Transportation Research Part C 129 (2021) 103236

Contents lists available at ScienceDirect

Transportation Research Part C

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

Compartmental model and fleet-size management for shared mobility systems with for-hire vehicles☆

Wen-Long Jin [a,*], Irene Martinez [b], Monica Menendez [c]

[a] Department of Civil and Environmental Engineering, California Institute for Telecommunications and Information Technology, Institute of Transportation Studies, 4000 Anteater Instruction and Research Bldg, University of California, Irvine, CA 92697-3600, United States
[b] Department of Civil and Environmental Engineering, Institute of Transportation Studies, 4000 Anteater Instruction and Research Bldg, University of California, Irvine, CA 92697-3600, United States
[c] Division of Engineering, New York University, Saadiyat Island, Abu Dhabi, Abu Dhabi 129188, United Arab Emirates

ARTICLE INFO

Keywords:
For-hire vehicles
Compartmental model
Demand and supply
Extra mileage ratio
Fleet-size management scheme
Well-defined condition
Optimal fleet-size cap

ABSTRACT

There have been conflicting results in the literature regarding the congestion impacts of shared mobility systems with for-hire vehicles (FHVs). To the best of our knowledge, there is no physically meaningful and mathematically tractable model to explain these conflicting results or devise efficient management schemes for such mobility systems. In this paper, we attempt to fill the gap by presenting a compartmental model for passenger trip and vehicle dynamics in shared mobility systems with FHVs and discussing the impacts of different fleet-size management schemes.

To develop the compartmental model, we first divide passenger trips into four compartments: planned, waiting, traveling, and completed. We describe the dynamics of the waiting trips by the point queue model, and those of the traveling trips by an extended bathtub model. As the traditional bathtub model for vehicular trips, the extended bathtub model is derived in a relative space with respect to individual trips' distances to their destinations. However, different from the traditional bathtub model, vehicular dynamics and trip dynamics in the extended bathtub model are not overlapping, as the dynamics of FHVs are controlled by the fleet-size management scheme; but they are related, as traveling trips travel with occupied FHVs, and empty FHVs supply seats to waiting trips. Within this modeling framework, the matching process between waiting passengers and FHVs is modeled at the aggregate level, such that the passenger trip flow from the waiting compartment to the traveling compartment equals the minimum of the waiting trips' demand of seats and the supply of seats determined by the completion rate of traveling trips and the fleet-size management scheme. In addition to the pooling ratio, the deadhead miles, the detour miles caused by pooling services, and other extra miles associated with the matching process are captured by another exogenous parameter, namely, the extra mileage ratio. With these assumptions and simplifications, the resulting compartmental model is a deterministic, coupled queueing model, which can be written as a system of differential equations. We also present the sufficient and necessary condition on the fleet-size management scheme for the model to be well-defined.

With the parsimonious, closed-form compartmental model, we demonstrate theoretically that limiting the wait time leads to a fleet-size management scheme equivalent to that of the privately

☆ This article belongs to the Virtual Special Issue on IG005584: VSI:ISTTT24.

* Corresponding author.
E-mail address: wjin@uci.edu (W.-L. Jin).

https://doi.org/10.1016/j.trc.2021.103236
Received 22 December 2020; Received in revised form 13 April 2021; Accepted 23 May 2021
Available online 11 June 2021

operated vehicles (POVs), i.e., the POV scheme. In such a system, the completion rate depends on the extra trip mileage ratio, as well as the pooling ratio. With 100% autonomous FHVs, the optimal fleet size that minimizes the total costs occurs at the maximum flow-rate and the free-flow speed. With mixed POVs and FHVs, we extend the compartmental model and numerically solve for the optimal fleet sizes under different market penetration rates. This study reconciles the conflicting results in the literature. We find that, with a low pooling ratio, the overall system's performance can be deteriorated or improved, depending on the fleet-size management scheme: with the POV scheme, the system could become more congested; but with an appropriate fleet-size cap, the system's performance can be substantially improved. A major policy implication of this study is that implementing a cap for the FHV fleet size is a viable measure to mitigate the congestion effects of extra deadhead and detour miles caused by FHVs.

1. Introduction

The core task of a mobility system is to accommodate various types of passenger and freight trips. Different from the traditional mobility system served by privately operated vehicles (POVs), in which travelers do not have to wait for their designated vehicles to be dispatched, recently a new type of shared mobility systems have been realized through for-hire vehicles (FHVs) managed by transportation network companies (TNCs). Such systems do require travelers to wait while TNCs dispatch FHVs to pick them up. Shared mobility systems have attracted much attention from the general public for the convenience they offer and the job opportunities they provide. It is believed that such a transportation mode could be even more prevalent when autonomous vehicles are incorporated into the future smart mobility systems (Hensher, 2018; Hyland and Mahmassani, 2018).

In the literature, many studies focus on the design of shared mobility systems with FHVs by solving dial-a-ride or vehicle-dispatch problems so as to reduce TNCs' operational costs and improve the service quality for travelers. The solutions to these problems are subject to constraints in the congestion condition of the road network and travel demands determined by travelers' choice behaviors in departure times, modes, and so on. See Mourad et al. (2019) for a comprehensive review for such studies. In contrast, another important problem is related to understanding the impacts of FHVs on the whole transportation system's performance and devising the corresponding system-level management schemes to improve the whole system's performance. In particular, FHVs could impact the mobility, public transit ridership, car ownership, and so on. This study is concerned with the mobility impacts of FHVs.

At the system level, there have been conflicting results in the literature regarding the impacts on mobility of shared mobility systems with FHVs. Through agent-based simulation studies, (Fagnant and Kockelman, 2014; Martinez et al., 2015) showed that FHVs can substantially relieve congestion of the transportation network with a fixed fleet size. Also through simulations, (Bischoff and Maciejewski, 2016; Fagnant and Kockelman, 2018) concluded that the current transportation system with POVs can be completely substituted by a shared mobility system with autonomous FHVs, at a rate of 8 to 10 POVs being replaced by a single FHV. These simulation-based studies report that wait times are reasonably low, but the total vehicles miles travelled (VMT) are longer than those by POVs, due to the deadhead and detour miles associated with FHVs. In general, these studies seem to support TNCs' claim that FHVs can help to relieve congestion. In contrast, recent empirical studies (Schaller, 2017) point out that, in reality, FHVs have imposed negative impacts on all the aspects related to congestion, private car ownership, and public transit usage: "trips, mileage and the number of vehicles" have all been increased in the central business district area of New York city, but the increases vary considerably with the time of the day. In San Francisco, FHVs are estimated to account for 25% of the total vehicle hour delay (Castiglione et al., 2018). In NYC Taxi and Limousine Commission (2019), it was observed that, "relying on short wait times to keep demand high, the companies have saturated the market with vehicles, which currently spend 41% of their time in the core cruising around without passengers." This suggests that both the existing fleet-size management schemes employed by the TNCs and extra deadhead and detour mileage of FHVs might be the cause of worsening congestion.

Shared mobility systems are quite complex with many stakeholders and can be studied from many different perspectives, such as the TNCs' operational perspectives, the traffic agencies' system perspective, the user perspective, etc. To the best of our knowledge, however, there is no physically meaningful and mathematically tractable model that (i) captures the impacts on congestion of fleet-size management schemes, pooling ratios, and extra deadhead and detour miles of FHVs, (ii) reconciles the conflicting results regarding FHVs' congestion effects in the literature, and (iii) provides guidelines to efficiently manage the future smart mobility systems. In this paper, we attempt to fill the gap by introducing a unified analytical approach to model and manage shared mobility systems with FHVs with respect to the whole system's mobility performance.

To that end, we adopt a congestion dynamics perspective from the traffic system agencies. In particular, we propose a simple model to describe the congestion dynamics in a shared mobility system and then develop effective fleet-size management schemes at the network level. We follow the same spirit of the compartmental models in epidemiology and introduce a closed-form model for such a complex system (Godfrey, 1983). Based on the following four assumptions and simplifications that are closely related to the specific characteristics of a shared mobility system with FHVs, the compartmental model can be written as a system of differential equations. First, we take a compartmental view of passenger trips, by dividing them into four compartments: planned, waiting, traveling, and completed. Second, the dynamics of the waiting trips are captured by the point queue model (Vickrey, 1969; Jin, 2015), and those of the traveling trips by an extended bathtub model, which can be considered a network queue model. As the traditional bathtub model of vehicular trips (Vickrey, 2020; Jin, 2020), the extended bathtub model is also derived in a relative space with respect to individual trips' distances to their destinations. But vehicular dynamics and trip dynamics are not overlapping, as the dynamics of FHVs are

controlled by the fleet-size management scheme; but they are related, as traveling trips travel with occupied FHVs, and empty FHVs supply seats to waiting trips. Third, the matching process between waiting passengers and FHVs is modeled at the aggregate level, such that the passenger trip flow from the waiting compartment to the traveling compartment is equals the minimum of the waiting trips' demand of seats and the supply of seats determined by the completion rate of the traveling trips and the fleet-size management scheme. Here, the demand of the waiting compartment is defined as that of the point queue as in Jin (2015), and a new definition of the supply will be introduced to capture the impacts of both the traveling compartment and the fleet-size management scheme. The aggregate model for calculating the matching (transition) rate between two compartments can be considered as an extension of the Cell Transmission Model for calculating the transmission rate between two cells (Daganzo, 1994). Finally, two exogenous parameters are introduced: the pooling ratio and the extra mileage ratio, which captures the deadhead miles of empty FHVs, the detour miles with FHVs' pooling services, and other extra miles associated with the matching process. Moreover, we present the sufficient and necessary condition on the fleet-size management scheme for the model to be well-defined. With the proposed model, we then examine the impacts of the fleet-size management scheme currently employed by many TNCs, which aims to limit travelers' wait times. We also examine the impacts of a fleet-size management scheme, which introduces a cap of the FHV fleet size introduced by many cities.

The contributions of this study are fourfold. First, we present a novel compartmental model by coupling the point queue model and an extended bathtub model to account for both passenger trip and vehicle dynamics in a shared mobility system with FHVs. Second, we rigorously define the current TNCs' fleet-size management schemes and analyze their impacts on the overall mobility performance of a shared mobility system. Third, we define an optimal fleet-size management scheme and, by comparing it with TNCs' schemes, prove that a cap for the number of FHVs as proposed by many cities is a viable option to limit traffic congestion. Fourth, we extend the compartmental model for the mixed mobility system with both FHVs and POVs and numerically solve the optimal fleet sizes under different market penetration rates.

The rest of the article is organized as follows. In Section 2 we discuss the compartmental modeling framework for shared mobility systems with FHVs and present the definition for the model to be well-defined. In Section 3 we present the mathematical model for both trip and vehicle dynamics in a shared mobility system with FHVs and present the condition on the fleet-size management scheme for the model to be well-defined. In Section 4 we present two feedback fleet-size management schemes aiming to limit the wait time employed by TNCs. In Section 5 we formulate and solve the solution for the optimal fleet-size cap assuming user equilibrium departure time choice. In Section 6 we extend the model and optimal fleet-size management scheme for a mixed mobility system with both POVs and FHVs. In Section 7 we conclude the study with discussions on the policy implications of this study and potential extensions of the methodology.

2. Compartmental modeling framework for shared mobility systems

For the readers' convenience, a list of notations is given in Table 1.

Compartmental models (or modifications thereof) have been widely used in epidemiology for capturing the transmission of viruses in a large population (Godfrey, 1983). For example, in the classic susceptible-infectious-recovered (SIR) model of epidemic dynamics, the whole population is divided into three compartments: susceptible, infectious, and recovered. The state of each compartment is described by the number of people, which can change dynamically. The transition rate from one compartment to another is determined by the corresponding state variables; such deterministic, aggregate transition rates approximately capture the detailed spatial-temporal interactions among different groups of people. As a result, each compartment's dynamics are described by a simple ordinary or partial differential equation, and the resulting models are mathematically tractable. Theoretically, the SIR model has led to invaluable analytical insights, including the definition of the basic reproduction number. Practically, with well calibrated parameters, such a model can be used to predict epidemic dynamics and devise large-scale mitigation schemes.

