On-Demand Ride-Matching in a Spatial Model with Abandonment and Cancellation

Guangju Wang, Hailun Zhang, Jiheng Zhang
Department of Industrial Engineering and Decision Analytics
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong S.A.R., China
gwangal@connect.ust.hk, hzhangaq@connect.ust.hk, jiheng@ust.hk

Ride-hailing platforms such as Uber, Lyft, and DiDi coordinate supply and demand by matching passengers and drivers. The platform has to promptly dispatch drivers when receiving requests, since otherwise passengers may lose patience and abandon the service by switching to alternative transportation methods. However, having less idle drivers results in a possible lengthy pick-up time, which is a waste of system capacity and may cause passengers to cancel the service after they are matched. Due to complex spatial and queueing dynamics, the analysis of the matching decision is challenging. In this paper, we propose a spatial model to approximate the pick-up time based on the number of waiting passengers and idle drivers. We analyze the dynamics of passengers and drivers in a queueing model where the platform can control the matching process by setting a threshold on the expected pick-up time. Applying fluid approximations, we obtain accurate performance evaluations and an elegant optimality condition, based on which we propose a policy that adapts to time-varying demand.

Key words: on-demand matching; ride-hailing platform; fluid approximation

1. Introduction

The on-demand ride-hailing market has experienced rapid growth in recent years. With Uber operating in over 700 cities around the world (Uber (2018b)) and DiDi-Chuxing dominating the ride-sharing market within China, ride-hailing platforms have profoundly changed the way people travel. On-demand ride-hailing service has obvious advantages over traditional taxi service. As the centralized controller, a ride-hailing platform connects between the drivers (supply) and the passengers (demand) by collecting the real-time locations of drivers and responding to passengers' requests, so that neither the drivers nor the passengers need to wander in the streets in the hope of finding each other. In this process, however, the platform faces the challenge of how to match idle drivers and requesting passengers.

To understand how matching affects a ride-hailing system and its users, we start with a discussion on two important considerations in the matching decision. The first is the *pick-up distance*, which is the distance between the idle driver and the requesting passenger. During pick-up, passengers must wait at a prescribed location (the pick-up location), and drivers must drive to this exact location without making any profit. Therefore, a long pick-up distance not only harms the user experience but also wastes system capacity. The other consideration is passenger patience as reflected by *abandonment* and *cancellation*. A passenger may *abandon* the service if the platform does not assign a driver promptly. Even after assignment, the passenger still has the option of *cancelling* the trip if the pick-up takes too long. While abandonment leads to a loss of business, cancellation is even worse as it also wastes driver effort and consequently the platform capacity. Forbidding cancellation during pick-up may put passengers off requesting for service in the first place and may, in turn, damage the platform's reputation. Some platforms impose a small amount of cancellation fee. For example, DiDi charges a small cancellation fee if a passenger cancels the service three minutes after a driver is assigned DiDi (2019) and Uber applies a similar rule (Uber (2018a)). However, the purpose such a penalty is primarily to prevent malicious attacks and irresponsible requests for services.

Abandonments can be reduced by assigning a driver promptly, and cancellations can be mitigated by shortening the pick-up distances. Thus, a simple greedy matching policy would always assign the "nearest" idle driver to the passenger making the request. However, this simple greedy policy is sub-optimal because it fails to consider the future dynamics of the system. For example, upon a request for a service, the nearest driver may in fact be far away from the pick-up point. But after a short while, other idle drivers may appear who happen to be closer to the pick-up location than the first driver, or a new request may be received that is closer than the earlier request to the first driver. A similar phenomenon is referred to as the "Wild Goose Chase" by Castillo et al. (2017). Another reason for the sub-optimality of this greedy policy is an imbalance in supply and demand, as explained in Ozkan and Ward (2016). In this paper, we focus on finding a guiding principle for designing a matching policy to optimize the performance of the system. Our analysis shows that, when pricing is exogenously given, a fine-tuned matching algorithm can benefit passengers, drivers, and the platform simultaneously.

In a typical ride-hailing system, a driver can be in one of three states, *idle*, *assigned* and *busy*. *Assigned* refers to the state where the driver is on the way to pick up a passenger and *busy* is the state where the driver has already picked up the passenger and is heading towards the destination. Correspondingly, there are three states for passengers, *requesting*, *waiting for pick-up* and *on trip*, where the last two states are coupled with the *assigned* and *busy* states of drivers. The analysis and optimization of a matching policy require incorporating the queueing dynamics of the drivers and passengers in different states. For this purpose, we require a spatial model that is able to sufficiently capture the pick-up time under the queueing dynamics. The greatest challenge is that the dimensionality of the system explodes when we incorporate the spatial interaction into the queueing model. Note that platforms like DiDi and Uber usually have over tens of thousands of drivers and passengers in major cities. Therefore, it is impractical to build a model that keeps track of all pairwise distances between passengers and drivers. Instead, we propose a spatial model that assumes the average pick-up time to be

