12-2024

An aggregate matching and pick-up model for mobility-on-demand services

Xinwei LI

Jintao KE

Hai YANG

Hai WANG
Singapore Management University, haiwang@smu.edu.sg

Yaqian ZHOU

An aggregate matching and pick-up model for mobility-on-demand services

Xinwei Li[a], Jintao Ke[b,*], Hai Yang[c,d], Hai Wang[e], Yaqian Zhou[f]

[a]*School of Economics and Management, Beihang University, Beijing, China*
[b]*Department of Civil Engineering, the University of Hong Kong, Hong Kong, China*
[c]*Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology, Hong Kong, China*
[d]*Intelligent Transportation Thrust, The Hong Kong University of Science and Technology (Guangzhou), China*
[e]*School of Computing and Information Systems, Singapore Management University, Singapore*
[f]*School of Economics and Business Administration, Chongqing University, Chongqing, China*

Abstract

This paper presents an Aggregate Matching and Pick-up (AMP) model to delineate the matching and pick-up processes in mobility-on-demand (MoD) service markets by explicitly considering the matching mechanisms in terms of matching intervals and matching radii. With passenger demand rate, vehicle fleet size and matching strategies as inputs, the AMP model can well approximate drivers' idle time, passengers' waiting time for matching and pick-up by considering batch matching in a stationary state. Properties of the AMP model are then analyzed, including the relationship between passengers' waiting time and drivers' idle time, and their changes with market thickness which is measured in terms of the passenger arrival rate (demand rate) and the number of active vehicles in service (supply). The model can also unify several prevailing inductive and deductive matching models used in the literature and spell out their specific application scopes. In particular, when the matching radius is sufficiently small, the model reduces to a Cobb-Douglas type matching model proposed by Yang and Yang (2011) for street-hailing taxi markets, in which the matching rate depends on the pool sizes of waiting passengers and idle vehicles. With a zero matching interval and a large matching radius, the model reduces to Castillo model (Castillo et al., 2017) based on an instant matching mechanism, or a bottleneck type queuing model, in which passengers' matching time is derived from a deterministic queue at a bottleneck with the arrival rate of idle vehicles as its capacity and waiting passengers as its customers. When both the matching interval and matching radius are relatively large, the model also reduces to the bottleneck type queuing model. The performance of the proposed AMP model is verified with simulation experiments.

Keywords: mobility-on-demand, market thickness, matching mechanism, waiting time

*Corresponding author
Email address: kejintao@hku.hk (Jintao Ke)

1. Introduction

Mobility-on-demand (MoD) services offered by platforms such as Uber, Lyft, and DiDi have undergone rapid growth in recent years and attracted considerable attention from researchers in the fields of transportation, operations research, economics, and computer sciences. The massive amount of real-time data collected from passengers and drivers through smartphones offer a plethora of novel instruments, such as surge pricing (Zha et al., 2016, Cachon et al., 2017, Yang et al., 2020b); ride-pooling (Ke et al., 2020a, Jacob and Roet-Green, 2021); driver relocation subsidies (Zhu et al., 2021); and on-demand order dispatching (Xu et al., 2018, Lyu et al., 2019), allowing MoD platforms to improve operational efficiency. Despite their success in business, MoD services have also raised some controversial issues, such as their negative impacts and unfair advantages with respect to conventional street-hailing taxi market (Nie, 2017) and public transit (Hall et al., 2018); congestion externalities caused by both in-service and idle vehicles (Erhardt et al., 2019, Ke et al., 2020b, Vignon et al., 2021, Diao et al., 2021); the pros and cons of competition between MoD platforms (Mo et al., 2020, Zhang and Nie, 2021); concerns on labor elasticities and driver's welfare (Sun et al., 2019); and the necessity of appropriate government regulations (Zha et al., 2016, Parrott and Reich, 2018, Li et al., 2019, Yu et al., 2020, Ke et al., 2021a). Readers may refer to a recent review by Wang and Yang (2019).

To address these issues, researchers have developed a variety of mathematical models to delineate the complex and intriguing relationships between platform decision variables (such as price and wage) and endogenous system variables (such as passenger waiting time and effective demand rate and supply). Typically, the core of these models includes a matching model to approximate the matching frictions between idle vehicles and waiting passengers, which also describes an essential feature that distinguishes MoD service markets from other transportation markets. Many of these models have their roots in the models developed for conventional street-hailing taxis, because the MoD service market and street-hailing taxi market share many common features. For example, Yang and Yang (2011) propose a Cobb-Douglas type matching model to describe the matching frictions between vacant taxis and unserved customers in street-hailing taxi market, and the model is later adopted by Zha et al. (2016) to model MoD service markets.

However, unlike conventional street-hailing taxi market in which passengers and drivers search and meet each other physically on the streets, the MoD service market has two matching stages: online matching and physical matching. After making a request on a MoD digital platform, a passenger may be matched with an idle vehicle (i.e., online matching) after waiting for a certain amount of time in the virtual queue (hereinafter, the waiting time for online matching is termed as matching time). Afterward, the driver that is dispatched to the passenger moves to the passenger's origin or a designated spot to pick up the passenger (i.e., physical matching, and hereinafter, the time for physical matching is termed as pick-up time), and then starts the delivery service. At any given moment, any active driver is in one of three states—idle, on the way to pick up a passenger (i.e., pick-up), or on the way to deliver a passenger to their destination (i.e., delivery). These features cannot be fully captured in legacy models developed for the taxi markets, and great

challenges have to be addressed for modeling both online and physical matching in MoD service markets.

As reviewed later, in the context of MoD service markets, inductive or deductive approaches are typically used to approximate matching frictions in aggregate models. Inductive approaches directly assume a type of hypothesized matching model (such as the Cobb-Douglas matching model) to characterize the matching frictions without using a specific matching mechanism or considering its micro foundations. In contrast, deductive approaches first assume a physical matching process as the micro foundations and then deduce the corresponding matching frictions. These inductive and deductive approaches have their own application scopes. For example, the Cobb-Douglas type matching model (Yang and Yang, 2011, Zha et al., 2016) is more suitable in the scenario in which the platform only matches a passenger with a driver who is in his/her proximity (like a street-hailing taxi market). In contrast, the model developed by Castillo et al. (2017) (hereinafter, it is called as Castillo model for convenience) relies on the assumption that the platform will immediately dispatch the nearest idle vehicle to a passenger who generates a ride request, regardless of how far the driver is from the passenger.

This study proposes an Aggregate Matching and Pick-up (AMP) model to describe a stationary matching process in the MoD service markets. In contrast to the existing matching models, our AMP model explicitly incorporates the platform matching strategy that can be generally articulated in terms of the matching interval (the time interval over which the waiting passengers and idle drivers are accumulated and then subjected to peer-to-peer matching) and the matching radius (the maximum allowable pick-up distance, within which waiting passengers and idle drivers can be matched). These two macro-parameters capture essentially the critical choice in matching mechanism implemented in actual MoD platforms. It is shown that some prevailing matching models in the literature can be viewed as special cases of our general AMP model. The properties of the AMP model are explored and the following intriguing findings are made:

- Passengers' matching time and drivers' matching time are negatively correlated;
- When passenger demand rate and vehicle fleet size increase proportionally, passengers' matching time and the expected pick-up time become shorter while drivers' matching time becomes longer;
- The AMP model reduces to a Cobb-Douglas type matching model (Yang and Yang, 2011) when the platform sets a small matching radius to avoid distant matching between waiting passengers and idle vehicles;
- The AMP model reduces to the Castillo model under a matching mechanism with a large matching radius and the density of idle drivers is much higher than that of waiting passengers; and
- The AMP model reduces to a bottleneck type queuing model when the platform sets a large matching radius and the density of waiting passengers is much higher than that of idle

drivers. In this case, passengers' matching time is determined from a deterministic queue at a bottleneck with the arrival rate of idle vehicles as its capacity and waiting passengers as its customers. In addition, passengers' pick-up time is inversely proportional to the square root of the number of waiting passengers.

Our study also reveals the application scopes of various existing matching models. This can help researchers choose appropriate matching models according to matching strategies and market conditions, assist the platforms in adjusting matching interval and radius to enhance system efficiency.

The rest of the paper is organized as follows. Section 2 reviews the major matching models for MoD service markets in the literature. Section 3 develops the AMP model that describes passengers' matching time, pick-up time and drivers' idle time as functions of market inputs (demand rate and supply) and matching strategy (matching interval and matching radius). Section 4 materializes the AMP model on the basis of a physical matching process. Section 5 examines the impacts of market thickness on matching frictions. Section 6 examines the model properties and discusses the situations in which the AMP model reduces to existing specific matching models. Section 7 conducts numerical experiments using an agent-based simulator to validate the theoretical findings and demonstrate the application scopes of the various existing matching models. Section 8 concludes the paper with future directions of research.

2. A review of matching models

This section introduces some popular matching models for MoD services in the literature, which are derived through either inductive or deductive approaches.

2.1. Inductive approaches

The simplest inductive approach is a perfect matching model, which assumes that the number of matched driver-passenger pairs equals the minimum of passenger demand and participating driver capacity. When demand exceeds capacity, demand is randomly rationed: some demand is unserved and all participating drivers serve one unit of demand. In contrast, when capacity exceeds demand, capacity is randomly rationed: All demand is served, but participating drivers only use a proportion of their capacity. Due to its neat formulation and ease of derivation, this matching model has been adopted in several recent studies on surge pricing (Cachon et al., 2017, Hu et al., 2021); government regulation (Yu et al., 2020); and electric vehicle subsidies (Mo et al., 2020). However, this model fails to describe the spatial matching frictions between passengers and drivers, and is more suitable for a special point-meeting marketplace (such as taxi stations) in which arriving passengers and drivers either get immediate matching or exit the system.

Another inductive approach is the Cobb-Douglas type matching model, which is a widely used model in economics (Varian, 1992) and is introduced to spell out the bilateral searching frictions in street-hailing taxi markets by Yang et al. (2010) and Yang and Yang (2011). This model assumes

that the matching rate (i.e., number of matched driver-passenger pairs per unit time) is given by a function of the number of waiting passengers (customers, m_c) and the number of idle vehicles (m_v) as follows:

$$M(m_c, m_v) \propto (m_c)^{\alpha_1}(m_v)^{\alpha_2} \tag{1}$$

where α_1 and α_2 represent the elasticity of the matching rate with respect to m_c and m_v, respectively. In a stationary state, m_c equals the product of passengers' arrival rate and their waiting time, and m_v equals the product of idle vehicles' arrival rate and their idle time. The meeting rate also equals passengers' arrival rate and idle vehicles' arrival rate in a stationary state. The market is said to exhibit increasing, constant, and decreasing returns to scale if $\alpha_1 + \alpha_2 > 1$, $\alpha_1 + \alpha_2 = 1$, and $\alpha_1 + \alpha_2 < 1$, respectively. Readers may refer to Yang and Yang (2011) for a detailed description.

The Cobb-Douglas type matching model is widely adopted in the literature on MoD services. For example, Zha et al. (2016) use the model to analyze the MoD service market in both monopoly and duopoly scenarios. Wigand et al. (2020) use it to approximate the meeting between passengers and drivers to study how autonomous vehicles can revolutionize MoD service markets. Wang et al. (2019) formulate a stochastic model for an MoD service system in which the pick-up rate is a Cobb-Douglas type model of the pool sizes of requesting passengers and idle vehicles.

2.2. Deductive approaches

Deductive approaches derive the matching formulas based on some assumptions of passengers' and drivers' arrivals as well as matching mechanisms. For example, Arnott (1996) studies a dispatching taxi system in which the taxi firm dispatches the nearest idle taxi to a customer who has requested a ride. This study finds that the expected passengers' waiting time is inversely proportional to the square root of the density of idle taxis in a dispatching taxi system, while under a cruising taxi system, waiting time is inversely proportional to the density of vacant taxis (Beesley and Glaister, 1983). Castillo et al. (2017) extend Arnott's model (Arnott, 1996) to endogenize passenger requests, driver labor supply, and platform pricing to investigate an MoD service system based on a first-come-first-serve dispatch protocol. A wild-goose-chase (WGC) phenomenon is observed whereby drivers may spend substantial time traveling to pick up distant passengers and finds that surge pricing can help prevent the system from falling into the WGC. This deductive approach can well capture the pick-up phase of vehicles, which is one important characteristic that distinguishes MoD service systems (or e-hailing/dispatching taxi systems) from street-hailing taxi systems. Some follow-up studies further extend the aggregate model to examine geometric matching and spatial pricing (Zha et al., 2018b); ride-pooling services (Ke et al., 2020a,b); dynamic waiting (Yan et al., 2019); government regulations (Vignon et al., 2021, Ke et al., 2021a); competition and substitution between MoD services and public transit services (Ke et al., 2021b); etc. By assuming that the distribution of passengers follows a spatial Poisson process, Xu et al. (2017) offer an analytical approximation for passengers' matching time, drivers' idle time and the expected pick-up time. Their model reduces to Castillo model under some extreme conditions.

Starting from a physical passenger–driver matching process model, Zhang et al. (2019) propose a unified cumulative distribution model of passengers' expected pick-up time for both street-hailing and e-hailing taxi markets. Their model, as a deductive approach, can be reformulated as an analytical Cobb-Douglas production model with specific parameters. It shows that the street-hailing taxi market displays increasing returns to scale (i.e., $\alpha_1 + \alpha_2 = 2$) and the e-hailing taxi market has constant returns to scale (i.e., $\alpha_1 + \alpha_2 = 1$). They also report some empirical evidence in which the returns to scale are 1.6 for street-hailing taxis and close to 1 for e-hailing taxis.

In addition, there is a rich body of literature on the use of queuing models to approximate customers' queuing time in MoD service markets. Based on different assumptions regarding customers' arrivals and service providers (drivers), researchers have used a variety of queuing models in MoD service markets, including the M/M/1 queue (Guo et al., 2018); M/M/k queue (Taylor, 2018, Bai et al., 2019, Feng et al., 2020); M/G/k queue (Li et al., 2019); Erlang loss system (Hu and Zhou, 2020); G/G/k queue (Wang and Odoni, 2016, Chen and Wang, 2018b, Wang et al., 2019); Erlang C formula (Benjaafar et al., 2021); and double-ended queue (Chen and Wang, 2018a, Xu et al., 2020). In particular, Benjaafar et al. (2021) argue that passengers' expected queuing time is a function of both the difference between supply and demand and the ratio of demand to supply. By implementing a series of approximations for the M/G/k queue, Li et al. (2019) find that customers' total waiting time (sum of queuing and pick-up time) is inversely proportional to the number of idle vehicles.

It is noteworthy that Yang et al. (2020a) first study the joint impacts of matching interval and matching radius on the main system performance measures in MoD service markets. However, their model simply focuses on passengers' waiting time and drivers' idle time in one batch matching interval, with the number of waiting passengers and idle drivers at the start of the considered interval as given. By contrast, this paper focuses on a stationary process of matching in which the variables do not change over a certain time horizon, and passenger arrival rate and vehicle fleet size are treated as exogenous inputs of the matching model. Then, the AMP model can approximate passengers' total expected waiting time and drivers' total expected idle time over a sequence of intervals in a stationary process, and characterize the complex endogenous relationships between endogenous variables (such as passengers' and drivers' expected waiting time) and model inputs (including passenger arrival rate, vehicle fleet size, matching interval and matching radius) in both the online and physical matching stages.

3. A general AMP model

This section presents the AMP model to delineate how to approximate passengers' and drivers' total expected waiting time with a given matching strategy and market condition in a stationary process. To avoid confusion, we define a passenger's total waiting time as the sum of the passenger's matching time (passenger waiting time for matching or order confirmation after order placement) and pick-up time (passenger waiting time for pick-up upon order confirmation), and define a driver's total waiting time as the sum of the driver's matching time (driver idle time waiting for an

order assignment after completion of the last ride) and pick-up time. It is worth mentioning that our AMP model can be incorporated into a large number of demand-supply equilibrium models, for almost all of them involve some types of matching process. In this case, the demand rate can be given as a function of the trip fare set by the platform and passengers' waiting time in both virtual and physical stages of matching. The driver supply can be given as a function of the wage per ride paid by the platform and the vehicles' utilization rate, which is related to drivers' waiting time. Passengers' and drivers' waiting times are, in turn, affected by passenger demand rate and driver supply, as characterized by the AMP model. In other words, the AMP model proposed in this study can be used as a backbone of models built for equilibrium analyses of MoD service markets.

3.1. General model setup

Let Q denote the passenger demand rate of MoD services (measured in arrival rate) and N the number of active vehicles/drivers in service in the examined area. Both Q and N are treated as exogenous inputs of the AMP model and their combination measures the market thickness. Note that the inputs of the AMP model also include the matching strategy defined below. In general, passengers and drivers can be matched in two ways: instantly or in batch. The former matches a passenger with the closest idle vehicle within a certain pick-up distance (termed the matching radius, r), immediately as the passenger raises his/her order. The latter waits for a certain time interval (termed the matching interval, τ) to accumulate more driver-passenger pairs for better matching through, for example, a bipartite graph matching algorithm. Clearly, if $\tau = 0$, we have instant matching; in other words, instant matching can be viewed as a special case of batch matching. We thus use the pair (r, τ) to denote a platform's matching strategy. It is worth noting that, different from some existing studies (Li and Netessine, 2020, Qin et al., 2021) allowing dynamic adjustment of matching time interval, we assume a stationary equilibrium state in which the platform's matching strategy (r, τ) remains constant within a certain time horizon spanned over a few batch matching intervals.

We consider a stationary process in which all randomly arriving passengers will eventually receive ride services with a ride time t, measured as a fraction of an hour. Let m_c and m_v denote the masses of the two pools of waiting passengers and idle vehicles, respectively, at the end of each matching interval and right before the next matching decision. Here and in what follows, the subscript "$_c$" stands for passengers (customers) and the subscript "$_v$" stands for vehicles (drivers). Then the expected number of successfully matched driver-passenger pairs and the expected pick-up distance, denoted by $M(m_c, m_v, r)$ and $L(m_c, m_v, r)$, can be expressed as a function of these two masses and the matching radius. The expected number of passengers left after execution of one batch matching is given by $m_c - M(m_c, m_v, r)$, and the expected number of new passengers arriving in the next interval is τQ. Then the expected number of passengers for the next batch matching is $m'_c = m_c - M(m_c, m_v, r) + \tau Q$. In a stationary process, we have $m'_c = m_c$, giving

rise to

$$M(m_c, m_v, r) = \tau Q \tag{2}$$

Note that we are considering batch matching by assuming $M(m_c, m_v, r)$ (termed as pairing function) as a function of the two masses of waiting passengers and idle vehicles at the instant of matching, both masses depend on the matching interval over which they are accumulated. This also implies that the matching rate $M(m_c, m_v, r)/\tau$ is equal to the arrival rate of passengers Q, which is consistent with the stationary condition in Yang and Yang (2011). The pairing function $M(m_c, m_v, r)$ is a key component in our AMP model to derive the total passenger waiting time for given demand rate Q and supply N.

Denote by p_c and p_v the matching probability of passengers and drivers, respectively. p_c is the proportion of passengers matched with idle vehicles in the mass of waiting passengers and p_v is the proportion of idle vehicles matched with waiting passengers in the mass of idle vehicles. At the execution of each batch matching, $p_c = M(m_c, m_v, r)/m_c$ and $p_v = M(m_c, m_v, r)/m_v$. In a stationary process, the matching times of passengers and drivers are denoted by w_c and w_v and can be estimated by

$$w_c = \frac{\tau}{2}p_c + \frac{3}{2}\tau p_c(1-p_c) + \frac{5}{2}\tau p_c(1-p_c)^2 + ... \tag{3}$$

$$w_v = \frac{\tau}{2}p_v + \frac{3}{2}\tau p_v(1-p_v) + \frac{5}{2}\tau p_v(1-p_v)^2 + ... \tag{4}$$

where $\frac{\tau}{2}$ indicates the expected matching time of passengers/drivers who arrive in one interval and get matched at the execution time of that interval (i.e., the end of the interval); $\frac{3}{2}\tau$ refers to the expected waiting of passengers/drivers who arrive in one interval and get matched at the end of the next interval, so on and so forth. In view of $p_c \in [0,1]$ and $p_v \in [0,1]$, by using the summation of the series formula, we can obtain

$$w_c = \left(\frac{1}{p_c} - \frac{1}{2}\right)\tau = \left[\frac{m_c}{M(m_c, m_v, r)} - \frac{1}{2}\right]\tau \tag{5}$$

$$w_v = \left(\frac{1}{p_v} - \frac{1}{2}\right)\tau = \left[\frac{m_v}{M(m_c, m_v, r)} - \frac{1}{2}\right]\tau \tag{6}$$

Clearly, $w_c \geq \frac{\tau}{2}$, $w_v \geq \frac{\tau}{2}$. We can also find that $w_c \to \frac{\tau}{2}$ as $p_c \to 1$ and $w_v \to \frac{\tau}{2}$ as $p_v \to 1$ (all arriving passengers/drivers can be successfully matched in the arrival interval without being carried over to the next batch); $w_c \to \infty$ as $p_c \to 0$ and $w_v \to \infty$ as $p_v \to 0$ (passengers/drivers can hardly be matched, so they wait for an extremely long matching time or idle time). By combining Eq. (2), Eq. (5), and Eq. (6) we can further obtain the following relationship between the masses of waiting passengers and drivers (m_c, m_v) and their matching time (w_c, w_v) and arrival rate Q:

$$m_c = \left(w_c + \frac{\tau}{2}\right)Q \tag{7}$$

$$m_v = \left(w_v + \frac{\tau}{2}\right)Q \tag{8}$$

At the beginning of a matching interval, $m_c - \tau Q$ passengers remain waiting in the system because $M(m_c, m_v, r) = \tau Q$ passenger-driver pairs are matched. At the end of a matching interval

and before the next matching decision, the mass of waiting passengers is m_c because τQ passengers have steadily arrived during the matching interval. Therefore, the queuing length of passengers is $\frac{1}{2}(m_c + m_c - \tau Q) = m_c - \frac{\tau}{2}Q$ in a matching interval; similarly, the queue length of drivers is $m_v - \frac{\tau}{2}Q$. By reformulating Eqs. (7) and (8), we have $m_c - \frac{\tau}{2}Q = w_c Q$ and $m_v - \frac{\tau}{2}Q = w_v Q$. It is consistent with Little's law, in which the average mass of waiting passengers $m_c - \frac{\tau}{2}Q$ (or drivers $m_v - \frac{\tau}{2}Q$) equals the product of the matching time w_c (or w_v) and the arrival rate of passengers Q (which is also equal to the arrival rate of idle vehicles). The term $-\frac{\tau}{2}Q$ at the left-hand side (LHS) in $m_c - \frac{\tau}{2}Q = w_c Q$ (or $m_v - \frac{\tau}{2}Q = w_v Q$) formula is due to the discrete batch matching setting; that is, the batch matching is only executed at the end of each matching interval.