Shared mobility systems are also quite complex and can be studied from many different perspectives. These include the economic perspective, in terms of market equilibrium with respect to demand and supply of ridesourcing services (see a comprehensive literature review in Wang and Yang (2019)), or the TNCs' operational perspectives of the demand of seats, supply of seats and matching process between both (see a comprehensive literature review (Agatz et al., 2012)). In this study we take a congestion dynamics perspective from the traffic system agencies. To that end, we propose a simple model that describes the congestion dynamics in a shared mobility system. In turn, this allows us to develop effective fleet-size management schemes at the network level. Following the same spirit of the compartmental models in epidemiology, we introduce four assumptions and simplifications to reflect specific characteristics of a shared mobility system with FHVs.

2.1. Compartmental view of passenger trips in different mobility systems

For the purpose of simplicity, we lump together all trips in a mobility system, regardless of their exact locations, and divide them into different compartments, depending on their stages. This compartment view can be applied to a traditional mobility system with only POVs, a shared mobility system with 100% autonomous FHVs, or a mixed mobility system with both POVs and FHVs.

With only POVs, the passenger trips can be divided into three sequential compartments: planned (planned but have not started), traveling (started but not completed), and completed. The three compartments are illustrated in Fig. 1(a), in which the total number of

Table 1
List of notations

Variables	Definitions
$B(t)$	Average remaining distance of travelling trips at t
$\widetilde{B}(t)$	Average entering trips' distances at t
C	Road capacity
$E(t)$	Cumulative entering flow at t
$F(t)$	Cumulative boarding flow at t
$G(t)$	Cumulative completed flow at t
$H(\cdot)$	Heaviside function
$K(t,x)$	Number of active trips at t with a remaining distance not smaller than x
L	Total lane-miles of a network
N	Total number of trips
$V(\rho)$	Network speed-density relation
$d(t)$	Demand rate of seats by the waiting trips at t
$e(t)$	Entering rate at t
$f(t)$	Boarding rate at t
$g(t)$	Completion rate at t
$s(t)$	Supply rate of seats to the waiting trips at t
$v(t)$	Average travel speed at t
x	POV (privately operated vehicle) trip distance
y	FHV (for-hire vehicle) trip distance
$z(t)$	Characteristic travel distance at t
$\Upsilon(t)$	Wait time for a traveler boarding at t
$\Phi(t,x)$	Proportion of trips with distances not smaller than x at t
$\widetilde{\Phi}(t,x)$	Proportion of the entering trips with distances not smaller than x at t
$\delta(t)$	Waiting queue size at t
$\lambda(t)$	Number of traveling trips at t
$\varphi(t)$	Number of empty seats at t
ϕ	Total cost function
$\rho(t)$	Per-lane vehicle density
ρ^*	Fleet-size cap for FHVs
ρ_j	Jam density
$\tau(z)$	Characteristic travel time with a characteristic travel distance of z
π	Pooling ratio
ξ	Extra trip mileage ratio
γ	Cost coefficient for the traveling time
β	Cost coefficient for the wait time
ν	Penalty of late arrivals
μ	Penalty for early arrivals
η_m	Mode choice ratio for mode m, where $m = \{F,P\}$

trips during a study period is denoted by N,[1] the cumulative entering flow from the planned compartment to the traveling compartment at t by $E(t)$, and the cumulative completed flow by $G(t)$. Thus, the numbers of planned, traveling, and completed trips at t are, respectively, $N-E(t)$,$E(t)-G(t)$, and $G(t)$. The rates of change in time of $E(t)$ and $G(t)$ are, respectively, $e(t)$ and $g(t)$, which represents the entering and completion rates of trips, respectively: $e(t)=\dot{E}(t)$ and $g(t)=\dot{G}(t)$ with the dot for the derivative with respect to time.

When all trips are served by autonomous FHVs, in contrast, a traveler starts his or her trip by placing a request with a TNC but has to wait for the assignment of an FHV and for the vehicle to pick him/her up. Thus, there is an additional compartment of waiting trips. In such a system, there can be congestion in both the waiting and traveling compartments during a peak period. Since TNCs monitor wait times in real time and dispatch FHVs dynamically to "ensure short wait times and spur demand" (NYC Taxi and Limousine Commission, 2019), it is also important to model the dynamics of waiting trips. Hence, in a shared mobility system with FHVs, the passenger trips are divided into planned, waiting, traveling, and completed (PWTC) compartments. The four compartments are illustrated in Fig. 1(b), in which the cumulative entering flow from the planned compartment to the waiting compartment at t is denoted by $E(t)$, and the cumulative boarding flow from the waiting compartment to the traveling compartment by $F(t)$. Thus, the number of waiting and traveling trips at t are, respectively, $E(t)-F(t)$ and $F(t)-G(t)$. The rates of change in time of $F(t)$ and $G(t)$ are, respectively, $f(t)$ and

[1] In this study we focus on the real-time operation and management of the shared mobility system. Thus, we assume that the total travel demand is constant.

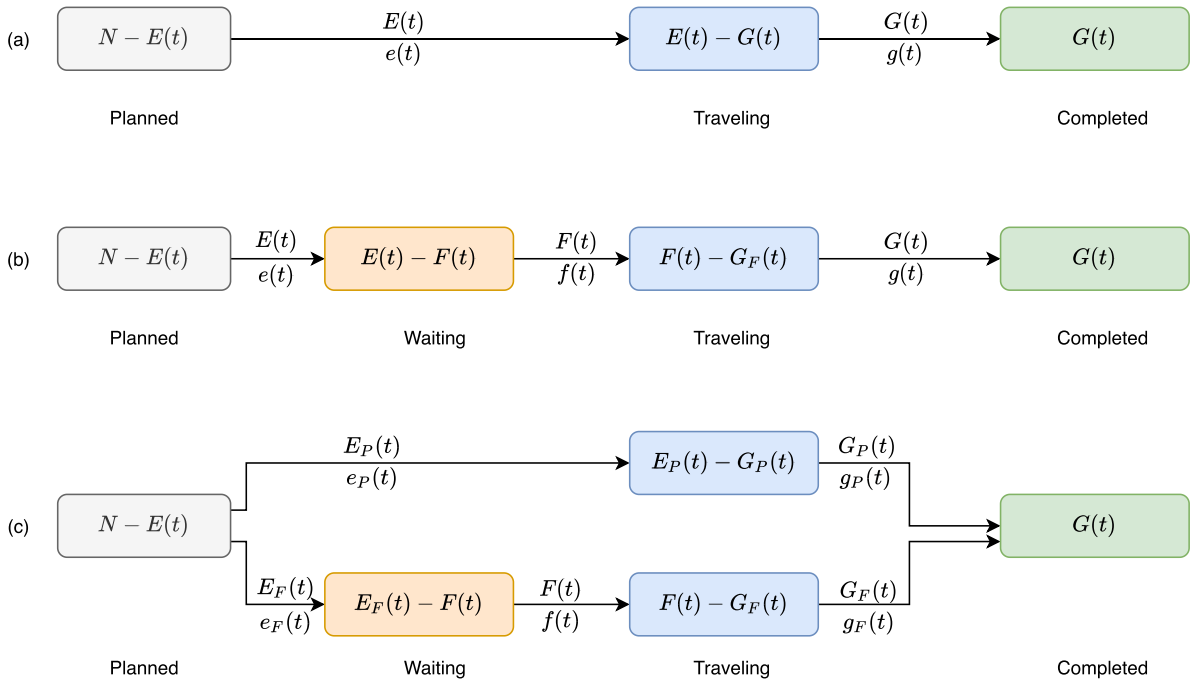

Fig. 1. Compartments of passenger trips in different mobility systems and trip flows between compartments: (a) Three compartments in a mobility system with POVs; $e(t)$ and $g(t)$ represent the entering and completion rates of trips, respectively, and $E(t)$ and $G(t)$ are the corresponding cumulative flows; (b) Four compartments in a shared mobility system with FHVs; $f(t)$ represents the matching (boarding) rate of trips, and $F(t)$ is the corresponding cumulative flow; (c) Compartments in a mixed mobility system with both POVs and FHVs.

$g(t)$, which represents the boarding and completion rates of trips. Notice that both $F(t)$ and $G(t)$ could be observed in practice, or predicted with the compartmental model if $E(t)$ is given as an input. Given their importance for developing a feedback fleet-size management scheme, and the fact that observing them can be quite expensive, here we use the short-term prediction by the compartmental model to develop the schemes discussed in Sections 4 and Section 5.

Fig. 1(c) illustrates the compartments in a mixed mobility system with both POVs and FHVs, where the subscripts P and F represent the two types of vehicles respectively. Here the total entering and completion rates are given by $e(t) = e_P(t) + e_F(t)$ and $g(t) = g_P(t) + g_F(t)$, respectively. The total entering and completion cumulative flows are given by $E(t) = E_P(t) + E_F(t)$ and $G(t) = G_P(t) + G_F(t)$, respectively.

2.2. Deterministic queueing models of waiting and traveling trips

For the compartments in Fig. 1, we assume that the total number of trips, N, and the entering flows and flow-rates ($E(t)$, and $e(t)$) are given. That is, we do not consider departure time choice or induced and suppressed demand caused by traffic congestion in a road network. Therefore, the dynamics of planned trips are known. In addition, the completion rate of trips is assumed to be determined by traveling trips. Therefore, the dynamics of passenger trips are purely determined by those in the waiting (if existing) and traveling compartments.

For the traveling trips served by POVs, illustrated in Fig. 1(a) or (c), we can directly apply the traditional bathtub model for vehicular trips in Vickrey (1991), Vickrey (2020), Jin (2020), since passenger trips overlap with vehicular trips (assuming an occupancy of 1). In this model, trip dynamics are described in a relative space, with respect to individual trips' distances to their destinations. As the relative space is independent of the absolute network topology, individual trips' origins, destinations, routes, and links are implicit, and different trips' trajectories can be described in a unified space-time domain. Two further assumptions are made: under the "bathtub" assumption, vehicles' speeds are the same at a time instant and irrelevant to the their exact locations in a network; under the network fundamental diagram assumption, the vehicle speed at a time instant is determined by the vehicle density. Then the trip dynamics at the aggregate level can be described by a differential equation in terms of the number of trips with a remaining trip distance. In such a bathtub model, the demand pattern is given by the entering rate of trips as well as the distribution of entering trips with respect to their distances. For a general distribution of trip distances, the bathtub model can be written as a partial differential equation (Jin, 2020); but when the trip distances follow a time-independent negative exponential distribution, it becomes Vickrey's bathtub model, which can be simplified as an ordinary differential equation (Vickrey, 1991; Vickrey, 2020). In a sense, the bathtub model can also be considered a network queue model in the relative space, where the queue size is the number of traveling trips, denoted by $\lambda(t)$. Such a model is mathematically more tractable than traditional network traffic flow models depending on the network topology, which lead to a large number of differential equations, depending on the number of origins, destinations, routes, and links. At

the same time, this model still captures the impacts of not only the entering rates of trips, but also the distribution of trip distances. Therefore, the bathtub model strikes a balance between mathematical tractability and physical realism. Therefore, we attempt to use the bathtub model to describe the traveling trip dynamics in a shared mobility system with FHVs.

However, for traveling trips served by FHVs, as illustrated in Fig. 1(b) or (c), we cannot directly apply the bathtub model, since the passenger trips and the vehicular trips can evolve separately. On one hand, with the pooling service, one FHV can carry multiple passengers at the same time, or can be traveling empty. On the other hand, the discrepancies among passengers' locations and FHVs' locations as well as the delays when matching passengers and FHVs can lead to extra mileage compared with each passenger's trip distance. Thus, the bathtub model needs to be extended for traveling trips served by FHVs. In particular, the passenger trips are still described by the bathtub model, but the dynamics of FHVs are separately modeled. In the extended bathtub model, passengers' trip distances are modified to incorporate the deadhead miles as well as the extra miles caused by re-routing with pooling services.

