p \cdot o \cdot W^{p'} \end{aligned} \tag{15b}$$

[5] This is because the driver revenue per ride R is identical for both services. Since $F^s > F^p$, the conclusion follows.

where $\eta_1 \geqslant 0$ is the Lagrangian multiplier associated with the commission cap.

Now, assuming the first-best is attained by the regulation, does there exist $\nu_1 \geqslant 0$ so that Eq. (14) holds? By analyzing the system of Eqs. (14) and (15), the answer to this question can be reduced to determining the conditions under which:

$$\nu_1 = \left(\lambda^s + \lambda^p \cdot o\right) \cdot \frac{\left[\tau^{int} + \frac{C(N)}{N} - C'\left(N\right)\right] \cdot \left(w^m + w^r\right) \cdot \bar{\lambda} - S^{-1}\left(N\right) \cdot w^I}{\tau^b \cdot w^I} \geqslant 0 \tag{16}$$

where $\bar{\lambda} = \frac{\lambda^s + \lambda^p}{\lambda^s + \lambda^p \cdot o}$. It then follows that the first-best can be replicated with a cap and toll if:

$$\bar{\pi} = \tau^{int} \cdot \left(w^m + w^r\right) \cdot \bar{\lambda} - S^{-1}\left(N\right) \cdot w^I \geqslant \left[C'\left(N\right) - \frac{C(N)}{N}\right] \cdot \left(w^m + w^r\right) \cdot \bar{\lambda} \tag{17}$$

Now, $\tau^{int} \cdot (w^m + w^r) \cdot \bar{\lambda}$ represents the platform's revenue per customer (solo and pooled) served; $\left[C'\left(N\right) - \frac{C(N)}{N}\right] \cdot (w^m + w^r) \cdot \bar{\lambda}$ represents the opportunity cost of equalizing drivers' benefit to their average cost rather than their marginal cost; and $S^{-1}(N) \cdot w^I$ represents the cost of an idle vehicle per customer served. Thus, Eq. (17) simply indicates that the first-best can be replicated when the platform's revenues are sufficiently high to cover the economic cost of its drivers. Can this condition be satisfied, regardless of the composition of the driver pool?

When drivers are homogeneous,[6] this is a slightly weaker condition than requiring that the platform's profits be positive since, as Eq. (8) indicates, profits include a pooling-service-related cost that does not appear in Eq. (17). Thus, under the assumption of homogeneous drivers, as long as the first-best is sustainable, the proposed regulation will be effective.

Additionally, Eq. (16), does not depend on η_1. Thus, any desired second-best equilibrium can be replicated by a regulation with a cap $\widehat{P}^s > 0$ and an appropriate toll. This is especially important when the first-best is not sustainable ($\widehat{P}^s < 0$) and the cap must be increased beyond its first-best level to ensure platform operation.

When drivers are heterogeneous, assuming the cost function $C(\cdot)$ is convex, then the right hand side of Eq. (17) is positive. Then, that the first-best is sustainable may or may not be sufficient to guarantee that the policy can replicate the regulator's objective.

In practice, collecting a toll on ridesourcing vehicles can be challenging, especially if there is no preexisting mechanism to toll other vehicles. However, imposing a congestion fee per use time on each rider is more easily implementable (cities and states already impose multiple fees on riders). From Eq. (14), we can easily deduce that the appropriate fee structure per rider is as follows:

$$f^s = \tau^b \tag{18}$$

$$f^p = \tau^b \cdot C^o \tag{19}$$

where f^s and f^p denote the fee imposed on solo and pooled riders, respectively. We note that $f^s > f^p$ since $C^o < \frac{1}{o} < 1$. Additionally, $\frac{f^s}{f^p} = \frac{1}{C^o}$ so that the ratio between the single fee and the pooling fee is not linear in occupancy. It will rather depend on the extent of congestion and its costs on society.

4.3. Commission cap only

Zha et al. (2016) explained that the commission cap incentivizes the monopolist to serve a higher demand than it otherwise would. When congestion is taken into account, maximizing demand served might involve maximizing occupancy for the platform, thus alleviating congestion and achieving the regulator's objective. Thus, a toll might not be needed once a commission cap is imposed.

To analyze such a regulation, it is convenient to introduce $\widehat{\tau}^N \in [0, \tau^b]$, the toll imposed by the regulator on each ridesourcing vehicle. Then, Eq. (14) becomes:

$$F^s = -\left[\lambda^s - \nu_1 + o \cdot \lambda^p\right] \cdot \frac{1}{\Lambda'} + \left[C'\left(N\right) + \tau^{int} + \widehat{\tau}^N\right] \cdot \left(w^m + w^r\right) + \left[S^{-1}\left(N\right) + \widehat{\tau}^N\right] \cdot w^I \cdot \frac{\nu_1}{\lambda^s + \lambda^p} - \tau^{int} \cdot \frac{\nu_1}{\lambda^s + \lambda^p \cdot o} \cdot \left(w^m + w^r\right) \tag{20a}$$

$$\frac{F^p}{o} = -\frac{\lambda^s - \nu_1 + \lambda^p \cdot o}{\Lambda'} + \left[\left[C'\left(N\right) + \tau^{int} + \widehat{\tau}^N\right] \cdot \left(w^m + w^r\right) + \left[S^{-1}\left(N\right) + \widehat{\tau}^N\right] \cdot w^I \cdot \frac{\nu_1}{\lambda^s + \lambda^p} - \tau^{int} \cdot \frac{\nu_1}{\lambda^s + \lambda^p \cdot o} \cdot \left(w^m + w^r\right)\right] \cdot C^o + \beta \cdot \lambda^p \cdot o \cdot W^{p'} \tag{20b}$$

Then, a solution to the monopoly problem that satisfies Eq. (15) can be obtained if:

[6] That is, $C'(N) = \frac{C(N)}{N} = c$ and $S^{-1}(N) = c$.

$$\nu_1 = \left(\lambda^s + \lambda^p \cdot o\right) \cdot \frac{\left[\tau^{int} + \frac{C(N)}{N} - C'\left(N\right) + \tau^b - \widehat{\tau}^N\right] \cdot \left(w^m + w^r\right) \cdot \bar{\lambda} - S^{-1}\left(N\right) \cdot w^l}{\widehat{\tau}^N \cdot w^l} \geqslant 0 \tag{21}$$

From Eq. (21), we note the following:

- When $\widehat{\tau}^N = \tau^b$, we recover the commission cap and toll policy from Section 4.2;
- For any toll $\widehat{\tau}^N \in (0, \tau^b)$, the first-best can be replicated if:

$$\bar{\pi} = \left(\tau^{int} + \tau^b - \widehat{\tau}^N\right) \cdot \left(w^m + w^r\right) \cdot \bar{\lambda} - S^{-1}\left(N\right) \cdot w^l \geqslant \left[C'\left(N\right) - \frac{C(N)}{N}\right] \cdot \left(w^m + w^r\right) \cdot \bar{\lambda} \tag{22}$$

  Here, $(\tau^b - \widehat{\tau}^N) \cdot (w^m + w^r) \cdot \bar{\lambda} > 0$ represents the additional revenue for the platform due to bearing only a fraction of the congestion externality it imposes on the background traffic. The interpretation of Eq. (22) is then similar to that of Eq. (17).
- Using our approach, it is not possible to recover ν_1 when $\widehat{\tau}^N = 0$. However, we note that, as $\widehat{\tau}^N \to 0, \nu_1 \to \infty$ i.e., ν_1 becomes more and more positive. This suggests that it is always possible to find a smaller toll $\widehat{\tau}^N$ such that the first-best (or another targeted equilibrium) can be replicated.

From this last point above, it appears that the selection of $\widehat{\tau}^N$ does not have a significant impact on the ability of the cap to bring the system to an efficient equilibrium. In fact, the commission cap alone might be able to achieve the desired effect. We will verify this intuition using numerical examples in Section 5. In practice, the regulating authority might still, however, choose $\widehat{\tau}^N > 0$ in order to meet other objectives: improving road infrastructure, satisfying special interests, etc.

5. Numerical experiments

Our proposed policies in Section 4 assume that users are identical in their value of time and that pooling users experience no detour time. In this section, we relax this assumption and apply our proposed policies to the full model described in Eq. (1). In this context, the social welfare maximization problem now becomes:

$$\begin{aligned} W = \max_{\substack{F^s \geqslant 0,\, F^p \geqslant 0 \\ R \geqslant 0}} & \overbrace{U^0 \cdot \left(\lambda^s + \lambda^p \cdot o\right)}^{\text{Total trip utility}} - \overbrace{\lambda^0 \cdot \int_{\beta_2}^{\beta_1} \beta \cdot \left(w^m + w^r\right) \cdot G'\left(\beta\right) \cdot d\beta}^{\text{Solo time cost}} \\ & \overbrace{-\lambda^0 \cdot \int_{\underline{\beta}}^{\beta_3} \beta \cdot \left(w^p + w^m + w^r + \Delta w\right) \cdot G'\left(\beta\right) \cdot d\beta}^{\text{Pooled time cost}} \\ & \underbrace{-R \cdot \left(\lambda^s + \lambda^p\right)}_{\text{Driver cost}} - \underbrace{\gamma^b \cdot \lambda^b \cdot w^{rb}}_{\text{Background traffic}} \end{aligned} \tag{FB}$$

where $\underline{\beta}$ is the lower bound of the support of G; and U^0 is the utility of completing the trip for an individual customer. Unlike in the homogeneous value of time case, however, analyzing the above system analytically for policy insights is substantially difficult. Thus, we turn instead to numerical experiments, for which we adopt the following functional forms:

$$D^m\left(N^l\right) = A \cdot N^{l\,-\alpha} \tag{23a}$$

$$C(N) = c \cdot N \tag{23b}$$

$$W^p\left(\widetilde{\lambda}^p\right) = \frac{1 - \exp(-\delta \cdot \phi \cdot \widetilde{\lambda}^p)}{\delta \cdot \widetilde{\lambda}^p} \tag{23c}$$

$$O(\widetilde{\lambda}^p) = 2 - \exp(\delta \cdot \phi \cdot \widetilde{\lambda}^p) \tag{23d}$$

$$\Delta D\left(\widetilde{\lambda}^p\right) = \frac{B}{\widetilde{\lambda}^p} \tag{23e}$$

$$V\left(\theta \cdot N + N^b\right) = V^0 - V^c \cdot \left(\theta \cdot N + N^b\right) \tag{23f}$$

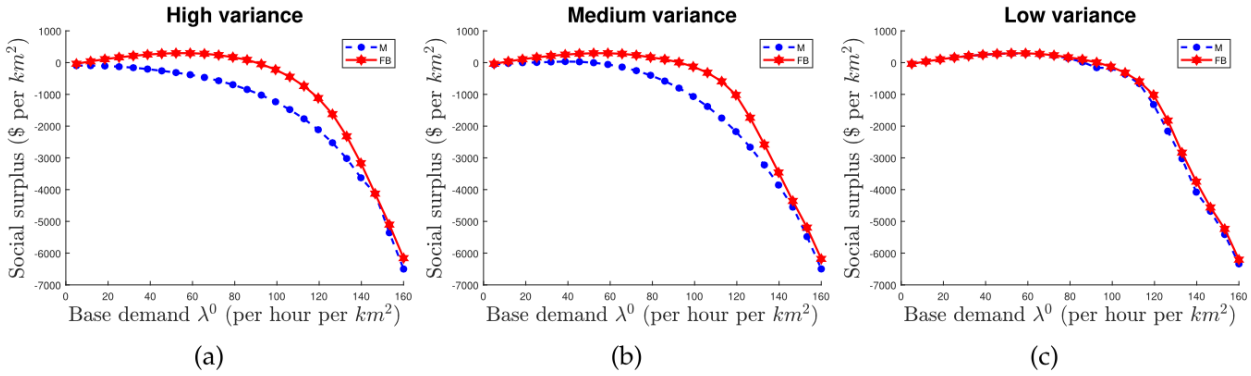

Fig. 1. Social surplus comparisons.