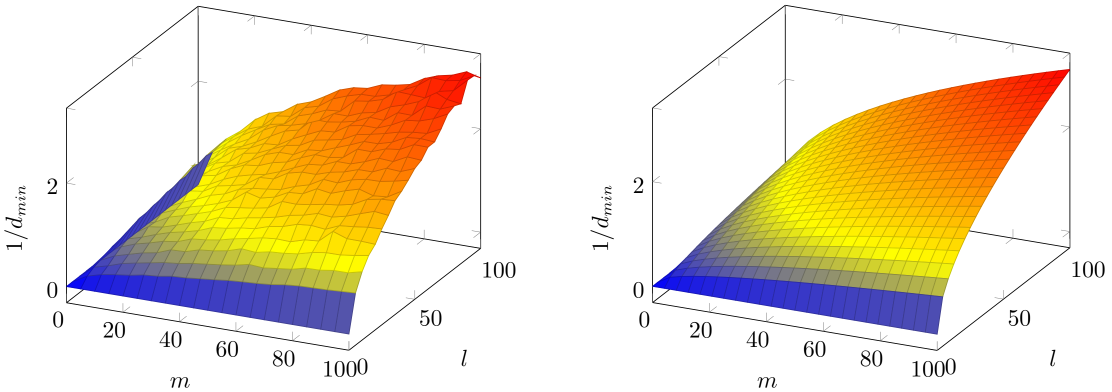
$$\frac{1}{\text{Average Pick-up Time}} = C\left(\#\text{ of } \textit{requesting} \text{ passengers}\right)^{\alpha_1} \times \left(\#\text{ of } \textit{idle} \text{ drivers}\right)^{\alpha_2}, \qquad (1)$$

for some constants $C,\ \alpha_1, \alpha_2 \in (0, \infty)$. The right-hand side of (1) shares the same form as the Cobb-Douglas production function, which was first introduced by Coma and Douglas (1928) and has been widely used to model quantitative relationships when a response depends on two or more factors. For example, the Cobb-Douglas production function has been used in the analysis of two-sided markets by Rochet and Tirole (2003) and in the taxi and ride-hailing market by Yang et al. (2010), Wang et al. (2016) and Xu et al. (2019).

On a specific map, the average pick-up time primarily depends on the number of *requesting* passengers and the number of *idle* drivers. Moreover, the average pick-up time should be shorter if either of the two factors increases. The parameters $C,\ \alpha_1, \alpha_2$ provide the model with substantial richness to accommodate different maps. To give a quick and simple justification, we randomly sample *idle* drivers and *requesting* passengers on a square map with Euclidian distance, and plot the change in average minimum pair-wise distance $d_{\min}$[1] with the two factors in Figure 1(a), along with the proposed spatial model (1) in Figure 1(b). The two graphs look similar and only differ by 1.7% on average. Further numerical justification is provided in Appendix B. On real city maps, the parameters C and α_1, α_2 can be estimated based on real data. For details, see Appendix B and Section 6.1.

In practice, ride-hailing platforms usually control the pick-up time by setting a maximum allowed pick-up distance, which is variously referred to as the *matching radius* (Xu et al. (2019)), *maximum dispatch radius* (Castillo et al. (2017)), or *response cap* (Feng et al. (2017)). When a passenger requests a ride, the platform tries to find an available driver who is within the matching radius of the pick-up location. If there is no available driver within this range, the passenger has to wait. Castillo et al. (2017) and Feng et al. (2017) both argued that such a threshold on the pick-up distance could improve the system performance. According to Castillo et al. (2017), Uber caps the pick-up distance as part of their matching algorithm. Throughout the paper, without preference, refer to the maximum allowed pick-up distance as the matching radius. In our queueing model, we model the matching radius as a threshold on the pick-up rate. To the best of our knowledge, we believe our work breaks ground by formally researching the choice of matching radius and obtain tractable operational guidelines.

[1] It reduces to the average pick-up time if we assume the speed of each driver is one.



(a) Sample Average

(b) Spatial Model

Figure 1 Comparison between the numerical experiment and the spatial model. ((a) Average taken over 5000 samples. (b) Spatial model 1 with $C = 0.031$ and $\alpha_1 = \alpha_2 = 0.5$. m: number of requesting passengers, l: number of idle drivers);

With the help of the spatial model, we are able to characterize the dynamics of the drivers (in states *idle*, *assigned* and *busy*) and passengers (in states *requesting*, *waiting for pick-up* and *on trip*) using a stochastic model which is formulated in Section 3. Since the stochastic process is prohibitively difficult to analyze, we study the fluid counterpart of the process and prove that the fluid process converges to its equilibrium, which serves as an accurate approximation to the system performances. Based on the fluid approximation, we formulate an optimization problem to maximize the throughput of the system. Our most important finding is identifying an optimality condition, which can be simply stated as

$$\alpha_1 \frac{\#\text{ cancellations}}{\#\text{ abandonments}} + \alpha_2 \frac{\#\text{ assigned drivers}}{\#\text{ idle drivers}} = 1, \tag{2}$$

where α_1 and α_2 are the parameters of the spatial model (1). The optimality condition has two main advantages. The first is its simplicity. Note that the first fraction in (2) involves quantities cumulated over a time interval, which can be estimated by counting the accumulative number of abandonments and cancellations over a short time interval. To calculate the second fraction in (2), all we need to do is count the number of *idle* and *assigned* drivers in the system. The second advantage of the optimality condition is that it does not involve too many parameters. To apply this optimality condition, the platform only has to estimate the spatial parameters α_1 and α_2. There is no need to estimate any other parameters, for example, the spatial parameter C, arrival rate, abandonment rate, and cancellation rate.

These advantages lead to easy-to-implement design principles of a matching policy. Essentially, the platform only needs to set the matching radius so that the optimality condition (2) holds. Intuitively, the matching radius serves as an adjustable variable so that the platform can balance the tradeoff between matching drivers and passengers promptly and waiting to make better matchings. Since the optimality condition does not depend on the arrival rate, we design a *self-adaptive* matching policy that can automatically adjust itself in response to time-varying arrival rate. In Section 6.3, we perform numerical experiments of the self-adaptive policy in a simulated ride-hailing system with the aim of approximating the practical situation and demonstrate the efficacy of the policy.