Combining Eqs. (7) and (8), yields

$$w_v = \frac{m_v}{m_c}\left(w_c + \frac{\tau}{2}\right) - \frac{\tau}{2} \tag{9}$$

In addition, in a stationary state, given the vehicle conservation condition whereby fleet size equals the sum of the average idle vehicles, picking-up vehicles and occupied vehicles in the delivery phase, we have

$$N = \frac{1}{2}(m_v - \tau Q + m_v) + Q w_p + tQ = m_v + Q\left[\frac{L(m_c, m_v, r)}{v} + t - \frac{\tau}{2}\right] \tag{10}$$

where ride time t of passengers is assumed to be constant, $w_p = L(m_c, m_v, r)/v$ is the expected pick-up time, and v is the vehicle moving speed, which is generally assumed to be constant. The term $\frac{1}{2}(m_v - \tau Q + m_v)$ represents the average mass of idle vehicles, wherein the number of waiting vehicles at the start of each matching interval is $m_v - \tau Q$, and at the end of each matching interval, it reaches m_v. By summing the passengers' matching time w_c and the expected pick-up time w_p, we obtain the total waiting time of passengers, denoted by W, as follows:

$$W = \left[\frac{m_c}{M(m_c, m_v, r)} - \frac{1}{2}\right]\tau + \frac{L(m_c, m_v, r)}{v} \tag{11}$$

For given constants t and v, from Eq. (2) and Eq. (10) we can obtain m_c and m_v as functions of market inputs (Q, N) and matching strategy (r, τ). Consequently, by substituting the expressions m_c and m_v into Eq. (11), we can obtain the total waiting time of passengers as a function of (Q, N) and (r, τ), which is given below after omitting the constants t and v:

$$W = W(Q, N, r, \tau) \tag{12}$$

It is noteworthy that Eq. (12) does not necessarily have an explicit formula; instead, the two intermediate variables m_c and m_v are given as the implicit solutions of a system of equations consisting of Eq. (2) and Eq. (10), and thus W is given by a function of these two intermediate variables.

To summarize, the AMP model for a stationary MoD service market is

$$\begin{cases} M(m_c, m_v, r) = \tau Q & \text{(13a)} \\ N = m_v + Q\left(w_p + t - \dfrac{\tau}{2}\right) & \text{(13b)} \\ m_c = \left(w_c + \dfrac{\tau}{2}\right) Q & \text{(13c)} \\ m_v = \left(w_v + \dfrac{\tau}{2}\right) Q & \text{(13d)} \\ w_p = \dfrac{L(m_c, m_v, r)}{v} & \text{(13e)} \end{cases}$$

All the variables related to the matching process, i.e., M, m_c, m_v, w_c, w_v and w_p, can be obtained with given market inputs (Q, N) and matching strategy (r, τ). Namely, this matching model delineates the intricate relationships between endogenous variables and model inputs (Q, N, r, τ) in both virtual and physical matching stages. Moreover, by integrating this matching model into a specific demand-supply equilibrium framework and constructing a bi-level optimization model, we can also derive the state-dependent optimal matching strategy. For example, as the supply and demand states vary across peak and off-peak hours within a day, we can use the bi-level optimization model to determine the optimal matching intervals from hour to hour.

3.2. General model properties

Since $m_v = \frac{M}{p_v} = \frac{\tau Q}{p_v} \geq \tau Q$, based on the AMP model, the average mass of idle vehicles is at least $\frac{\tau}{2}Q$. To guarantee the system stability and obtain valuable insights of the AMP model, the following two common assumptions are introduced and used henceforth.

Assumption 1. *The vehicle fleet size N is larger than the minimum consumed supply $tQ + \frac{\tau}{2}Q$, i.e., $N > tQ + \frac{\tau}{2}Q$.*

Assumption 2. *Given a matching strategy (τ, r), the pairing function $M(m_c, m_v, r)$ increases with the mass of waiting passengers m_c (with given m_v) and idle drivers m_v (with given m_c) and $M(m_c, m_v, r) \leq \min\{m_c, m_v\}$.*

Assumption 1 is a widely used assumption in the literature, which assumes a supply slack (difference between the total supply N and the minimum consumed supply $tQ + \frac{\tau}{2}Q$. Then the feasible range of m_v is $\left[\tau Q, N - tQ + \frac{\tau}{2}Q\right]$). Assumption 2 is also intuitive: the larger the mass of waiting passengers and idle drivers, the larger the number of matched driver-passenger pairs in each time interval. Based on these assumptions, the relationships between m_c and m_v and between w_c and w_v are established below.

Proposition 1. *Given any matching strategy (τ, r) and passenger arrival rate Q,*

1. *the mass of waiting passengers m_c decreases with the mass of idle drivers m_v, and vice versa;*
2. *passengers' matching time w_c decreases with drivers' matching time w_v, and vice versa.*

Since at the stationary state, m_c and m_v satisfy $M(m_c, m_v, r) = \tau Q$, then, m_c can be regarded as a function of m_v for given r, τ and Q, i.e., $m_c \coloneqq m_c(m_v)$. Define

$$w_p(m_c(m_v), m_v, r) = \frac{L(m_c(m_v), m_v, r)}{v} \tag{14}$$

when m_v approaches its minimum at τQ, $p_v = \frac{M}{m_v} = \frac{\tau Q}{m_v}$ should approach 1, which implies m_c should be extremely large. In this situation, an idle vehicle is surrounded by lots of waiting passengers, and if it is matched with the nearest waiting passenger, the expected pick-up distance is nearly zero. Then the following mild Assumption 3 about the pick-up time function in Eq. (14) is introduced.

Assumption 3. *With a given matching strategy (r, τ) and passenger arrival rate Q, $w_p(m_c(m_v), m_v, r)$ in Eq. (14) is a continuous function of m_v and satisfies $w_p(m_c(m_v), m_v, r) \to 0$ when $m_v \to \tau Q$.*

With Assumptions 1-3, the existence of solutions for the AMP model (13) is guaranteed as follows:

Proposition 2. *For any given (Q, N, r, τ), there exists a solution (m_c^*, m_v^*) of the AMP model (13).*

4. A physical process for model realization and properties

This section substantializes the AMP model by considering a realistic physical matching process, under which the assumptions of the AMP model in Section 3 are satisfied.

4.1. Physical process

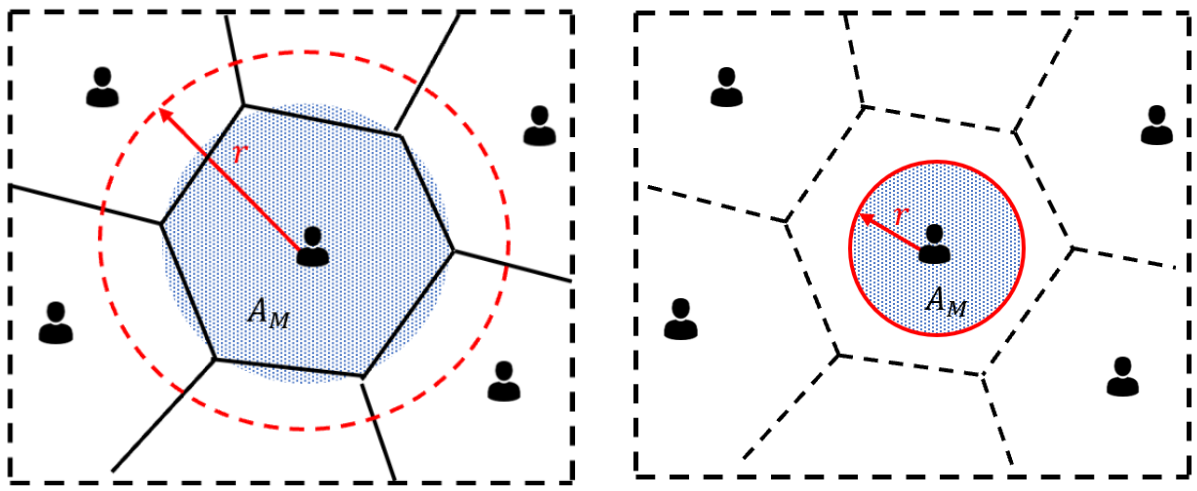
Denote by A the total area of the service region. Then the densities of waiting passengers and idle vehicles in the service region are given by $\rho_c = m_c/A$ and $\rho_v = m_v/A$, respectively. We assume that waiting passengers are uniformly located in the service region and the stochasticity of passengers' location is not considered for simplicity. The spatial distribution of idle vehicles is assumed following a spatial Poisson point process (Chiu et al., 2013). For analytical tractability, we consider a simplified matching procedure in which the platform matches a waiting passenger to his/her closest idle vehicle within the matching radius.

For any waiting passenger, the probability of getting matched with a nearby idle vehicle and the expected distance to the closest idle vehicle depend on the densities of waiting passengers and idle vehicles and the matching radius. To estimate the pick-up distance between the matched pairs of waiting passengers and idle vehicles, Xu et al. (2017) introduce the notion of *dominant zone*, which is inspired by the well-known Voronoi Diagram (Voronoi, 1908). The dominant zone of each waiting passenger refers to a neighboring area within which the distance from any point to this passenger is shorter than that to any other waiting passenger. Then each passenger will be matched with the closest idle vehicle within the dominant zone of that passenger. This assumption of the matching mechanism successfully reflects the competition between waiting passengers in

"catching" idle vehicles. However, it fails to capture the influences of the matching radius within which only a passenger and an idle vehicle can be paired. Specifically, when the matching radius is small, some idle vehicles in the dominant zone of a passenger may not be feasible for matching since they are out of the matching radius. To address this issue, Yang et al. (2020a) extend the concept of *dominant zone* to *matching area*, which is the intersection of the dominant zone and the matchable area constrained by the matching radius. In this work, we adopt the assumption of Yang et al. (2020a). Since the area of each passenger's dominant zone equals $(\rho_c)^{-1} = A/m_c$, we further assume that the shape of the dominant zone of each passenger is approximated by a circle with a radius of $\sqrt{(\rho_c)^{-1}/\pi}$. Then the matching area A_M—the area around a waiting passenger in which they can be matched with an idle vehicle—is given by

$$A_M = \min\{(\rho_c)^{-1}, \pi r^2\} \tag{15}$$

where πr^2 is the area of the circle centered at each waiting passenger with a radius equal to the matching radius r (as illustrated in Figure 1).



(a) Matching area governed by the density of waiting passengers.

(b) Matching area governed by the the matching radius.

Figure 1: Illustrations of the matching area.

For spatial Poisson distribution, the probability that there are n idle vehicles within the matching area of each waiting passenger can be written as $P\{n\} = \frac{1}{n!}\exp(-A_M\rho_v)\cdot(-A_M\rho_v)^n$. During each matching interval, a waiting passenger is matched with an idle vehicle if at least one idle vehicle is within the passenger's matching area. Then the probability of each passenger being

successfully matched, i.e., the matching probability p_c, is given as follows[1]:

$$p_c = 1 - P\{0\} = 1 - \exp(-A_M \rho_v) \tag{16}$$

The pairing function $M\left(m_c, m_v, r\right)$ is given by

$$M(m_c, m_v, r) = m_c \left[1 - \exp(-A_M \rho_v)\right] \tag{17}$$

It is worth mentioning that, when $\left(\rho_c\right)^{-1} \leq \pi r^2$, we have $M\left(m_c, m_v, r\right) = m_c \left[1 - \exp\left(-m_v/m_c\right)\right]$, this pairing function is consistent with the widely used matching function in the economic literature, such as Petrongolo and Pissarides (2001) and Buchholz (2022), which is derived from the well-known urn-ball matching problem first formulated in Butters (1977) and Hall (1979).

Eqs. (5)–(6) and (17) yield the following specific form of the matching time of passengers and drivers:

$$w_c = \left[\frac{1}{1 - \exp(-A_M \rho_v)} - \frac{1}{2}\right] \tau \tag{18}$$

$$w_v = \left\{\frac{\rho_v}{\rho_c \left[1 - \exp(-A_M \rho_v)\right]} - \frac{1}{2}\right\} \tau \tag{19}$$

The expected pick-up time can also be deduced based on the above-stated assumptions regarding uniformly located waiting passengers and spatial Poisson distributed idle vehicles. We denote by x the distance of the unmatched passenger to his/her closest idle vehicle with a cumulative distribution function $H(\cdot)$ and density function $h(\cdot)$. When the distribution of idle vehicles follows a spatial Poisson distribution, we obtain $H(\cdot)$ and $h(\cdot)$ as follows ((Chiu et al., 2013)):

$$H(x) = 1 - \exp(-\pi x^2 \rho_v), 0 \leq x \leq \sqrt{\frac{A_M}{\pi}} \tag{20}$$

$$h(x) = 2\pi x \rho_v \exp(-\pi x^2 \rho_v) \tag{21}$$

where $\sqrt{\frac{A_M}{\pi}}$ is the approximated radius of the matching area A_M. Then the expected pick-up time w_p can be approximated as

$$\begin{aligned} w_p(m_c, m_v, r) =& \frac{L(m_c, m_v, r)}{v} \cong \frac{\zeta \int_0^{\sqrt{\frac{A_M}{\pi}}} x h\left(x\right) dx}{v H\left(\sqrt{\frac{A_M}{\pi}}\right)} \\ =& \frac{\zeta \left[\frac{\operatorname{erf}(\sqrt{A_M \rho_v})}{2\sqrt{\rho_v}} - \sqrt{\frac{A_M}{\pi}} \cdot \exp(-A_M \rho_v)\right]}{v \left[1 - \exp(-A_M \rho_v)\right]} \end{aligned} \tag{22}$$

[1]We do not take into account the "secondary effects" that may actually underestimate the matching probability when the matching area is governed by the density of waiting passengers. It happens when no idle vehicle is in one passenger's dominant zone while two (or more) idle vehicles are located in his/her adjacent passenger's dominant zone, so that these two passengers can all be matched if the matching radius is sufficiently large.

where $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \mathrm{d}t$ is a Gaussian error function, ζ is a detour ratio (Yang et al., 2018), i.e., the distance of actual road distance to straight line distance, and v is the speed of the vehicle. Clearly, w_p in Eq. (22) depends on the densities of waiting passengers $\rho_c = m_c/A$ and idle vehicles $\rho_v = m_v/A$ and matching radius r, which can also be written as $w_p(\rho_c, \rho_v, r)$ when the area of the service region A is given. This form of pick-up time function and its variants were also used by Arnott (1996), Xu et al. (2017), Zha et al. (2018b) and Yang et al. (2020a).

The densities of waiting passengers and idle vehicles are given by

$$\rho_c = \frac{m_c}{A} = \frac{M}{p_c A} = \frac{\tau Q}{\left[1 - \exp\left(-A_M \rho_v\right)\right] A} \tag{23}$$

$$\rho_v = \frac{m_v}{A} = \frac{N - w_p(\rho_c, \rho_v, r)Q - tQ + \frac{\tau}{2}Q}{A} \tag{24}$$

Therefore, when the market inputs (Q, N) and matching strategy (r, τ) are given, ρ_c and ρ_v can be determined by Eq. (23) and (24). Then ρ_c, ρ_v, w_c, w_v and w_p of the AMP model under this model realization is

$$\begin{cases} \rho_c = \dfrac{\tau Q}{\left[1 - \exp\left(-A_M \rho_v\right)\right] A} & \text{(25a)} \\ \rho_v = \dfrac{N - w_p(\rho_c, \rho_v, r)Q - tQ + \frac{\tau}{2}Q}{A} & \text{(25b)} \\ w_c = \left[\dfrac{1}{1 - \exp(-A_M \rho_v)} - \dfrac{1}{2}\right] \tau & \text{(25c)} \\ w_v = \left\{\dfrac{\rho_v}{\rho_c \left[1 - \exp(-A_M \rho_v)\right]} - \dfrac{1}{2}\right\} \tau & \text{(25d)} \\ w_p = \dfrac{\zeta \left[\frac{\text{erf}(\sqrt{A_M \rho_v})}{2\sqrt{\rho_v}} - \sqrt{\frac{A_M}{\pi}} \cdot \exp(-A_M \rho_v)\right]}{v \left[1 - \exp(-A_M \rho_v)\right]} & \text{(25e)} \end{cases}$$

4.2. Specific model properties

We now examine the specific properties of the AMP model under the physical process introduced above. From Eq. (17), it is easy to see that the matched driver-passenger pairs in a matching interval $M(m_c, m_v, r)$ increases with m_c and m_v, i.e., $\frac{\partial M(m_c, m_v, r)}{\partial m_c} > 0$ and $\frac{\partial M(m_c, m_v, r)}{\partial m_v} > 0$, which satisfies Assumption 2. Then, we have the following Corollary 1 to show the relationship between m_c and m_v, and between w_c and w_v.

Corollary 1. *Given any matching strategy (τ, r) and passenger arrival rate Q,*

1. *The mass of waiting passengers m_c is a monotonically decreasing and convex function of the mass of idle drivers m_v, and vice versa.*

2. *Drivers' matching time w_v is a monotonically decreasing and convex function of passengers' matching time w_c, and vice versa.*

Corollary 1 indicates that w_c is not only negatively correlated with w_v, but also a convex function of w_v in the entire feasible domain. This corollary is consistent with the findings made without considering a matching radius in Xu et al. (2017). In this sense, the proposed model verifies their findings in a more general setting with a matching radius.

Before analyzing the existence of solutions of Eqs. (25), we provide the following lemma on w_p.

Lemma 1. *Let $m_c(m_v)$ be the mass of waiting passengers that satisfies $M(m_c, m_v, r) = m_c[1 - \exp(-A_M \rho_v)] = \tau Q$, then $w_p(m_c(m_v), m_v, r)$ is a continuous function of m_v and $w_p(m_c(m_v), m_v, r) \to 0$ when $m_v \to \tau Q$.*

Lemma 1 shows that this model realization satisfies Assumption 3, then according to Proposition 2 the existence of solutions of Eq. (25) is guaranteed as shown in the following Corollary 2.

Corollary 2. *For any given (Q, N, r, τ), there exists a solution (ρ_c^*, ρ_v^*) to the system of Eqs. (25).*

Let ϕ denote $\frac{\pi r^2 \tau Q}{A}$ to simplify the expression, which represents the number of new arrival passengers density within the maximum matching radius in each matching interval. Given that $\rho_c \geq \frac{\tau Q}{A}$, the condition $\phi < 1$ indicates that $\pi r^2 < \frac{A}{\tau Q} \leq \frac{1}{\rho_c}$, which implies that the matching area $A_M = \pi r^2$. When $\frac{1}{\rho_c} = \pi r^2$, from Eqs. (25a)–(25b), we have $\phi < 1$ and

$$N - Q\frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{-\ln(1-\phi)}\right)}{2\sqrt{\frac{-\ln(1-\phi)}{\pi r^2}}} - r\cdot(1-\phi)\right]}{v\cdot\phi} - tQ + \frac{\tau}{2}Q = \frac{-A\cdot\ln(1-\phi)}{\pi r^2} \tag{26}$$

Therefore, given N and Q, the matching strategy (r, τ) satisfying Eq. (26) will divide the plane of (r, τ) into two subareas: one is a matching area governed by the matching radius, and the other is governed by the density of waiting passengers. Furthermore, as shown in Corollary 3 below, the range of the solutions of passenger and driver densities can be further determined by comparing the values on the left and right hand sides of Eq. (26).

Corollary 3. *For any given (Q, N, r, τ),*

1. *when $\phi < 1$ and*

$$N - Q\frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{-\ln(1-\phi)}\right)}{2\sqrt{\frac{-\ln(1-\phi)}{\pi r^2}}} - r\cdot(1-\phi)\right]}{v\cdot\phi} - tQ + \frac{\tau}{2}Q \geq \frac{-A\cdot\ln(1-\phi)}{\pi r^2} \tag{27}$$

   there exists at least one solution (ρ_c^*, ρ_v^*) such that $\rho_c^* \leq \frac{1}{\pi r^2}$ to the system of Eqs. (25a)–(25b);

2. *when $\phi \geq 1$ or when $\phi < 1$ and*

$$N - Q\frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{-\ln(1-\phi)}\right)}{2\sqrt{\frac{-\ln(1-\phi)}{\pi r^2}}} - r\cdot(1-\phi)\right]}{v\cdot\phi} - tQ + \frac{\tau}{2}Q < \frac{-A\cdot\ln(1-\phi)}{\pi r^2} \tag{28}$$

   there exists at least one solution (ρ_c^*, ρ_v^*) such that $\rho_c^* > \frac{1}{\pi r^2}$ to the system of Eqs. (25a)–(25b).

Corollary 3 indicates that under condition $\phi < 1$ and Condition (27), there exists a solution such that the matching area is governed by the matching radius, i.e., $A_M = \pi r^2$; under condition $\phi < 1$ and Condition (28) or $\phi \geq 1$ there exists a solution such that the matching area is governed by the density of waiting passengers, i.e., $A_M = (\rho_c)^{-1}$.

We are now ready to present the following monotonicity results to show the impacts of the matching strategy (r, τ).

Proposition 3. *With any given market inputs (Q, N):*

1. *If the matching area is governed by the matching radius, i.e., $A_M = \pi r^2$,*
   1) *for given τ, when r increases, then w_p increases, ρ_v decreases, w_v decreases, but ρ_c, w_c and p_c are not necessarily monotonic.*
   2) *for given r, when τ increases, then ρ_v increases, w_v decreases, p_c increase, but ρ_c, w_c and w_v are not necessarily monotonic.*
2. *If the matching area is governed by the density of waiting passengers, i.e., $A_M = (\rho_c)^{-1}$,*
   1) *for given τ, then ρ_c, ρ_v, p_c, w_p, w_c and w_v are independent of r;*
   2) *for given r, when τ increases, then w_p decreases, ρ_v increases, but w_v, w_c and p_c are not necessarily monotonic.*

This proposition implies that when the matching radius is smaller than a threshold, it influences the outputs of matching models, including passengers' matching time and pick-up time, densities of waiting passengers and idle vehicles, and drivers' matching time. As the matching radius increases, the pick-up time will increase while drivers' matching time will decrease. However, when the matching radius is larger than the threshold, the outputs of matching models are irrelevant to it but only relevant to the matching interval. As the matching interval increases, the pick-up time will decrease.