Similarly, we lump together all waiting trips, and their exact locations in the road network are not explicitly tracked. With this simplification, we can describe their dynamics with Vickrey's point queue model (Vickrey, 1969; Jin, 2015), where the state variable is the number of waiting trips, i.e., the waiting queue size, denoted by $\delta(t)$. Following (Jin, 2015), we can define the demand rate from the point queue by $d(t)$, which can be determined by the waiting queue size and the entering rate. At the aggregate level, the demand rate represents the number of seats needed per unit time to accommodate the waiting passengers.

Within the aforementioned point queue and bathtub models, the waiting and traveling compartments for the shared mobility system with FHVs can be viewed as a tandem of two deterministic queues (Newell, 1982): a waiting queue plus a traveling queue. Fig. 2 illustrates the tandem of both waiting and traveling queues, in which $\delta(t)$ and $\lambda(t)$ are the waiting and traveling queue sizes, respectively, and $E(t), F(t)$, and $G(t)$ are the arrival and departure curves for the two queues.

2.3. Aggregate model of the matching process

In reality, the boarding rate, $f(t)$, is determined by the matching rate between waiting travelers and available FHVs. Such a dynamic matching problem is quite complex and relatively new in the literature (Agatz et al., 2011; Agatz et al., 2012; Yang et al., 2020). In addition, TNCs' pricing schemes and other social-economic incentives influence drivers' choice behaviors, which in turn determine the availability of FHVs in a road network (Wang and Yang, 2019). But in this study we aim to approximately calculate the real-time boarding rate at the aggregate level, such that the resulted compartmental model is still mathematically tractable.

First, we examine the two types of mobility systems in terms of demand and supply. In the traditional mobility system served by POVs, travelers and POVs are integrated into traveler-vehicle units and pose a demand in terms of road space onto the infrastructure directly (Downs, 2004), as shown in Fig. 3(a). In contrast, for the shared mobility system served by FHVs, travelers and FHVs are separated. Thus, the mobility system has a two-layered structure, as shown in Fig. 3(b), where trips are served by FHVs, and FHVs share the road infrastructure (Lam et al., 1999; Shaheen et al., 2018). Consequently, in the shared mobility system, there are two-layers of demand and supply: on the first layer, the travelers impose a demand in terms of trips on FHVs, which supply seats; on the second layer, the vehicles impose a demand in terms of road space on the infrastructure, which supplies the road capacity. In this two-layered shared mobility system, the immediate stakeholders include passengers, drivers, mobility service vehicles, mobility service providers, and traffic system agencies. Moreover, there are complicated interactions and conflicts between the demand and the supply in both layers. For example, the supply of seats could be improved by providing more vehicles, but this imposes more demand in road space on the road network and therefore leads to more congestion. Conceptually, this demonstrates that the fleet-size management scheme is critical for the overall performance of a shared mobility system.

In this study, we assume that the rate of change in the number of FHVs is determined by a deterministic and aggregate fleet-size management scheme, which depends on TNCs' matching algorithms, traffic management agencies' regulation policies, and the availability of the drivers of FHVs. In particular, TNCs can move an FHV into or out of a network, traffic management agencies can impose a limit on the number of FHVs in a network, and the driver of an FHV can choose to cruise on roads or leave roads and wait. Here we assume that the economic and other incentives are sufficiently high such that the number of FHVs is as large as determined by the fleet-size management scheme.

Mathematically, we can determine the supply rate of seats, $s(t)$, from the number of FHV seats, the number of traveling passenger trips, as well as the completion rate of passenger trips. Then we calculate the boarding rate, $f(t)$, as the minimum of the demand rate of seats imposed by the waiting compartment, $d(t)$, and the supply rate of seats, $s(t)$. This aggregate model approximates the matching process between passengers and FHVs.

The matching process among waiting passengers and available FHV seats could take up to several minutes, depending on passengers' ride-sharing choices, and the relative locations of the seats and waiting passengers. For example, it is possible that a waiting passenger may not be matched with the first available FHV seat, if they are too far away, or the FHV does not satisfy the passenger's preferences in ride-sharing, driver rating, vehicle type, and so on. Therefore, the aggregate matching model best applies to a relatively dense network with large numbers of passengers and FHVs, where the waiting passengers and available FHV seats can be matched almost instantaneously. If the waiting passengers and available FHV seats cannot be matched instantaneously, there can be excess deadhead miles, which will be discussed in the following subsection.

2.4. Pooling ratio and extra mileage ratio

We assume that, with the pooling services (e.g., UberPool, Lyft Shared/Lyft Line), one FHV serves on average $1+\pi$ travelers at the same time, where $\pi \geqslant 0$ is called hereafter the pooling ratio. Thus, the total number of FHV seats is the product of the number of FHVs

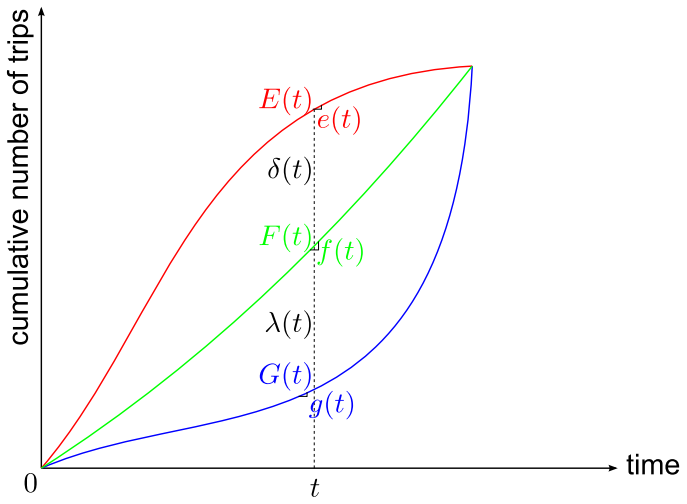

Fig. 2. Illustration of the tandem waiting and traveling queues in a shared mobility system with FHVs.

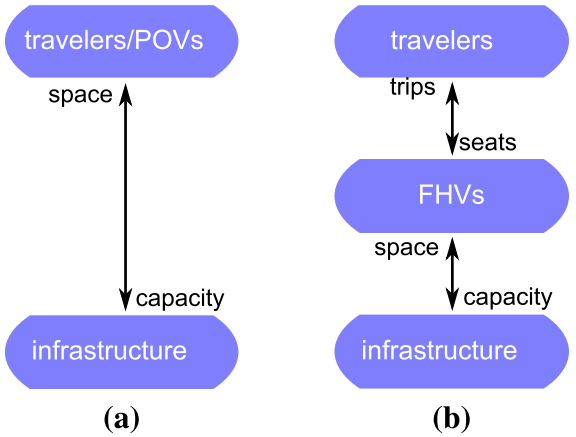

Fig. 3. Demand and supply in different mobility systems with (a) privately operated vehicles (POVs) and (b) for-hire vehicles (FHVs).

and $1+\pi$. Intuitively, the pooling ratio depends on the fares charged by TNCs, passengers' socioeconomic characteristics, TNCs' matching methods, and other factors. Here, we assume that π is exogenously observed or given. In particular, we can calculate the average value of π by comparing the total actual trip miles of all travelers and the total non-empty vehicle miles traveled, since the ratio of the former to the latter equals $1+\pi$.

When waiting passengers and available FHVs cannot be matched instantaneously, when the matched FHV and passenger are a distance apart from each other, and when an FHV is removed from a newtork after dropping off the passengers, there can induce some deadhead miles. In addition, FHVs may have to take detours to facilitate ride-sharing. Intuitively, these extra miles traveled by FHVs will increase traffic congestion and are equivalent to extra passenger trips' distances. Therefore, we introduce another exogenous parameter, the average extra mileage ratio for all trips, denoted by ξ. In this sense, a trip whose POV distance[2] is x has an FHV distance of $y=(1+\xi)x$ in the shared mobility system served by FHVs. In other words, a passenger trip with a length of x leads to a vehicular trip with a length of $(1+\xi)x$. ξ is affected by the distribution of both the demand and the supply, the size of the network, and the matching algorithm, among other factors. Here, we define it exogenously. We can estimate ξ by comparing the total vehicle-miles traveled of FHVs and the total POV trip distances, since the ratio of the total vehicle-miles traveled to the total POV trip distances equals $\frac{1+\xi}{1+\pi}$.

3. Compartmental model for shared mobility systems with for-hire vehicles

Here, we present a complete compartmental model for shared mobility systems with for-hire vehicles, illustrated in Fig. 1(b), by deriving two differential equations for the dynamics of waiting and traveling trips.

[2] Here the POV distance of a trip is assumed to be that served by a POV.

From the definitions of the variables, we have the following relations:

$$f(t) = \dot{F}(t), \tag{1a}$$

$$g(t) = \dot{G}(t). \tag{1b}$$

In addition, the waiting queue size is given by

$$\delta(t) = E(t) - F(t), \tag{1c}$$

and the number of traveling trips by

$$\lambda(t) = F(t) - G(t). \tag{1d}$$

The unknown variables are $F(t)$,$G(t)$,$f(t)$,$g(t)$,$\delta(t)$, and $\lambda(t)$, among which only two are independent due to the four relations in (1).

3.1. Point queue model for waiting trip dynamics

From (1), we have the following differential equation for the dynamics of the waiting queue size:

$$\dot{\delta}(t) = e(t) - f(t), \tag{2}$$

where $e(t)$ is given, but $f(t)$ unknown.

For the waiting trips, the demand rate of seats is given by (Jin, 2015)

$$d(t) = \frac{\delta(t)}{\epsilon} + e(t), \tag{3}$$

where ϵ is an infinitesimal positive number and equals the time-step size, Δt, in the discrete version.[3] Thus, the demand of seats during the time interval between t and $t+\Delta t$ is $\delta(t) + e(t)\Delta t$, which equals the waiting queue size plus the number of entering trips. That is, $d(t)$ is the number of travelers that are waiting to board FHVs per unit time.

For a road network with L lane-miles, the number of FHVs, i.e., the fleet size, at t is $L\rho(t)$, where $\rho(t)$ is the density of FHVs (unit: vehicles per mile per lane), and the number of seats is $(1+\pi)L\rho(t)$. We denote the number of empty (unassigned) seats by $\varphi(t)$, which is given by

$$\varphi(t) = (1+\pi)L\rho(t) - \lambda(t). \tag{4}$$

When $\varphi(t) > 0$, the average number of travelers on each FHV is fewer than $1+\pi$; and when $\varphi(t) = 0$, all FHVs are fully utilized. By fully utilized we mean that FHVs are all either in use (with passenger(s) inside), or on route to pick up someone (assigned), or in the process of being removed from a network after dropping off passengers. From t to $t+\Delta t$, the number of available seats equal $\varphi(t)$ plus those recently vacated by completed trips, $g(t)\Delta t$, and those due to the net change of FHVs in the system, $(1+\pi)L\dot{\rho}(t)\Delta t$. If we define by $s(t)$ the supply rate of available seats at t, which is the maximum number of travelers that FHVs can accommodate per unit time, then we have

$$s(t) = \frac{\varphi(t)}{\epsilon} + g(t) + (1+\pi)L\dot{\rho}(t). \tag{5}$$

Here $g(t)$ is determined by the traveling trip dynamics discussed in the following subsection, and $\dot{\rho}(t)$ is determined by the fleet-size management schemes. Depending on whether FHVs are added to or removed from a road network, $\dot{\rho}(t)$ can take any sign. When $\dot{\rho}(t) = 0$, the number of FHVs is constant over time.[4]

Then the boarding rate is given by the minimum of $d(t)$ and $s(t)$. From (3) and (5) we have

$$f(t) = \min\{d(t), s(t)\} = \min\left\{\frac{\delta(t)}{\epsilon} + e(t), \frac{\varphi(t)}{\epsilon} + g(t) + (1+\pi)L\dot{\rho}(t)\right\}. \tag{6}$$

Note that in this point queue model, (2) with (6), the supply rate, $s(t)$, varies in time. Therefore, it is different from the point queue of a single bottleneck considered in Vickrey (1969), where the capacity is constant.

3.2. Bathtub model for traveling trip dynamics

For the shared mobility system with FHVs, we assume the following network fundamental diagram for the speed-density relation

[3] Mathematically, the compartmental model is well-defined in the hyperreal domain. A rigorous mathematical treatment can be done in nonstandard analysis (Robinson, 1996), but it is beyond the scope of the paper.