The rest of the paper is organized as follows. Section 2 reviews the related literature. In Section 3, we formulate a stochastic model for the dynamics of drivers and passengers in different states. We construct and analyze the corresponding fluid model in Section 4. The optimization problem is analyzed in Section 5. In Section 6, we perform numerical experiments of our matching policy. Section 7 concludes the paper with discussions and possible directions of future research.

2. Literature Review

There is an emerging stream of literature on on-demand services and the sharing economy, covering a wide range of topics, including but not limited to performance analysis, control mechanisms, and social welfare. Here we review some closely related topics such as dispatch control, service system with state-dependent rates, two-sided matching, and pricing.

Routing and dispatch control of on-demand ride-hailing platforms has drawn attention from the operations research field in recent years. Ozkan and Ward (2016) study revenue-maximizing state-independent dispatch control by solving a minimum cost flow problem in the fluid limit. Iglesias et al. (2016) consider centralized matching and repositioning decisions in the context of a closed network model. Braverman et al. (2019) model the system as a closed queueing network of all vehicles and propose optimal routing decisions based on the corresponding fluid model. Afeche et al. (2018) study the problem of matching demand (riders) with self-interested capacity (drivers) over a spatial network using a two-location fluid model. Banerjee et al. (2018) consider the design of state-dependent controls for a closed queueing network model and develop matching policies with exponential-decay demand-dropping probability as the number of drivers scales up. In contrast to their research directions, we center on formulating formulate a queueing system that couples the closed queueing network of drivers and the open queueing network of passengers under the proposed spatial model. We develop an approximation to such a queueing system to show insights and construct optimal control policies.

Our paper is also related to the literature that studies service systems with state-dependent processing rates (for some examples, see Mandelbaum and Pats (1995), Mandelbaum et al. (1998) and Powell and Schultz (2004).) Chan et al. (2014) investigate an Erlang-R service system in which the service rate can be sped up whenever congestion is above a certain threshold and Dong et al.

(2015) study a service system in which agents will reduce their service rate as the system's workload increases. In their work, the state-dependent service rates are due to the system congestion levels, while in our model the pick-up rate depends on both the number of requesting passengers and the number of idle drivers.

The ride-hailing model is an example of a two-sided market, pioneered by several recent works. Hu and Zhou (2018) model dynamic matching between the demand and supply of heterogeneous types, providing sufficient conditions such that the optimal policy follows a priority hierarchy. Baccara et al. (2018) consider vertically different preferences of agents and show that the optimal mechanism always matches congruent pairs immediately while holding to a stock of incongruent pairs up to a certain threshold. A common method to control dynamic matching systems is the market thickening approach. In ride-hailing systems, the thickness is represented by the number of *requesting* passengers and that of *idle* drivers. We capture how the thickness affects the dynamics of the system using the spatial model (1). In the queueing literature, Gurvich and Ward (2014) consider a matching system where jobs of multiple classes arrive to the system dynamically and prove the asymptotic optimality of a discrete review matching policy as the arrival rates of the jobs increase (i.e., a large market assumption). Akbarpour et al. (2017) consider a dynamic matching network where jobs may abandon if matching takes too long. They prove that if the system controller can identify which jobs are about to abandon and prioritize them, then the thickness of the market becomes valuable. Our results not only reveal that market thickness benefits the system but also provide quantitative insights into the optimal thickness of the market.

There is a large amount of research on ride-sharing platforms using pricing as a leverage to optimize the system performance (see Banerjee et al. (2015), Gurvich et al. (2019), Hu and Zhou (2019), Bimpikis et al. (2019), Taylor (2018), Benjaafar et al. (2018) and Bai et al. (2018), to name a few). Banerjee et al. (2016) and Ozkan (2018) deliberate the joint pricing and optimization decisions in a ride-sharing market. Castillo et al. (2017) and Besbes et al. (2018b) look into the role of surge pricing in ride-hailing systems. Our work focuses on the matching decisions of the platform and takes the pricing scheme as exogenous. The performance analysis in our work has the potential to be applied to the study of the pricing problem. For example, one possibility is to search for a joint optimal control and pricing policy.

The most related research to ours are Feng et al. (2017), Besbes et al. (2018a), Korolko et al. (2018) and Cheng et al. (2019), who develop models based on the fact that the total time a passenger spends in the system consists of three parts: waiting time, pick-up time and on-trip time. Feng et al. (2017) propose an approximation scheme and use a capped matching mechanism to improve system efficiency. They also derive heuristic methods to calculate a near-optimal cap. In our paper, we explicitly characterize the optimal matching radius based on a fluid approximation scheme and propose a feedback-based algorithm that approaches the optimal control. Besbes et al. (2018a) develop a Markovian queueing model that expresses the service rate as a function of the average pick-up time and on-trip time, where the pick-up time depends on the number of idle drivers and waiting passengers. Our spatial model can be viewed as an extension of the model proposed by Besbes et al. (2018a). In our model, we explicitly consider the three states of drivers, i.e., idle, assigned, and busy. Korolko et al. (2018) decompose the driver states similarly and study the joint optimization of dynamic pricing and dynamic waiting with data from Uber in a static equilibrium model. With the state decomposition, we approximate the long-run behavior of drivers in different states and propose optimization methods accordingly.