5. Impact of market thickness

Market thickness in terms of the size of the two pools of demand and supply is an important factor that governs market frictions (Frechette et al., 2019). The above proposed AMP model explicitly captures market frictions by considering passenger and driver spatial distributions and thus contrasts with the conventional queuing models applied in MoD service markets, which give the same waiting time due to the same traffic intensity when the demand and supply are changed proportionally.

To examine the impact of market thickness on market frictions, we now scale up or down both passenger demand rate Q and driver supply N by the same factor κ in the AMP model. In this case, the utilization rate of vehicles $\frac{tQ}{N}$ does not change. However, passengers' and drivers' matching time and the expected pick-up time may change. With the above model realization, the returns to scale

property of the driver-passenger pairing function (defined in Eq. (17)) together with the matching time function with respect to the scaling factor can be established in the following Proposition 4.

Proposition 4. *Given the matching strategy (τ, r), when scaling up passenger demand rate Q and driver supply N with the same scaling factor $\kappa > 1$, we have*

1. *Given the matching radius r, the driver-passenger pairing function $M(m_c, m_v, r)$ (defined in Eq. (17)) exhibits constant return to scale when $\frac{1}{\rho_c} \leq \pi r^2$, and exhibits increasing return to scale when $\frac{1}{\rho_c} > \pi r^2$. Moreover, $M(m_c, m_v, r)$ satisfies $M(m_c, \kappa m_v, r) \leq \kappa M(m_c, m_v, r)$.*

2. *The expected pick-up time w_p and passgeners' matching time w_c decrease with κ.*

3. *Drivers' matching time w_v increases with κ.*

Proposition 4 shows that when the number of waiting passengers m_c and drivers m_v in the pool increases proportionally, the number of matched driver–passenger pairs $M(m_c, m_v, r)$ increases proportionally or more than proportionally. This is because the expected distance between a waiting passenger and his/her closest idle vehicle becomes shorter when m_c and m_v in the pool are increased. Then, when the maximum matching radius is not binding, the matched driver-passenger pairs $M(m_c, m_v, r)$ has constant return to scale for the waiting passengers are assumed uniformly located and the spatial distribution of idle vehicles follows a spatial Poisson point process. When the maximum matching radius is binding, some waiting passengers who are unmatchable in the absence of idle vehicles within the matching radius can now be matched. In this case, the number of matched driver-passenger pairs or the pairing function has an increasing return to scale. This result is also supported by some relevant studies, such as Yang et al. (2014), Zhang et al. (2019) and Wei et al. (2022). When only the density of idle vehicles increases while the number of waiting passengers keeps unchanged, although the expected pick-up distance becomes shorter, the number of matched driver–passenger pairs satisfies $M(m_c, \kappa m_v, r) \leq \kappa M(m_c, m_v, r)$ for $\kappa > 1$, because not all increased drivers are matched.

With these propositions of the pairing function $M(m_c, m_v, r)$, we can find that, when passenger demand rate Q and driver supply N increase with the same scaling factor κ and $\kappa > 1$, the market will have more waiting passengers and idle vehicles, then the expected pick-up time will be shorter, or w_p decreases with the scaling factor. Because the proportion of time spent by each driver in the delivery phase does not change, a shorter time spent in the pick-up phase also implies a larger proportion of time in the idle phase, or the matching time w_v increases with the scaling factor. A larger w_v indicates that the mass of idle vehicles m_v increases by more than the matching rate $\frac{M}{\tau} = Q$. Also, because the pairing function $M(m_c, m_v, r)$ has a constant/increasing return to the scale, the mass of waiting passengers m_c should increase less than the matching rate. Therefore, the proportional increase in passenger demand rate Q and vehicle fleet size N will decrease passengers' matching time.

6. Special cases

When the density of idle drivers is much higher than that of waiting passengers in the matching pools, i.e., $\rho_v \gg \rho_c$, it is called the case that supply dominates the demand or ρ_v is regarded as a dominant supply. Otherwise, demand dominates supply or ρ_c is a dominant demand if $\rho_c \gg \rho_v$. In this section, we investigate several special cases: (1) a matching mechanism with an extremely small matching radius, which resembles a street-hailing taxi market; (2) batch matching with a large matching radius and a dominant supply; and (3) batch matching with a large matching radius and a dominant demand. We analyze and present the properties of the AMP model under these special cases.

6.1. Batch matching with a small matching radius

When the matching radius is small, the matching area is dominated by the matching radius rather than the density of waiting passengers; i.e., $A_M = \pi r^2$.

Proposition 5. *When the matching radius r is extremely small, we have $w_c \approx \frac{\tau}{\rho_v \pi r^2} - \frac{\tau}{2}$, $w_v \approx \frac{\tau}{\rho_c \pi r^2} - \frac{\tau}{2}$ and $w_p \approx \frac{\zeta r}{v}$. Besides that, the matching rate $\frac{M}{\tau} \approx \frac{\pi r^2}{A\tau} m_c m_v$.*

In view of Proposition 5, the matching rate can be rewritten as a Cobb-Douglas type matching model as follows:

$$\frac{M}{\tau} = Q \approx \frac{\pi r^2}{A\tau} (m_c)^{\alpha_1} (m_v)^{\alpha_2} \tag{29}$$

with the return-to-scale factors $\alpha_1 = \alpha_2 = 1$, which indicates that the matching model exhibits increasing returns to scale ($\alpha_1 + \alpha_2 > 1$). This is because when the matching radius r is extremely small, the market resembles a street-hailing taxi market, where a passenger and an idle vehicle can be matched only upon physical encounter. Since customers' origins and destinations are uniformly located, the mass of idle vehicles m_v is also uniformly located. With a given matching strategy and m_c, if the mass of idle vehicles doubles, the matching rate $\frac{M}{\tau} = Q$ (equivalent to the street-hailing market's meeting rate) also doubles, hence $\alpha_2 = 1$. Similarly, treating idle vehicles as the main agents in the matching strategy, given m_v, doubling the mass of passengers also doubles the matching rate $\frac{M}{\tau} = Q$, thus $\alpha_1 = 1$. This observation is also consistent with Zhang et al. (2019)'s findings of increasing returns to scale with $\alpha_1 + \alpha_2 = 2$ in the street-hailing market.

In addition, Proposition 5 indicates that when the matching radius is extremely small, the expected pick-up time w_p can be treated as an exogenous variable that is directly governed by the matching radius r; it is no longer endogenously related to the arrival rate of passengers, matching time of passengers and drivers.

It turns out that this special case is a variant of the model proposed by Yang and Yang (2011) and Zha et al. (2016) with $\alpha_1 = \alpha_2 = 1$, except that the ride time t in their model is replaced by the summation of t and the expected pick-up time $\frac{\zeta r}{v}$. From Eq. (11), we can obtain an explicit formula for the total waiting time of passengers W as follows:

$$W(Q, N, r_s, \tau) = \frac{A\tau}{\left[N - Q\left(t + \frac{\zeta r_s}{v} - \frac{\tau}{2}\right)\right] \pi r_s^2} - \frac{\tau}{2} + \frac{\zeta r_s}{v} \tag{30}$$

where the third argument r_s refers to an extremely small matching radius.

6.2. Batch matching with a large matching radius and a dominant supply

With a large matching radius, the matching area is governed by the density of waiting passengers; i.e., $A_M = {\rho_c}^{-1}$ and the driver-passenger pairing function $M(m_c, m_v, r) = m_c[1 - \exp(-\rho_v/\rho_c)]$. Based on Eq. (23) and Eq. (24), the endogenous variables of ρ_c and ρ_v can be solved by the following implicit equation:

$$-\rho_c \ln\left(1 - \frac{\tau Q}{\rho_c A}\right) = \frac{N}{A} - \frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)}\right)}{2\sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)}} - \sqrt{\frac{1}{\pi}}\left(1 - \frac{\tau Q}{\rho_c A}\right)\right]\sqrt{\rho_c}}{v\tau} - \frac{tQ}{A} + \frac{\tau Q}{2A} \tag{31}$$

and

$$\rho_v = -\rho_c \ln\left(1 - \frac{\tau Q}{\rho_c A}\right) \tag{32}$$

which are independent of the maximum matching radius r as indicated in Proposition 3.

Proposition 6. *When the matching area is governed by the density of waiting passengers and for a dominant supply with $\rho_v \gg \rho_c$, we have $w_c \approx \frac{\tau}{2}$, $w_v \approx \frac{\rho_v}{\rho_c}\tau - \frac{\tau}{2}$ and $w_p = \zeta/\left(2v\sqrt{\rho_v}\right)$.*

Proposition 6 indicates that, with a dominant supply, the competition among waiting passengers over an idle vehicle is negligible and almost all waiting passengers can be successfully matched in their arrival interval. However, the competition among idle vehicles is tough, they must wait for several batches before being matched with a waiting passenger. The expected pick-up time $w_p = \zeta/\left(2v\sqrt{\rho_v}\right)$ only depends on the density of idle vehicles and this result is consistent with the meeting distance formula proposed by Daganzo (1978), Arnott (1996) and Chen et al. (2019) for the transit, taxi and ride-hailing markets, respectively.

When $\rho_v \gg \rho_c$, $p_c \approx 1$, we have $\tau = \frac{M}{Q} = \frac{p_c m_c}{Q} \approx \frac{\rho_c A}{Q}$. The total waiting time of passengers can thus be written as

$$W\left(Q, N, r_l, \frac{\rho_c A}{Q}\right) = \frac{\tau}{2} + \frac{\zeta}{2v\sqrt{\rho_v}} \tag{33}$$

where the third and fourth arguments refer to a large matching radius and a matching interval approaching $\frac{\rho_c A}{Q}$, respectively. Clearly, in this special case, W reduces to Castillo model with an additional term $\frac{\tau}{2}$, which is due to discrete batch matching. From Eq. (10), we find that W can be solved by the following implicit equation with given (Q, N, r, τ):

$$N = \frac{A\zeta^2}{4v^2}\frac{1}{\left(W - \frac{\tau}{2}\right)^2} + \left(W - \frac{\tau}{2}\right)Q + tQ - \frac{\tau Q}{2} \tag{34}$$

6.3. Batch matching with a large matching radius and a dominant demand

In this special case, the matching area is also governed by the density of waiting passengers; i.e., $A_M = {\rho_c}^{-1}$ and ρ_c, ρ_v can be solved by Eqs. (31)–(32).

Proposition 7. *When the matching area is governed by the density of waiting passengers and for a dominant demand with $\rho_c \gg \rho_v$, we have $w_c \approx \left(\frac{m_c}{m_v} - \frac{1}{2}\right)\tau$, $w_v \approx \frac{\tau}{2}$ and $w_p \approx \frac{\zeta}{v\sqrt{\pi\rho_c}}$.*

Proposition 7 indicates that the pick-up time mostly relies on the denser side and is inversely proportional to the square root of its density (Xu et al., 2017). The coefficients of these two formulas differ, because the waiting passengers and idle vehicles are assumed to follow different distributions. In addition, with a dominant demand, idle vehicles can be successfully matched in the first matching interval. Then, the total waiting time of passengers W is given by

$$W\left(Q, N, r_l, \frac{\rho_c A}{Q}\right) = w_c + w_p = \frac{\rho_c A}{Q} - \frac{\tau}{2} + \frac{\zeta}{v\sqrt{\pi\rho_c}} \tag{35}$$

where the fourth arguments refer to a matching interval approaching $\frac{\rho_v A}{Q}$.

To summarize, the matching model in this special case has a few interesting features. First, the effective arrival rate of passengers Q (or matching rate) can be expressed as $\frac{1}{\tau}(m_c)^0(m_v)^1$, which is a Cobb-Douglas type matching function with $\alpha_1 = 0$ (for waiting passengers) and $\alpha_2 = 1$ (for idle vehicles), and implies that the sensitivity of matching rate to the number of waiting passengers is zero in the market scenarios with dominant demand. Second, since $w_c \approx \frac{m_c}{m_v/\tau} - \frac{\tau}{2}$, passengers' matching time is linearly proportional to the ratio of the number of waiting passengers to the arrival rate of idle vehicles when the matching interval is very short. This is analogous to the bottleneck model, such that the mean waiting time at the bottleneck (i.e., matching time of passengers) is equal to the queue lengths (i.e., number of waiting passengers in the pool) divided by the capacity of the bottleneck (i.e., arrival rate of idle vehicles in the pool). Third, the pick-up time is inversely proportional to the square root of the number of waiting passengers, which is contrasted to the assumption of Castillo model that the pick-up time is inversely proportional to the square root of the number of idle vehicles. This is because this bottleneck type queuing model is appropriate for the market scenarios with a large quantity of demand and limited supply, while Castillo model is suitable for cases with a larger number of idle vehicles and only one waiting passenger at any instant of matching.

6.4. Summary of the specific results

Passengers' and drivers' matching time and pick-up time in each special case are summarized in the following Table 1. Special cases 1-3 in Table 1 represent the matching scenarios analyzed in Subsections 6.1-6.3, respectively. The Cobb-Douglas type matching model in Table 1 is defined as $Q = a(m_c)^{\alpha_1}(m_v)^{\alpha_2}$.

Waiting times for both passengers and drivers are solvable with given (Q, N, r, τ), then we analyze under what matching strategies (r, τ) or market conditions (Q, N), these special cases will occur. The service rate of the MoD system is $\frac{N}{t}$ and the service intensity of these N drivers is $\frac{tQ}{N}$, which is a measure of vehicle utilization. Under the assumption $N > \left(t + \frac{\tau}{2}\right)Q$, the vehicle utilization $\frac{tQ}{N} \in \left(0, \frac{1}{1+\frac{\tau}{2t}}\right)$. As demonstrated in Table 1, three batch-matching special cases are analyzed, and the conditions of occurrence are summarized below (see the Appendix for a detailed expression):

Table 1: Summary of passengers' and drivers matching time and pick-up time under these three special cases.

Special cases	r	**Relationship between ρ_v and ρ_c**	A_M	w_c	w_v	w_p	**Matching model**
1	$\to 0$	undetermined	πr^2	$\frac{\tau}{\rho_v \pi r^2} - \frac{\tau}{2}$	$\frac{\tau}{\rho_c \pi r^2} - \frac{\tau}{2}$	$\frac{\zeta r}{v}$	A Cobb-Douglas type matching model with $\alpha_1 = 1$, $\alpha_2 = 1$
2	$\to \infty$	$\rho_v \gg \rho_c$	$\frac{1}{\rho_c}$	$\frac{\tau}{2}$	$\frac{\rho_v}{\rho_c}\tau - \frac{\tau}{2}$	$\frac{\zeta}{2v\sqrt{\rho_v}}$	Castillo model: $W = \frac{\tau}{2} + \frac{\zeta}{2v\sqrt{\rho_v}}$
3	$\to \infty$	$\rho_c \gg \rho_v$	$\frac{1}{\rho_c}$	$\frac{\rho_c}{\rho_v}\tau - \frac{\tau}{2}$	$\frac{\tau}{2}$	$\frac{\zeta}{v\sqrt{\pi\rho_c}}$	A bottleneck type queuing model: $W = \frac{\rho_c}{\rho_v}\tau + \frac{\zeta}{v\sqrt{\pi\rho_c}} - \frac{\tau}{2}$

- When the platform sets a small matching radius, the AMP model reduces to Special case 1, which is the Cobb-Douglas type matching model developed by Yang and Yang (2011).

- When the platform sets a large matching radius 1) with an extremely small matching interval and sufficient supply; or 2) with the vehicle utilization approaching 0, the AMP model reduces to Special case 2, which is Castillo model.

- When the platform sets a large matching radius 1) with an extremely small matching interval and limited supply; or 2) with the vehicle utilization approaching $\frac{1}{1+\frac{\tau}{2t}}$, the AMP model reduces to Special case 3. This bottleneck type queuing model has not been discussed in the literature but has intuitive assumptions and settings. As mentioned in Section 6.3, this model assumes the pick-up time is inversely proportional to the square root of the number of waiting passengers, and the matching time of passengers is derived from a deterministic queue at a bottleneck with a capacity equal to the arrival rate of idle vehicles and queue length equal to the number of waiting passengers.

In addition, as analyzed in Subsections 6.2 and 6.3, we have $M(m_c, m_v, r) = m_c[1 - \exp(-\rho_v/\rho_c)] \approx \min(m_c, m_v)$ when $\rho_v \gg \rho_c$ or $\rho_c \gg \rho_v$ and $(\rho_c)^{-1} < \pi r^2$. This indicates that Special cases 2 and 3 can also be reduced to the perfect matching model, in which the matching rate is given by the minimum of the arrival rate of waiting passengers and idle vehicles (Cachon et al., 2017, Yu et al., 2020).

7. Numerical studies and model validation

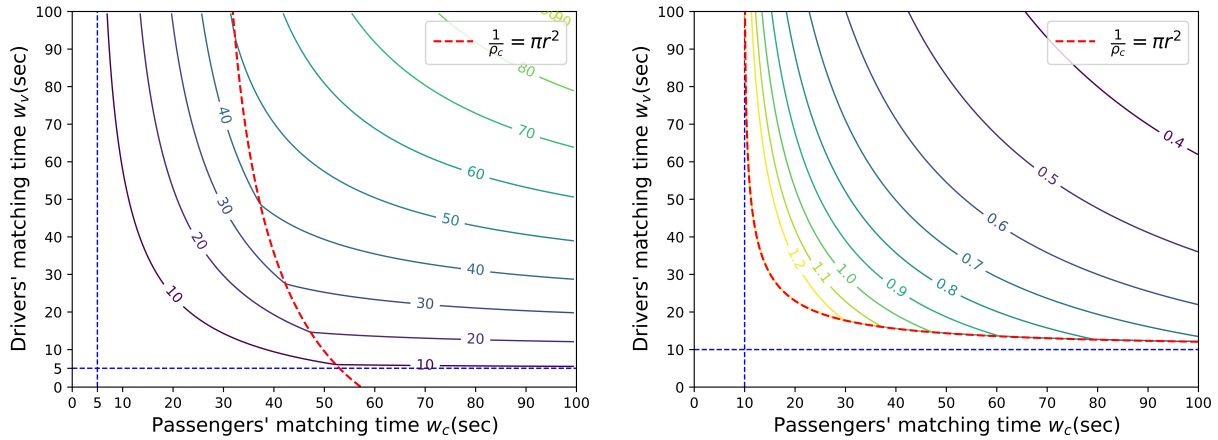
In this section, we conduct numerical experiments to investigate how the matching strategy (τ, r) and the exogenous variables (Q, N) affect system performance measures to illustrate the properties of the AMP model. A set of simulation studies are designed in Subsections 7.2-?? to validate the AMP model and investigate its application scope. Specifically, we examine how well specific matching models, including the Cobb-Douglas type model, Castillo model and the

bottleneck type queuing model, approximate the proposed AMP model in diverse scenarios of demand and supply and matching strategies.

7.1. Sensitivity analyses of matching strategies and market thickness

7.1.1. Relationships between passengers' and drivers' matching time

Figure 2 examines the relationship between passengers' matching time w_c and drivers' matching time w_v with passengers' demand rate $Q = 10,000$ person/h and the area is $A = 500$ km^2. In Figure 2a, the matching radius is fixed to 1 km. In the area on the left of the dashed red line, $A_M = \pi r^2$; in the area on the right, $A_M = \frac{1}{\rho_c}$. Figure 2a shows that drivers' matching time w_v is a monotonically decreasing convex function of passengers' matching time w_c regardless of the matching interval τ. In addition, w_c and w_v are always greater than $\frac{\tau}{2}$. In Figure 2b, the matching interval is 20 sec, w_c and w_v are always greater than $\frac{\tau}{2}$, i.e., 10 sec. Below the dashed red line, $A_M = \frac{1}{\rho_c}$. From Eq. (25), when $\pi r^2 > \frac{1}{\rho_c}$, we have, $w_v = \left(w_c + \frac{\tau}{2}\right) \ln\left(\frac{w_c+\frac{\tau}{2}}{w_c-\frac{\tau}{2}}\right) - \frac{\tau}{2}$, which overlaps with the dashed red line, and w_c and w_v are independent of the matching radius as stated in Proposition 3. Above the dashed red line, $A_M = \pi r^2$, and we can see that w_v is a monotonically decreasing convex function of passengers' matching time w_c regardless of the matching radius r. Therefore, Figure 2 demonstrates that, given the matching strategy (τ, r), drivers' matching time w_v is a monotonically decreasing convex function of passengers' matching time w_c and vice versa, which is consistent with Proposition 1. Besides that, if the matching interval is relatively large (as shown in Figure 2a) or if the matching radius is relatively large (as shown in Figure 2b), the matching area is governed by the density of waiting passengers, i.e., $A_M = \frac{1}{\rho_c}$. Otherwise, the matching area is governed by the matching radius, i.e., $A_M = \pi r^2$.



(a) Relationships between w_c and w_v under different τ(sec) (Note: the value of τ is displayed on curves).

(b) Relationships between w_c and w_v under different r (km)(Note: the value of r is displayed on curves).

Figure 2: Relationships between w_c and w_v.

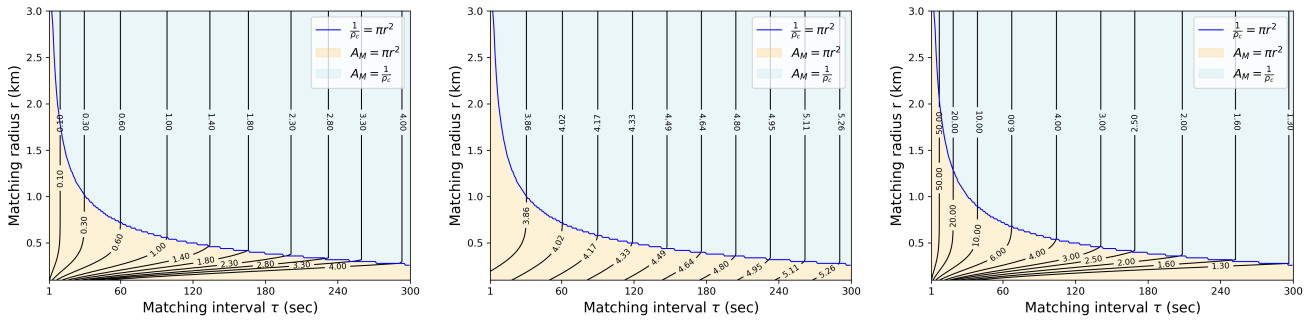
In the experiments illustrated in Figure 3-Figure 4, we study an area of 100 km^2 with a vehicle speed equal to 40 km/h and passenger ride time equal to $\frac{1}{6}$ h. Passengers' demand rate $Q = 3,600$ person/h, the fleet size $N = 1,000$ veh and detour ratio $\zeta = \frac{4}{\pi}$.