[4] Note that, in contrast, the fleet size in a bus or metro system can be relatively constant during the peak periods.

$$v(t) = V(\rho(t)), \tag{7}$$

where $v(t)$ is the average travel speed in the network. Here we assume that the boarding and alighting times and the impacts of other factors on the speed are already captured by the speed-density relation. We denote by $z(t)$ the characteristic travel distance of the whole system:

$$z(t) = \int_0^t v(s)ds. \tag{8}$$

For a signalized network, we use the following speed-density relation, which leads to a trapezoidal flow-density relation (Jin and Yu, 2015; Ambühl et al., 2020):

$$V(\rho) = \frac{C}{\rho} \cdot \min\left\{\frac{\rho}{\rho_{c1}}, 1, \frac{\rho_j - \rho}{\rho_j - \rho_{c2}}\right\}, \tag{9}$$

where C is the road network capacity, determined by the green ratio and other signal settings, ρ_j is the jam density, and ρ_{c1} and ρ_{c2} are two critical densities. When ρ is between ρ_{c1} and ρ_{c2}, the flow-rate is C. For $\rho < \rho_{c1}$, traffic is under-saturated; for $\rho \in [\rho_{c1}, \rho_{c2}]$, traffic is saturated; and for $\rho > \rho_{c2}$, traffic is over-saturated. The corresponding flow-density relation is illustrated in Fig. 4(a).

We denote the proportion of the trips entering the travel compartment at t with the POV distances not smaller than x by $\widetilde{\Phi}(t,x)$. Then the distribution of the FHV distances is $\widetilde{\Phi}\left(t, \frac{y}{1+\xi}\right)$. If the average POV distance of trips entering the travel compartment at t is $\widetilde{B}(t)$, the average FHV distance is $(1+\xi)\widetilde{B}(t)$. We further denote by $K(t,y)$ the number of traveling trips at t with a remaining FHV distance not smaller than y. Then, the number of traveling trips is $\lambda(t) = K(t,0)$; and from the conservation of trips with a remaining FHV distance not smaller than y we have the following generalized bathtub model for traveling trip dynamics (Jin, 2020)

$$\frac{\partial}{\partial t}K(t,y) - V(\rho(t))\frac{\partial}{\partial y}K(t,y) = f(t)\widetilde{\Phi}\left(t, \frac{y}{1+\xi}\right), \tag{10}$$

whose discrete version is $K(t+\Delta t, y) = K(t, y + v(t)\Delta t) + f(t)\widetilde{\Phi}(t, \frac{y}{1+\xi})\Delta t$. Here the completion rate is

$$g(t) = V(\rho(t))\frac{\partial}{\partial y}K(t,0). \tag{11}$$

Further from (1) we can solve the traveling trips, $\lambda(t)$, by

$$\dot{\lambda}(t) = f(t) - g(t), \tag{12}$$

which can also be derived from (10) by setting $y = 0$. Assuming the network is initially empty, we have the integral version of the generalized bathtub model:

$$\lambda(t) = \int_0^t f(s)\widetilde{\Phi}\left(s, \frac{z(t) - z(s)}{1+\xi}\right)ds. \tag{13}$$

It is obvious that $\lambda(t)$ is guaranteed to be non-negative for non-negative $f(t)$.

If we assume that all trips' POV distances follow the same time-independent negative exponential distribution:

$$\widetilde{\Phi}(t,x) = e^{-\frac{x}{\widetilde{B}}}, \tag{14}$$

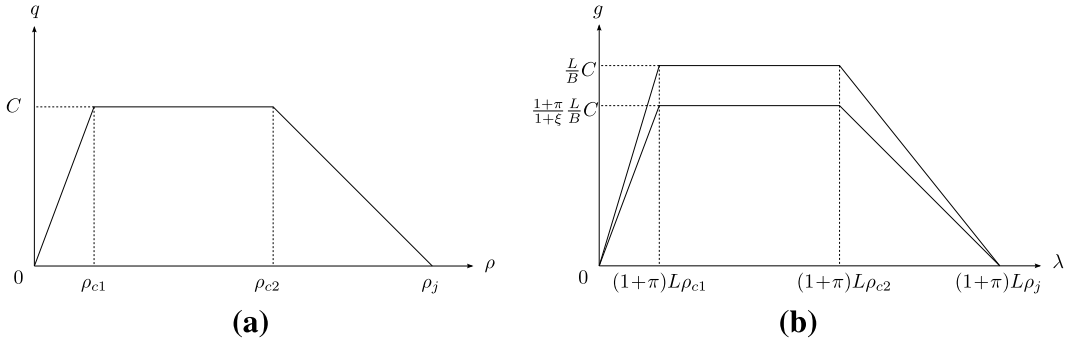

Fig. 4. Trapezoidal network fundamental diagram: (a) Relationship between FHV flow-rate and FHV density; (b) Relationship between trip completion rate and number of traveling trips.

where B is the average POV trip distance; then the average FHV trip distance is $(1+\xi)B$. In this case, $K(t,y)=\lambda(t)e^{-\frac{y}{(1+\xi)B}}$ and $\lambda(t)$ satisfies the following ordinary differential equation

$$\dot{\lambda}(t)=f(t)-\frac{1}{(1+\xi)B}V(\rho(t))\lambda(t), \tag{15}$$

which is an extension of the original bathtub model by (Vickrey, 1991; Vickrey, 2020; Small and Chu, 2003; Daganzo, 2007). Here

$$g(t)=\frac{1}{(1+\xi)B}V(\rho(t))\lambda(t). \tag{16}$$

Correspondingly, (13) can be simplified as

$$\lambda(t)=\int_0^t f(s)e^{-\frac{z(t)-z(s)}{(1+\xi)B}}ds. \tag{17}$$

When all trips have the same POV distance of B, the generalized bathtub model is the basic bathtub model considered in Arnott et al. (2016), Arnott and Buli (2018), Jin (2020). In this case,

$$\widetilde{\Phi}(t,x)=H(B-x)=\begin{cases}0, & x>B;\\ 1, & x\leqslant B,\end{cases} \tag{18}$$

where $H(\cdot)$ is the Heaviside function. Since the remaining FHV distance of any traveling trip $y\leqslant(1+\xi)B$, (10) can be simplified as ($y\in[0,(1+\xi)B]$):

$$\frac{\partial}{\partial t}K(t,y)-V(\rho(t))\frac{\partial}{\partial y}K(t,y)=f(t). \tag{19}$$

Correspondingly, the integral forms can be simplified as

$$\lambda(t)=F(t)-F(\tau(z(t)-(1+\xi)B)), \tag{20}$$

where $\tau(z)$ is the characteristic travel time and an inverse function of $z(t)$ for $v(t)>0$ (Jin, 2020).

3.3. Complete PWTC compartmental model

Combining (2), (6), and (12), we have the following differential version of the PWTC compartmental model of shared mobility systems with FHVs:

$$\dot{\delta}(t)=\max\Big\{-\frac{\delta(t)}{\epsilon},e(t)-\frac{\varphi(t)}{\epsilon}-g(t)-(1+\pi)L\dot{\rho}(t)\Big\}, \tag{21a}$$

$$\dot{\lambda}(t)=\min\Big\{\frac{\delta(t)}{\epsilon}+e(t)-g(t),\frac{\varphi(t)}{\epsilon}+(1+\pi)L\dot{\rho}(t)\Big\}, \tag{21b}$$

where the number of unused seats $\varphi(t)$ can be written in terms of the two state variables as in (4), and the completion rate $g(t)$ depends on the number of traveling trips at t as in (11), in which $\frac{\partial}{\partial y}K(t,0)$ is determined by (10). From (4) we have $\dot{\varphi}(t)=(1+\pi)L\dot{\rho}(t)-\dot{\lambda}(t)$, which combined with (21b) leads to

$$\dot{\varphi}(t)=\max\Big\{-\frac{\varphi(t)}{\epsilon},-\frac{\delta(t)}{\epsilon}-e(t)+g(t)+(1+\pi)L\dot{\rho}(t)\Big\}. \tag{22}$$

Therefore, for a general trip distance distribution, the PWTC compartmental model comprises of five equations, (4), (10), (11), and (21), with six unknown variables, $\delta(t),\lambda(t),\varphi(t),g(t),K(t,y)$, and $\rho(t)$. Among the five equations, (10) is a partial differential equation, (21) are two ordinary differential equations, and (4) and (11) are two algebraic equations. Here $\rho(t)$ and $\dot{\rho}(t)$ are given by the fleet-size management scheme; therefore, they the control variables in the sense of control theory (Aström and Murray, 2008). Hence the PWTC compartmental model is an open-loop system. Notice that increasing the empty seats ($\varphi(t)=(1+\pi)L\rho(t)-\lambda(t)$) increases the supply $s(t)$ and, therefore, reduces the waiting queue size, but also reduces the speed $v(t)$ and, therefore, increases the number of traveling trips. Thus, there is a trade-off that needs to be considered for the fleet-size management.

For the PWTC compartmental model, $\delta(0)$ and $\lambda(0)$ are given by the initial conditions, and $e(t)$ by the boundary conditions. We have the following definition.

Definition 3.1. A PWTC compartmental model for the shared mobility system with FHVs is well-defined if $f(t),g(t),\delta(t),\lambda(t)\geqslant 0$ for $e(t),\delta(0),\lambda(0)\geqslant 0$.

The definition is self-evident for a physically meaningful model.

Lemma 3.2. *In the PWTC compartmental model, $\delta(t)$ and $\varphi(t)$ are always non-negative for $\delta(0), \varphi(0) \geqslant 0$. In addition, when $\varphi(0) = 0$ and the fleet-size management scheme satisfies*

$$(1+\pi)L\dot{\rho}(t) \leqslant e(t) + \frac{\delta(t)}{\epsilon} - g(t) - \frac{\varphi(t)}{\epsilon}, \tag{23}$$

the FHVs are fully utilized with $\varphi(t) = 0$.

Proof. Discretizing (21a) with $\epsilon = \Delta t$ we have $\delta(t+\Delta t) = \delta(t) + \max\{-\delta(t), \cdots\} = \max\{0, \cdots\} \geqslant 0$. By induction, we can see that, if $\delta(0) \geqslant 0$, we have $\delta(t) \geqslant 0$ at any time t. Similarly, from the discrete version of (22), we can prove $\varphi(t) \geqslant 0$ at any time t.

When (23) is satisfied, from (22) we have $\dot{\varphi}(t) = -\frac{\varphi(t)}{\epsilon}$, which leads to $\varphi(t+\Delta t) = 0$. Thus $\varphi(t) = 0$ for any t, given $\varphi(0) = 0$. □

Lemma 3.2 implies that if the drivers of some empty FHVs decide to cruise in the network, then (23) is violated, and they become under-utilized. In contrast, if unassigned FHVs leave the roads, (23) is satisfied, and all FHVs are fully utilized. Therefore, any fleet-size management schemes satisfying (23) would prevent unassigned FHVs from contributing to congestion; in this study, we focus on such schemes.

Theorem 3.3. *If and only if the fleet-size management scheme satisfies*

$$(1+\pi)L\dot{\rho}(t) \geqslant -\frac{\varphi(t)}{\epsilon} - g(t), \tag{24}$$

$s(t), f(t), \lambda(t), \rho(t), g(t)$ are all non-negative for $e(t), \lambda(0) \geqslant 0$. That is, (24) is the sufficient and necessary condition for the PWTC compartmental model of shared mobility systems with FHVs to be well-defined in the sense of Definition 3.1. Intuitively, (24) makes sense because the maximum number of seats that can be taken out of the system equals $\frac{\varphi(t)}{\epsilon} + g(t)$.

Proof. From the definition of $s(t)$ in (5), the condition on the fleet-size management scheme, (24), leads to $s(t) \geqslant 0$.

Then from (6) we have $f(t) \geqslant 0$, since $s(t) \geqslant 0$, and $\delta(t) \geqslant 0$ from Lemma 3.2.

From (13), $\lambda(t) \geqslant 0$, since $f(t) \geqslant 0$.