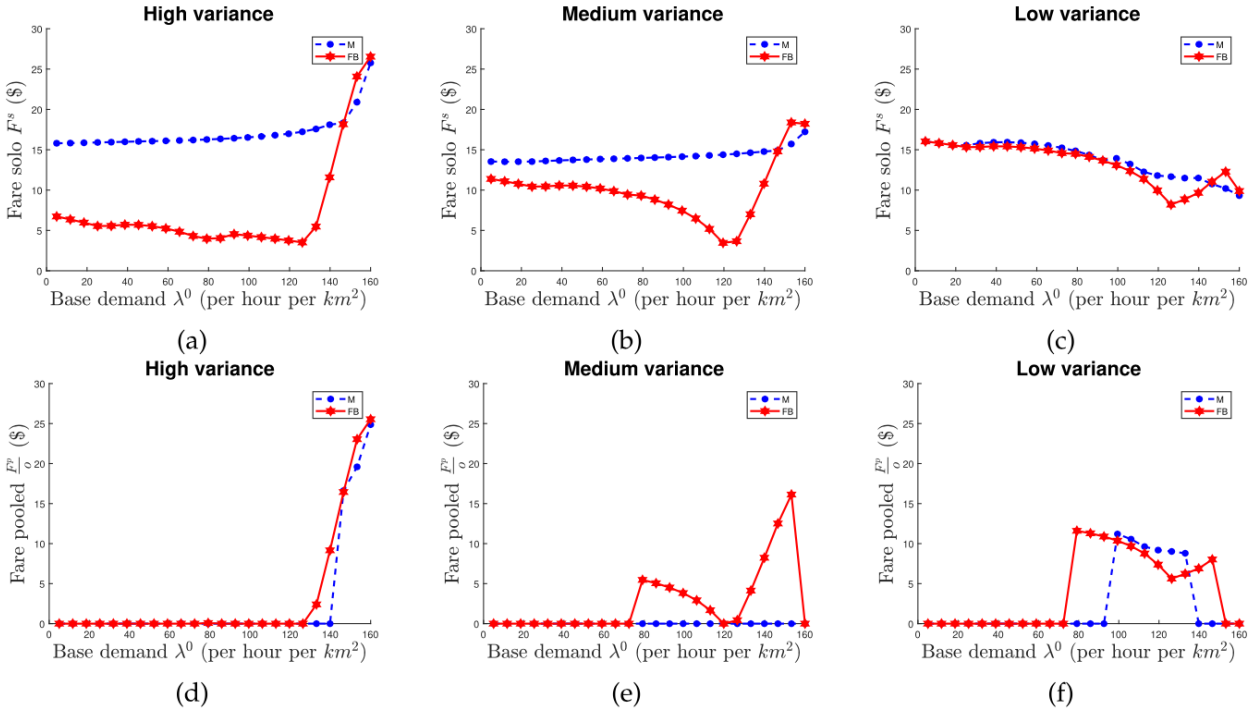

Fig. 2. Fare comparisons.

$$G\left(\beta\right) = \frac{\beta - \underline{\beta}}{\overline{\beta} - \underline{\beta}} \tag{23g}$$

$$C^o = \frac{1 - \exp(-\delta \cdot \phi \cdot \widetilde{\lambda}^p) \cdot \delta \cdot \phi \cdot \widetilde{\lambda}^p}{o} \tag{23h}$$

We assume, as in Korolko et al. (2018), that the maximum occupancy is two and that the probability of being paired is constant. Then, the resulting pairing time and occupancy functions are given in Eqs. (23c). Furthermore, we assume that β follows a uniform distribution on $\left(\underline{\beta}, \overline{\beta}\right)$. Eq. (23a) is a classical representation of the pickup distance between a randomly chosen driver and their closest unmatched customer (Zha et al., 2018b; Korolko et al., 2018; Castillo, 2018). We also borrow from Korolko et al. (2018) for information on δ. The data and parameters used are presented in Appendix A.

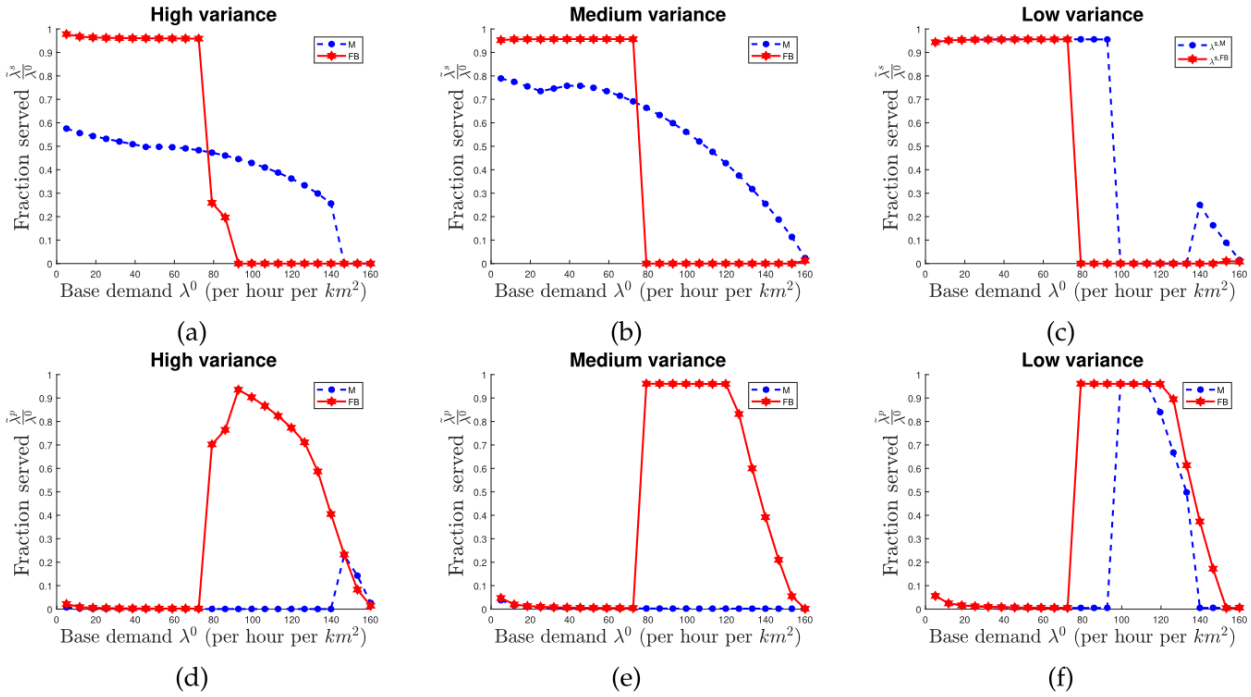

Fig. 3. Demand rate comparisons.

5.1. *Comparative analysis*

In this section, we compare the first-best and monopoly outcomes. To tease out the effects moving away from the homogeneous value of time assumption, we also vary the variance of the value of time distribution and investigate how it affects both outcomes.[7] Especially, we consider the producer and social surpluses, the distribution of rides across the two services and the congestion effect.

As Fig. 1 shows, the monopolist's behavior results in an inefficient outcome. This inefficient outcome is, however, not driven by congestion, since, as shown in Fig. 4, traffic speed in the unregulated scenario is similar to that in the first-best scenario when congestion is high. Thus, most inefficiencies are the results of the monopolist's exercise of market power, as shown by the fares in Fig. 2. As the variance of the value of time distribution increases, so does the difference between the *laissez-faire* outcome and the first-best outcome. This is due to the fact that, when variance is high, the monopolist can easily maximize profits by catering to high value of time customers. However, as variance decreases, the ridesourcing service must become less niche to survive and maximize its profits. This is best seen in Fig. 3: as variance in the population decreases, total demand served (across both services) increases. Such a pattern can also easily be understood in terms of price elasticity: as variance in the population increases, the price elasticity decreases, thus making it easier for the firm to charge higher fares (Fig. 2).