Figure 3 depicts the influence of matching strategy (τ, r) on system performance. The blue line is obtained by solving Eqs. (23)–(24). The two-dimensional space of the matching interval and matching radius is divided into two areas by the blue line: above it, the matching area is governed by the density of waiting passengers; i.e., $A_M = \frac{1}{\rho_c}$; below it, the matching area is governed by the matching radius, i.e., $A_M = \pi r^2$.

We first examine the influences of the matching interval τ. Figures 3a-3b indicate that given a matching radius r, the densities of waiting passengers and drivers increase with the matching interval τ, since a larger matching interval leads to an accumulation of more waiting passengers and drivers. Higher densities of passengers and drivers also result in a shorter pick-up time, as shown in Figure 3g. Figure 3e and Figure 3f show that passengers' matching time w_c and drivers' matching time increases with the matching interval τ for the length of each batch matching interval is longer. Then from Eq. (5) and Eq. (6), we can find that m_c and m_v increases faster than $\frac{\tau}{2}Q$. Figure 3c indicates ρ_v/ρ_c decreases with the matching interval τ, which implies ρ_c increases faster than ρ_v with the matching interval τ. Interestingly, the matching interval τ has different impacts on the matching probability of passenger p_c in different matching area scenarios as displayed in Figure 3d. When $A_M = \pi r^2$, a passenger's matching area is governed by the matching radius, then, given the matching radius, he/she can be matched is only determined by the density of idle vehicles. Since ρ_v increases with τ, the matching probability of passengers also increases with τ. In contrast, if $A_M = \frac{1}{\rho_c}$, the matching probability of passengers is determined by the ratio of the density of idle drivers to the density of waiting passengers ρ_v/ρ_c. Since ρ_c increases faster than ρ_v, the matching probability of passengers decreases with the matching interval τ. We also investigate the influence of the matching radius r. In Figure 3, we find that the density of waiting passengers ρ_c, the density of idle drivers ρ_v, passengers' matching time w_c and drivers' matching time w_v decrease with the matching radius r. Conversely, the ratio of the density of idle drivers to the density of waiting passengers ρ_v/ρ_c, the matching probability of passengers p_c, and expected pick-up time w_p increase with matching radius r.

We next examine the impacts of the exogenous inputs Q and N on system performance in Figure 4. In this experiment, the matching interval and matching radius are fixed as $\tau = 10$ sec and $r = 1$ km, respectively. Given Q, when N is in the shaded area with the color as the legend color of Q, the matching area is governed by the density of waiting passengers (e.g., the shaded green area shows that when $Q = 9.0 \times 10^3$ person/h and N is in the shaded green area, the matching area is governed by the density of waiting passengers). The left border of the shaded area is the lower bound of the fleet size; i.e., $N = Q\left(t + \frac{\tau}{2}\right)$. Along the right border of the shaded area, fleet size N satisfies $\frac{1}{\rho_c} = \pi r^2$, which is the upper bound of the fleet size that makes $A_M = \frac{1}{\rho_c}$. As shown in Figure 4a and 4b, passengers' matching time w_c increases with passengers' demand rate Q and decreases with vehicle fleet size N, while drivers' matching time w_v exhibits the opposite trend. Since a larger Q or a smaller N (a larger N or a smaller Q) implies more competition among passengers (drivers), this causes a longer passengers' (drivers') matching time. However, the expected pick-up time w_p is not monotonic with passengers' demand rate Q or drivers' supply

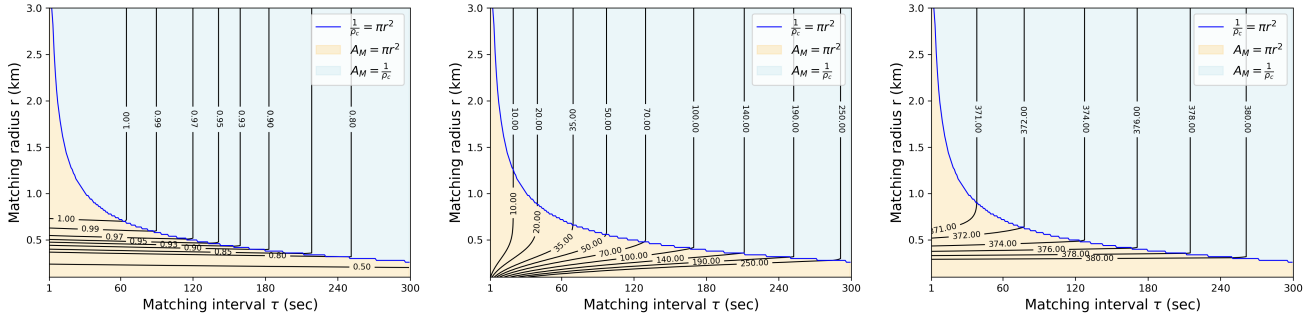
N, as shown in Figure 4c. The reason is that the density of waiting passengers ρ_c and idle vehicles ρ_v have the opposite monotonicity with respect to N or Q.



(a) Density of waiting passengers ρ_c (person/km^2).

(b) Density of idle drivers ρ_v (veh/km^2).

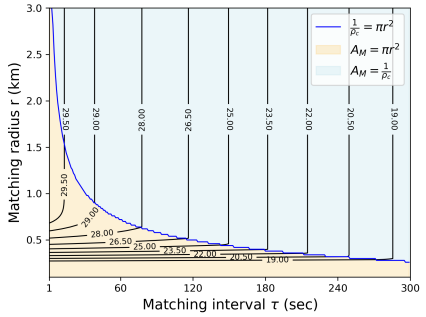
(c) Ratio of the density of idle drivers to the density of waiting passengers ρ_v/ρ_c.



(d) Matching probability of passengers p_c.

(e) Passengers' matching time w_c (sec).

(f) Drivers' matching time w_v (sec).

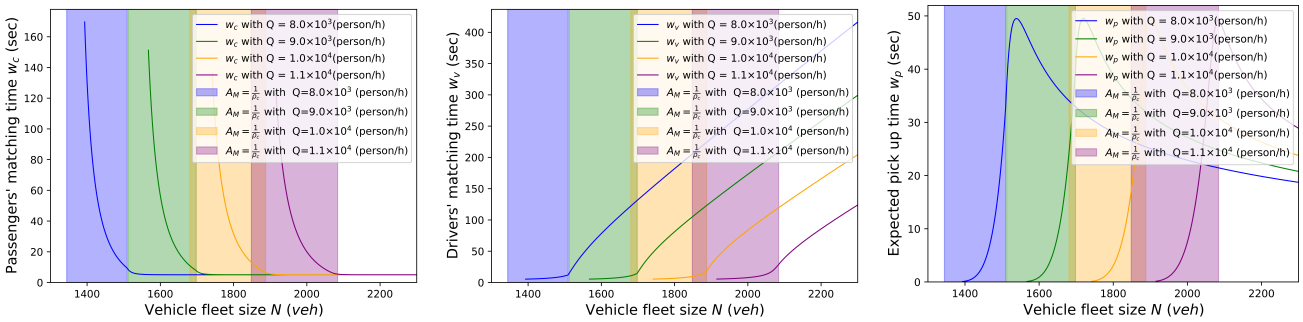


(g) Expected pick-up time w_p (sec).

Figure 3: Influences of the matching strategy (τ, r)

7.1.2. Effects of the scaling factor

The impacts of the scaling factor of passenger demand rate and vehicle fleet size are examined in Figure 5 and Figure 6. In this experiment, the matching interval is fixed as $\tau = 10$ sec, the initial vehicle fleet size N and the initial passenger demand rate Q are 200 veh and 1000 person/h, respectively. The matching radius is $r = 1$ km in Figure 5a and is $r = 3$ km in Figure 5b. Then,



(a) Passengers' matching time w_c. (b) Drivers' matching time w_v. (c) Expected pick-up time w_p.

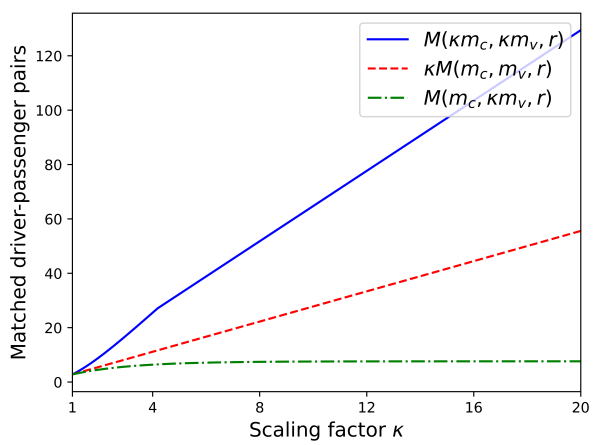
Figure 4: Influences of the market condition (Q, N).

as demonstrated in Figure 3, the matching area is governed by the matching radius in Figure 5a and governed by the density of waiting passengers in Figure 5b. Figure 5 shows that the pairing function has an increasing return to scale when $A_M = \pi r^2$, while has a constant return to scale when $A_M = \frac{1}{\rho_c}$. It indicates that when the number of waiting passengers and idle drivers increases proportionally, the number of matched driver-passenger pairs increases more than proportionally due to the matching restrictions caused by the maximum matching radius r. However, when only the number of idle drivers increases, the number of matched driver-passenger pairs increases less than proportionally. Therefore, Proposition 4.1 is verified.

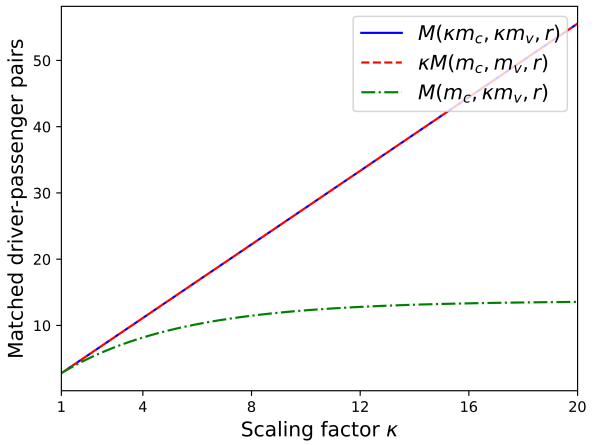
In Figure 6 the vehicle fleet size N and the passenger demand rate Q are scaled proportionally. The matching radius is fixed as $r = 1$ km. As shown in Figure 6a, the expected pick-up time w_p and passengers' matching time w_c decrease with the scaling factor while drivers' matching time w_v increases with the scaling factor. Besides that, when the scaling factor is relatively large, e.g., when $\kappa > 16$, passengers' matching time will reach its minimum value, i.e., half of the matching interval. These findings are consistent with Proposition 4. Furthermore, Figure 6b indicates when scaling up passenger demand rate Q and driver supply N with the same scaling factor $\kappa > 1$, w_c, w_p, W cannot reduce proportionally to $\frac{1}{\kappa}$ of their initial values (represented by w_c^1, w_p^1, W^1, respectively, in Figure 6b). Even though this result is difficult to prove theoretically, we have verified this result with different (Q, N, r, τ) settings.

7.2. Model validation with a simulation study

In this subsection, a set of simulation studies based on a comprehensive agent-based simulator is conducted to evaluate how well the AMP model fits the matching outcomes, which are treated as a proxy for the real markets. Vehicles move at a speed equivalent to 1 cell edge length per second, corresponding to 40km/h. Passenger request times follow a uniform distribution, and both the origin and destination locations of the requests are uniformly distributed within the service

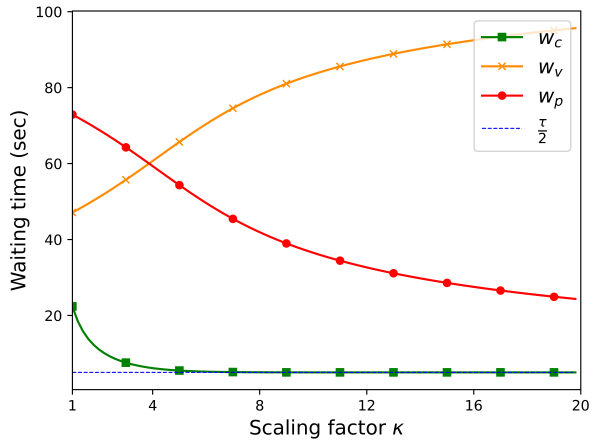


(a) Increasing return to scale of $M\left(m_c, m_v, r\right)$ when $A_M = \pi r^2$.

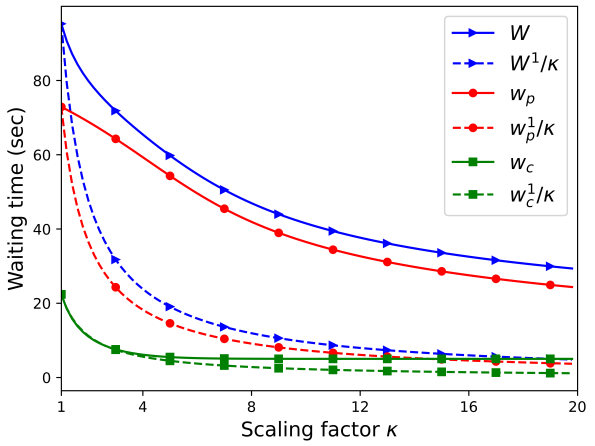


(b) Constant return to scale of $M\left(m_c, m_v, r\right)$ when $A_M = \frac{1}{\rho_c}$.

Figure 5: Return to scale of $M\left(m_c, m_v, r\right)$.



(a) Changes in w_c, w_v and w_p with the scaling factor of the market.



(b) Compare w_c, w_p and W with $\frac{w_c^1}{\kappa}$, $\frac{w_p^1}{\kappa}$ and $\frac{W^1}{\kappa}$.

Figure 6: Influences of the scaling factor κ of the market condition (Q, N).

area. The trip length for passengers is uniformly distributed, with a mean of $\frac{1}{6}$ h [2]. To enhance the realism of our simulations, passengers' cancellation behavior is incorporated and passengers' maximum endurable matching time is set as 300 sec. The matching process of the simulator is summarized as the following Algorithm 1 and the simulation length is 5 h. The drivers' average matching time is used as a condition for judging whether a stationary state is reached, and it will be calculated at the end of each batch matching interval. Take the market condition where $Q = 3,600$ person/h, $N = 1,000$ veh and the matching strategy as $\tau = 5$ sec, $r = 2$ km for an example, Figure 7 illustrates the evolution of drivers' matching times and the progression towards the stable state. Given that customers' trip time is a stochastic variable, drivers' matching time also exhibits stochastic behavior, illustrated by the blue line. However, we observe that the rolling mean of w_v, utilizing a rolling window of 250, stabilizes around $9,000$ seconds, equivalent to 2.5 hours, as indicated by the red line. Therefore, to ensure the simulation reaches a stationary state, we only take the average matching time of passengers and drivers in the last, i.e., the 5^{th} simulated hour for subsequent analysis.

By implementing various matching strategies (as shown in Figure 8) and adjusting for different

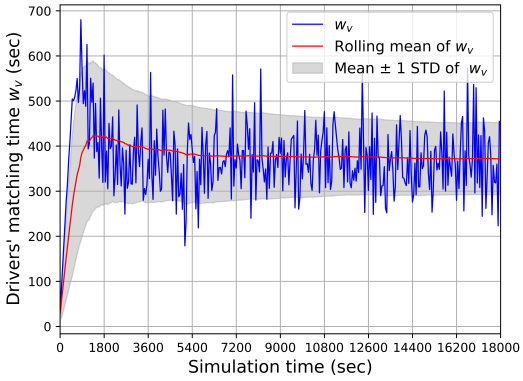


Figure 7: The process of reaching stable state of drivers' matching time (w_v).

market conditions (as illustrated in Figure 9), we are able to simulate and analyze the market performance under diverse scenarios. The mean absolute percentage error (MAPE) is used to evaluate and investigate the AMP model discussed in Section 4. MAPE is calculated by

$$MAPE = \left| \frac{\hat{x} - x}{x} \right| \cdot 100\% \tag{36}$$

where $\hat{x}$ is the estimated label, which is generated by the AMP model (18)–(22), and x is the true label outputted by the simulator. In Figure 8b-8d and Figure 9b-9d, the size of the points represents the value of MAPE of the estimated labels.

With $Q = 3,600$ person/h and $N = 1,000$ veh, Figure 8 shows the performance when the matching radius $r > 0.1$ km and the matching interval $\tau < 50$ sec, MAPE of passengers' matching

[2]The mean value of the Manhattan distance between two uniformly distributed points in a square is $\frac{2}{3}$ grid edge length, then the mean trip time is $\frac{2}{3} \times 10/40\text{h} = \frac{1}{6}\text{h}$

Algorithm 1 Simulator for an MoD services market with batch matching

Input: Information of the passenger requests (request time, coordinates of the trip's origin and destination), drivers' initial coordinates, distributions of passengers' maximum endurable matching time

Output: w_c, w_v, w_p and passengers' cancellation rate

Initialize: Set all drivers are available
for simulation time $h <$ simulation length T **do**
  Request generation: The new requests are added to the list of unserved requests
  Queue abandonment: Passengers whose accumulated matching time exceeds their maximum endurable matching time abandon the queue and are removed from the waiting list
  Completed trips update: Update the status of drivers who completes their trips from "occupied" to "idle" and add them to the list of idle vehicles
  Idle vehicle cruise: For drivers in an 'idle' status, they will cruise along their current direction until they reach the boundary of the area, at which point they will randomly select a new direction from the available options for cruising
  for $h = n \cdot \tau$, $n = 1, 2, 3 \ldots$ **do**
    Batch matching: Conduct batch matching between the lists of unserved requests and idle vehicles with pick-up distance within the matching radius, following the total pick-up minimization rule
  end for
  Update matching outcomes: The matched requests are removed from the list of unserved passenger requests; the status of the matched drivers is updated from "idle" to "occupied"
  Update occupied drivers' state: A matched driver will remain in the 'occupied' state for a duration equal to the sum of the Manhattan distances between its initial position (x_o^v, y_o^v) and the matched passenger's position (x_o^c, y_o^c), the matched passenger's position and the destination position (x_d^c, y_d^c)[a]. After this duration, the driver's position is updated to (x_d^c, y_d^c).
end for

[a] Since the driver's speed is one cell edge length per second.

time w_c, drivers' matching time w_v and expected pick-up time w_p is less than 10%. While when the matching radius is very small (e.g., $r < 0.5$ km) or the matching interval is very long ($\tau > 300$ sec), passengers' cancellation rate and MAPE of each estimated label are relatively large. Since passengers' abandonment is not considered in our theoretical model, it may result in a large MAPE of the estimated label when some passengers cancel orders in the simulation. Considering that the MoD platforms usually set $r > 1$ km and $\tau < 10$ sec, we can justify that the AMP model is a good estimation of passengers' and drivers' matching time and expected pick-up time in MoD services.

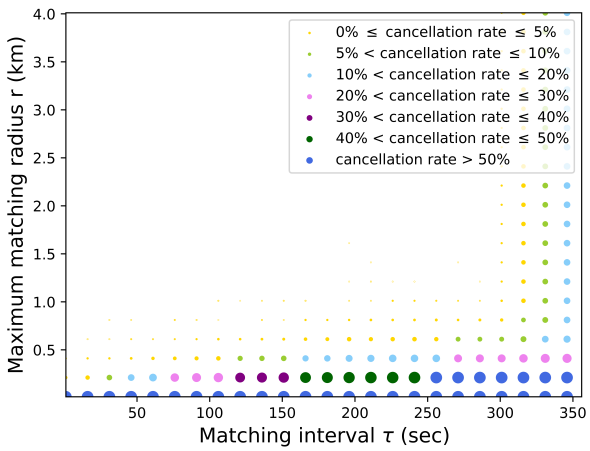
Figure 9 shows the performance when the matching radius $r = 3$ km and matching interval $\tau = 2$ sec. Each point above the red line (which is the line of $N = Q\left(t + \frac{\tau}{2}\right)$) in Figure 9 is a feasible combination of $(Q,\ N)$. When the points are around the red line, the vehicle supply is limited and the market is under-supply; when the points are close to the upper left corner, the demand is limited and the market is over-supply; in other cases, the market can be regarded as a balanced market. From Figure 9, we can find that the AMP model can estimate the matching process of MoD services well with MAPE less than 10% in a balanced market. In an under-supply market, the cancellation rate of passengers increases due to a long passenger matching time, resulting in an increase of MAPE of the estimated labels. In an over-supply market, drivers' matching time w_v estimated by the AMP model can be extremely long (longer than $10,000$ sec), which is not acceptable in reality, and its MAPE in these scenarios is also large.

Figure 8 and Figure 9 indicate that even passengers' cancellation is not explicitly incorporated into the AMP model, it is suitable for portraying the matching process of MoD services under commonly used matching strategies and balanced supply-demand conditions.