From (4), we have $(1+\pi)L\rho(t) = \lambda(t) + \varphi(t) \geqslant 0$, since $\lambda(t) \geqslant 0$ and $\varphi(t) \geqslant 0$ from Lemma 3.2.

From the definition of $K(t,y)$, we can see that $\frac{\partial}{\partial y}K(t,0) \geqslant 0$ when $\lambda(t) \geqslant 0$. Thus, from (11), we have $g(t) \geqslant 0$.

On the other hand, if (24) is violated, then $s(t) < 0$, and $f(t) < 0$.

Therefore, (24) is the sufficient and necessary condition for the PWTC compartmental model to be well-defined. □

In the special case of a time-independent negative exponential distribution of trip distances, $g(t)$ is determined by (16). Then, the PWTC compartmental model can be further simplified as

$$\dot{\delta}(t) = \max\left\{-\frac{\delta(t)}{\epsilon}, e(t) - \frac{(1+\pi)L\rho(t) - \lambda(t)}{\epsilon} - \frac{V(\rho(t))}{(1+\xi)B}\lambda(t) - (1+\pi)L\dot{\rho}(t)\right\}, \tag{25a}$$

$$\dot{\lambda}(t) = \min\left\{\frac{\delta(t)}{\epsilon} + e(t) - \frac{V(\rho(t))}{(1+\xi)B}\lambda(t), \frac{(1+\pi)L\rho(t) - \lambda(t)}{\epsilon} + (1+\pi)L\dot{\rho}(t)\right\}. \tag{25b}$$

In this case, the PWTC compartmental model is a system of ordinary differential equations, with two variables, $\delta(t)$ and $\lambda(t)$, and one control variable, $\rho(t)$. For simplicity, hereafter we only consider this special case.

4. Feedback fleet-size management schemes to limit waiting queue sizes

In this section, we consider two fleet-size management schemes aiming to limit the waiting queue size, as a proxy for the wait times. The first one is the POV scheme, in which FHVs are operated as POVs. The second one is an approximation of the current TNCs' fleet-size management scheme, aiming to guarantee short wait times. We refer to this second scheme as short waiting queue scheme.

4.1. The POV scheme

When FHVs are operated as POVs, a traveler is immediately assigned to an FHV when s/he sends a request (i.e. zero wait time), and the FHV is removed from the network once the trip is completed. However, the assigned FHV may need to travel a certain distance to pick up the user, which is captured by ξ. Thus, we obtain the following feedback fleet-size management scheme

$$(1+\pi)L\dot{\rho}(t) = \min\left\{\left(1+\pi\right)L\frac{\rho_j - \rho(t)}{\epsilon}, e(t) + \frac{\delta(t)}{\epsilon} - g(t) - \frac{\varphi(t)}{\epsilon}\right\}. \tag{26}$$

The first term on the right-hand side of (26) guarantees that $\rho(t) \leqslant \rho_j$ with ρ_j as the jam density. The second term suggests that we add as many as FHVs to accommodate the demand in seats, $e(t) + \frac{\delta(t)}{\epsilon}$, and remove as many as FHVs for completed trips and empty seats, $g(t) + \frac{\varphi(t)}{\epsilon}$. Here we assume that $e(t)$ could be predicted (e.g., based on historical data or some type of real-time information) to use (26).

Notice that if $e(t)$ is unknown, another integral controller is

$$(1+\pi)L\dot{\rho}(t) = \min\left\{\left(1+\pi\right)L\frac{\rho_j-\rho(t)}{\epsilon}, \alpha\delta(t) - g(t) - \frac{\varphi(t)}{\epsilon}\right\} \tag{27}$$

with $\alpha \in (0, \frac{1}{\epsilon}]$. This fleet-size management scheme still satisfies which also satisfies (23), (24), and $\rho(t) \leqslant \rho_j$, but is less efficient in the sense that the waiting queue size approaches zero slower.

Clearly, (26) satisfies (24), and the closed-loop control system, (21) and (26), is well-defined, according to Theorem 3.3. (26) also satisfies (23), and, from Lemma 3.2 we have $\varphi(t) = 0$, and all FHVs are fully utilized. In addition, we have the following theorem, whose proof is straightforward and omitted.

Theorem 4.1. *In the closed-loop control system, (21) and (26), the waiting queue size is zero with $\delta(t) = 0$. Therefore, with the POV scheme, the closed-loop control system is equivalent to the bathtub model:*

$$\dot{\lambda}(t) = e(t) - g(t), \tag{28}$$

where $g(t)$ is given by (11). In this system, the only state variable is $\lambda(t)$. Therefore, the shared mobility system with FHVs operates like the traditional mobility system with POVs, with the pooling ratio π and the extra mileage ratio ξ.

4.2. The short waiting queue scheme

One may argue that the TNCs would not use the POV scheme or ensure zero waiting queue size. However, in order to remain competitive with POVs which have no wait time, TNCs do attempt to ensure "short wait times to keep demand high" (NYC Taxi and Limousine Commission, 2019). Since ensuring a low enough wait time can be considered equivalent to ensuring short enough queue length on the waiting compartment, we add FHVs at the rate of $\max\{0, e(t) + \frac{\delta(t)-\delta_0}{\epsilon}\}$ with $\delta_0 > 0$ and remove vehicles at the rate of $g(t) + \frac{\varphi(t)}{\epsilon}$. We then obtain the following feedback fleet-size management scheme:

$$(1+\pi)L\dot{\rho}(t) = \min\left\{\left(1+\pi\right)L\frac{\rho_j-\rho(t)}{\epsilon}, \max\left\{0, e(t) + \frac{\delta(t)-\delta_0}{\epsilon}\right\} - g(t) - \frac{\varphi(t)}{\epsilon}\right\}, \tag{29}$$

which also satisfies (23), (24), and $\rho(t) \leqslant \rho_j$. Thus from Lemma 3.2 we have $\varphi(t) = 0$, and the FHVs are fully utilized as in the POV scheme. The closed-loop system formed by (21) and (29) for $\rho(t) \leqslant \rho_j$ can be re-written as

$$\dot{\delta}(t) = \max\left\{-\frac{\delta(t)}{\epsilon}, \min\left\{e(t), \frac{\delta_0-\delta(t)}{\epsilon}\right\}\right\}, \tag{30a}$$

$$\dot{\lambda}(t) = \max\left\{0, e(t) + \frac{\delta(t)-\delta_0}{\epsilon}\right\} - g(t). \tag{30b}$$

The closed-loop system reaches an equilibrium when $\dot{\delta}(t) = 0, \dot{\lambda}(t) = 0$, and $\dot{\rho}(t) = 0$. In an equilibrium state, therefore, $\delta(t) = \delta_0, \lambda(t) = \lambda_0$, and $\rho(t) = \rho_0$. From (30b) we have $g(t) = e(t)$; from (29) we have $\varphi(t) = 0$; i.e., there are no empty seats. Further, from (12) we have $f(t) = g(t)$. Thus, the equilibrium solution is almost the same as that for the POV scheme, except that the waiting queue size is non-zero, but δ_0. Therefore, both zero or short waiting queue schemes lead to the POV-like scheme, in which the waiting queue size is fixed, $\varphi(t) = 0$, and $f(t) = e(t)$.

Around the equilibrium, with a small perturbation in the three variables, denoted by $\widetilde{\delta}(t) = \delta(t) - \delta_0, \widetilde{\lambda}(t) = \lambda(t) - \lambda_0$, and $\widetilde{\rho}(t) = \rho(t) - \rho_0$, then we have from (30) and (29)

$$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\delta}(t) \approx -\frac{\widetilde{\delta}(t)}{\epsilon}, \tag{31}$$

$\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\lambda}(t) \approx \frac{\widetilde{\delta}(t)}{\epsilon}$, and $(1+\pi)L\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{\rho}(t) \approx \frac{\widetilde{\delta}(t)}{\epsilon}$. From (31), we can see that the waiting queue size is indeed finite-time stable, since a small disturbance, $\widetilde{\delta}(t)$, recovers to 0 in one time-step. Hence, the traveling trips and the fleet size are also finite-time stable.

4.3. System performance with the POV-like fleet-size management schemes

In this subsection we analyze the performance of the closed-loop system with the POV scheme and negative exponential distribution of trip distances. For such case, the closed-loop control system can be simplified as

$$\dot{\lambda}(t) = e(t) - \frac{1}{(1+\xi)B}V\left(\frac{\lambda(t)}{(1+\pi)L}\right)\lambda(t). \tag{32}$$

For the trapezoidal network fundamental diagram, as shown in Fig. 4(a), the corresponding relationship between the completion rate, $g(t) = \frac{1}{(1+\xi)B}V\left(\frac{\lambda(t)}{(1+\pi)L}\right)\lambda(t)$, and the number of traveling trips, $\lambda(t)$, is shown in Fig. 4(b). From the figure, we can see that the pooling

ratio and the extra mileage ratio change the maximum completion rate from $\frac{L}{B}C$ to $\frac{1+\pi}{1+\xi}\frac{L}{B}C$ and the trip capacity from C to $\frac{1+\pi}{1+\xi}C$; this is true for any network fundamental diagram. Evidently, if the increase in the pooling ratio overcompensates for the increase in the mileage ratio, i.e., $\pi > \xi$, the maximum completion rate of trips increases. Otherwise, it decreases.

In addition, (32) can be re-written as

$$\dot{\lambda}(t) = e(t) - \frac{1+\pi}{1+\xi}\frac{L}{B}C \cdot \min\left\{\frac{\lambda(t)}{(1+\pi)L\rho_{c1}}, 1, \frac{(1+\pi)L\rho_j - \lambda(t)}{(1+\pi)L(\rho_j - \rho_{c2})}\right\}.$$

For a constant entering rate, $e(t) = e$, we have the following observations:

- When $e < \frac{1+\pi}{1+\xi}\frac{L}{B}C$, the system can have two equilibrium points: $\lambda(t) = \lambda_1 = (1+\xi)\frac{e}{C}B\rho_{c1}$, and $\lambda_2 = (1+\pi)L\rho_j - (1+\xi)\frac{e}{C}B(\rho_j - \rho_{c2})$. It is easy to show that the first under-saturated equilibrium point is stable, but the second over-saturated one is unstable, and a small increase in the number of trips leads the system to the gridlock state.
- When $e = \frac{1+\pi}{1+\xi}\frac{L}{B}C$, the system has many possible equilibrium points between $(1+\pi)L\rho_{c1}$ and $(1+\pi)L\rho_{c2}$.
- When $e > \frac{1+\pi}{1+\xi}\frac{L}{B}C$, or equivalently,

$$(1+\xi)Be > (1+\pi)LC, \tag{33}$$

the system converges to the gridlock state with $\lambda = (1+\pi)L\rho_j$ and $v = 0$. The left-hand side of (33) represents the demand, and the right-hand side the supply (Jin, 2020). In this sense, the deadhead and detour miles effectively increase the demand by ξ.

Therefore, with the POV scheme, the shared mobility system is easier to get gridlocked, when the extra mileage ratio ξ is larger than the pooling ratio π.

For the short waiting queue scheme, $\delta(t) = \delta_0$ during the peak period, and the compartmental model is also equivalent to the bathtub model, (28). The above results also apply.

Therefore, if the fleet-size management scheme only aims to limit the wait time, as in (26) or (29), the shared mobility system with FHVs operates like the traditional mobility system with HOVs, except with the pooling ratio of π, and the extra mileage ratio of ξ. Furthermore, if $\xi > \pi$, such a shared mobility system with FHVs is more congested than the traditional mobility system with HOVs. That is, FHVs can lead to more congestion, even if all POVs are replaced by FHVs. Conceptually, this explains why more congestion has been observed after the introduction of FHVs (Schaller, 2018).

5. Improved fleet-size management scheme with capping

In the preceding section, we showed that, if we only attempt to minimize the wait times, the shared mobility system leads to similar or worse congestion patterns than the ones caused by POVs when the pooling ratio is smaller than the extra mileage ratio. In addition, we demonstrated that the system could become gridlocked with the POV-like schemes, since the maximum value of $\rho(t)$ is the jam density, ρ_j. That is, there is no cap on the FHV fleet size. In this section, we analyze the effect of capping the fleet size based on the compartmental model.