3. Model

We formulate the stochastic model for a ride-hailing system. Assume there are in total K drivers in the system. Let $(Z_0(t), Z_1(t), Z_2(t))$ denote the number of (*idle*, *assigned*, *busy*) drivers at time t. The drivers in different states form a closed queueing network with

$$Z_0(t) + Z_1(t) + Z_2(t) = K. \tag{3}$$

Passengers who cannot be matched with an idle driver immediately upon arrival join the pool of *requesting* passengers (*waiting pool* in short), whose size is denoted by $Q(t)$. We refer to number of requesting passengers as the pool size for simplicity. Passengers are impatient and will abandon from the *waiting pool* when their patience runs out. In addition, while waiting for a pick-up, passengers may also lose patience and cancel the service. After being picked up, passengers will stay in the system until the end of their journey. Upon completion of the service, passengers leave the system and drivers become idle again. The ride-hailing system is essentially a coupled model of an open queueing network of passengers in different states and a closed queueing network of drivers, as the number of passengers *waiting for pick-up* equals the number $Z_1(t)$ of *assigned* drivers and the number of passengers *on trip* equals the number $Z_2(t)$ of *busy* drivers. Figure 2 depicts the partially coupled queueing model. To formalize the stochastic queueing dynamics, we need to introduce additional notations and assumptions.

Passengers arrive to the system according to a Poisson process $A(t)$ with rate $\tilde{\lambda}$. An arriving passenger is willing to wait for a random amount of time, independent of all other random variables and exponentially distributed with rate θ_0, to be matched with a driver. Those whose patience runs out before they are matched abandon the system. Denote by $R_0(t)$ the number of abandonments by time t. After being matched with a driver, passengers are willing to wait for another random amount of time, independent of all other random variables and exponentially distributed with rate θ_1, for pick-up. Passengers will cancel the service and leave the system if their patience exhausts

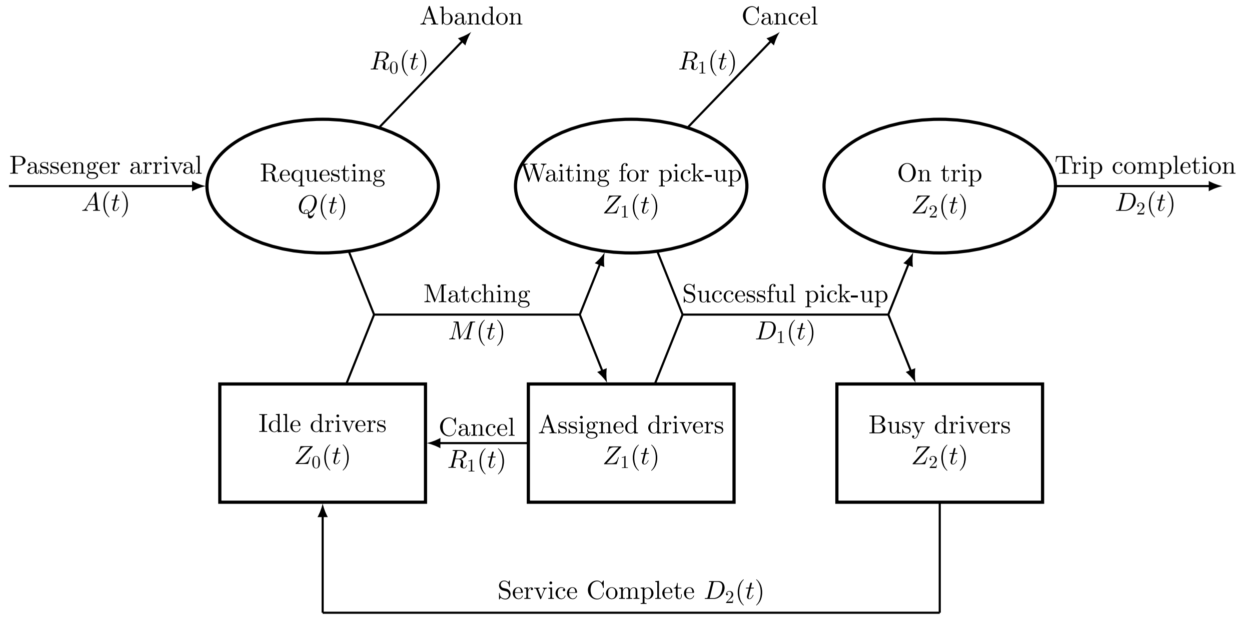


Figure 2 On-demand Ride-hailing System.

before they are picked up by their assigned drivers. Denote by $R_1(t)$ the number of cancellations by time t. If an *assigned* driver arrives at the pick-up location before the passenger's patience time runs out, we call it a successful pick-up. Let $D_1(t)$ be the number of successful pick-ups by time t. We assume that each trip, during which passengers are not allowed to abandon, takes a random amount of time, independent of all other random variables and exponentially distributed with rate μ_2. Let $D_2(t)$ be the number of completed trips by time t. The detailed mathematical formulations of all these processes are presented in Appendix A. With the introduced notations, we can describe the system dynamics of the stochastic model underlying the ride-hailing system as follows:

$$Q(t) = Q(0) + A(t) - R_0(t) - M(t), \tag{4}$$

$$Z_0(t) = Z_0(0) + R_1(t) + D_2(t) - M(t), \tag{5}$$

$$Z_1(t) = Z_1(0) + M(t) - D_1(t) - R_1(t), \tag{6}$$

$$Z_2(t) = Z_2(0) + D_1(t) - D_2(t), \tag{7}$$

where $M(t)$ is the total number of matchings made by time t. Note that the exponential distribution assumption of a passenger's patience and service time is made to facilitate modeling system states as counting processes. What remains to be defined is the matching process between passengers and drivers, which we will formulate after we introduce the spatial model.