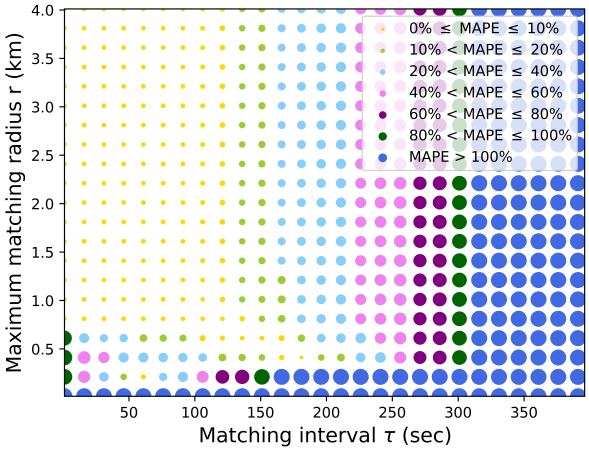
Figure 10-Figure 12 illustrate how well the specific matching models in different special cases discussed in Section 6 approximate the AMP model and the best-fit matching model under different matching strategies and market conditions[3]. The blue line represents results obtained from the AMP model by Eq. (25) under various matching conditions (Q, N) and matching strategies (τ, r). The green, yellow, and brown lines, respectively, represent results obtained by substituting w_c, w_v and w_p in Eq. (25) with the corresponding parameters in Table 1 under the Cobb-Douglas type, Castillo, and bottleneck type queuing models. In the case with an extremely small matching radius ($r = 0.1$ km) that governs the matching area, Figure 10 reveals that the Cobb-Douglas matching model is closest to the AMP model, which is consistent with the analysis in Subsection 6.1. It is established in Figure 10a-10c under different matching strategies with a given market condition, and established in Figure 10d-10f under different market conditions with a given matching strategy. In addition, the Cobb-Douglas matching model is the best-fit model with the simulation among these specific models (Cobb-Douglas type matching model, Castillo model and the bottleneck type queuing model).

Figure 11 and Figure 12 depict the results in the case with a large matching radius (10 km)

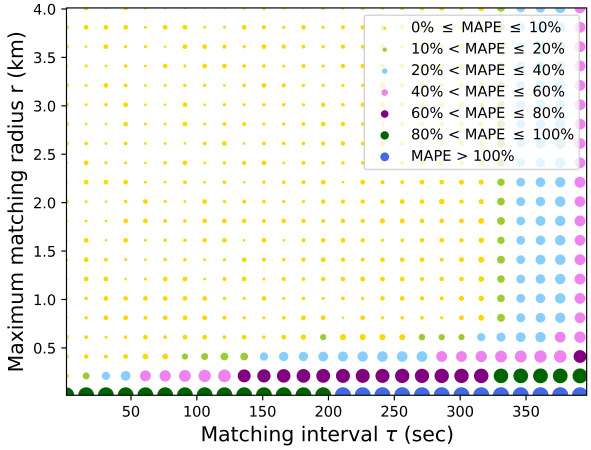
[3]The results of the Cobb-Douglas type matching model in Figure 11f, 12c, 12f, and the results of the bottleneck type queuing model is not depicted in Figure 10b-10c, 10f, 11f for its huge deviation from the real data.



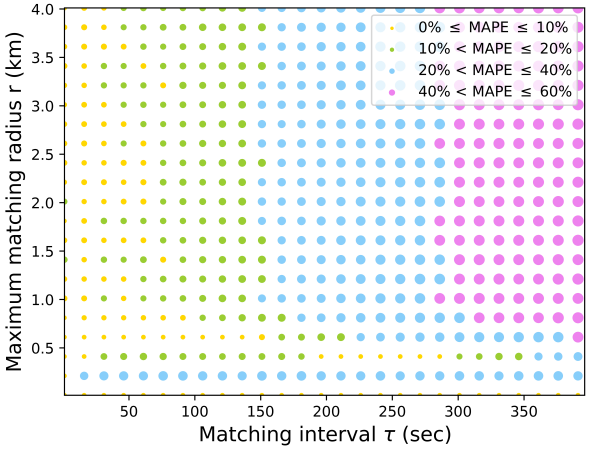
(a) Cancellation rate of the simulation.



(b) MAPE of passengers' matching time (w_c).

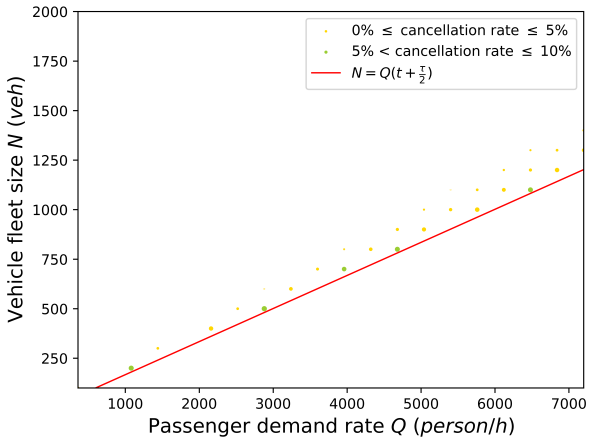


(c) MAPE of drivers' matching time (w_v).

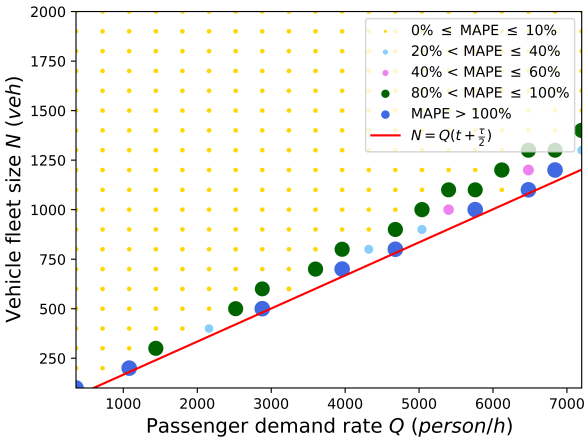


(d) MAPE of expected pick-up time (w_p).

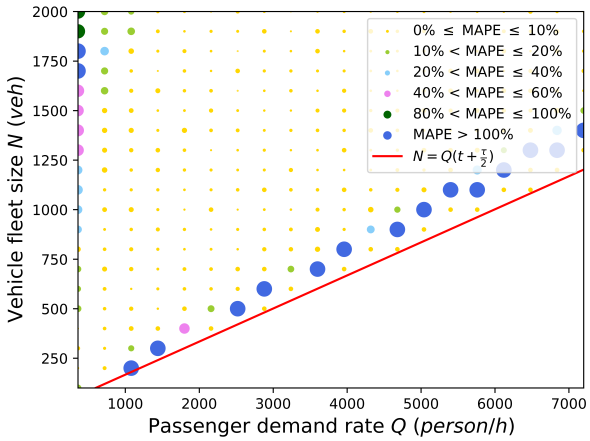
Figure 8: Evaluation of the general matching model under different matching strategies (τ, r) (Note: the size of the points represents the value of cancellation rate or MAPE of the estimated labels).



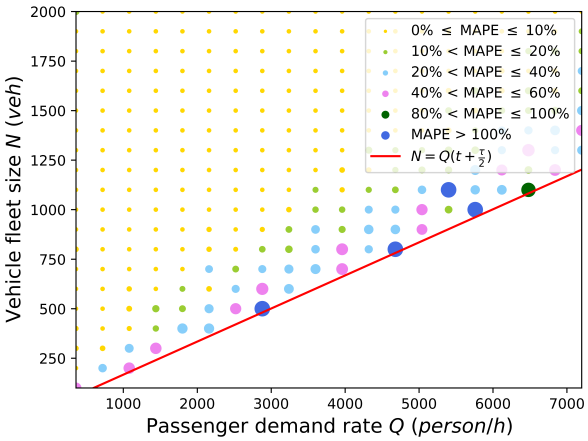
(a) Cancellation rate of the simulation.



(b) MAPE of passengers' matching time (w_c).

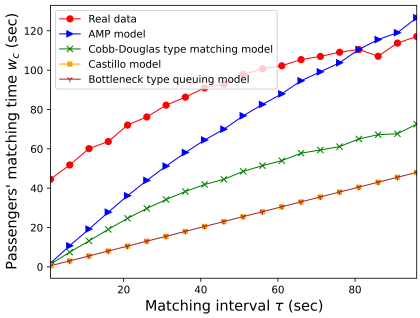


(c) MAPE of drivers' matching time (w_v).

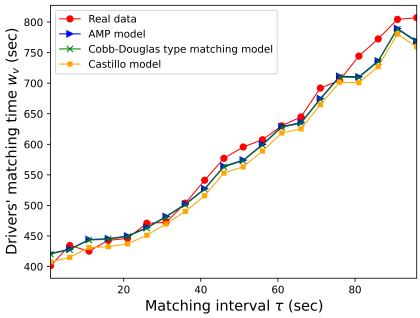


(d) MAPE of expected pick-up time (w_p).

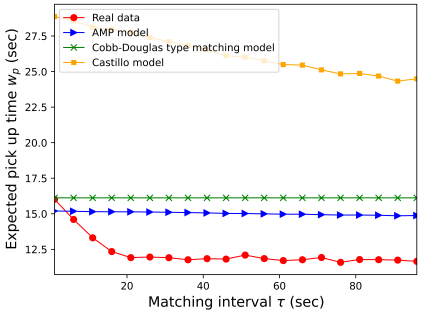
Figure 9: Evaluation of the general matching model under different market conditions (Q, N) (Note: the size of the points represents the value of cancellation rate or MAPE of the estimated labels).



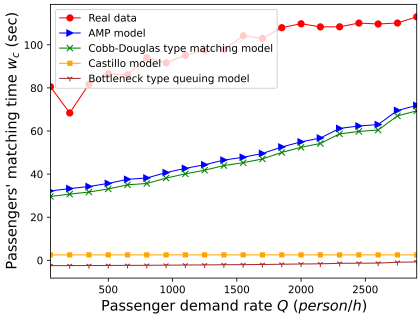
(a) w_c under different matching strategies with $Q = 3600$ person/h, $N = 1000$ veh and $r = 0.1$ km.



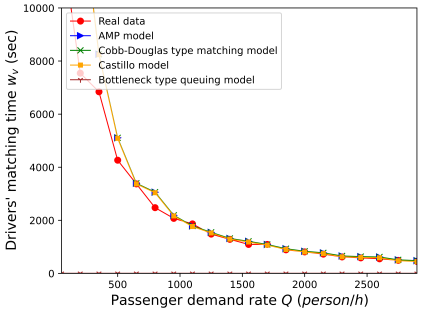
(b) w_v under different matching strategies with $Q = 3600$ person/h, $N = 1000$ veh and $r = 0.1$ km.



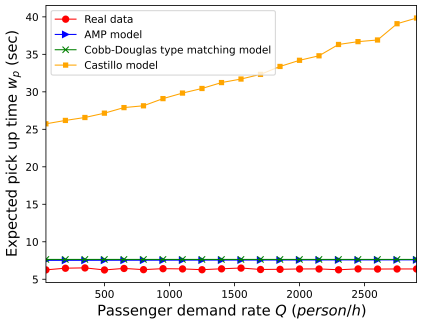
(c) w_p under different matching strategies with $Q = 3600$ person/h, $N = 1000$ veh and $r = 0.1$ km.



(d) w_c under different demand rate with $N = 500$ veh, $\tau = 5$ sec and $r = 0.1$ km.



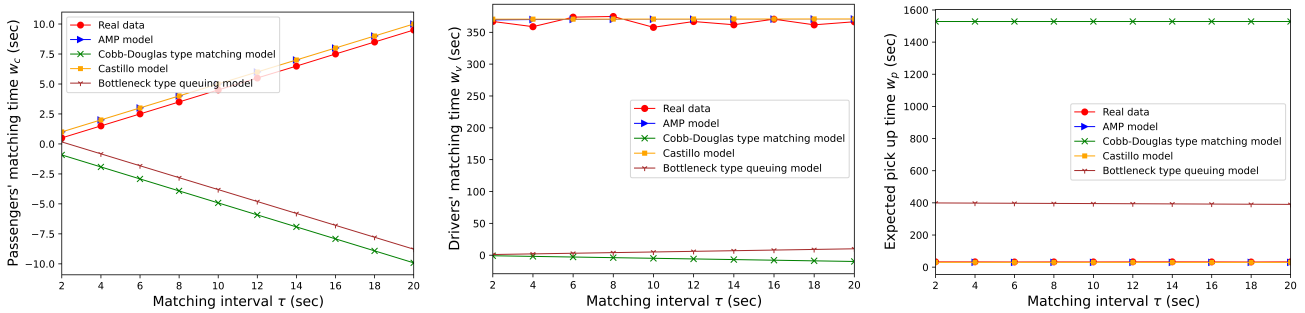
(e) w_v under different demand rate with $N = 500$ veh, $\tau = 5$ sec and $r = 0.1$ km.



(f) w_p under different demand rate with $N = 500$ veh, $\tau = 5$ sec and $r = 0.1$ km.

Figure 10: Batch matching with a small matching radius.

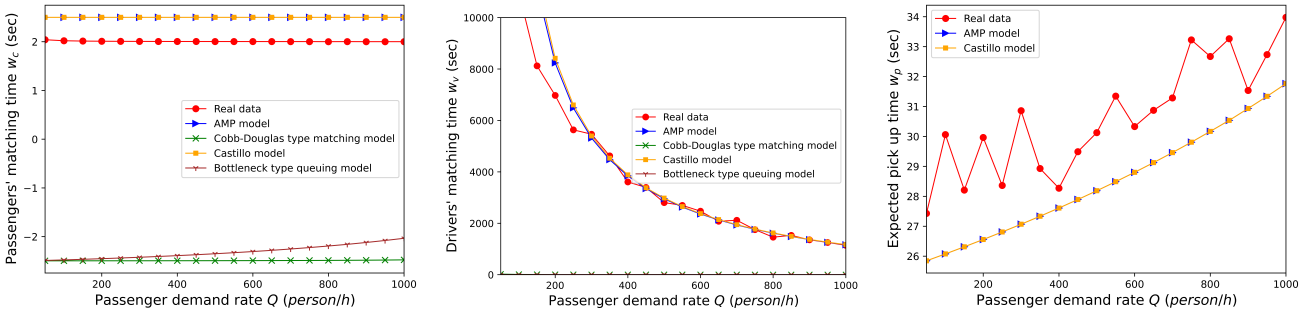
and the matching area governed by the density of waiting passengers. In Figure 11a-11c, the ratio of the density of idle drivers to the density of waiting passengers (ρ_v/ρ_c) is greater than 20 as indicated in Figure 3c. In Figure 11d-11f, the service intensity $\frac{tQ}{N} < \frac{1}{3}$, which implies a dominate supply. Therefore, in the scenario with a large matching radius and a dominant supply displayed in Figure 11, the Castillo model is very close to the AMP model and is the best-fit model among these specific models.



(a) w_c under different matching strategies with $Q = 3600$ person/h, $N = 1000$ veh and $r = 10$ km.

(b) w_v under different matching strategies with $Q = 3600$ person/h, $N = 1000$ veh and $r = 10$ km.

(c) w_p under different matching strategies with $Q = 3600$ person/h, $N = 1000$ veh and $r = 10$ km.



(d) w_c under insufficient demand rate with $N = 500$ veh, $\tau = 5$ sec and $r = 10$ km.

(e) w_v under insufficient demand rate with $N = 500$ veh, $\tau = 5$ sec and $r = 10$ km.

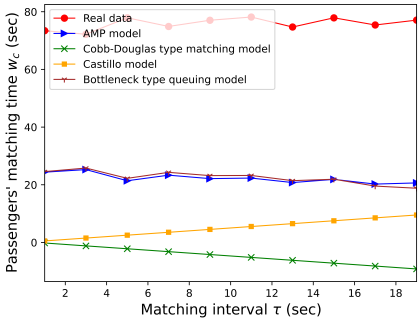
(f) w_p under insufficient demand rate with $N = 500$ veh, $\tau = 5$ sec and $r = 10$ km.

Figure 11: Batch matching with a large matching radius and a dominant supply.

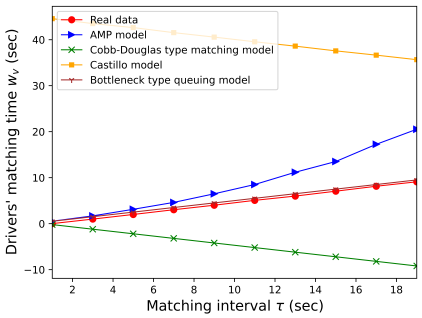
In contrast, the vehicle supply is insufficient in Figure 12. The value $\frac{N-tQ+\frac{\tau}{2}Q}{Q}$ approaches the matching interval τ in Figure 12a-12c, and the market intensity $\frac{tQ}{N}$ is close to $\frac{1}{1+\frac{\tau}{2t}}$ in Figure 12d-12f, which all indicate $\rho_c \gg \rho_v$ as discussed in Subsection 6.4. Therefore, in the scenario with a large matching radius and a dominant demand revealed in Figure 12, the bottleneck type queuing model comes close to the AMP model and is the best-fit model among these specific models.

8. Conclusions

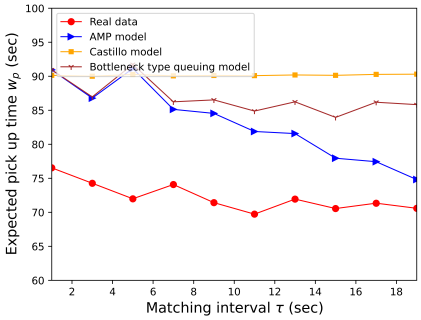
This study introduces an AMP model to characterize the matching process of MoD service markets under different supply and demand conditions and matching strategies in terms of matching interval and matching radius. The existence and uniqueness of the solution of the AMP model,



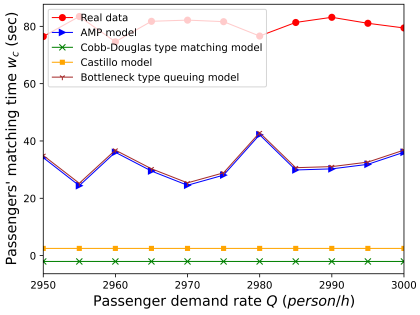
(a) w_c under different matching strategies with $Q = 3600$ person/h, $N = 620$ veh and $r = 10$ km.



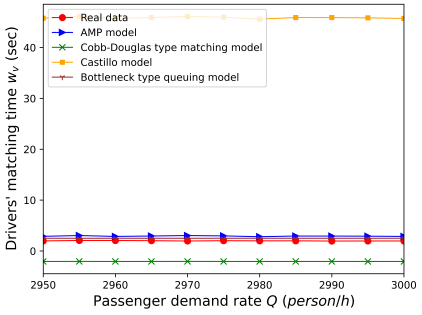
(b) w_v under different matching strategies with $Q = 3600$ person/h, $N = 620$ veh and $r = 10$ km.



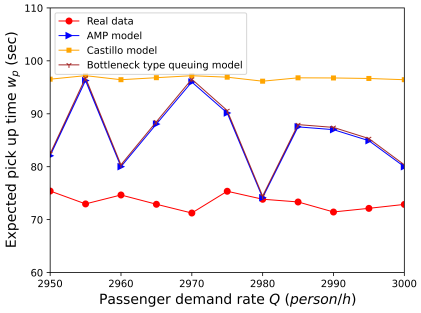
(c) w_p under different matching strategies with $Q = 3600$ person/h, $N = 620$ veh and $r = 10$ km.



(d) w_c under insufficient supply with $N = 500$ veh, $\tau = 5$ sec and $r = 10$ km.



(e) w_v under sufficient supply with $N = 500$ veh, $\tau = 5$ sec and $r = 10$ km.



(f) w_p under sufficient supply with $N = 500$ veh, $\tau = 5$ sec and $r = 10$ km.

Figure 12: Batch matching with a large maximum matching radius and a dominant demand.

the relationship between passengers' matching time and drivers' matching time, and their changes with the scaling factor of passenger demand rate and vehicle fleet size are analyzed. We prove that, without assuming a specific form of the pick-up time function, passengers' matching time and drivers' matching time are always negatively correlated. It is also interesting to find that the expected pick-up time and passengers' matching time decreases but drivers' matching time increases when passenger demand rate and vehicle fleet size are scaled up proportionally.

We show that the AMP model can reduce to some specific matching models which are widely used in the literature. Specifically, if the matching radius is small, the AMP reduces to a form similar to the Cobb-Douglas type matching model developed by Yang and Yang (2011) for street-hailing taxi market. This is reasonable, since an MoD services system with a small matching radius resembles a street-hailing taxi system in which a passenger can hail a taxi in view. If the matching radius is large and the supply is sufficient, the AMP model reduces to the model developed by Castillo et al. (2017), which assumes that a passenger will be immediately matched to the nearest idle vehicle no matter how far the driver is from the passenger. If the matching radius is large and the demand is sufficient, the AMP model reduces to a bottleneck type queuing model, in which passengers' matching time can be approximated by a deterministic queuing model at the bottleneck and their pick-up time is inversely proportional to the square root of the number of waiting passengers. Numerical studies and an agent-based simulation are conducted to demonstrate the application scope of these matching models (namely, how well the models approximate the simulated reality) under different supply-demand conditions and matching strategies.

The AMP model for MoD service markets unifies some existing matching models developed under specific assumptions and thus helps to discern their application scopes. It has potential applications to capture the effects of market fragmentation and thickness on matching frictions in the situation with multiple MoD platforms competing with each other, and the situation with competitive platforms and a third-party integrator that allows passengers to hail rides from multiple platforms through one single integrator (Zhou et al., 2022). Moreover, the AMP model with given demand rate and supply as input can also be applied to study time-of-day MoD services with period-specific demand rate and supply updated using a rolling horizon approach (Yang et al., 2005, Zha et al., 2018a).

Acknowledgments

This work is supported by grants from NSFC of China under Project No. 72001014, grants from Hong Kong Research Grants Council under projects HKUST16208619 and HKU15209121, a grant from NSFC/RGC Joint Research Scheme under project N_HKUST627/18 (NSFC-RGC 71861167001), and Lee Kong Chian Fellowship awarded to the fourth author by Singapore Management University.

References

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

J. Bai, K. C. So, C. S. Tang, X. Chen, and H. Wang. Coordinating supply and demand on an on-demand service platform with impatient customers. *Manufacturing & Service Operations Management*, 21(3):556–570, 2019.

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

S. Benjaafar, J.-Y. Ding, G. Kong, and T. Taylor. Labor welfare in on-demand service platforms. *Manufacturing & Service Operations Management*, 2021.

N. Buchholz. Spatial equilibrium, search frictions, and dynamic efficiency in the taxi industry. *The Review of Economic Studies*, 89(2):556–591, 2022.

G. R. Butters. Equilibrium distributions of sales and advertising prices. *The Review of Economic Studies*, 44(3):465–491, 1977.

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

J. C. Castillo, D. Knoepfle, and G. Weyl. Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 241–242, 2017.

H. Chen, K. Zhang, M. Nie, and X. Liu. A physical model of street ride-hail. *Available at SSRN 3318557*, 2019.

Y. Chen and H. Wang. Why are fairness concerns so important? lessons from a shared last-mile transportation system. *Lessons from a Shared Last-Mile Transportation System (April 25, 2018)*, 2018a.