5.1. Feedback fleet-size management scheme with capping

Intuitively, if all trips are served by FHVs, it is possible to prevent the occurrence of gridlock or over-saturation by limiting the number of FHVs; i.e., by introducing a cap, ρ^*, such that

$$\rho(t) \leqslant \rho^*. \tag{34}$$

This insight has been verified by simulations in Fagnant and Kockelman (2014), Martinez et al. (2015), Bischoff and Maciejewski (2016), which showed that it is possible to serve all trips with a much smaller fleet size than that of the POVs, as long as we allow for some wait times.

To minimize the total cost, when the fleet size is below the cap ρ^* the desired waiting queue size should be zero. Therefore, we revise (26) as

$$(1+\pi)L\dot{\rho}(t) = \min\left\{\left(1+\pi\right)L\frac{\rho^* - \rho(t)}{\epsilon}, e(t) + \frac{\delta(t)}{\epsilon} - g(t) - \frac{\varphi(t)}{\epsilon}\right\}, \tag{35}$$

which satisfies (23), (24), and $\rho(t) \leqslant \rho_j$. Thus from Lemma 3.2 we have $\varphi(t) = 0$, and the FHVs are fully utilized. The closed-loop system can be re-written as

$$\dot{\delta}(t) = \max\left\{-\frac{\delta(t)}{\epsilon}, e(t) - g(t) - (1+\pi)L\frac{\rho^* - \rho(t)}{\epsilon}\right\}, \tag{36a}$$

$$\dot{\lambda}(t) = \min\left\{\frac{\delta(t)}{\epsilon} + e(t) - g(t), (1+\pi)L\frac{\rho^* - \rho(t)}{\epsilon}\right\}. \tag{36b}$$

With constant entering rate $e(t) = e$, this system can have the following two types of equilibrium states:

1. In the first type of equilibrium states, the fleet size is smaller than the cap; i.e., $\rho(t) = \rho < \rho^*$. We have from (35) that $\delta(t) = \delta = 0$ and $f(t) = g(t) = e$. This equilibrium is similar to the one discussed in Section 4.2 and, thus, finite-time stable.
2. In the second type of equilibrium states, the fleet size equals the cap; i.e., $\rho(t) = \rho^*$. This occurs when the travel demand is high with $e > f(t) = g(t)$, and the waiting queue size keeps increasing. Thus, (35) leads to

$$(1+\pi)L\dot{\rho}(t) = (1+\pi)L\frac{\rho^* - \rho(t)}{\epsilon}.$$

Thus, $\rho(t)$ and $\lambda(t) = (1+\pi)L\rho(t)$ are exponentially stable around the equilibrium state. Note that the waiting queue size, $\delta(t)$, keeps increasing.

5.2. Method for determining the optimal fleet-size cap

In the traditional mobility system with POVs, which can be described by Vickrey's bathtub, the system can reach a departure time user equilibrium during the peak period. As shown in Small and Chu (2003), the network needs to be saturated or even over-saturated in order for traveling costs to compensate the differences in the scheduling costs. Thus, the POV-like schemes without capping would lead to worse congestion with $\xi > \pi$ in the departure time user equilibrium.

Here, we propose to determine the optimal fleet-size by minimizing the total cost in the departure time user equilibrium during the peak period. For simplicity, we assume that the fleet size equals the maximum value of $\rho(t) = \rho^*$ during the whole peak period. Obviously, all FHVs should be fully utilized with $\lambda(t) = (1+\pi)L\rho^*$ is constant, and the traveling compartment is in a steady state, where the completion rate $g(t) = g^*$ is constant:

$$g^* = \frac{1+\pi}{1+\xi}\frac{L}{B}V(\rho^*)\rho^*. \tag{37}$$

The travel speed is the same for all the trips $V(\rho^*)$, and the average traveling time is $\frac{B}{V(\rho^*)}$. In addition, (36) can be simplified as Vickrey's point queue model (Vickrey, 1969; Jin, 2015)

$$\dot{\delta}(t) = \max\left\{-\frac{\delta(t)}{\epsilon}, e(t) - g^*\right\}, \tag{38}$$

where $e(t)$ is determined by the departure time choice and the waiting queue size is the only state variable.

We consider the following cost function for travelers

$$\phi(t) = \beta\Upsilon(t) + \gamma\frac{B}{V(\rho^*)} + \mu\max\left\{t^* - \frac{B}{V(\rho^*)} - t, 0\right\} + \nu\max\left\{t - t^* + \frac{B}{V(\rho^*)}, 0\right\}, \tag{39}$$

where $\Upsilon(t) = \frac{\delta(t-\Upsilon(t))}{g^*}$ is the wait time for a traveler boarding at t, t^* the ideal arrival time, β the cost coefficient for the wait time, γ the cost coefficient for the traveling time, μ the penalty for early arrivals, and ν the penalty for late arrivals. In the departure time user equilibrium, the waiting cost and the early and late arrival penalties are balanced, and all travelers have the same total cost:

$$\phi^* = \frac{N}{(\frac{1}{\mu}+\frac{1}{\nu})g^*} + \gamma\frac{B}{V(\rho^*)}. \tag{40}$$

The solutions are illustrated in Fig. 5. Notice that these cumulative flow functions are only an approximation. They are accurate when the system reaches the equilibrium with $L\rho^*$ vehicles. However, at the beginning (during congestion formation) or at the end (during congestion alleviation) the FHVs fleet size is $\rho(t) < \rho^*$, and the constant flow is only an approximation.

We define the following optimization problem to find the optimal fleet-size cap:

$$\min_{\rho^*}\phi^* = \frac{N}{(\frac{1}{\mu}+\frac{1}{\nu})g^*} + \gamma\frac{B}{V(\rho^*)}. \tag{41}$$

Theorem 5.1. *For the trapezoidal network fundamental diagram defined in (9), the optimal FHV density is* ρ_{c1}.

Proof. When $\rho \leqslant \rho_{c1}$, a larger ρ leads to a larger g^* but the same speed; thus the cost at ρ_{c1} is the smallest. When $\rho \in [\rho_{c1}, \rho_{c2}]$, increasing

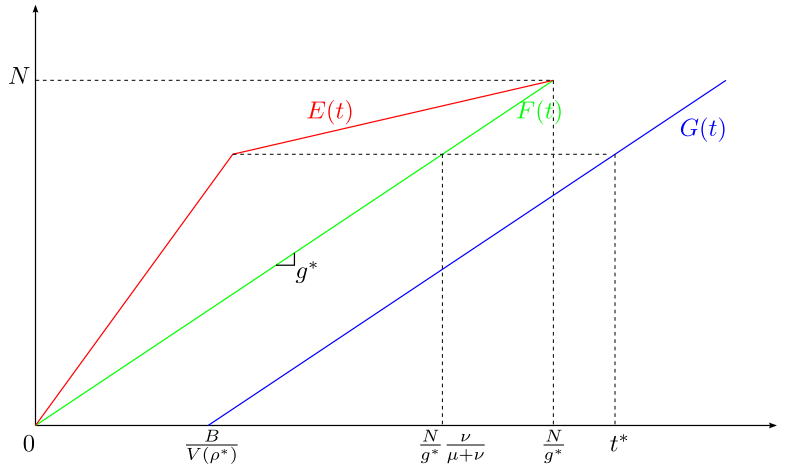

Fig. 5. Solution of the departure time user equilibrium in the shared mobility system with a fixed number of FHVs.

ρ leads to the same g^* but smaller speeds, and, thus larger costs. When $\rho \geqslant \rho_{c2}$, increasing ρ leads to smaller g^* and lower speeds, and, thus larger costs. Therefore, the cost at ρ_{c1} is the smallest. □

For example, the total lane-miles of New York City were about 19,000 in 2018 [5], and the number of registered vehicles was about five million in 2016 [6]. That is, if all trips are served by FHVs, Theorem 5.1 shows that the optimal number of vehicles is about one million (assuming $\rho_j = 230$ vpmpl, and $\rho_{c1} = 0.25\rho_j$). As it is not viable economically to hire one million drivers to serve the city's travel demand, it can only occur in the era of autonomous vehicles for 100% of the trips to be served by FHVs. However, it is viable to replace five million POVs by one million autonomous FHVs. Therefore, it is reasonable to assume that "the number of FHVs is as large as determined by the fleet-size management scheme", as in Section 2.3.

Theorem 5.1 is not only relevant for policy making, but it also reveals an important property of the optimal FHV fleet size when considering trapezoidal network fundamental diagrams. The optimal FHV fleet size is independent of the extra mileage ratio, ξ. This parameter does affect the total cost for the users, since higher deadhead and detour miles reduce the completion rate of trips (37) but not the optimal size of the fleet. In other words, ρ^* is given by the capacity of the network, and these results hold independently of any approximation we might use to describe the deadhead and detour miles.

For any general flow-density relationship, the solution of the optimization problem (41) cannot be in the over-saturated regime, where both the completion rate and the speed decrease as a function of density. Therefore, the feedback fleet-size management scheme based on the optimal fleet-size cap can help to prevent over-saturation or hypercongestion (Small and Chu, 2003). Notice that μ and γ do not influence the solution of ρ^* for a trapezoidal flow-density relationship; however, these parameters may play a role for other flow-density shapes, e.g., the Greenshields shape (Vickrey, 1991; Vickrey, 2020).

Note that the optimal fleet-size cap is obtained based on the assumption of a constant fleet size in the departure time user equilibrium. Such a solution is attractive, since, practically, such a cap is relatively easy to implement by traffic system agencies. We believe that the total travel cost in such a user equilibrium should be smaller than that in the POV user equilibrium or other user equilibria; but a rigorous proof is out of the scope of this study, since no simple analytical or numerical solution methods are available for such user equilibria in the literature.

In (39), the cost is linearly proportional to the wait time $\Upsilon(t)$. In reality, it is possible that, especially when travelers expect relatively low wait times, the cost increases nonlinearly (e.g., quadratically) in the wait time, then we can relax the constraint on the wait time or the waiting queue by revising (29) as

$$(1+\pi)L\dot{\rho}(t) = \min\left\{\left(1+\pi\right)L\frac{\rho^* - \rho(t)}{\epsilon}, \max\left\{0, e(t) + \frac{\delta(t)-\delta_0}{\epsilon}\right\} - g(t) - \frac{\varphi(t)}{\epsilon}\right\}, \tag{42}$$

which can lead to more congestion but lower wait times. That is, it is possible that the optimal cap $\rho^* > \rho_{c,1}$ in this case. In addition, if considering the initiation and dissipation stages, there can be an optimal $\delta_0 > 0$.

6. Compartmental model and fleet-size management for mixed mobility systems

It is estimated that "FHVs make up nearly 30% of all traffic" in downtown Manhattan (NYC Taxi and Limousine Commission, 2019).

[5] See https://www1.nyc.gov/office-of-the-mayor/news/254-18/pave-baby-pave-mayor-de-blasio-record-5-000-lane -miles-city-roadways-have-been-repaved#/0

[6] See https://www.statista.com/statistics/196061/number-of-registered-automobiles-in-new-york/

In this section, we extend the compartmental model for such a mixed system and discuss the corresponding fleet-size management problem.

6.1. Coupled compartmental model for mixed mobility systems with POVs and FHVs

For the mixed mobility system with two modes, FHVs and POVs, we extend the compartmental model for the FHV trip dynamics, and apply the bathtub model for the POV trip dynamics (Jin, 2020), as depicted in Fig. 1b. Both FHVs and POVs share the same road network; thus, the two dynamics are coupled together.

In the following, the variables for the FHV trips are denoted by subscript F, and those for the POV trips by subscript P. The total number of trips is $N = N_F + N_P$, where N_F and N_P are respectively the total number of trips for FHVs and POVs. Fig. 6 illustrate the coupled queuing systems for both modes. Notice that the left tandem queue in Fig. 6 is similar to that in Fig. 2, but now the completion rate of trips served by FHVs, $g_F(t)$, depends on the speed of the system which has both POVs and FHVs.