As far as the trip distribution is concerned, Fig. 3 shows that, at the first-best, we can discern two regimes. In the first, as base demand increases, the fraction of demand served by the solo service is non-increasing while that served by the pooling service increases. This is because, when demand is low, frictions due to pooling cannot be easily overcome. However, as demand increases, economies of scale lead to a reduction in pooling and detour time costs, thus making pooling the more efficient option to serve the demand. In the second regime, as congestion increases, pooling (and the ridesourcing service overall) become less desirable and the service is eventually eliminated. Under the monopolist, the share of solo trips is non-increasing, just as in the first-best case. However, when variance is high, pooling is only provided when congestion is very high: because it manages less demand than under the first-best, the monopolist is able to maintain a good quality of service for its high value of time customers longer while getting adequately compensated. When variance is low, the distribution pattern under the monopolist becomes similar to that under the first-best, with a caveat: when congestion is very high, the monopolist still provides the solo service, since it does not have to bear its full contribution to congestion.

Interestingly, regardless of the value of time distribution, beyond a certain point, as the base demand increases, the lesser the discrepancy between the first-best and the monopoly. This is because, similar to the findings of Xu et al. (2017), the interests of the platform and the planner become more aligned. As demand rises, both actors are looking to manage their fleet more efficiently in order to serve the growing demand. Thus, both the fleet size and the utilization rates increase (Fig. 4). However, as congestion becomes more

[7] The effect of the mean of the distribution was also studied. However, the findings were as expected and do not add much regulatory insights to what can be gleaned from studying the variance. Higher value of time implies lower profits for the platform since it needs to hire a larger number of drivers. Lower value of time implies that pooling becomes preferable and that profits are easier to generate.

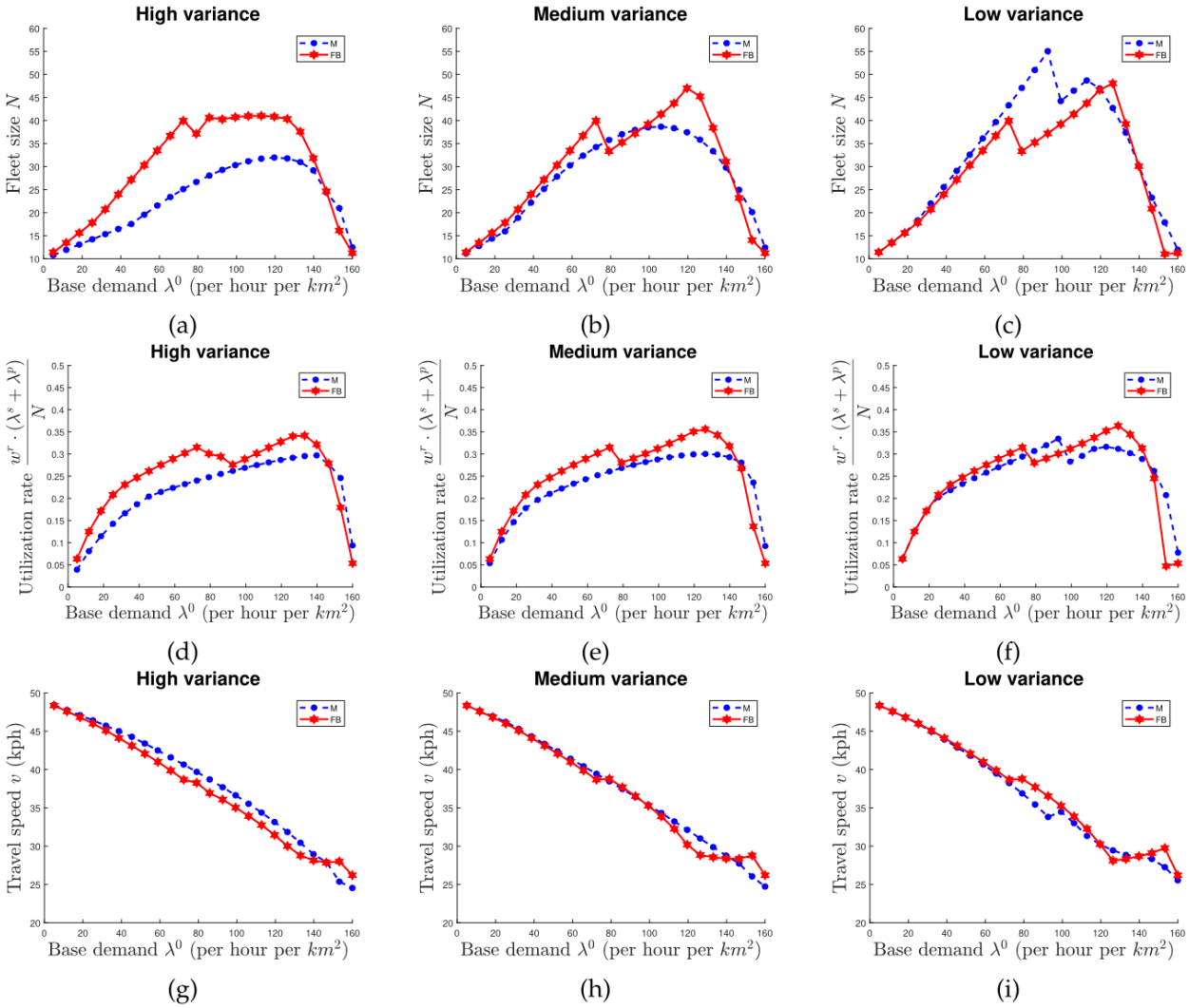

Fig. 4. Fleet size, utilization rate and traffic speed comparisons.