Y. Chen and H. Wang. Pricing for a last-mile transportation system. *Transportation Research Part B: Methodological*, 107:57–69, 2018b.

S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

C. F. Daganzo. An approximate analytic model of many-to-many demand responsive transportation systems. *Transportation Research*, 12(5):325–333, 1978.

M. Diao, H. Kong, and J. Zhao. Impacts of transportation network companies on urban mobility. *Nature Sustainability*, pages 1–7, 2021.

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

G. Feng, G. Kong, and Z. Wang. We are on the way: Analysis of on-demand ride-hailing systems. *Manufacturing & Service Operations Management*, 2020.

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

P. Guo, C. S. Tang, Y. Tang, and Y. Wang. Gender-based operational issues arising from on-demand ride-hailing platforms: Safety concerns, service systems, and pricing and wage policy. *Service Systems, and Pricing and Wage Policy (September 28, 2018)*, 2018.

J. D. Hall, C. Palsson, and J. Price. Is uber a substitute or complement for public transit? *Journal of Urban Economics*, 108:36–50, 2018.

R. E. Hall. A theory of the natural unemployment rate and the duration of employment. *Journal of monetary economics*, 5(2):153–169, 1979.

B. Hu, M. Hu, and H. Zhu. Surge pricing and two-sided temporal responses in ride hailing. *Manufacturing & Service Operations Management*, 2021.

M. Hu and Y. Zhou. Price, wage, and fixed commission in on-demand matching. *Available at SSRN 2949513*, 2020.

J. Jacob and R. Roet-Green. Ride solo or pool: Designing price-service menus for a ride-sharing platform. *European Journal of Operational Research*, 2021.

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

J. Ke, H. Yang, and Z. Zheng. On ride-pooling and traffic congestion. *Transportation Research Part B: Methodological*, 142:213–231, 2020b.

J. Ke, X. Li, H. Yang, and Y. Yin. Pareto-efficient solutions and regulations of congested ride-sourcing markets with heterogeneous demand and supply. *Available at SSRN 3773481*, 2021a.

J. Ke, Z. Zhu, H. Yang, and Q. He. Equilibrium analyses and operational designs of a coupled market with substitutive and complementary ride-sourcing services to public transits. *Transportation Research Part E: Logistics and Transportation Review*, 148:102236, 2021b.

J. Li and S. Netessine. Higher market thickness reduces matching rate in online platforms: Evidence from a quasiexperiment. *Management Science*, 66(1):271–289, 2020.

S. Li, H. Tavafoghi, K. Poolla, and P. Varaiya. Regulating tncs: Should uber and lyft set their own rules? *Transportation Research Part B: Methodological*, 129:193–225, 2019.

G. Lyu, W. C. Cheung, C.-P. Teo, and H. Wang. Multi-objective online ride-matching. *Available at SSRN 3356823*, 2019.

D. Mo, J. Yu, and X. M. Chen. Modeling and managing heterogeneous ride-sourcing platforms with government subsidies on electric vehicles. *Transportation Research Part B: Methodological*, 139:447–472, 2020.

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

J. A. Parrott and M. Reich. An earnings standard for new york city's app-based drivers. *New York: The New School: Center for New York City Affairs*, 2018.

B. Petrongolo and C. A. Pissarides. Looking into the black box: A survey of the matching function. *Journal of Economic literature*, 39(2):390–431, 2001.

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

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

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

H. R. Varian. *Microeconomic analysis*. Number 338.5 V299m 1992. WW Norton, 1992.

D. A. Vignon, Y. Yin, and J. Ke. Regulating ridesourcing services with product differentiation and congestion externality. *Transportation Research Part C: Emerging Technologies*, 127:103088, 2021.

G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les parallélloèdres primitifs. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 1908(134):198–287, 1908.

G. Wang, H. Zhang, and J. Zhang. On-demand ride-matching in a spatial model with abandonment and cancellation. *Available at SSRN 3414716*, 2019.

H. Wang and A. Odoni. Approximating the performance of a "last mile" transportation system. *Transportation Science*, 50(2):659–675, 2016.

H. Wang and H. Yang. Ridesourcing systems: A framework and review. *Transportation Research Part B: Methodological*, 129:122–155, 2019.

S. Wei, S. Feng, J. Ke, and H. Yang. Calibration and validation of matching functions for ride-sourcing markets. *Communications in Transportation Research*, 2:100058, 2022.

K.-M. Wigand, T. Brandt, and D. Neumann. The effect of autonomous vehicles on consumer welfare in ride-hailing markets. *Available at SSRN*, 2020.

Z. Xu, Y. Yin, and L. Zha. Optimal parking provision for ride-sourcing services. *Transportation Research Part B: Methodological*, 105:559–578, 2017.

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

Z. Xu, Y. Yin, and J. Ye. On the supply curve of ride-hailing systems. *Transportation Research Part B: Methodological*, 132:29–43, 2020.

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

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

H. Yang, M. Ye, W. H.-C. Tang, and S. C. Wong. A multiperiod dynamic model of taxi services with endogenous service intensity. *Operations research*, 53(3):501–515, 2005.

H. Yang, C. W. Leung, S. C. Wong, and M. G. Bell. Equilibria of bilateral taxi–customer searching and meeting on networks. *Transportation Research Part B: Methodological*, 44(8-9):1067–1083, 2010.

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

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

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

T. Yang, H. Yang, S. C. Wong, and N. N. Sze. Returns to scale in the production of taxi services: an empirical analysis. *Transportmetrica A: Transport Science*, 10(9):775–790, 2014.

J. J. Yu, C. S. Tang, Z.-J. Max Shen, and X. M. Chen. A balancing act of regulating on-demand ride services. *Management Science*, 66(7):2975–2992, 2020.

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

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

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

K. Zhang and Y. M. Nie. Inter-platform competition in a regulated ride-hail market with pooling. *Transportation Research Part E: Logistics and Transportation Review*, 151:102327, 2021.

K. Zhang, H. Chen, S. Yao, L. Xu, J. Ge, X. Liu, and M. Nie. An efficiency paradox of uberization. *Available at SSRN 3462912*, 2019.

Y. Zhou, H. Yang, J. Ke, H. Wang, and X. Li. Competition and third-party platform-integration in ride-sourcing markets. *Transportation Research Part B: Methodological*, 159:76–103, 2022.

Z. Zhu, J. Ke, and H. Wang. A mean-field markov decision process model for spatial-temporal subsidies in ride-sourcing markets. *Transportation Research Part B: Methodological*, 150:540–565, 2021.

Appendix A. Mathematical proofs

Proof for Proposition 1

Suppose $(m_{c,1}, m_{v,1})$, $(m_{c,2}, m_{v,2})$ satisfy $M(m_{c,1}, m_{v,1}, r) = \tau Q$, $M(m_{c,2}, m_{v,2}, r) = \tau Q$ and $m_{c,1} > m_{c,2}$. Then, we have $M\left(m_{c,2}, m_{v,2}, r\right) = M\left(m_{c,1}, m_{v,1}, r\right) > M\left(m_{c,2}, m_{v,1}, r\right)$. Therefore, $m_{v,2} > m_{v,1}$. This indicates that m_c decreases with m_v and vice versa.

From Eq. (7) and Eq. (8), w_c and m_c are positively correlated, w_v and m_v are positively correlated, then w_c decreases with w_v and vice versa.

This completes the proof. ■

Proof for Proposition 2

Define $h\left(m_v\right) = m_v + \left[w_p\left(m_c\left(m_v\right), m_v, r\right) + t - \frac{\tau}{2}\right] Q - N$. Then there exists a m_v satisfying Eq.(13), which is equivalent to that there existing a $m_v \in \left[\tau Q, N - tQ + \frac{\tau}{2}Q\right]$ such that $h\left(m_v\right) = 0$. With the assumption of $\lim_{m_v \to \tau Q} w_p\left(m_c\left(m_v\right), m_v, r\right) = 0$, we have

$$\lim_{m_v \to \tau Q} h\left(m_v\right) = \frac{\tau}{2}Q + tQ - N + \lim_{m_v \to \tau Q} w_p\left(m_c\left(m_v\right), m_v, r\right) Q = \frac{\tau}{2}Q + tQ - N \leq 0$$

and

$$h\left(N - tQ + \frac{\tau}{2}Q\right) = w_p\left(m_c\left(N - tQ + \frac{\tau}{2}Q\right), N - tQ + \frac{\tau}{2}Q, r\right) Q \geq 0$$

Since $w_p\left(m_c\left(m_v\right), m_v, r\right)$ is continuous in m_v, $h\left(m_v\right)$ is also continuous in m_v, based on the intermediate value theorem, there exists at least one $m_v^* \in \left[\tau Q, N - tQ + \frac{\tau}{2}Q\right]$ such that $h\left(m_v^*\right) = 0$, which means that there exists at least one solution $\left(m_c\left(m_v^*\right), m_v^*\right)$ of the general matching model that satisfies Eq. (13).

This completes the proof. ■

Proof for Corollary 1

First, we prove the relationship between m_v and m_c. Based on the relationship between $\left(\rho_c\right)^{-1}$ and πr^2, we have the following two cases.

1) When $\left(\rho_c\right)^{-1} \leq \pi r^2$, we have $m_c \geq \frac{A}{\pi r^2}$, and $A_M = \left(\rho_c\right)^{-1} = \frac{A}{m_c}$, based on Eq.(17), we present the mass of idle drivers m_v as a function of the mass of waiting passengers m_c, i.e.,

$$m_v = -m_c \ln\left(1 - \frac{\tau Q}{m_c}\right) \tag{A.1}$$

The first- and second-order derivatives of m_v with respect to m_c can be derived as

$$\frac{\mathrm{d}m_v}{\mathrm{d}m_c} = \ln\left(1 + \frac{\tau Q}{m_c - \tau Q}\right) - \frac{\tau Q}{m_c - \tau Q}$$

$$\frac{\mathrm{d}^2 m_v}{\left(\mathrm{d}m_c\right)^2} = \frac{\left(\tau Q\right)^2}{m_c\left(m_c - \tau Q\right)^2}$$

Since $\ln\left(1 + x\right) - x < 0$ when $x > 0$, we have $\frac{\mathrm{d}m_v}{\mathrm{d}m_c} \leq 0$ and $\frac{\mathrm{d}^2 m_v}{\left(\mathrm{d}m_c\right)^2} \geq 0$ and the equal sign holds if and only if $m_c = \tau Q$. This indicates that m_v is a monotonically decreasing convex function of m_c, and vice versa.

2) When $\left(\rho_c\right)^{-1} > \pi r^2$, we have $m_c < \frac{A}{\pi r^2}$ and $A_M = \pi r^2$, based on Eq. (17), we can get

$$m_v = -\frac{A}{\pi r^2} \ln\left(1 - \frac{\tau Q}{m_c}\right) \tag{A.2}$$

The first- and second-order derivatives of m_v with respect to m_c can be derived as

$$\frac{\mathrm{d}m_v}{\mathrm{d}m_c} = -\frac{A}{\pi r^2} \frac{\tau Q}{m_c\left(m_c - \tau Q\right)}$$

$$\frac{\mathrm{d}^2 m_v}{\left(\mathrm{d}m_c\right)^2} = \frac{A}{\pi r^2}\left[\frac{1}{\left(m_c - \tau Q\right)^2} - \frac{1}{{m_c}^2}\right] > 0$$

This indicates that m_v is a monotonically decreasing convex function of m_c, and vice versa.

Then we prove m_v is a monotonically decreasing convex function of m_c, and vice versa over the entire feasible domain.

3) Since A_M is a continuous function of m_c, in the two cases above, we can see m_v can also be regarded as a continuous function of m_c. Define $g_1(m_c) = -\frac{A}{\pi r^2}\ln\left(1-\frac{\tau Q}{m_c}\right)$ and $g_2(m_c) = -m_c\ln\left(1-\frac{\tau Q}{m_c}\right)$, from the proof in 1) and 2), we have that $g_1(m_c)$ and $g_2(m_c)$ are monotonically decreasing convex functions of m_c. In addition, $g_1(m_c) > g_2(m_c)$ when $m_c < \frac{A}{\pi r^2}$; $g_2(m_c) > g_1(m_c)$ when $m_c > \frac{A}{\pi r^2}$. Then m_v can be represented by the following $g(m_c)$

$$m_v \coloneqq g(m_c) = \begin{cases} g_1(m_c), & m_c < \frac{A}{\pi r^2} \\ g_2(m_c), & m_c \geq \frac{A}{\pi r^2} \end{cases}. \tag{A.3}$$

Suppose $m_{c,1} < \frac{A}{\pi r^2}$ and $m_{c,2} \geq \frac{A}{\pi r^2}$, $0 < \lambda < 1$. If $\lambda m_{c,1} + (1-\lambda) m_{c,2} < \frac{A}{\pi r^2}$, we have

$$\begin{aligned} g(\lambda m_{c,1} + (1-\lambda) m_{c,2}) &= g_1(\lambda m_{c,1} + (1-\lambda) m_{c,2}) < \lambda g_1(m_{c,1}) + (1-\lambda) g_1(m_{c,1}) \\ &< \lambda g_1(m_{c,1}) + (1-\lambda) g_2(m_{c,2}) = \lambda g(m_{c,1}) + (1-\lambda) g(m_{c,2}) \end{aligned}$$

If $\lambda m_{c,1} + (1-\lambda) m_{c,2} \geq \frac{A}{\pi r^2}$, we have

$$\begin{aligned} g(\lambda m_{c,1} + (1-\lambda) m_{c,2}) &= g_2(\lambda m_{c,1} + (1-\lambda) m_{c,2}) < \lambda g_2(m_{c,1}) + (1-\lambda) g_2(m_{c,2}) \\ &< \lambda g_1(m_{c,1}) + (1-\lambda) g_2(m_{c,2}) = \lambda g(m_{c,1}) + (1-\lambda) g(m_{c,2}) \end{aligned}$$

Therefore, $g(m_c)$ is a monotonically decreasing convex function of m_c.

Combining the proofs above, we can conclude that m_v is a monotonically decreasing convex function of m_c, and vice versa.

Second, we prove the relationship between w_c and w_v. Since

$$\frac{\mathrm{d}w_v}{\mathrm{d}w_c} = \frac{\mathrm{d}\left(\frac{m_v}{Q} - \frac{\tau}{2}\right)}{\mathrm{d}\left(\frac{m_c}{Q} - \frac{\tau}{2}\right)} = \frac{\mathrm{d}m_v}{\mathrm{d}m_c}$$

$$\frac{\mathrm{d}^2 w_v}{(\mathrm{d}w_c)^2} = \frac{\mathrm{d}^2\left(\frac{m_v}{Q} - \frac{\tau}{2}\right)}{\left[\mathrm{d}\left(\frac{m_c}{Q} - \frac{\tau}{2}\right)\right]^2} = \frac{\mathrm{d}^2 m_v}{(\mathrm{d}m_c)^2}$$

It can be concluded that w_v is a monotonically decreasing and convex function of w_c, and vice versa.

This completes the proof. ■

Proof for Lemma 1

When $m_v = \tau Q$, if $m_c < \frac{A}{\pi r^2}$, from Eq. (17) we have $\tau Q = m_c\left[1-\exp\left(-\frac{\tau Q \pi r^2}{A}\right)\right] < m_c\left[1-\exp\left(-\frac{\tau Q}{m_c}\right)\right]$, then $\exp\left(-\frac{\tau Q}{m_c}\right) + \frac{\tau Q}{m_c} < 1$. Since $\exp(-x) + x > 1$, when $x > 0$, which is contradicted with $\exp\left(-\frac{\tau Q}{m_c}\right) + \frac{\tau Q}{m_c} < 1$, we can get m_c cannot be less than $\frac{A}{\pi r^2}$ and $m_c \geq \frac{A}{\pi r^2}$.

Then we have $\tau Q = m_c\left[1-\exp\left(-\frac{\tau Q}{m_c}\right)\right]$, which is equivalent to $\exp\left(-\frac{\tau Q}{m_c}\right)+\frac{\tau Q}{m_c}=1$. This equation holds if and only if m_c approaches $+\infty$.

When $m_v = \tau Q$ and $m_c \to +\infty$, based on Taylor expansion, we thus conclude that $\operatorname{erf}\left(\sqrt{\rho_v A_M}\right) = \operatorname{erf}\left(\sqrt{m_v/m_c}\right) \approx \frac{2}{\sqrt{\pi}}\sqrt{m_v/m_c}$, $\exp\left(-\rho_v A_M\right) = \exp\left(-m_v/m_c\right) \approx 1 - m_v/m_c$. From Eq. (22) we conclude that $w_p = \frac{\zeta}{v}\sqrt{\frac{A}{\pi m_c}}$, which approaches 0.

This completes the proof. ■

Proof for Corollary 3

Based on the values on the left and right sides of Eq. (26), we have the following two cases.

1) When $\pi r^2 \le (\rho_c)^{-1}$, Eq. (23) and Eq. (24) can be reduced to

$$\rho_c = \frac{\tau Q}{\left[1-\exp\left(-\pi r^2 \rho_v\right)\right] A} \tag{A.4}$$

$$\rho_v = \left\{ N - Q\left[\frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{\pi r^2 \rho_v}\right)}{2\sqrt{\rho_v}} - r\cdot\exp\left(-\pi r^2\rho_v\right)\right]}{v\left[1-\exp\left(-\pi r^2\rho_v\right)\right]} + t - \frac{\tau}{2}\right]\right\}/A \tag{A.5}$$

The condition $\pi r^2 \le (\rho_c)^{-1}$ requires $\rho_c = \frac{\tau Q}{[1-\exp(-\pi r^2\rho_v)]A} \le \frac{1}{\pi r^2} \Leftrightarrow \rho_v \ge -\frac{\ln(1-\phi)}{\pi r^2}$ and $\phi \le \frac{\tau Q}{\rho_c A} = \frac{M}{m_c} \le 1$. Therefore, when there exists a $\rho_v \ge -\frac{\ln(1-\phi)}{\pi r^2}$ satisfying Eq. (A.5), we can also get the corresponding solution ρ_c satisfies $\rho_c < \frac{1}{\pi r^2}$ by Eq. (A.4).

Define $h_1\left(\rho_v\right) = \rho_v - \left\{ N - Q\left[\frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{\pi r^2 \rho_v}\right)}{2\sqrt{\rho_v}} - r\cdot\exp\left(-\pi r^2\rho_v\right)\right]}{v\left[1-\exp\left(-\pi r^2\rho_v\right)\right]} + t - \frac{\tau}{2}\right]\right\}/A$. Then there exists a ρ_v satisfying Eq. (A.5), which is equivalent to that there existing a $\rho_v \ge -\frac{\ln(1-\phi)}{\pi r^2}$ such that $h\left(\rho_v\right) = 0$. Condition (27) implies $h_1(-\frac{\ln(1-\phi)}{\pi r^2}) \le 0$. In addition, when $\rho_v \to \frac{N-tQ+\frac{\tau}{2}Q}{A}$, $h_1\left(\frac{N-tQ+\frac{\tau}{2}Q}{A}\right) = Qw_p/A \ge 0$. Since $h_1\left(\rho_v\right)$ is continuous with respect to $\rho_v \in \left[-\frac{\ln(1-\phi)}{\pi r^2}, \frac{N-tQ+\frac{\tau}{2}Q}{A}\right]$, based on the intermediate value theorem, there exists at least one $\rho_v \ge -\frac{\ln(1-\phi)}{\pi r^2}$ such that $h_1\left(\rho_v\right) = 0$, which means that there exists at least one solution (ρ_c^*, ρ_v^*) such that $\rho_c \le \frac{1}{\pi r^2}$ and $A_M = \pi r^2$ that satisfies Eqs. (23)–(24) simultaneously.

2) When $\pi r^2 > (\rho_c)^{-1}$, Eq. (23) and Eq. (24) can be reduced to

$$\rho_v = -\rho_c \ln\left(1-\frac{\tau Q}{\rho_c A}\right) \tag{A.6}$$

$$-\rho_c \ln\left(1-\frac{\tau Q}{\rho_c A}\right) = \frac{N}{A} - \frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)}\right)}{2\sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)}} - \sqrt{\frac{1}{\pi}\left(1-\frac{\tau Q}{\rho_c A}\right)}\right]\sqrt{\rho_c}}{v\tau} - \frac{tQ}{A} + \frac{\tau Q}{2A} \tag{A.7}$$

When $\phi \ge 1$, since $\frac{\tau Q}{\rho_c A} = \frac{M}{m_c} \le 1$, we can get that $\frac{1}{\rho_c} = \frac{A}{m_c} \le \frac{A}{\tau Q} \le \pi r^2$ and $A_M = \frac{1}{\rho_c}$.

Let

$$h_2(\rho_c) = -\rho_c \ln\left(1-\frac{\tau Q}{\rho_c A}\right) - \left\{ \frac{N}{A} - \frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)}\right)}{2\sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)}} - \sqrt{\frac{1}{\pi}}\left(1-\frac{\tau Q}{\rho_c A}\right)\right]\sqrt{\rho_c}}{v\tau} - \frac{tQ}{A} + \frac{\tau Q}{2A} \right\},$$

then there exists a ρ_c satisfying Eq. (A.7), which is equivalent to that there exists a $\rho_c > \max\left(\frac{1}{\pi r^2}, \frac{\tau Q}{A}\right)$ such that $h_2(\rho_c) = 0$. When $\rho_c \to \frac{\tau Q}{A}$, we have $h_2(\rho_c) = +\infty$, and Condition (28) implies $h_2\left(\frac{1}{\pi r^2}\right) > 0$, thus $h_2\left(\max\left(\frac{1}{\pi r^2}, \frac{\tau Q}{A}\right)\right) > 0$. When $\rho_c \to +\infty$, we have $h_2(+\infty) = -\left(N - tQ + \frac{\tau Q}{2}\right)/A < 0$. Since $h_2(\rho_c)$ is continuous with respect to $\rho_c \in \left(\max\left(\frac{1}{\pi r^2}, \frac{\tau Q}{A}\right), +\infty\right)$, based on the intermediate value theorem, there exists at least one $\rho_c \in \left(\max\left(\frac{1}{\pi r^2}, \frac{\tau Q}{A}\right), +\infty\right)$ such that $h_2(\rho_c) = 0$, which means that there exists at least one solution (ρ_c, ρ_v) such that $\rho_c > \frac{1}{\pi r^2}$ and $A_M = (\rho_c)^{-1}$ that satisfies Eqs. (23)–(24) simultaneously.