We assume that both types of vehicles have the same contribution to traffic dynamics, and the speed-density relation is

$$v(t) = V(\rho_P(t) + \rho_F(t)). \tag{43}$$

For simplicity, we assume that all trips' distances follow the same negative exponential distribution with the average distance B. Thus, the average distance of FHV trips is $(1+\xi)B$ to account for the deadhead and detour miles; but the average POV trip distance is still B.

- Each POV is assumed to serve only one traveler with $\lambda_P(t) = L\rho_P(t)$. Notice that in many places, POV drivers might also be encouraged to carpool. Therefore, this could be included in the formulation as $\lambda_P(t) = (\widehat{\pi}+1)L\rho_P(t)$, where $\widehat{\pi}$ is the average passengers pooled in POVs. The dynamics of the POV trips can be described by Vickrey's bathtub model (Vickrey, 1991; Vickrey, 2020):

$$\dot{\lambda}_P(t) = e_P(t) - \frac{1}{B}v(t)\lambda_P(t). \tag{44}$$

- The pooling ratio of FHVs is still denoted by π. The dynamics of the FHV trips can be described by the modified version of (25):

$$\dot{\delta}(t) = \max\Big\{-\frac{\delta(t)}{\epsilon}, e_F(t) - \frac{\varphi(t)}{\epsilon} - \frac{v(t)}{(1+\xi)B}\lambda_F(t) - (1+\pi)L\dot{\rho}_F(t)\Big\}, \tag{45a}$$

$$\dot{\lambda}_F(t) = \min\Big\{\frac{\delta(t)}{\epsilon} + e_F(t) - \frac{v(t)}{(1+\xi)B}\lambda_F(t), \frac{\varphi(t)}{\epsilon} + (1+\pi)L\dot{\rho}_F(t)\Big\}. \tag{45b}$$

6.2. Fleet-size management schemes for FHVs in a mixed mobility system

In the two-commodity system, we assume that the fleet size of POVs cannot be controlled, but can manage the fleet size of FHVs.

According to the analyses in Section 4, if the fleet-size management scheme aims to limit the wait time as in (29), then FHVs operate like POVs, and the compartmental model for the FHV trip dynamics can be approximated by a bathtub model. If $\xi > \pi$, introducing FHVs worsens traffic congestion in the whole road network.

Therefore, we apply the same fleet-size management scheme with capping as in (35):

$$(1+\pi)L\dot{\rho}_F(t) = \min\Big\{\Big(1+\pi\Big)L\frac{\rho_F^* - \rho_F(t)}{\epsilon}, e_F(t) + \frac{\delta(t)}{\epsilon} - g_F(t) - \frac{\varphi(t)}{\epsilon}\Big\}, \tag{46}$$

where $g_F(t) = \frac{v(t)}{(1+\xi)B}\lambda_F(t)$. Here the cap on the FHV fleet size is $L\rho_F^*$. Similarly to Section 5, the proposed fleet-size management scheme leads to $\varphi(t) = 0$.

The boundary conditions on $e_F(t)$ and $e_P(t)$, and the initial conditions on $\delta(0), \rho_F(0)$, and $\rho_P(0)$ are assumed to be given. In particular, we are interested in finding ρ_F^*, such that the total time for travelers is minimized:

$$\min_{\rho_F^*}\phi = \int_0^T (E(t) - G_F(t) - G_P(t))dt. \tag{47}$$

The total time can be calculated as the sum of the area between $E_F(t)$ and $G_F(t)$ and the area between $E_P(t)$ and $G_P(t)$ in Fig. 6. Assuming the value of time, c, is the same for all users and for the waiting and traveling compartments, then $c \cdot \phi$ represents the total cost for the travellers.

Intuitively, with different FHV fleet-sizes, the POVs could change their departure times accordingly, and travelers could also change their modes. To include these choice behaviors endogenously, one needs to solve the departure time user equilibrium for both modes and also consider fares for FHVs, car ownership and operation costs, even though the trip dynamics model in the preceding subsection

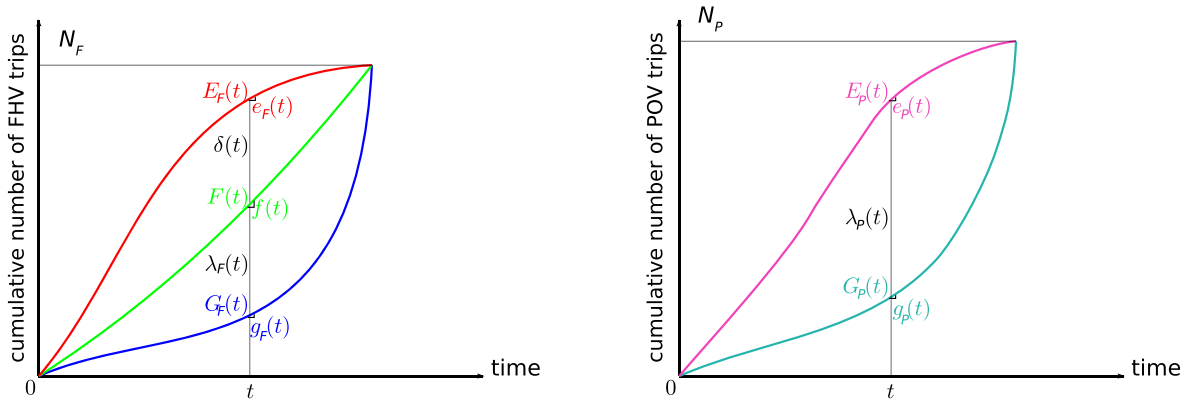

Fig. 6. Illustration of the two queuing systems in mixed environment with both FHVs and POVs. This queuing system is equivalent to the compartmental representation in Fig. 1(b).

directly applies. Such comprehensive treatments are beyond the scope of this study. Therefore, here we consider these choice behaviors exogenously[7], by assuming that $e_F(t) = \eta_F e(t)$ and $e_P(t) = \eta_P e(t)$ with given values $\eta_F + \eta_P = 1$. Note that η_P is at most the ratio of passengers that own a vehicle, while η_F represents the passengers without a POV that have to rely on FHVs to travel plus the passengers that own a POV and still decide to use FHVs. In a real application, we could obtain the demand for each mode from empirical data. To demonstrate the methodology to obtain the optimal fleet cap under a mixed environment, here we assume a trapezoidal demand. Starting from 0 veh/h, with $e(t)$ increasing for 1 h, staying stable for 1 h at e_{max}, and then decreasing again to 0 veh/h over a 1 h period. Thus, the total number of trips considered is $N = 2e_{max}$. Then, we can define the maximum demand ratio as:

$$R = e_{max}\frac{(1+\xi)B}{(1+\pi)LC}, \tag{48}$$

which depends on the maximum completion rate of trips of FHVs, as illustrated in Fig. 4(b). For $R \leq 1$ there is no need to cap the FHV fleet size, since the demand of the road network (the product of the trip rate and the effective trip distance) does not exceed the supply (the product of the road capacity per-lane, the total lane miles, and the effective vehicle occupancy). Thus we only consider cases with $R > 1$.

Even with the aforementioned simplifications, (47) cannot be analytically solved, and we resort to numerical solutions. The coupled compartmental model for a mixed mobility system can be numerically solved by the Euler forward method with $\epsilon = \Delta t$. Fig. 7 shows the total time of travellers [in minutes] as a function of the FHVs cap for different values of R and different mode choice ratios η_F. The analysis is normalized in terms of city size L and for a normalized network fundamental diagram, by C and ρ_j. In this particular case, we have chosen $\rho_{c,1} = 0.2\rho_j$ and $\rho_{c,2} = 0.5\rho_j$. From the figures, we want to highlight the following observations. The total travel time is unimodal or flat in the fleet-size cap, for each demand R and mode split considered. Thereafter, there is a unique value or interval of ρ_F^* for the minimum total cost. Caps that are either too small or too large are very detrimental to the system in terms of total cost. This demonstrates the necessity of introducing a proper fleet-size cap for FHVs, even in the mixed system.

7. Conclusion

In this paper, we presented a compartmental modeling framework to describe trip dynamics and traffic congestion in shared mobility systems with FHVs (for-hire vehicles), in which trips are categorized into four compartments: planned, waiting, traveling, and completed. We then applied the point queue model and the bathtub model to describe the dynamics of the waiting and traveling trips, respectively. Both models are coupled together by the supply of seats resulting from the traveling compartment and the fleet-size management scheme. We presented the sufficient and necessary condition on the fleet-size management scheme for the model to be well-defined, as well as the condition for FHVs to be fully utilized. In addition, we introduced two parameters to capture the pooling ratio and the extra mileage ratio caused by the FHV's deadhead and detour miles. The resulting model is a system of ordinary and partial differential equations, which can also be considered a tandem fluid queueing model.

We then discussed two types of feedback fleet-size management schemes for the shared mobility system with 100% autonomous FHVs. In all these fleet-size management strategies, the compartmental model is well-defined, and the FHVs are fully utilized, i.e. the vehicles are either carrying a passenger, assigned to pick up a passenger, or in process of being removed from the system. The first type of fleet-size management strategies are referred to as POV-like schemes, and aim to limit the waiting queue size without imposing any caps on the FHV fleet size. We demonstrated that FHVs operate like POVs, and the compartmental model is almost the same as the

[7] In reality, it is possible that passengers without POVs or viable public transportation have to rely on FHVs to travel, and passengers with POVs do not use FHVs at all. In this case, mode choice is irrelevant.

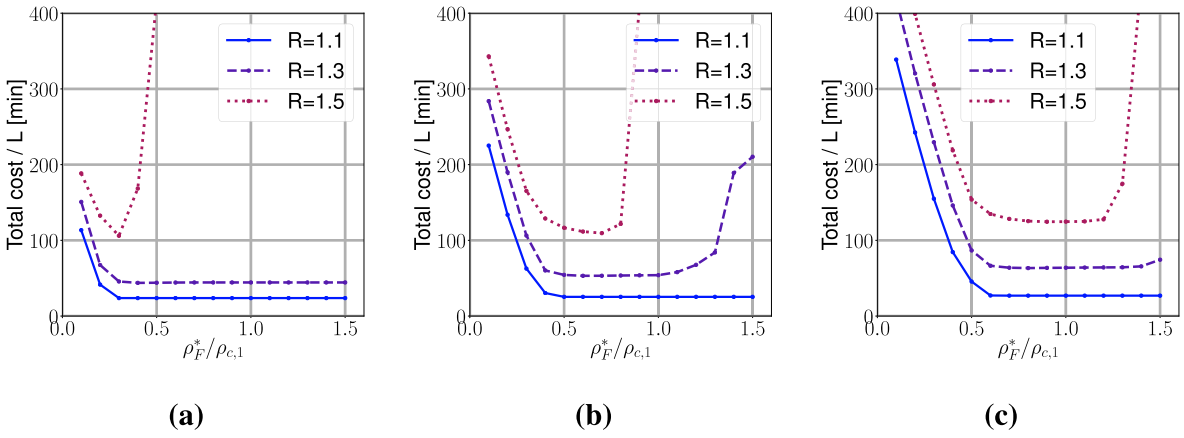

Fig. 7. Impacts of different fleet-size caps on the total time in a mixed mobility system with $\pi = 0.2$ and $\xi = 0.5$: (a) $\eta_F = 0.25$, (b) $\eta_F = 0.4$ and (c) $\eta_F = 0.55$.

bathtub model for the traveling trips. When the extra mileage ratio is larger than the pooling ratio, the POV-like schemes lead to worse congestion in the road network. In the second type of fleet-size management schemes, a cap is introduced for the FHV fleet size. We demonstrated that the compartmental model is approximately the same as the point queue model for the waiting trips. Moreover, the optimal fleet-size cap can be derived by minimizing the total traveler's cost in the departure time user equilibrium. In particular, an optimal cap can prevent the occurrence of over-saturated traffic (or hypercongestion) and, therefore, could potentially improve the safety and environmental impacts of the transportation system. Finally, we presented coupled compartmental models for mixed mobility systems with both POVs and FHVs, and numerically demonstrated the necessity of a FHV fleet-size cap. In a given city, the proposed methodology can be applied to determine the optimal cap that leads to thelowest total cost, if we are able to observe the demand pattern for private cars and for-hire vehicles.