The Spatial Model. For a passenger-driver pair matched at time t, we model the pick-up time as a random variable following an exponential distribution with rate $\mu_1(t)$. One can think of the reciprocal of this rate as the average pick-up time. Based on the spatial model introduced in equation (1), we make the Assumption 1.

Assumption 1 **(Spatial Model)**. *If a passenger and a driver are matched at time t, the pick-up of this specific passenger-driver pair takes a random amount of time following an exponential distribution with rate*

$$\mu_1(t) = \tilde{C}(Q(t))^{\alpha_1}(Z_0(t))^{\alpha_2}, \tag{8}$$

for some constants $\tilde{C},\ \alpha_1, \alpha_2 \in (0, \infty)$. Conditioning on $Q(t)$ and $Z_0(t)$, the pick-up time of this specific passenger-driver pair is independent of all other random variables.

We refer to the above defined $\mu_1(t)$ as the pick-up rate. The intuition behind Assumption 1 is that the pick-up time is shorter when there is a higher density of requesting passengers and idle drivers on the map. For notational simplicity, we write the pick-up rate as a function of time t. It should be stated that that the pick-up rate depends on the number of *requesting* passengers and number of *idle* drivers in the system. With wide-ranging parameters C, α_1 and α_2, the spatial model can be applied in different scenarios, as indicated in Appendix B. The prescriptive analysis in Section 5 shows that in our model, α_1 and α_2 are the only parameters required for optimization.

The spatial model extends the assumption of Besbes et al. (2018a), who consider a non-idling matching policy and approximate the *mean pick-up time*, with our notations, by $C(Q(t) \vee Z_0(t) \vee 1)^{-1/2}$. Note that in a non-idling policy, at most one of $Q(t)$ and $Z_0(t)$ can be positive. Therefore, by assuming a non-idling policy and setting $\alpha_1 = \alpha_2 = 0.5$, the spatial model (8) is consistent with the assumption of Besbes et al. (2018a). Another difference is that Besbes et al. (2018a) assume that the service rate depends on the current state of the system, while in our case the pick-up rate of any passenger-driver pair depends on the state at the time when the matching is made. Korolko et al. (2018) adopt a similar assumption by considering the expected pick-up time of a single passenger as the number of idle drivers changes (e.g. $\mu_1(t) = \tilde{C}Z_0(t)^{\alpha_2}$). They also verify that, by relaxing the range of α_2 to be $(0, 1)$, the model fits better to real Uber data.

The Matching Process. Now we are ready to define the matching process $M(t)$. We assume that the platform applies a threshold-based matching policy, motivated by the *matching radius* which regulates the pick-up distance. In our model, where distance is not directly modeled, the platform controls the pick-up time by setting a threshold μ_1 on the pick-up rate. When $\mu_1(t) < \mu_1$, the platform does not make any matchings and will let the number of requesting passengers and idle drivers increase. As soon as $\mu_1(t)$ exceeds the threshold μ_1, the platform deems the pick-up rate satisfactory and makes as many matchings as possible until $\mu_1(t)$ falls below μ_1 again. Without loss of generality, we assume $\mu_1(0) < \mu_1$ since otherwise the platform makes a positive number of matchings at time 0 and the pick-up rate would immediately drops below μ_1. We use $M_{\mu_1}(t)$ instead of $M(t)$ to emphasize its dependence on μ_1. Since a matching can only happen when a passenger arrives (a jump in the process $A(t)$) to the system or a driver becomes idle (a jump in the process $R_1(t)$ or $D_2(t)$), we can formulate the matching process as follows:

$$M_{\mu_1}(t) = \int_0^t \mathbf{1}_{\{\mu_1(t)\geq\mu_1, Z_0(s-)>0\}} dA(s) + \int_0^t \mathbf{1}_{\{\mu_1(t)\geq\mu_1, Q(s-)>0\}} d(R_1(s)+D_2(s)), \tag{9}$$

where $Q(s-)$ and $Z_0(s-)$ are the left limits. Note that this matching process directly affects the cancellation process $R_1(t)$ and the successful pick-up process $D_1(t)$. To avoid interrupting the flow of presentation, we postpone the detailed formulation of the complete stochastic model to Appendix A.

In summary, we have introduced a stochastic model for the ride-hailing application with primitive parameter settings $\phi = (K, \tilde{\lambda}, \mu_1, \mu_2, \theta_0, \theta_1, \tilde{C}, \alpha_1, \alpha_2)$ where μ_1 is also the decision variable. Assume that the fixed fee charged per trip is p_f and that the price charged per unit of time is p_s. The expected revenue can then be written as

$$\mathbb{E}\left(p_s \int_0^t Z_2(s)ds + p_f D_2(t)\right). \tag{10}$$

In our work, the prices p_f, p_s and how the platform and drivers split the revenue are assumed to be given and we focus our analysis on how the matching decision, i.e. the threshold μ_1, affects the system performance. In this sense, our analytical results can be applied with any pricing policy, and any revenue allocation between the platform and the drivers, such as the ones in Taylor (2018) and Bai et al. (2018).