This completes the proof. ■

Before proving Proposition 3 and 4, we provide the following Lemma 2 on the pick-up function in Eq. (22).

Lemma 2. *The pick-up time function w_p in Eq. (22) has the following properties:*

1. *w_p is continuous and decreasing in m_c (in m_v) with given m_v(with given m_c), i.e., $\frac{\partial w_p}{\partial m_c} \le 0$ and $\frac{\partial w_p}{\partial m_v} \le 0$.*
2. *Let $m_c(m_v)$ be the mass of waiting passengers that satisfies $M(m_c, m_v, r) = m_c\left[1-\exp\left(-A_M\rho_v\right)\right] = \tau Q$, then $w_p(m_c(m_v), m_v, r)$ is a continuous function that first increases and then decreases in m_v.*

Proof for Lemma 2

1) Define $\rho_v A_M = x$,

$$f(x) = \frac{\left[\frac{\operatorname{erf}\left(\sqrt{x}\right)}{2\sqrt{x}} - \sqrt{\frac{1}{\pi}}\cdot\exp\left(-x\right)\right]}{1-\exp\left(-x\right)} \tag{A.8}$$

and

$$g(x) = \sqrt{x}f(x) = \sqrt{x}\frac{\left[\frac{\operatorname{erf}\left(\sqrt{x}\right)}{2\sqrt{x}} - \sqrt{\frac{1}{\pi}}\cdot\exp\left(-x\right)\right]}{1-\exp\left(-x\right)} \tag{A.9}$$

we can get $w_p = \frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{\rho_v A_M}\right)}{2\sqrt{\rho_v}} - \sqrt{\frac{A_M}{\pi}}\cdot\exp\left(-\rho_v A_M\right)\right]}{v\left[1-\exp\left(-\rho_v A_M\right)\right]} = \frac{\zeta}{v}\sqrt{A_M}\frac{\left[\frac{\operatorname{erf}\left(\sqrt{\rho_v A_M}\right)}{2\sqrt{\rho_v A_M}} - \sqrt{\frac{1}{\pi}}\cdot\exp\left(-\rho_v A_M\right)\right]}{1-\exp\left(-\rho_v A_M\right)} = \frac{\zeta}{v}\sqrt{A_M}f(x) = \frac{\zeta}{v}\frac{1}{\sqrt{\rho_v}}g(x)$. We obtain that $f(x)$ is a decreasing and $g(x)$ is an increasing function for $x > 0$ as shown in the following Figure A.1a–A.1b. Since $\frac{\partial x}{\partial m_c} = \rho_v\frac{\partial A_M}{\partial m_c} = 0$ when $A_M = \pi r^2$, and $= -\frac{m_v}{(m_c)^2}$ when $A_M = \frac{A}{m_c}$, we have $\frac{\partial x}{\partial m_c} \le 0$. Besides that $\frac{\partial x}{\partial m_v} = \frac{A_M}{A} > 0$, then we can obtain $\frac{\partial w_p}{\partial m_c} = \frac{\zeta}{v}\frac{1}{\sqrt{\rho_v}}\frac{\partial g(x)}{\partial x}\frac{\partial x}{\partial m_c} \le 0$ and $\frac{\partial w_p}{\partial m_v} = \frac{\zeta}{v}\sqrt{A_M}\frac{\partial f(x)}{\partial x}\frac{\partial x}{\partial m_v} < 0$.

2) First, we have

$$\frac{\mathrm{d}w_p\left(m_c\left(m_v\right),m_v,r\right)}{\mathrm{d}m_v}=\frac{\partial w_p}{\partial m_c}\frac{\mathrm{d}m_c}{\mathrm{d}m_v}+\frac{\partial w_p}{\partial m_v} \tag{A.10}$$

Based on the relationship between $(\rho_c)^{-1}$ and πr^2, we have the following two cases.

i. When $\phi<1$ and $(\rho_c)^{-1}\leq\pi r^2$, we have $\tau Q\leq m_v\leq-\frac{A}{\pi r^2}\ln\left(1-\phi\right)$, $A_M=(\rho_c)^{-1}=\frac{A}{m_c}$ and $x=\rho_v A_M=\frac{m_v}{m_c}$. From the proof in Corollary 1 and Eq. (17), we have

$$\frac{\mathrm{d}m_c}{\mathrm{d}m_v}=\frac{1}{-\ln\left(1-\frac{\tau Q}{m_c}\right)-\frac{\tau Q}{m_c-\tau Q}}=\frac{1}{\frac{m_v}{m_c}+1-\exp\left(\frac{m_v}{m_c}\right)}=\frac{1}{x+1-\exp\left(x\right)}$$

Then

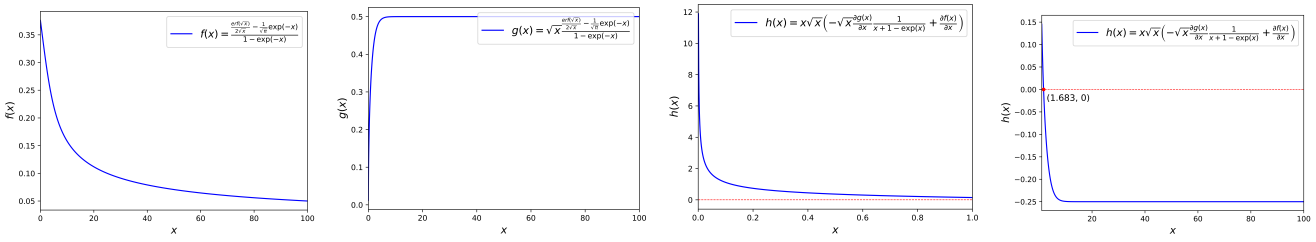
$$\begin{aligned}\frac{\mathrm{d}w_p(m_c(m_v),m_v,r)}{\mathrm{d}m_v}&=\frac{\zeta}{v}\frac{1}{\sqrt{\rho_v}}\frac{\partial g(x)}{\partial x}\frac{\partial x}{\partial m_c}\frac{\mathrm{d}m_c}{\mathrm{d}m_v}+\frac{\zeta}{v}\sqrt{\frac{A}{m_c}}\frac{\partial f(x)}{\partial x}\frac{\partial x}{\partial m_v}\\&=\frac{\zeta}{v}\sqrt{A}\frac{1}{m_v\sqrt{m_v}}x\sqrt{x}\left(-\sqrt{x}\frac{\partial g(x)}{\partial x}\frac{1}{x+1-\exp(x)}+\frac{\partial f(x)}{\partial x}\right)\end{aligned}$$

Define $h\left(x\right)=x\sqrt{x}\left(-\sqrt{x}\frac{\partial g(x)}{\partial x}\frac{1}{x+1-\exp(x)}+\frac{\partial f(x)}{\partial x}\right)$, as shown in Figure A.1c-A.1d, there is only one solution $x^*=1.683$ such that $h\left(x^*\right)=0$ and when $0<x<x^*$, $h\left(x\right)>0$, when $x>x^*$, $h\left(x\right)<0$. Since $\frac{\mathrm{d}m_c}{\mathrm{d}m_v}<0$, we can get $x=\frac{m_v}{m_c}$ increases with m_v. Let m_v^* is the mass of idle vehicles such that $\frac{m_v^*}{m_c(m_v^*)}=x^*$, then we have, when $m_v<m_v^*$, $\frac{\mathrm{d}w_p(m_c(m_v),m_v,r)}{\mathrm{d}m_v}>0$ and when $m_v>m_v^*$, $\frac{\mathrm{d}w_p(m_c(m_v),m_v,r)}{\mathrm{d}m_v}<0$.

ii. When $\phi<1$ and $(\rho_c)^{-1}\geq\pi r^2$, we have $m_v\geq-\frac{A}{\pi r^2}\ln\left(1-\phi\right)$, and $A_M=\pi r^2$. Then from Eq. (A.10) we can get $\frac{\partial w_p}{\partial m_c}=\frac{\zeta}{v}\frac{1}{\sqrt{\rho_v}}\frac{\partial g(x)}{\partial x}\frac{\partial x}{\partial m_c}=0$ and $\frac{\mathrm{d}w_p(m_c(m_v),m_v,r)}{\mathrm{d}m_v}=\frac{\partial w_p}{\partial m_v}<0$. Therefore, when $m_v\geq-\frac{A}{\pi r^2}\ln\left(1-\frac{\tau Q\pi r^2}{A}\right)$, $w_p\left(m_c\left(m_v\right),m_v,r\right)$ decreases in m_v.

iii. When $\phi\geq1$, from the proof in Corollary 1, we have $A_M=(\rho_c)^{-1}=\frac{A}{m_c}$ and $x=\rho_v A_M=\frac{m_v}{m_c}$. Based on the proof in i we can get $w_p\left(m_c\left(m_v\right),m_v,r\right)$ first increases then decreases in m_v.

From the above proof, we have $w_p\left(m_c\left(m_v\right),m_v,r\right)$ first increases then decreases in m_v.



(a) $f\left(x\right)$. (b) $g\left(x\right)$. (c) $h\left(x\right)$ when $0<x\leq1$. (d) $h\left(x\right)$ when $x>1$.

Figure A.1: The curve of $f\left(x\right)$, $g\left(x\right)$ and $h\left(x\right)$.

This completes the proof. ■

Proof for Proposition 3

1) When the matching area is governed by the matching radius, i.e., $A_M = \pi r^2 \leq (\rho_c)^{-1}$, we have the solution of Eqs. (25a)–(25b), ρ_c and ρ_v are solved by Eq. (A.4) and Eq. (A.5) and $\rho_v \geq -\frac{1}{\pi r^2}\ln(1-\phi)$ with $\phi = \frac{\pi r^2 \tau Q}{A}$.

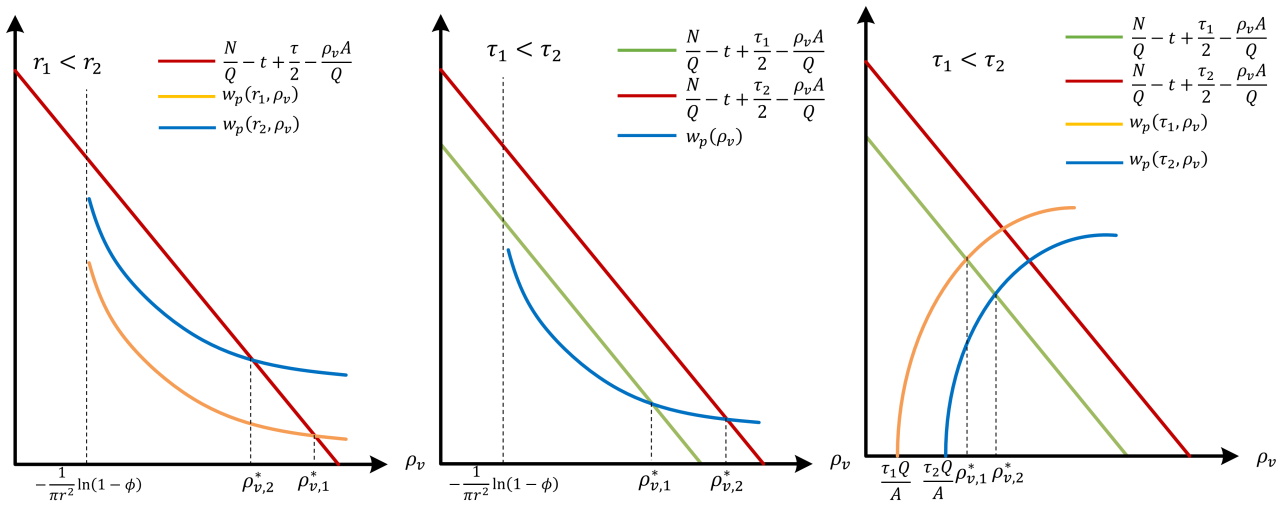
When τ is given, define $w_p(r,\rho_v) = \frac{\zeta\left[\frac{\mathrm{erf}\left(\sqrt{\pi r^2 \rho_v}\right)}{2\sqrt{\rho_v}} - r\cdot\exp\left(-\pi r^2\rho_v\right)\right]}{v[1-\exp(-\pi r^2\rho_v)]}$. We have $w_p(r,\rho_v) = \frac{\zeta}{v}\frac{1}{\sqrt{\rho_v}}g(\pi r^2\rho_v)$ where $g(x)$ is defined in Eq. (A.9) is an increasing function as shown in Figure A.1b. Therefore, for a given ρ_v, when r increases, $w_p(r,\rho_v)$ increases, i.e., $w_p(r_1,\rho_v) > w_p(r_2,\rho_v)$ for $\forall r_1 > r_2$. Besides that $w_p(r,\rho_v) = \frac{\zeta}{v} r f(\pi r^2\rho_v)$ with $f(x)$ is defined in Eq. (A.8) is a decreasing and convex function as shown in Figure A.1a, we have $w_p(r,\rho_v)$ is also a decreasing and convex function with rho_v for any given r. Since the solution of the aggregate matching model (25) (ρ_v^*, ρ_c^*) satisfies $w_p(r,\rho_v^*) = \frac{N}{Q} - t + \frac{\tau}{2} - \frac{\rho_v^* A}{Q}$, i.e., ρ_v^* is the intersection of line $y = \frac{N}{Q} - t + \frac{\tau}{2} - \frac{\rho_v^* A}{Q}$ and the curve $w_p(r,\rho_v)$, therefore, when r increases, ρ_v^* decreases and w_p increases, which is shown in Figure A.2a. Based on Eq. (19), we have $w_v = \left(\frac{\rho_v}{\tau Q} - \frac{1}{2}\right)\tau$, then given τ, we can also have that w_v decreases with r. However, the monotonicity of $\pi r^2\rho_v$ with r is undetermined, thus the monotonicity of $p_c = 1-\exp(-\pi r^2\rho_v)$, $\rho_c = \frac{\tau Q}{p_c}$ and $w_c = \left(\frac{1}{p_c} - \frac{1}{2}\right)\tau$ with r are undetermined, either.

When r is given, then w_p is only related to ρ_v and can be defined as $w_p(\rho_v) = \frac{\zeta}{v}\sqrt{\pi} r f(\pi r^2\rho_v)$, which is a decreasing and convex function of ρ_v. Since the solution of the aggregate matching model (25) (ρ_v^*, ρ_c^*) satisfies $w_p(\rho_v^*) = \frac{N}{Q} - t + \frac{\tau}{2} - \frac{\rho_v^* A}{Q}$, i.e., ρ_v^* is the intersection of line $y = \frac{N}{Q} - t + \frac{\tau}{2} - \frac{\rho_v^* A}{Q}$ and the curve $w_p(\rho_v^*)$. Therefore, when τ increases, ρ_v^* increases and w_p decreases, which is shown in Figure A.2b. Since $p_c = 1-\exp(-\pi r^2\rho_v)$, we have p_c increase with τ. However, the monotonicity of $\rho_c = \frac{\tau Q}{A(1-\exp(-\pi r^2\rho_v))}$, $w_c = \left(\frac{1}{p_c} - \frac{1}{2}\right)\tau$ and $w_v = \left(\frac{\rho_v}{\tau Q} - \frac{1}{2}\right)\tau$ with τ is undetermined.

2) When the matching area is governed by the density of waiting passengers, i.e., $A_M = (\rho_c)^{-1} \leq \pi r^2$, the solution of Eqs. (25a)–(25b), ρ_c and ρ_v are solved by Eq. (A.6) and Eq. (A.7) and and $\rho_v \leq -\frac{1}{\pi r^2}\ln(1-\phi)$. Eq. (A.7) shows that ρ_c is only determined by τ and independent of r. In addition, based on Eq. (A.6), ρ_v is also only determined by τ. Since $p_c = 1-\exp(-\rho_v/\rho_c)$, $w_p = \frac{\zeta\left[\frac{\mathrm{erf}\left(\sqrt{\rho_v/\rho_c}\right)}{2\sqrt{\rho_v}} - \frac{1}{\sqrt{\pi\rho_c}}\cdot\exp(-\rho_v/\rho_c)\right]}{v[1-\exp(-\rho_v/\rho_c)]}$, $w_c = \left(\frac{1}{p_c} - \frac{1}{2}\right)\tau = \left(\frac{m_c}{M} - \frac{1}{2}\right)\tau = \frac{\rho_c A}{Q} - \frac{\tau}{2}$ and $w_v = \left(w_c + \frac{\tau}{2}\right)\ln\left(\frac{w_c+\frac{\tau}{2}}{w_c-\frac{\tau}{2}}\right) - \frac{\tau}{2}$, we have that p_c, w_p, w_c and w_v are determined by τ and independent of r.

Based on the above analysis, we define $w_p(\tau,\rho_v) = \frac{\zeta\left[\frac{\mathrm{erf}\left(\sqrt{\rho_v/\rho_c}\right)}{2\sqrt{\rho_v}} - \frac{1}{\sqrt{\pi\rho_c}}\cdot\exp(-\rho_v/\rho_c)\right]}{v[1-\exp(-\rho_v/\rho_c)]} = \frac{\zeta}{v}\frac{1}{\sqrt{\rho_v}}g\left(\frac{\rho_v}{\rho_c}\right)$, where ρ_c is determined by Eq. (A.6). Since when ρ_v is given, the function of $M(\rho_c,\rho_v,r) = \rho_c A\left[1-\exp\left(-\frac{\rho_v}{\rho_c}\right)\right]$ increases with ρ_c. Therefore, given ρ_v, when τ increase, we have that ρ_c also increases. Since $g(x)$ is an increasing and function, and when τ

increase and ρ_v is given, we have $\frac{\rho_v}{\rho_c}$ decreases, then we can get $w_p(\tau_1, \rho_v) > w_p(\tau_2, \rho_v)$ for $\forall \tau_1 < \tau_2$. Since the solution of the general matching model (25) (ρ_v^*, ρ_c^*) satisfies $w_p(\tau, \rho_v^*) = \frac{N}{Q} - t + \frac{\tau}{2} - \frac{\rho_v^* A}{Q}$, i.e., ρ_v^* is the intersection of line $y = \frac{N}{Q} - t + \frac{\tau}{2} - \frac{\rho_v^* A}{Q}$ and the curve $w_p(\tau, \rho_v^*)$, therefore, when τ increases, ρ_v^* increases and w_p decreases, which is shown in Figure A.2c. However, the monotonicity of ρ_c, p_c, w_c and w_v are undertermined.



(a) $w_p(r, \rho_v)$. (b) $w_p(\rho_v)$. (c) $w_p(\tau, \rho_v)$.

Figure A.2: An illustration of the proof for Proposition 3.

Proof for Proposition 4.1

1) Based on the value of ρ_c, we have the following three cases.

   i. When $\rho_c \geq \frac{1}{\pi r^2}$, we have $\frac{1}{\kappa\rho_c} < \frac{1}{\rho_c} \leq \pi r^2$. Then $M(\kappa m_c, \kappa m_v, r) = \kappa m_c \left[1 - \exp\left(-\frac{\kappa\rho_v}{\kappa\rho_c}\right)\right] = \kappa m_c \left[1 - \exp\left(-\frac{\rho_v}{\rho_c}\right)\right] = \kappa M(m_c, m_v, r)$.

   ii. When $\frac{1}{\kappa\pi r^2} < \rho_c < \frac{1}{\pi r^2}$, we have $\frac{1}{\kappa\rho_c} < \pi r^2 < \frac{1}{\rho_c}$. Then $M(\kappa m_c, \kappa m_v, r) = \kappa m_c \left[1 - \exp\left(-\frac{\kappa\rho_v}{\kappa\rho_c}\right)\right] = \kappa m_c \left[1 - \exp\left(-\frac{\rho_v}{\rho_c}\right)\right] > \kappa m_c [1 - \exp(-\pi r^2 \rho_v)] = \kappa M(m_c, m_v, r)$.

   iii. When $\rho_c \leq \frac{1}{\kappa\pi r^2}$, we have $\pi r^2 \leq \frac{1}{\kappa\rho_c} < \frac{1}{\rho_c}$. Then $M(\kappa m_c, \kappa m_v, r) = \kappa m_c [1 - \exp(-\pi r^2 \kappa \rho_v)] > \kappa m_c [1 - \exp(-\pi r^2 \rho_v)] = \kappa M(m_c, m_v, r)$.

   Combining the three cases, we can conclude that $M(m_c, m_v, r)$ has a constant/increasing return to scale.

2) Since $\kappa M(m_c, m_v, r) - M(m_c, \kappa m_v, r) = \kappa m_c [1 - \exp(-\rho_v A_M)] - m_c [1 - \exp(-\kappa\rho_v A_M)]$. Let $f(\kappa) = \kappa [1 - \exp(-a)] - [1 - \exp(-\kappa a)]$, $a > 0$, $\kappa > 1$. We have $f'(\kappa) = 1 - \exp(-a) - a\exp(-\kappa a)$ and $f''(\kappa) = a^2 \exp(-\kappa a) > 0$. Therefore, $f'(\kappa)$ is an increasing function and reaches its minimum at $\kappa = 1$, i.e., $f'(\kappa) \geq f'(1) = 1 - \exp(-a) - a\exp(-a)$. Define

$g(a) = 1 - \exp(-a) - a\exp(-a)$, we can get $g'(a) = a^2\exp(-a) > 0$, which indicates $g(a) \geq g(0) = 0$. Hence, $f'(\kappa) \geq f'(1) = g(a) \geq g(0) = 0$ and $f(\kappa)$ is an increasing function. Then we can get $f(\kappa) \geq f(1) = 0$ and $\kappa M(m_c, m_v, r) \geq M(m_c, \kappa m_v, r)$.

This completes the proof. ■

Proof for Proposition 4.2 and 4.3

From the proof of Corollary 1, it is easy to get that with a given matching strategy (τ, r) and passenger arrival rate κQ, for any $m_v^\kappa \in [\kappa\tau Q, \kappa N - \kappa t Q + \kappa\frac{\tau}{2}Q]$ there is a unique m_c^κ such that $M(m_c^\kappa, m_v^\kappa, r) = \kappa\tau Q$, and we define this unique m_c^κ as $m_c^\kappa(m_v^\kappa)$. When the vehicle fleet size is κN, we have $m_v^\kappa \in \left[\kappa\tau Q, \kappa N - \kappa t Q + \frac{\kappa\tau}{2}Q\right]$ and $\frac{m_v^\kappa}{\kappa} \in \left[\tau Q, N - tQ + \frac{\tau}{2}Q\right]$. Also define $w_p^\kappa(x) = w_p(m_c^\kappa(\kappa x), \kappa x, r)$, where $x = \frac{m_v^\kappa}{\kappa} \in \left[\tau Q, N - tQ + \frac{\tau}{2}Q\right]$ and $w_p(m_c^\kappa(\kappa x), \kappa x, r)$ is defined in Eq. (22). Based on Lemma 2, we have $w_p^\kappa(x)$ is a continuous and first increasing and then decreasing function in x.