This study helps to reconcile the conflicting results regarding the congestion relieving effects of FHVs in the literature. We find that, when the pooling ratio is smaller than the extra mileage ratio, the overall system's performance can deteriorate with the POV-like schemes employed by TNCs that aim to limit wait times. However, the system can prevent over-saturation or hypercongestion with a fleet-size cap. A major policy implication of this study is that capping the FHV fleet size is a viable measure to mitigate the congestion effects of extra deadhead and detour miles caused by FHVs. If all trips are served by FHVs in the future, the optimal cap can be obtained by solving the departure time user equilibrium at a single bottleneck as in (41). We have shown that this optimal cap is unique for the trapezoidal network fundamental diagram in Fig. 4a) and and does not depend on the deadhead and detour miles when the total cost is linear in the wait time. Thus, for real-world implementations, it is important to use real data to calibrate the network fundamental diagram including, especially, the critical density ρ_{c1}, which reflects the optimal fleet size according to Theorem 5.1. For a mixed system, however, we need to consider the mode choice among POVs and FHVs as well as the departure time choice in both modes to obtain more realistic FHV fleet-size caps, which can be the subject of future analytical and numerical studies. Empirically, it will also be important to observe, estimate, and calibrate the distribution patterns in space and time of multi-modal demands.

We would like to emphasize that, even though the compartmental model in this work is built on several simplifications and assumptions, it has led to valuable theoretical insights on the impacts of different fleet-size management schemes on the system performance. Practically, with well calibrated parameters, such a model can be used to predict congestion dynamics and devise large-scale mitigation schemes. The compartmental model has very low data requirements and computational costs. This is in the same spirit of other compartmental models, e.g., SIR model used by epidemiologists. The parsimonious nature of our model is inherited from that of the point queue model and the bathtub model.

The compartmental model provides a simple framework for evaluating the impacts of the pooling ratio and the extra mileage ratio. They both affect the maximum trip completion rate of a road network, as shown in Fig. 4(b). Therefore, it is important to calibrate both parameters with empirical data. In addition, to improve the system performance, the TNCs should increase the pooling ratio and reduce the deadhead and detour miles of FHVs. In practice, these parameters can be time-dependent, affected by demand patterns, traffic conditions, TNCs' matching algorithms, drivers' decisions, and other factors. The corresponding compartmental model can be numerically solved to evaluate their impacts on the system performance and develop efficient fleet-size management schemes. In reality, there can be delays in the matching process as well as the fleet-size management scheme; thus, more realistically, the resulted compartmental model should include delay-differential equations. However, they are much more challenging than the ordinary differential equations; thus, they are left for future research.

Fares, congestion pricing, and other prices are important when we explicitly consider FHV drivers' choice to provide services or not, and passengers' mode choice behavior. For the former, we assume in this study that the economic and other incentives are sufficiently high such that the number of FHVs is as large as determined by the fleet-size management scheme. In the future, we will consider how to manage the fleet size with economic measures, including fares and congestion pricing. For the latter, we either assume 100% autonomous for-hire vehicles without mode choice in Sections 4 and 5, or assume exogenous demands for different modes in a

mixed mobility system considered in Section 6. In the future, we will explicitly consider choice among for-hire vehicles, public transportation, and other modes. For such studies, we need to include fares, costs for purchasing and maintaining cars, congestion pricing, and other costs.

In the future we are also interested in studying how different operational matching algorithms, economic incentives, and heterogeneous spatial distributions of trips would affect the pooling ratio and the extra mileage ratio in the model and, therefore, the congestion dynamics in the network. A better understanding of these impacts would help us to devise better matching algorithms, economic measures, and fleet-size management schemes to optimize the system performance with respect to congestion and environmental impacts as well as the benefits of all related stakeholders. The model proposed in this paper could easily served as the foundation and first building block for all those extensions. Its parsimonious nature not only allows us to derive general insights quite easily, but to add future modules as needed. Moreover, the proposed model can be used with minimal extensions to study a future mobility service, where FHVs are autonomous and the number of FHVs in the system is controlled directly by the TNCs or traffic management agencies.

In the future, one can also extend the compartmental modeling framework for parking systems (Cao and Menendez, 2015) and other multi-modal shared mobility systems, including subway and bus systems (Loder et al., 2017). For different systems, the compartments may be different, and the trip dynamics models (e.g., point queue and bathtub models) could also be different. With such compartmental models, one can further consider the planning, operation, and control strategies of transit agencies (Ibarra-Rojas et al., 2015; Desaulniers and Hickman, 2007) as well as congestion pricing and other control measures. Notice that congestion pricing on POVs would affect the departure time and mode choices. Thus, the compartmental models can also be used to develop such strategies.

CRediT authorship contribution statement

Wen-Long Jin: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing - original draft, Reviewing, Writing - review & editing. **Irene Martinez:** Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - review & editing. **Monica Menendez:** Funding acquisition, Investigation, Validation, Writing - review & editing.

Acknowledgments

The first author would like to thank his brother, Wen-bin Jin, for helpful discussions in November, 2018 that stimulated this and other related research. We would also like to sincerely thank Prof. Richard Arnott from UC Riverside for his detailed and constructive comments and suggestions, which have helped to substantially improve the presentation of the article. The comments and suggestions of several anonymous reviewers are greatly appreciated. W.-L. Jin would like to acknowledge the support of NSF CMMI#1952241 “SCC-PG: Addressing Unprecedented Community-Centered Transportation Infrastructure Needs and Policies for the Mobility Revolution”. M. Menendez wants to acknowledge the support of the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001 and by the Swiss Re Institute under the Quantum Cities™ initiative.

References

Agatz, N., Erera, A., Savelsbergh, M., Wang, X., 2012. Optimization for dynamic ride-sharing: A review. Eur. J. Oper. Res. 223, 295–303. https://doi.org/10.1016/j.ejor.2012.05.028.

Agatz, N.A., Erera, A.L., Savelsbergh, M.W., Wang, X., 2011. Dynamic ride-sharing: A simulation study in metro Atlanta. Transp. Res. Part B: Methodol. 45 ‘, 1450–1464. https://doi.org/10.1016/j.trb.2011.05.017.

Ambühl, L., Loder, A., Bliemer, M.C., Menendez, M., Axhausen, K.W., 2020. A functional form with a physical meaning for the macroscopic fundamental diagram. Transp. Res. Part B 137, 119–132. https://doi.org/10.1016/j.trb.2018.10.013 advances in Network Macroscopic Fundamental Diagram (NMFD) Research.

Arnott, R., Buli, J., 2018. Solving for equilibrium in the basic bathtub model. Transp. Res. Part B 109, 150–175.

Arnott, R., Kokoza, A., Naji, M., 2016. Equilibrium traffic dynamics in a bathtub model: A special case. Econ. Transp. 7, 38–52.

Aström, K., Murray, R., 2008. Feedback systems: an introduction for scientists and engineers. Princeton Univ Pr.

Bischoff, J., Maciejewski, M., 2016. Simulation of City-wide Replacement of Private Cars with Autonomous Taxis in Berlin. Procedia Comput. Sci. 83, 237–244. https://doi.org/10.1016/j.procs.2016.04.121.

Cao, J., Menendez, M., 2015. System dynamics of urban traffic based on its parking-related-states. Transp. Res. Part B 81, 718–736. https://doi.org/10.1016/j.trb.2015.07.018.

Castiglione, J., Cooper, D., Sana, B., Tischler, D., Chang, T., Erhardt, G.D., Roy, S., Chen, M., Mucci, A., 2018. TNCs & Congestion. Technical Report. SF County Transportation Authority.

Daganzo, C.F., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with hydrodynamic theory. Transp. Res. Part B 28, 269–287.

Daganzo, C.F., 2007. Urban gridlock: Macroscopic modeling and mitigation approaches. Transp. Res. Part B 41, 49–62.

Desaulniers, G., Hickman, M.D., 2007. Public transit. Handbooks Oper. Res. Manage. Sci. 14, 69–127.

Downs, A., 2004. Still Stuck in Traffic: Coping With Peak-Hour Traffic Congestion. Brookings Institution Press.

Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. Transp. Res. Part C 40, 1–13.

Fagnant, D.J., Kockelman, K.M., 2018. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. Transportation 45, 143–158.

Godfrey, K., 1983. Compartmental Models and Their Application. Academic Press.

Hensher, D.A., 2018. Tackling road congestion – what might it look like in the future under a collaborative and connected mobility model? Transp. Policy 66, A1–A8. https://doi.org/10.1016/j.tranpol.2018.02.007 http://www.sciencedirect.com/science/article/pii/S0967070X17308867.

Hyland, M., Mahmassani, H.S., 2018. Dynamic autonomous vehicle fleet operations: Optimization-based strategies to assign AVs to immediate traveler demand requests. Transp. Res. Part C 92, 278–297. https://doi.org/10.1016/j.trc.2018.05.003.

Ibarra-Rojas, O.J., Delgado, F., Giesen, R., Muñoz, J.C., 2015. Planning, operation, and control of bus transport systems: A literature review. Transp. Res. Part B: Methodol. 77, 38–75.

Jin, W.L., 2015. Point queue models: A unified approach. Transp. Res. Part B 77, 1–16.
Jin, W.L., 2020. Generalized bathtub model of network trip flows. Transp. Res. Part B 136, 138–157.
Jin, W.L., Yu, Y., 2015. Performance analysis and signal design for a stationary signalized ring road. arXiv preprint arXiv:1510.01216.
Lam, W.H., Cheung, C.Y., Lam, C., 1999. A study of crowding effects at the hong kong light rail transit stations. Transp. Res. Part A 33, 401–415.
Loder, A., Ambühl, L., Menendez, M., Axhausen, K.W., 2017. Empirics of multi-modal traffic networks – Using the 3D macroscopic fundamental diagram. Transp. Res. Part C 82, 88–101. https://doi.org/10.1016/j.trc.2017.06.009.
Martinez, L.M., Correia, G.H., Viegas, J.M., 2015. An agent-based simulation model to assess the impacts of introducing a shared-taxi system: an application to Lisbon (Portugal). J. Adv. Transp. 49, 475–495.
Mourad, A., Puchinger, J., Chu, C., 2019. A survey of models and algorithms for optimizing shared mobility. Transp. Res. Part B: Methodol. 123, 323–346. https://doi.org/10.1016/j.trb.2019.02.003.
Newell, G., 1982. Applications of queueing theory, vol. 733. Chapman and Hall New York.
NYC Taxi and Limousine Commission, 2019. Improving Efficiency and Managing Growth in New York's For-Hire Vehicle Sector. Technical Report. New York City Taxi and Limousine Commission and Department of Transportation.
Robinson, A., 1996. Non-standard analysis. Princeton University Press.
Schaller, B., 2017. Empty seats, full streets: Fixing manhattan's traffic problem. Schaller Consulting 1.
Schaller, B., 2018. The new automobility: Lyft, uber and the future of american cities. Schaller Consulting.
Shaheen, S., Totte, H., Stocker, A., 2018. Future of mobility white paper.
Small, K.A., Chu, X., 2003. Hypercongestion. J. Transp. Econ. Policy (JTEP) 37, 319–352.
Vickrey, W.S., 1969. Congestion theory and transport investment. In: The American Economic Review: Papers and Proceedings of the Eighty-first Annual Meeting of the American Economic Association, 59, pp. 251–260.
Vickrey, W.S., 1991. Congestion in midtown Manhattan in relation to marginal cost pricing. Technical Report. Columbia University.
Vickrey, W.S., 2020. Congestion in midtown Manhattan in relation to marginal cost pricing. Economics of Transportation 21, 100152. Co-edited by Richard Arnott and W.L. Jin.
Wang, H., Yang, H., 2019. Ridesourcing systems: A framework and review. Transp. Res. Part B: Methodol. 129, 122–155.
Yang, H., Qin, X., Ke, J., Ye, J., 2020. Optimizing matching time interval and matching radius in on-demand ride-sourcing markets. Transp. Res. Part B 131, 84–105. https://doi.org/10.1016/j.trb.2019.11.005.