4. Performance Approximation

Fluid approximation has been widely used in the queueing literature because of its efficiency and tractability. To find a proper fluid model to approximate the stochastic system, we first introduce a regime with a sequence of stochastic models indexed by the number of drivers N. In the Nth system, the primitive parameters are $\phi^N = (N, N\lambda, \mu_1, \mu_2, \theta_0, \theta_1, C/N^{\alpha_1+\alpha_2}, \alpha_1, \alpha_2)$. In this regime, the passenger arrival rate grows proportionally with the number of drivers while all other rate parameters remain in order 1. The pick-up rate in the Nth system is set to be $\mu_1(t) = C/N^{\alpha_1+\alpha_2}(Q(t))^{\alpha_1}(Z_0(t))^{\alpha_2}$ so that it remains in constant order across all systems in the regime. In other words, we implicitly enlarge the “map” as the scale of the system grows, since otherwise increasing the number of drivers would cause the pick-up time to converge to zero, which results in a degenerated limit that fails to capture the critical spatial factor. The design of the regime follows the principle that its corresponding fluid model can be used to approximate a particular system with primitive parameters $\phi = (K, \tilde{\lambda}, \mu_1, \mu_2, \theta_0, \theta_1, \tilde{C}, \alpha_1, \alpha_2)$, as introduced in Section 3. Notably, by setting $\lambda = \tilde{\lambda}/K$ and $C = \tilde{C}K^{\alpha_1+\alpha_2}$, we have $\phi^K = \phi$, meaning that the Kth system in the regime is exactly the system we introduced in Section 3.

We will demonstrate that in the proposed regime $\{\phi^N\}$, the scaled stochastic processes $(Q(t)/N, Z_0(t)/N, Z_1(t)/N, Z_2(t)/N)$ can be effectively approximated by the fluid model $(q(t), z_0(t), z_1(t), z_2(t))$ with corresponding parameters $\phi_f = (\lambda, \mu_1, \mu_2, \theta_0, \theta_1, C, \alpha_1, \alpha_2)$. In Section 4.1, we formally define the fluid model and demonstrate the efficacy of the approximation at process level. In Section 4.2, we show that the fluid model converges to a unique equilibrium and demonstrate that the equilibrium approximates the steady-state behavior of the stochastic model well.

4.1. Fluid Model

We introduce the fluid model with exogenous parameters $\phi_f = (\lambda, \mu_1, \mu_2, \theta_0, \theta_1, C, \alpha_1, \alpha_2)$, analogous to the stochastic model introduced in Section 3. The idea of constructing the fluid model is to replace the stochastic components with their average. For example, we replace the arrival process by a deterministic flow. In the fluid model, passengers and drivers should be thought of as flow that is not integer-valued. Let $q(t)$ and $z_0(t)$ be respectively the number of *requesting* passengers and the number of *idle* drivers at time t in the fluid model. Let $z_1(t)$ and $z_2(t)$ be the number of *assigned* drivers and the number of *busy* drivers respectively. In the regime where the size of the system, indicated by the number of drivers, scales up, what the fluid model can approximate is the normalized stochastic processes by the system size. Thus, the total amount of drivers in the fluid model is 1, i.e. for all time $t \geq 0$,

$$z_0(t) + z_1(t) + z_2(t) = 1. \tag{11}$$

To introduce the rest of the fluid dynamic equations, we replace the arrival process $A(t)$ by the deterministic fluid process λt. Regarding the definition of the stochastic abandonment processes $R_0(t)$ and completion process $D_2(t)$ in Appendix A (see (32) and (33)), their fluid counterparts can be written as

$$r_0(t) = \theta_0 \int_0^t q(s)ds, \tag{12}$$

$$d_2(t) = \mu_2 \int_0^t z_2(s)ds. \tag{13}$$

The fluid counterparts of the cancellation and successful pick-up processes are more complicated due to their dependency on the matching process. Let $m(t)$ be amount of matchings made up to time t, corresponding to the stochastic process $M(t)$. Further, let $d_1(t)$ be the total number of successful pick-ups and $r_1(t)$ be the total number of cancellations, up to time t. Analogous to the balance equations (4)–(7) in the stochastic system, the fluid processes should satisfy

$$q(t) = q(0) + \lambda t - \theta_0 \int_0^t q(s)ds - m(t), \tag{14}$$

$$z_0(t) = z_0(0) + r_1(t) + \mu_2 \int_0^t z_2(s)ds - m(t), \tag{15}$$

$$z_1(t) = z_1(0) + m(t) - d_1(t) - r_1(t), \tag{16}$$

$$z_2(t) = z_2(0) + d_1(t) - d_2(t). \tag{17}$$

The following equations formally define $d_1(t)$ and $r_1(t)$.

$$d_1(t) = \frac{\mu_1}{\mu_1 + \theta_1} z_1(0)(1 - e^{-(\mu_1+\theta_1)t}) + \int_0^t \frac{\mu_1(s)}{\mu_1(s) + \theta_1}(1 - e^{-(\mu_1(s)+\theta_1)(t-s)})dm(s), \tag{18}$$

$$r_1(t) = \frac{\theta_1}{\mu_1 + \theta_1} z_1(0)(1 - e^{-(\mu_1+\theta_1)t}) + \int_0^t \frac{\theta_1}{\mu_1(s) + \theta_1}(1 - e^{-(\mu_1(s)+\theta_1)(t-s)})dm(s). \tag{19}$$

The first term on the right-hand side of (18) is about the initial state, where $z_1(0)e^{-(\mu_1+\theta_1)t}$ is the number of drivers who remain in the *assigned* state at time t, with the rest transited either to the *idle* state due to cancellation or the *busy* state due to successful pick-up. Note that $\frac{\mu_1}{\mu_1+\theta_1}$ is the probability that a pick-up is successful for the *assigned* drivers at time 0. The second term calculates the total number of successful pick-ups resulting from passengers matched during the time interval $(0,t]$ by taking integrals from 0 to t. To see this, for the $dm(s)$ amount of passengers matched at time s, a proportion of $1 - e^{-(\mu_1(s)+\theta_1)(t-s)}$ have turned into either successful pick-ups or cancellations by time t. Note that $\frac{\mu_1(s)}{(\mu_1(s)+\theta_1)}$ is the probability of a successful picking-up if the passenger is matched at time s, thus the amount of successful pick-ups that are matched at time s is $\frac{\mu_1(s)}{\mu_1(s)+\theta_1}(1 - e^{-(\mu_1(s)+\theta_1)(t-s)})dm(s)$ at time t. Taking integrals from 0 to t leads to the second term on the right-hand side of (18). The explanations for (19) follows the same logic by replacing successful pick-ups with cancellations.