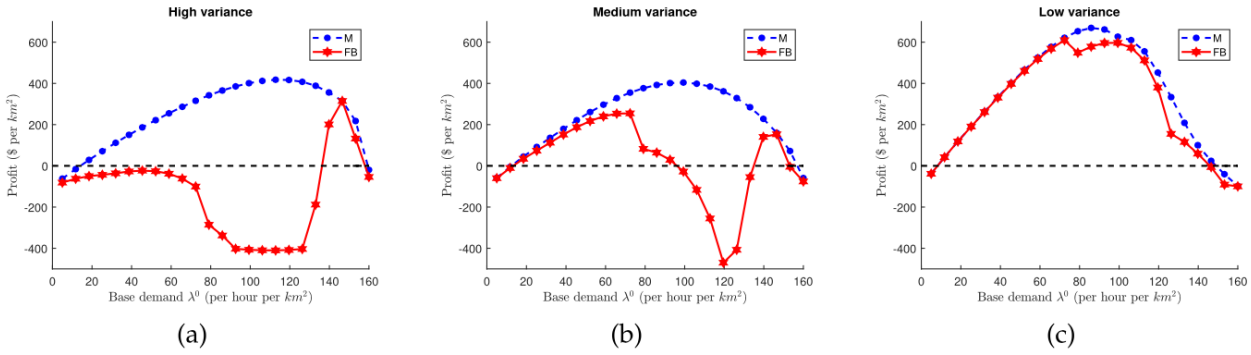

Fig. 5. Profit comparisons.

problematic, both under the first-best and the monopoly, the number of vehicles and their utilization rate decrease (Fig. 4). This is the result of two forces. Firstly, falling pooling demand served increases detour times. Secondly, the reduction in traffic speed increases pickup and detour times. Both forces thus contribute to reducing the attractiveness of the ridesourcing service while increasing the time drivers spend without customers on board.

Lastly, we consider the ridesourcing company's profits. As predicted in our analysis, under the first-best, the ridesourcing platform becomes sustainable in high congestion regimes (Fig. 5). Moreover, as variance in the value of time decreases, the range under which the service is sustainable increases. Intuitively, when variance is high, welfare can be increased mainly through expanding access, so

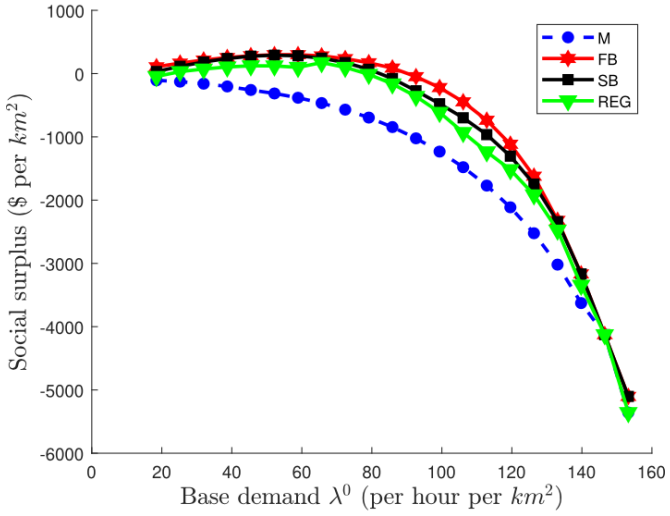

Fig. 6. Comparison of regulation results to first-best, second-best and monopoly.

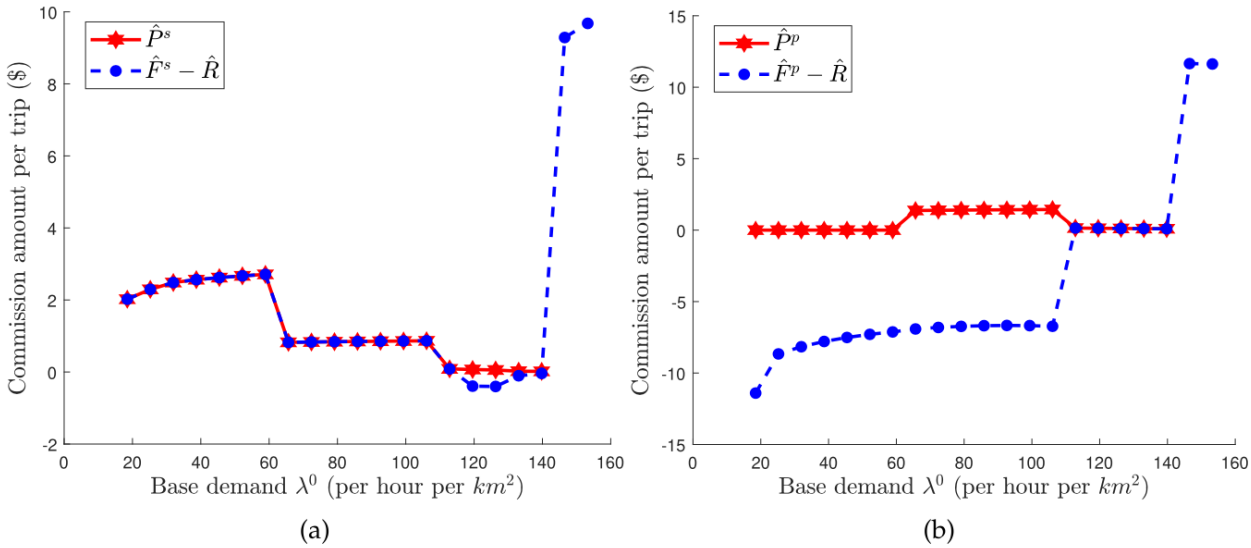

Fig. 7. Regulatory and realized commissions.

that effects of scale dominate. In that context, average costs and fares decrease with increasing demand, leading to a net loss for the platform. However, when variance is low, demand increases play a lesser role, since if one customers uses the platform, most of them likely also do. The main gains to be made are from efficiency in fleet and demand management, which accrue to the platform under the form of fares.

These numerical examples have a few implications. First, intervention by the planner to reduce congestion might not be necessary, since, as congestion increases, the platform acts similarly to the planner. Second, significant welfare gains can only be made when heterogeneity in the market served by the platform is high. Then, the main contribution from the planner is to increase demand served by expanding access to lower value of time customers. In that context, there may be limited gains to be made from focusing on ridesourcing-induced congestion, as pointed out by Tarduno (2021). Rather, applying commission caps with limited tolling should be the preferred regulatory strategy.

5.2. *Effects of proposed policies*

In order to evaluate whether our proposed policies can improve welfare relative to the monopoly solution, we can only consider cases under which the ride-sourcing market is sustainable under the first-best. Therefore, if the first-best is sustainable, it will be our regulatory target. Otherwise, we settle for a second-best in which the monopolist makes some profit. From the previous numerical examples, it is clear that when regulation is needed (i.e. in the high and medium variance cases), one will often have to settle for the second-best. Most importantly, we must determine how the choice of the caps P^s and P^p and of the toll τ^N is made. Our analysis in

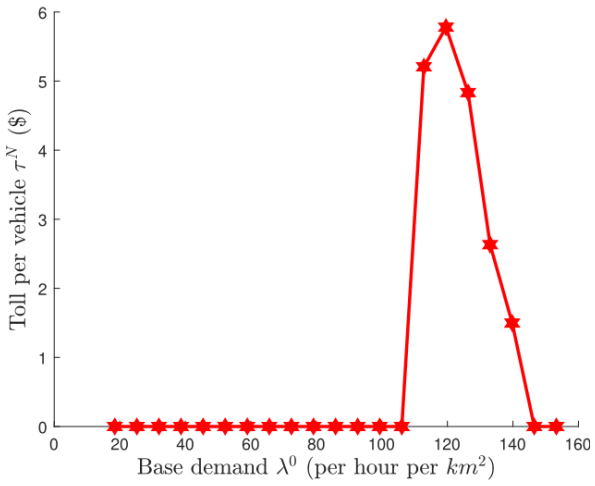

Fig. 8. Implemented toll.