Let $0 < \kappa_1 < \kappa_2$, we have $M(m_c^{\kappa_2}(\kappa_2 x), \kappa_2 x, r) = \kappa_2\tau Q = \frac{\kappa_2}{\kappa_1}\kappa_1\tau Q = \frac{\kappa_2}{\kappa_1}M(m_c^{\kappa_1}(\kappa_1 x), \kappa_1 x, r) \geq M(m_c^{\kappa_1}(\kappa_1 x), \kappa_2 x, r)$, the last inequality is due to $\kappa M(m_c, m_v, r) \geq M(m_c, \kappa m_v, r)$ for $\forall \kappa > 1$. Therefore, $m_c^{\kappa_2}(\kappa_2 x) \geq m_c^{\kappa_1}(\kappa_1 x)$. Since $\kappa_1 x < \kappa_2 x$, we can obtain $w_p^{\kappa_1}(x) = w_p(m_c^{\kappa_1}(\kappa_1 x), \kappa_1 x, r) > w_p(m_c^{\kappa_1}(\kappa_1 x), \kappa_2 x, r) \geq w_p(m_c^{\kappa_2}(\kappa_2 x), \kappa_2 x, r) = w_p^{\kappa_2}(x)$. It indicates $w_p^{\kappa_1}(x) > w_p^{\kappa_2}(x)$ for $\forall\, x \in \left[\tau Q, N - tQ + \frac{\tau}{2}Q\right]$ and $0 < \kappa_1 < \kappa_2$.

The solution of the aggregate matching model in Eq. (25) with passengers' demand rate being κQ and the vehicle supply being κN is defined as $m_v^{\kappa,*}$, then $m_v^{\kappa,*}$ satisfies $\kappa N = \kappa Q\left[w_p(m_c^\kappa(m_v^{\kappa,*}), m_v^{\kappa,*}, r) + t - \frac{\tau}{2}\right] + m_v^{\kappa,*}$. Define $f(x) = \frac{N}{Q} - t + \frac{\tau}{2} - \frac{x}{Q}$, we have that $\frac{m_v^{\kappa,*}}{\kappa} = x^{\kappa,*}$ is the intersection of $f(x)$ and $w_p^\kappa(x)$, i.e., $f(x^{\kappa,*}) = w_p^\kappa(x^{\kappa,*})$. Then for any $0 < \kappa_1 < \kappa_2$, we obtain that $f(x^{\kappa_2,*}) = w_p^{\kappa_2}(x^{\kappa_2,*}) < w_p^{\kappa_1}(x^{\kappa_1,*})$, as shown in Figure A.3. Besides that, we can get $x^{\kappa_2,*} > x^{\kappa_1,*}$, which implies $m_v^{\kappa_2,*} > \frac{\kappa_2}{\kappa_1}m_v^{\kappa_1,*}$. Therefore, the expected pick-up time w_p decreases with the scaling factor [4]. This completes the proof. ■ Since w_c, and w_v can be regarded as functions of (Q, N, τ, r), for ease of notation, we define $w_c^\kappa = w_c(\kappa Q, \kappa N, \tau, r)$, $w_v^\kappa = w_v(\kappa Q, \kappa N, \tau, r)$. Since $w_v = \frac{m_v}{Q} - \frac{\tau}{2}$, we have $w_v^{\kappa_2} = \frac{m_v^{\kappa_2,*}}{\kappa_2 Q} - \frac{\tau}{2} > \frac{\frac{\kappa_2}{\kappa_1}m_v^{\kappa_1,*}}{\kappa_2 Q} - \frac{\tau}{2} = \frac{m_v^{\kappa_1,*}}{\kappa_1 Q} - \frac{\tau}{2} = w_v^{\kappa_1}$. Therefore, drivers' matching time w_v increases with the scaling factor.

Since $M(m_c, m_v, r)$ has constant/increasing return to scale, and according to Proposition 4.1, we have $M\left(\frac{\kappa_2}{\kappa_1}m_c^{\kappa_1}(m_v^{\kappa_1,*}), \frac{\kappa_2}{\kappa_1}m_v^{\kappa_1,*}, r\right) \geq \frac{\kappa_2}{\kappa_1}M(m_c^{\kappa_1}(m_v^{\kappa_1,*}), m_v^{\kappa_1,*}, r) = \frac{\kappa_2}{\kappa_1}\kappa_1\tau Q = \kappa_2\tau Q = M(m_c^{\kappa_2}(m_v^{\kappa_2,*}), m_v^{\kappa_2,*}, r) > M\left(m_c^{\kappa_2}(m_v^{\kappa_2,*}), \frac{\kappa_2}{\kappa_1}m_v^{\kappa_1,*}, r\right)$. It indicates $m_c^{\kappa_2}(m_v^{\kappa_2,*}) < \frac{\kappa_2}{\kappa_1}m_c^{\kappa_1}(m_v^{\kappa_1,*})$. Then $w_c^{\kappa_2} = \frac{m_c^{\kappa_2}(m_v^{\kappa_2,*})}{\kappa_2 Q} - \frac{\tau}{2} < \frac{\frac{\kappa_2}{\kappa_1}m_c^{\kappa_1}(m_v^{\kappa_1,*})}{\kappa_2 Q} - \frac{\tau}{2} = \frac{m_c^{\kappa_1}(m_v^{\kappa_1,*})}{\kappa_1 Q} - \frac{\tau}{2} = w_c^{\kappa_1}$. Therefore, passengers' matching time w_c decreases with the scaling factor.

This completes the proof. ■

[4] Since the unstable solutions are not considered, this proof is still valid when the general matching model has multiple solutions.

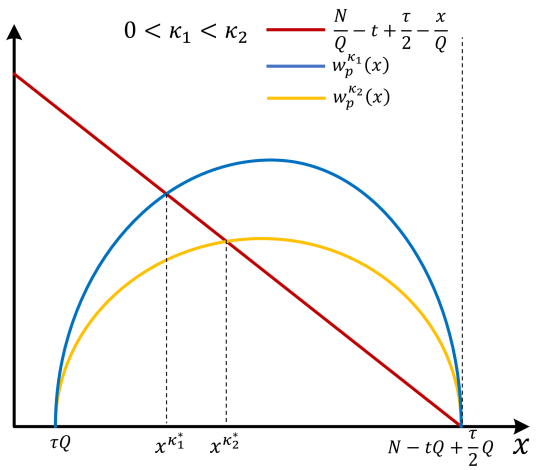


Figure A.3: An illustration for the proof of Proposition 4 with $0 < \kappa_1 < \kappa_2$.

Proof for Proposition 5

When $A_M = \pi r^2$, the expected number of matched driver-passenger pairs and pick-up time is given by

$$M(m_c, m_v, r) = m_c \left[1 - \exp\left(-\rho_v \pi r^2\right)\right] \tag{A.11}$$

$$w_p(m_c, m_v, r) = \frac{\zeta\left[\frac{\operatorname{erf}\left(\sqrt{\rho_v \pi r^2}\right)}{2\sqrt{\rho_v}} - r \cdot \exp\left(-\rho_v \pi r^2\right)\right]}{v\left[1 - \exp\left(-\rho_v \pi r^2\right)\right]} \tag{A.12}$$

Specifically, if r is extremely small, the two terms $\operatorname{erf}(\cdot)$ and $\exp(\cdot)$ can be Taylor expanded as the following series (since $\rho_v < \frac{N}{A}$, there must be an extremely small r such that $\rho_v \pi r^2$ is also extremely small):

$$\operatorname{erf}(\sqrt{\rho_v \pi r^2}) \approx 2\sqrt{\rho_v} r \tag{A.13}$$

$$\exp(-\rho_v \pi r^2) \approx 1 - \rho_v \pi r^2 \tag{A.14}$$

In this case, the matching time of passengers can be approximated by

$$w_c \approx \frac{\tau}{\rho_v \pi r^2} - \frac{\tau}{2} \tag{A.15}$$

and

$$w_v = \frac{m_v}{m_c}\left(w_c + \frac{\tau}{2}\right) - \frac{\tau}{2} \approx \frac{\tau}{\rho_c \pi r^2} - \frac{\tau}{2} \tag{A.16}$$

Since $\rho_v = m_v / A$, we further have

$$m_v\left(w_c + \frac{\tau}{2}\right) \approx \frac{A\tau}{\pi r^2} \tag{A.17}$$

In view of Eqs. (7) and (A.17), the matching rate can be rewritten as a Cobb-Douglas type matching model as follows:

$$Q = \frac{m_c}{w_c + \frac{\tau}{2}} \approx \frac{\pi r^2}{A\tau} m_c m_v \tag{A.18}$$

In addition, by substituting Eq.(A.13) and Eq. (A.14) into Eq. (A.12), we have

$$w_p = \frac{\zeta r}{v} \tag{A.19}$$

This completes the proof. ■

Proof for Proposition 6

When $A_M = {\rho_c}^{-1}$ we can get

$$w_c\left(\rho_c, \rho_v\right) = \left[\frac{1}{1-\exp\left(-\rho_v/\rho_c\right)} - \frac{1}{2}\right]\tau \tag{A.20}$$

and

$$w_p\left(\rho_c, \rho_v\right) = \frac{\zeta[\frac{\mathrm{erf}(\sqrt{\rho_v/\rho_c})}{2\sqrt{\rho_v}} - \sqrt{\frac{1}{\pi\rho_c}}\cdot\exp(-\rho_v/\rho_c)]}{v[1-\exp(-\rho_v/\rho_c)]} \tag{A.21}$$

For a dominant supply with $\rho_v \gg \rho_c$, we have

$$\sqrt{\frac{\rho_v}{\rho_c}} = \sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)} \approx +\infty; \mathrm{erf}(\sqrt{\frac{\rho_v}{\rho_c}}) \approx 1; \exp(-\frac{\rho_v}{\rho_c}) \approx 0 \tag{A.22}$$

Then the matching time for passengers and drivers can be obtained as follows:

$$w_c \approx \frac{\tau}{2} \tag{A.23}$$

$$w_v \approx \frac{\rho_v}{\rho_c}\tau - \frac{\tau}{2} \tag{A.24}$$

Based on Eq. (A.21), the expected pick-up time only depends on the density of idle vehicles as follows:

$$w_p = \zeta/\left(2v\sqrt{\rho_v}\right) \tag{A.25}$$

This completes the proof. ■

Proof for Proposition 7

When $A_M = {\rho_c}^{-1}$ and the density of waiting passengers dominates the density of idle vehicles, i.e., $\rho_c \gg \rho_v$, the two terms $\ln(\cdot)$ and $\mathrm{erf}(\cdot)$ in Eqs. (31)–(32) can be Taylor expanded as:

$$\ln\left(1-\frac{\tau Q}{\rho_c A}\right) \approx -\frac{\tau Q}{\rho_c A} \tag{A.26}$$

$$\mathrm{erf}\left(\sqrt{-\ln\left(1-\frac{\tau Q}{\rho_c A}\right)}\right) \approx \frac{2}{\sqrt{\pi}}\sqrt{\frac{\tau Q}{\rho_c A}} \tag{A.27}$$

From Eq. (32), we can obtain the density of idle vehicles, which is approximated by

$$\rho_v \approx \tau Q/A \tag{A.28}$$

The matching time of passengers can be approximated by

$$w_c = \left[\frac{1}{1-\exp\left(-\rho_v/\rho_c\right)} - \frac{1}{2}\right]\tau \approx \left(\frac{\rho_c A}{\tau Q} - \frac{1}{2}\right)\tau \approx \left(\frac{m_c}{m_v} - \frac{1}{2}\right)\tau \tag{A.29}$$

In addition, by substituting Eqs. (A.27) and (A.28) into Eq.(A.21), we can get

$$w_p \approx \frac{\zeta}{v\sqrt{\pi\rho_c}} \tag{A.30}$$

Based on Eqs. (9), (A.28), and (A.29), we can get

$$w_v = \frac{\rho_v}{\rho_c}\left(w_c + \frac{\tau}{2}\right) - \frac{\tau}{2} \approx \frac{\tau Q}{\rho_c A}\left(\frac{\rho_c A}{Q}\right) - \frac{\tau}{2} = \frac{\tau}{2} \tag{A.31}$$

Occurrence condition for Special case 2 and Special case 3

When the platform sets a large matching radius so that the matching area is governed by the density of waiting passengers, i.e., $A_M = \frac{1}{\rho_c}$.

1) First we prove that when adopts an instant matching, then the following relationship between the density of idle vehicles and the density of waiting passengers emerges: $\rho_v \gg \rho_c$ or $\rho_c \gg \rho_v$.

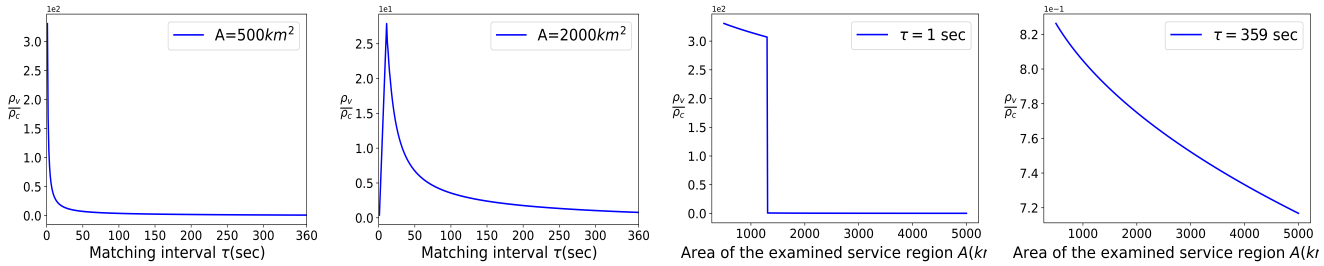
Based on Eq. (32), $\rho_v \gg \rho_c$ is equivalent to $\frac{\tau Q}{\rho_c A} = \frac{M}{m_c} \to 1$ and $\rho_c \gg \rho_v$ is equivalent to $\frac{\tau Q}{\rho_c A} = \frac{M}{m_c} \to 0$. Then it is to prove that when $\tau \to 0$, we can get $\frac{\tau Q}{\rho_c A} = p_c \to 0$ or 1.

Obviously, $p_c \in [0,1]$. Eq. (31) can be rewritten as

$$-\frac{\tau Q}{A}\frac{\ln\left(1-p_c\right)}{p_c} = \frac{N}{A} - \frac{\zeta\sqrt{\frac{Q}{A}}\left[\frac{\operatorname{erf}\left(\sqrt{-\ln(1-p_c)}\right)}{2\sqrt{-\ln(1-p_c)}} - \sqrt{\frac{1}{\pi}}\left(1-p_c\right)\right]}{v\sqrt{\tau p_c}} - \frac{tQ}{A} + \frac{\tau Q}{2A} \tag{A.32}$$

p_c can be solved by Eq. (A.32) and regarded as a function of τ.

Then we prove $\lim_{\tau\to 0} p_c \notin (0,1)$. Suppose $\lim_{\tau\to 0} p_c \in [\varepsilon_1, \varepsilon_2]$, ε_1 and ε_2 can be any positive number satisfying $0 < \varepsilon_1 < \varepsilon_2 < 1$. Then $\lim_{\tau\to 0} \ln\left(1-p_c\right)$ is a finite number. When $\tau \to 0$, from Eq. (A.32) we can get that $0 = -\infty$, which is incorrect. Therefore, when $\tau \to 0$, we have $p_c = 0$ or $p_c = 1$. Furthermore, when $\lim_{\tau\to 0} p_c = 0$, we can get $-\ln\left(1-p_c\right) \approx p_c \approx 0$ and $\frac{\operatorname{erf}\left(\sqrt{-\ln(1-p_c)}\right)}{2\sqrt{-\ln(1-p_c)}} \approx \sqrt{\frac{1}{\pi}}$. From Eq. (A.32) we have that $\frac{\tau Q}{A} = \frac{N}{A} - \frac{\zeta}{v}\sqrt{\frac{Q}{A\pi}}\sqrt{\frac{p_c}{\tau}} - \frac{tQ}{A} + \frac{\tau Q}{2A}$ and $p_c = \frac{\pi\tau}{AQ}\left[\frac{v}{\zeta}\left(N - \left(t+\frac{\tau}{2}\right)Q\right)\right]^2$. Then $\rho_c = \frac{1}{\pi}\left[\frac{\zeta Q}{v\left(N - tQ - \frac{\tau}{2}Q\right)}\right]^2$ and $\rho_v \approx \tau Q/A$. When $p_c = 1$, Eq. (31) and Eq. (32) indicate that $\rho_v = \frac{N}{A} - \frac{\zeta\frac{1}{2\sqrt{\rho_v}}\rho_c}{v\tau} - \frac{tQ}{A} + \frac{\tau Q}{2A} = \frac{N}{A} - \frac{\zeta\frac{1}{2\sqrt{\rho_v}}\cdot\frac{\rho_c A}{\tau Q}}{vA}Q - \frac{tQ}{A} + \frac{\tau}{2Q} = \frac{N}{A} - \frac{\zeta Q}{\sqrt{\rho_v}vA} - \frac{tQ}{A} + \frac{\tau Q}{2A}$. From this equation, we can get ρ_v. In the following Figure A.4, we set $N = 5{,}000$ veh, $Q = 10{,}000$ person/h, $v = 70$ km/h, and $t = 0.4$ h; Figures A.4a-A.4c shows that when $\tau \to 0$, it must be $\rho_v \gg \rho_c$ or $\rho_c \gg \rho_v$.



(a) Trend of ρ_v/ρ_c when $A = 500\text{km}^2$.

(b) Trend of ρ_v/ρ_c when $A = 2,000\text{km}^2$.

(c) Trend of ρ_v/ρ_c when $\tau = 1$ sec.

(d) Trend of ρ_v/ρ_c when $\tau = 359$ sec.

Figure A.4: Trends of ρ_v/ρ_c under different conditions.

2) Second, we prove when the vehicle utilization $\frac{tQ}{N} \to 0$, the density of idle vehicles must dominate the density of waiting passengers, i.e., $\rho_v \gg \rho_c$

Based on Eq. (A.32) we have $p_c \to 1 \iff \rho_v \gg \rho_c$, then it is to prove when $\frac{N}{t} \gg Q$, we have $p_c \to 1$.

When $\frac{N}{t} \gg Q$, from Eq. (A.32) we can obtain $-\frac{\tau Q}{A}\frac{\ln(1-p_c)}{p_c} + \frac{\zeta\sqrt{\frac{Q}{A}}\left[\frac{\text{erf}\left(\sqrt{-\ln(1-p_c)}\right)}{2\sqrt{-\ln(1-p_c)}} - \sqrt{\frac{1}{\pi}}(1-p_c)\right]}{v\sqrt{\tau p_c}} \to \infty$. Define $g\left(p_c\right) = \frac{\zeta\sqrt{\frac{Q}{A}}\left[\frac{\text{erf}\left(\sqrt{-\ln(1-p_c)}\right)}{2\sqrt{-\ln(1-p_c)}} - \sqrt{\frac{1}{\pi}}(1-p_c)\right]}{v\sqrt{\tau p_c}}$, we have that $g\left(p_c\right)$ is a continuous function and $g\left(0\right) = g\left(1\right) = 0$, therefore, there is a $p_c^* \in (0,1)$ such that $g\left(p_c\right)$ gets its maximum value and $g\left(p_c^*\right)$ is finite.

Furthermore, $-\frac{\tau Q}{A}\frac{\ln(1-p_c)}{p_c} + \frac{\zeta\sqrt{\frac{Q}{A}}\left[\frac{\text{erf}\left(\sqrt{-\ln(1-p_c)}\right)}{2\sqrt{-\ln(1-p_c)}} - \sqrt{\frac{1}{\pi}}(1-p_c)\right]}{v\sqrt{\tau p_c}} \leq -\frac{\tau Q}{A}\frac{\ln(1-p_c)}{p_c} + g\left(p_c^*\right)$ for any $p_c \in [0,1]$. Since $-\frac{\ln(1-p_c)}{p_c}$ is strictly increasing in $p_c \in [0,1]$ and $\lim_{p_c \to 1} -\frac{\ln(1-p_c)}{p_c} \to \infty$, we can obtain that only when $p_c \to 1$, we have $-\frac{\tau Q}{A}\frac{\ln(1-p_c)}{p_c} + \frac{\zeta\sqrt{\frac{Q}{A}}\left[\frac{\text{erf}\left(\sqrt{-\ln(1-p_c)}\right)}{2\sqrt{-\ln(1-p_c)}} - \sqrt{\frac{1}{\pi}}(1-p_c)\right]}{v\sqrt{\tau p_c}} \to \infty$, which indicates when $\frac{N}{t} \gg Q$, we have $\rho_v \gg \rho_c$.

3) Third, we prove when $\tau \approx \frac{N - tQ + \frac{\tau}{2}Q}{Q}$ or equivalent to when the utilization $\frac{tQ}{N} \to \frac{1}{1+\frac{\tau}{2t}}$ we have $\rho_c \geq \rho_v$.

When $\frac{tQ}{N} \to \frac{1}{1+\frac{\tau}{2t}}$, from Eq. (A.32) we have $-\frac{\tau Q}{A}\frac{\ln(1-p_c)}{p_c} + g\left(p_c\right) \to \frac{\tau Q}{A}$. Since $g\left(p_c\right) \geq 0$, and $-\frac{\ln(1-p_c)}{p_c}$ is strictly increasing in $p_c \in [0,1]$, we can get $-\frac{\tau Q}{A}\frac{\ln(1-p_c)}{p_c} + g\left(p_c\right) \geq -\frac{\tau Q}{A}\frac{\ln(1-p_c)}{p_c} \geq -\frac{\tau Q}{A}$ and the equality sign holds only when $p_c = 0$. Therefore, when $\frac{tQ}{N} \to \frac{1}{1+\frac{\tau}{2t}}$, we have $p_c = 0$ and $\rho_c \gg \rho_v$.

This completes the proof. ■