We now analyze how $m(t)$ is determined by our matching policy. Following Assumption 1, the fluid pick-up rate function is

$$\mu_1(t) = C(q(t))^{\alpha_1}(z_0(t))^{\alpha_2}. \tag{20}$$

To avoid tedious discussion on the initial condition, we make a simplifying assumption that for drivers in the *assigned* state at time 0, i.e., those counted in $z_1(0)$, the pick-up rate is μ_1. Analogous to the stochastic model, what the fluid match process $m(t)$ does is to keep the matching rate $\mu_1(t)$ under the threshold μ_1 by matching the minimum amount of passengers. Mathematically, we formulate the intuition above as

$$\mu_1(t) \leq \mu_1, \tag{21}$$

$$\int_0^t \mathbf{1}_{\{\mu_1(s)<\mu_1\}} dm(s) = 0. \tag{22}$$

Note that (22) requires that $m(t)$ stays constant, i.e., no passengers are matched, when $\mu_1(t) < \mu_1$. Moreover, (21) and (22) imply that $m(t)$ increases, i.e., passengers are matched, only when $\mu_1(t) =$

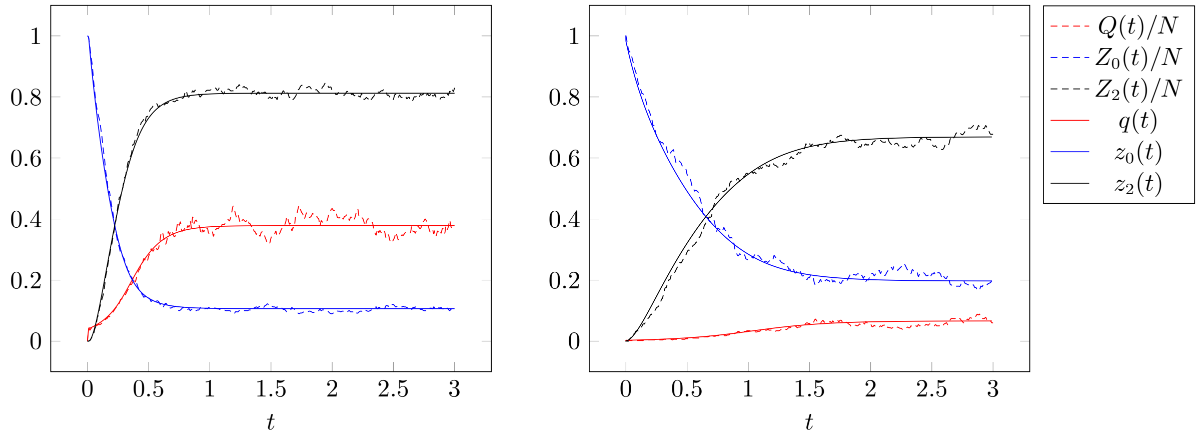
μ_1. In other words, in our fluid model, for all matched passenger-driver pair, the corresponding pick-up rate must be μ_1. With all the formulations introduced above, we can now define the fluid model.

Definition 1 (Fluid process). A process $x(t) = (q(t), z_0(t), z_1(t), z_2(t))$ is called the fluid process of the stochastic system with decision μ_1 if it satisfies (11)–(22).

Although the formulation does not give an analytic form of $m(t)$, Proposition 1 establishes the existence and uniqueness of the fluid process under the matching policy with any given threshold μ_1. The proof is postponed to Appendix D.

Proposition 1. *For any initial condition satisfying* $q(0), z_0(0), z_1(0), z_2(0) \geq 0$, $\mu_1(0) = C(q(0))^{\alpha_1}(z_0(0))^{\alpha_2} \leq \mu_1$ *and* $z_0(0) + z_1(0) + z_2(0) = 1$, *there exists a unique fluid process that satisfies* (12)–(22).

For a side-by-side comparison between the stochastic model and the fluid model, please see Table 4 in Appendix A. To illustrate how the fluid model approximates the stochastic one, Figure 3 plots the fluid process and the normalized stochastic process with size $N = 500$ under two set of parameters. Two observations can be made based on the figure. First, the fluid model is indeed an accurate approximation to the stochastic one at the process level. Second, both the scaled process and the fluid process seem to converge to some "equilibrium", which we will formally prove in the next section. Similar observations can be made under various parameter settings.



(a) $\lambda = 5$, $\mu_1 = 10$, $C = 50$, $\alpha_1 = \alpha_2 = 0.5$

(b) $\lambda = 2$, $\mu_1 = 5$, $C = 30$, $\alpha_1 = 0.3, \alpha_2 = 0.6$

Figure 3 **The scaled stochastic process** $\phi^N = (N, N\lambda, \mu_1, \mu_2, \theta_0, \theta_1, C/N^{\alpha_1+\alpha_2}, \alpha_1, \alpha_2)$ **(dashed lines) and the corresponding fluid process** $\phi_f = (\lambda, \mu_1, \mu_2, \theta_0, \theta_1, C, \alpha_1, \alpha_2)$ **(solid lines).**
Common parameters: $\mu_2 = 1$, $\theta_0 = 10$, $\theta_1 = 5$, $N = 500$.