Section 4 assumed a homogeneous value of time. When the population is heterogeneous in the value of time, selecting the optimal caps and tolls can be modeled as a bilevel program in which the social planner is at the upper-level and seeks to maximize welfare by choosing P^s, P^p and τ^N subject to a profit constraint. We present the problem and our solution method in detail in Appendix B.

Upon implementation, our strategy is able to substantially improve welfare (Fig. 6) and results in an outcome close to the second-best. The results suggest that significant welfare gains can be realized from our policy, especially when the welfare gap is significant.

Fig. 7 shows our chosen caps ($\widehat{P}^s$ and $\widehat{P}^p$) as well as the commissions implemented by the platform ($\widehat{F}^s - \widehat{R}$ and $\widehat{F}^p - \widehat{R}$) in response to the regulation. It is evident that, at most demand levels considered, only one cap is required for the solo service. However, in the highly congested regimes, the regulation is not needed, since, as discussed in Section 5.1, the monopolist's behavior aligns with the regulator's objective. It is interesting to note that, as base demand increases, the cap for the solo service is reduced while that for the pooling service increases.

Finally, Fig. 8 shows the regulatory tolls. We note that the tolls first increase with congestion but then decrease. This is in keeping with the fact that the monopolist's objective becomes closer to that of the regulator when congestion becomes very high.

6. Conclusion

In this paper, we present a model of the ridesourcing market with the presence of congestion externality, and the integration of the solo and pooling services. In order to derive analytical insights, we then consider a simplified version of our model with a homogeneous value of travel time. Analyzing the market equilibrium under both the first-best and the monopoly, we show that:

- under a socially optimal equilibrium and similar to the taxi market, the ridesourcing market may be sustainable when congestion is high;
- a monopoly platform internalizes part of the congestion externality its drivers impose but still employs larger number of vehicles than is socially efficient;
- a regulation coupling a single commission cap and a congestion toll (however small) can replicate any sustainable equilibrium when customers are homogeneous in their value of time;
- in the case that the collection of a toll is impractical, we derive a set of congestion fees to be collected directly from customers of each service. Interestingly, the ratio between the fee for the solo service and the pooling service does not vary linearly with occupancy.

We also briefly apply New York's congestion mitigation policies to our setting and show that, compared to our proposed regulations, they are redundant. In some cases, this redundancy could also potentially lead to inefficiencies. In order to understand how our policy performs in the more realistic setting of heterogeneous value of time, we perform numerical experiments. Our numerical results show that regulatory intervention is only warranted when the population is highly heterogeneous. In those circumstances, however, the main source of inefficiency may not necessarily be congestion and there are limited welfare gains to be made by focusing on that issue. Rather, maximizing demand served would be the best strategy for the regulator. We confirm our intuition by solving a Stackelberg game to choose optimal regulatory caps and tolls. These examples reveal that, when the welfare gap between the unregulated market and the first-best is the highest, it is optimal not to impose a toll but, rather, to impose low commission caps–with lowest caps on the solo service. Moreover, when tolls are applied, their value should decrease with the level of congestion, since the monopolist

naturally aligns with the regulator in these cases. Thus, since imposing commission caps is more parsimonious and justifiable with regards to addressing inefficiencies, we favor that approach to regulating the ridesourcing market under a monopoly. However, the regulator might still choose to impose a toll on traffic as a whole, rather than singling out ridesourcing vehicles.

While our assumption of a monopolistic market might make sense in certain contexts,[8] most other markets feature two or more companies competing for customers and drivers. Thus, a few questions may arise in that context. As shown in our analysis of the first-best, marginal cost pricing is not sustainable, except in highly congested instances. Thus, for more than one firm to subsist in a long-term equilibrium, competition must result in a non-efficient pricing pattern and/or significant product differentiation. Whether welfare will be higher than in the monopolist setting and closer to the first-best is, however, unclear. Zha et al. (2016) showed that, in a duopoly setting, welfare might be lower than in the monopoly setting when matching frictions are high or if market size is too small. The inclusion of congestion will likely increase matching frictions, though the extent to which this will degrade profitability is unclear. Additionally, since congestion enhances profitability, one might reasonably ask whether competition for drivers–driven by profit seeking–might worsen traffic conditions at peak times in a manner that is welfare degrading. In-depth analysis of these aspects is left to future work.

Customer characteristics indubitably vary across geographical locations. This leads to more opportunities for product differentiation but also raises the question of the fairness of regulations such as congestion pricing. Indeed, Uber has argued that ridesourcing regulations in NYC disproportionately hurt customers from low-income neighborhoods and from areas with poor transit access who disproportionately use their pooling service (Dobbs, 2019). Thus, taking into account these unintended effects might provide an opportunity for spatially differentiated tolling/subsidy strategies that we will investigate. Moreover, such a setting provides an opportunity to analyze the fleet management behavior of the platform as it must contend with the opportunity cost of providing service in one neighborhood as opposed to another. Here again, supplementing our work with an empirical basis might be appropriate and will be done in future work.

Ridesourcing companies have also been touting their potential to aid and complement public transit, thus making it more accessible. In select cities, for example, users are able to see public transit options alongside UberX and UberPool in the Uber app. Additionally, rides beginning or ending near public transit stations in select cities are now subsidized in an effort to address the first- and last-mile problem. This provides additional opportunities for congestion mitigation and increasing the share of pooled rides on the platform, but also creates additional modeling challenges.

Lastly, we have also taken the background traffic demand to be fixed, thus obviating the possibility that congestion management and social welfare might be better served by encouraging the use and aiding the efficiency of the ridesourcing service–rather than discouraging it. Our future work will seek to integrate these substitution effects to provide a more complete picture of the transportation conundrum faced by urban planners.

CRediT authorship contribution statement

Daniel A. Vignon: Methodology, Investigation, Writing - original draft. **Yafeng Yin:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Jintao Ke:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work described in this paper was partly supported by research grants from the National Science Foundation (CMMI-1854684 and CMMI-1904575).

Appendix A. Numerical examples

A.1. Demand data

Base demand λ^0 is obtained from Korolko et al. (2018). Background traffic demand is obtained by scaling λ^0 by 10^9 so that as λ^0 increases, so does congestion and the externality of ridesourcing (as would happen in practice). We show the base demand pattern in Fig. 9.

[8] Currently Didi Chuxing in some Chinese cities, for example.

[9] Our results show that, accross demand levels, the number of ride-sourcing vehicles N is *at least* 10% of the number of background vehicles N^b, which seems reasonable.

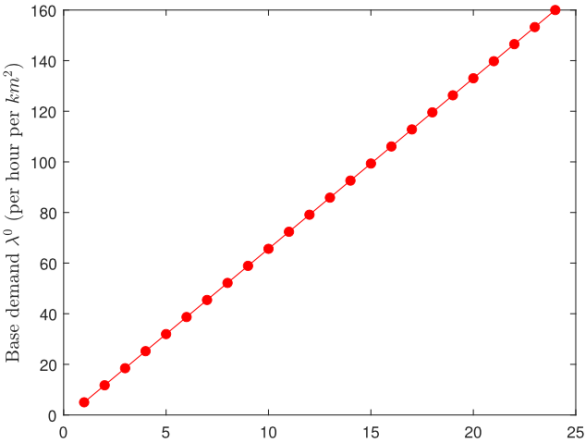

Fig. 9. Base demand and background demand data.

A.2. Parameter values

All the parameters (except the bounds on the support of the value of time distribution) and their values are described in Table 1. The parameter values used for the different value of time distributions are given in Table 2.

Table 1
Parameter values for numerical examples

Notation	Interpretation	Value
α	Meeting distance elasticity	0.5
ξ^s	Mode specific constant	0
γ	Correlation coefficient between customer and driver detour time	1
γ^b	Value of travel time for background traffic	$30\frac{\$}{\text{hr}}$
δ	Pairing probability	0.1
ϕ	Matching time window	5 min
θ	Marginal effect of ridesourcing vehicles on traffic	1
A	Scaling parameter for meeting time function	25
B	Scaling parameter for detour distance function	5
d^r	Average distance of ridesourcing trips	7 km
$d^{r,b}$	Average distance of background traffic trips	3.5 km
V^c	Slope of speed function	$-0.11\frac{\text{kph}\cdot\text{km}^2}{\text{veh}}$
V^0	Free-flow speed	50 kph
U^0	Trip utility	\$50
c	Cost per driver	\$10
μ^0	Cost of outside option	\$30

Table 2
Uniform distribution parameter values.

	$\bar{\beta}$	$\underline{\beta}$
High variance	75	5
Medium variance	60	20
Low variance	45	35

Appendix B. Finding regulatory policy

As outlined in Section 5, we consider a Stackelberg game between the planner (leader) and the monopolist (follower). The planner solves (REG) below[10]:

$$
\begin{aligned}
W = \max_{\substack{P^s \geqslant 0,\, F^s \geqslant 0,\\ P^p \geqslant 0,\, F^p \geqslant 0,\\ \tau^N \geqslant 0,\, R \geqslant 0}} \quad & U^0 \cdot \left(\lambda^s + \lambda^p \cdot o\right) - \lambda^0 \cdot \int_{\beta_2}^{\beta_1} \beta \cdot \left(w^m + w^r\right) \cdot G'\left(\beta\right) \cdot d\beta \\
& -\lambda^0 \cdot \int_{\underline{\beta}}^{\beta_3} \beta \cdot \left(w^p + w^m + w^r + \Delta w\right) \cdot G'\left(\beta\right) \cdot d\beta \\
& -R \cdot \left(\lambda^s + \lambda^p\right) + \tau^N \cdot N - \gamma^b \cdot \lambda^b \cdot w^{rb} \\
\text{s.t.} \quad & \left(F^s, F^p, R\right) = S\left(P^s, P^p, \tau^N\right)
\end{aligned}
\tag{REG}
$$

In the above, $S(P^s, P^p, \tau^N)$ is the best response function of the firm when subject to the regulation (P^s, P^p, τ^N). This best response is obtained by solving (M-CAPT) below:

$$
\begin{aligned}
\pi = \max_{\substack{F^s \geqslant 0,\, F^p \geqslant 0,\\ R \geqslant 0}} \quad & \left(F^s - R\right) \cdot \lambda^s + \left(F^p - R\right) \cdot \lambda^p \\
\text{s.t.} \quad & F^s - R \leqslant P^s \\
& F^p - R \leqslant P^p
\end{aligned}
\tag{M-CAPT}
$$

Moreover, the positivity constraint on P^s and P^p ensure that the regulation is implementable (the platform can generate a profit and thus operates). The positivity constraint on τ^N enforces a no-subsidy constraint on the outcome of the regulation.

As formulated, the problem is a bilevel program, a class of optimization problems difficult to solve. For our numerical experiments, we used a heuristic solution procedure. In the process, we solve (M-CAPT) with the current regulation policy (P^s_k, P^p_k, τ^N_k). Then, holding the firm's response constant, we solve (REG) and use the method of successive averages (MSA) to update the regulation policy for the next step. Since our algorithm may not necessarily converge and given that we do not consider whether the updated policy effectively increases welfare, we keep track of the best solution obtained up to the current step. Thus, after solving for the firm's best response to (P^s_k, P^p_k, τ^N_k), we compare the welfare resulting from the monopolist's response, W_k, to the maximum realized welfare up to k, W_{best}.

The above solution procedure yielded reasonably good solutions, as demonstrated in Fig. 6. The solutions may be further improved by applying a derivative-free method such as Nelder–Mead or pattern search, as the decision variables of REG are of a low dimension.

References

Agarwal, S., Mani, D., Telang, R., 2019. The Impact of Ride-hailing Services on Congestion: Evidence from Indian Cities. SSRN Scholarly Paper ID 3410623, Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3410623.

An, P., 2020. The Information's 411 — Uber Drowning in the Pool. The Information. https://www.theinformation.com/articles/the-informations-411-uber-drowning-in-the-pool.

Arnott, R., 1996. Taxi travel should be subsidized. J. Urban Econ. 40(3), 316–333. ISSN 00941190. https://linkinghub.elsevier.com/retrieve/pii/S0094119096900352.

Bellon, T., 2019. Uber to limit drivers' app access to comply with NYC regulation. Reuters. https://www.reuters.com/article/us-uber-new-york-idUSKBN1W12OV.

Beojone, C.V., Geroliminis, N., 2021. On the inefficiency of ride-sourcing services towards urban congestion. Transp. Res. Part C: Emerg. Technol. 124, 102890. ISSN 0968-090X. https://www.sciencedirect.com/science/article/pii/S0968090X20307907.

Castiglione, J., Cooper, D., Sana, B., Tischler, D., Chang, T., Erhardt, G.D., Roy, S., Chen, M., Mucci, A., 2018. TNCs and Congestion. Final Report, San Francisco County Transportation Authority. https://www.sfcta.org/sites/default/files/2019-05/TNCs_Congestion_Report_181015_Finals.pdf.

Castillo, J.C., 2018. Who benefits from surge pricing? SSRN Electron. J.. ISSN 1556-5068. https://www.ssrn.com/abstract=3245533.

Castillo, J.C., Weyl, E.G., 2018. Surge Pricing Solves the Wild Goose Chase. p. 29. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2890666.

Clewlow, R.R., Mishra, G.S., 2017. Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States. https://escholarship.org/uc/item/82w2z91j.

Daganzo, C.F., 1978. An approximate analytic model of many-to-many demand responsive transportation systems. Transp. Res. 12(5), 325–333. ISSN 0041-1647. http://www.sciencedirect.com/science/article/pii/0041164778900072.

Dobbs, C., 2019. TLC regulations risk leaving lower-income New Yorkers behind. https://www.uber.com/blog/new-york-city/tlc-regulation-risks/.

Erhardt, G.D., Roy, S., Cooper, D., Sana, B., Chen, M., Castiglione, J., 2019. Do transportation network companies decrease or increase congestion? Science Adv. 5(5), eaau2670. ISSN 2375-2548. https://advances.sciencemag.org/content/5/5/eaau2670.

N. Geroliminis and C.F. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. Transportation Research Part B: Methodological, 42(9):759–770, Nov. 2008. ISSN 01912615. https://linkinghub.elsevier.com/retrieve/pii/S0191261508000180.

[10] Note that the revenue from tolls needs to be added to the social welfare function. Otherwise there would be a missing transfer in our system.

Hampshire, R., Simek, C., Fabusuyi, T., Di, X., Chen, X., 2017. Measuring the Impact of an Unanticipated Disruption of Uber/Lyft in Austin, TX. SSRN Scholarly Paper ID 2977969, Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2977969.

Honan, K., 2019. Uber, Lyft Drivers Face Stiffer Regulations in New York City. Wall Street Journal. ISSN 0099-9660. https://www.wsj.com/articles/uber-lyft-drivers-face-stiffer-regulations-in-new-york-city-11560375449.

Ke, J., Yang, H., Li, X., Wang, H., Ye, J., 2020a. Pricing and equilibrium in on-demand ride-pooling markets. Transp. Res. Part B: Methodol. 139, 411–431. ISSN 0191-2615. http://www.sciencedirect.com/science/article/pii/S0191261520303611.

Ke, J., Yang, H., Zheng, Z., 2020b. On ride-pooling and traffic congestion. Transp. Res. Part B: Methodol. 142, 213–231. ISSN 0191-2615. http://www.sciencedirect.com/science/article/pii/S0191261520304094.

Korolko, N., Yan, C., Woodard, D., Zhu, H., 2018. Dynamic pricing and matching in ride-hailing platforms. SSRN Electron. J. ISSN 1556-5068. https://www.ssrn.com/abstract=3258234.

Li, S., Tavafoghi, H., Poolla, K., Varaiya, P., 2019. Regulating TNCs: Should Uber and Lyft set their own rules? Transp. Res. Part B: Methodol. 129, 193–225. ISSN 0191-2615. http://www.sciencedirect.com/science/article/pii/S0191261519300669.

Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. Transp. Policy 45, 168–178. ISSN 0967-070X. http://www.sciencedirect.com/science/article/pii/S0967070X15300627.

Schaller, B., 2018. The New Automobility: Lyft, Uber and the Future of American Cities. Technical report, Schaller Consulting. http://www.schallerconsult.com/rideservices/automobility.htm.

Spotwood, B., 2017. Not Surprisingly, UberPool Has Been A Major Loss Leader: SFist. https://sfist.com/2017/05/31/uber_pool_loses_much_money/.

Tarduno, M., 2021. The congestion costs of Uber and Lyft. J. Urban Econ. 103318. ISSN 0094-1190. http://www.sciencedirect.com/science/article/pii/S0094119020300899.

Uber Inc., 2020. Uber's Q4 2019 Earnings-Supplemental Data. https://s23.q4cdn.com/407969754/files/doc_financials/2019/q4/Quarterly-Earnings-Report-Q42019.pdf.

Wei, B., Saberi, M., Zhang, F., Liu, W., Waller, S.T., 2020. Modeling and managing ridesharing in a multi-modal network with an aggregate traffic representation: A doubly dynamical approach. Transp. Res. Part C: Emerg. Technol. 117, 102670. ISSN 0968-090X. https://www.sciencedirect.com/science/article/pii/S0968090X20305854.

Xu, Z., Yin, Y., Zha, L., 2017. Optimal parking provision for ride-sourcing services. Transp. Res. Part B: Methodol. 105, 559–578. ISSN 01912615. https://linkinghub.elsevier.com/retrieve/pii/S0191261517308962.

Yang, H., Ye, M., Tang, W.H., Wong, S., 2005. Regulating taxi services in the presence of congestion externality. Transp. Res. Part A: Policy Pract. 39(1), 17–40. ISSN 09658564. http://linkinghub.elsevier.com/retrieve/pii/S0965856404001053.

Yang, T., Yang, H., Wong, S.C., 2014. Taxi services with search frictions and congestion externalities. J. Adv. Transp. 48(6), 575–587. ISSN 2042-3195. https://onlinelibrary.wiley.com/doi/abs/10.1002/atr.1210.

Zha, L., Yin, Y., Du, Y., 2018a. Surge pricing and labor supply in the ride-sourcing market. Transp. Res. Part B: Methodol. 117, 708–722. ISSN 0191-2615. http://www.sciencedirect.com/science/article/pii/S0191261517307683.

Zha, L., Yin, Y., Xu, Z., 2018b. Geometric matching and spatial pricing in ride-sourcing markets. Transp. Res. Part C: Emerg. Technol. 92, 58–75. ISSN 0968-090X. http://www.sciencedirect.com/science/article/pii/S0968090X18305138.

Zha, L., Yin, Y., Yang, H., 2016. Economic analysis of ride-sourcing markets. Transp. Res. Part C: Emerg. Technol. 71, 249–266. ISSN 0968-090X. http://www.sciencedirect.com/science/article/pii/S0968090X16301188.

Zhang, K., Nie, M., 2019. To Pool or Not to Pool: Equilibrium, Pricing and Regulation. SSRN Scholarly Paper ID 3497808, Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3497808.