4.2. Equilibrium and Approximations

In this section, we first show that the fluid process converges to a unique equilibrium, and then demonstrate the accuracy of the approximation. The equilibrium also reveals how the decision variable μ_1 affects the system performances and paves the way for the optimization analysis in Section 5.

We start with some basic assumptions to avoid trivial or unrealistic cases. For a given fluid model with $\phi_f = (\lambda, \mu_1, \mu_2, \theta_0, \theta_1, C, \alpha_1, \alpha_2)$, the effective range for μ_1 is $(0, C(\frac{\lambda}{\theta_0})^{\alpha_1}]$. Note that any threshold μ_1 that satisfies $\mu_1 > C(\frac{\lambda}{\theta_0})^{\alpha_1}$ simply means no passengers will be matched, because

$$\mu_1(t) = C(q(t))^{\alpha_1}(z_0(t))^{\alpha_2} \leq C(\frac{\lambda}{\theta_0})^{\alpha_1},$$

where the inequality follows from $z_0(t) \leq 1$ and $q(t) \leq \lambda/\theta_0$ by (14). Another assumption we would like to make is $\theta_1 > \mu_2$. The interpretation of the assumption is that the average patience time for pick-up is shorter than the average travel time. The assumption is made on the average level and does not exclude the possibility that some passengers would wait longer than their travel time. Theorem 1 establishes the equilibrium and the convergence of the fluid model $x(t)$.

THEOREM 1 **(Convergence to the equilibrium)**. *For any fluid model with parameters $\phi_f = (\lambda, \mu_1, \mu_2, \theta_0, \theta_1, C, \alpha_1, \alpha_2)$ that satisfies $\theta_1 > \mu_2$ and $\mu_1 \in (0, C(\frac{\lambda}{\theta_0})^{\alpha_1}]$, the statements below hold.*

1. *There exists a unique solution $\bar{x} = (\bar{q}, \bar{z}_0, \bar{z}_1, \bar{z}_2) \geq 0$ to the following equations*

$$\lambda = q\theta_0 + z_1\theta_1 + z_2\mu_2, \tag{23}$$

$$z_1\mu_1 = z_2\mu_2, \tag{24}$$

$$1 = z_1 + z_2 + z_0, \tag{25}$$

$$\mu_1 = C(q)^{\alpha_1}(z_0)^{\alpha_2}. \tag{26}$$

2. *$\bar{x}$ is the unique equilibrium of the fluid model $x(t)$.*
3. *The fluid process converges to $\bar{x}$, i.e., $(q(t), z_0(t), z_1(t), z_2(t)) \to (\bar{q}, \bar{z}_0, \bar{z}_1, \bar{z}_2)$ as $t \to \infty$.*

The intuition behind (23)–(26) follows from the balance equations of the system in steady states. Equation (23) requires that the arrival rate of passengers be equal to the passenger outflow, which is the sum of passenger abandonments, passenger cancellations, and trip completions. In order for the number of busy drivers to stay stable, the rate of successful pick-up should be equal to the trip completion rate, as required in (24). The right-hand side of (26) follows from Assumption 1, which should be equal to the desired matching rate μ_1 in equilibrium. Equation (25) regulates the total number of drivers in the system.

Let $\mathbb{E}[Q]$, $\mathbb{E}[Z_0]$, $\mathbb{E}[Z_1]$ and $\mathbb{E}[Z_2]$ be the expectation of the steady states of stochastic processes $Q(t)$, $Z_0(t)$, $Z_1(t)$, $Z_2(t)$, respectively. Table 1 presents the 95% confidence interval of the above-mentioned quantities out of 4500 simulated samples after warm up and the corresponding parts in the fluid equilibrium. We vary the arrival rate λ to cover systems of different loads. As illustrated in Table 1, our approximation works well for large-scaled systems. We also observe that the accuracy of our approximation decreases when $\mathbb{E}[Q]$ or $\mathbb{E}[Z_0]$ is small primarily due to the *rounding loss.* See Appendix C for detailed discussion on the rounding loss.

λ	N	$E[Q]/N$	$E[Z_0]/N$	$E[Z_1]/N$	$E[Z_2]/N$
λ=0.5N	100	0.0082±0.0001	0.6918±0.0015	0.0226±0.0005	0.2856±0.0014
	500	0.0119±0.0001	0.7304±0.0008	0.0247±0.0002	0.2449±0.0007
	1000	0.0128±0.0001	0.7295±0.0005	0.0244±0.0002	0.2461±0.0005
	eq.	0.0136	0.7333	0.0242	0.2424
λ=2N	100	0.0789±0.0008	0.1213±0.0010	0.0774±0.0008	0.8013±0.0012
	500	0.0789±0.0004	0.1259±0.0005	0.0791±0.0004	0.7951±0.0006
	1000	0.0806±0.0002	0.1234±0.0004	0.0799±0.0002	0.7967±0.0004
	eq.	0.0806	0.1241	0.0796	0.7962
λ=10N	100	0.8653±0.0029	0.0088±0.0001	0.0740±0.0007	0.9172±0.0007
	500	0.8644±0.0013	0.0104±0.0001	0.0881±0.0004	0.9015±0.0004
	1000	0.867±0.0009	0.011±0.0001	0.0875±0.0003	0.9014±0.0003
	eq.	0.8652	0.0116	0.0899	0.8986

Table 1 A comparison of the fluid approximation with stochastic system in steady state.
$C=100$, $\theta_0=10, \theta_1=5$, $\mu_2=1$, $\mu_1=10$, $\alpha_1=\alpha_2=0.5$.

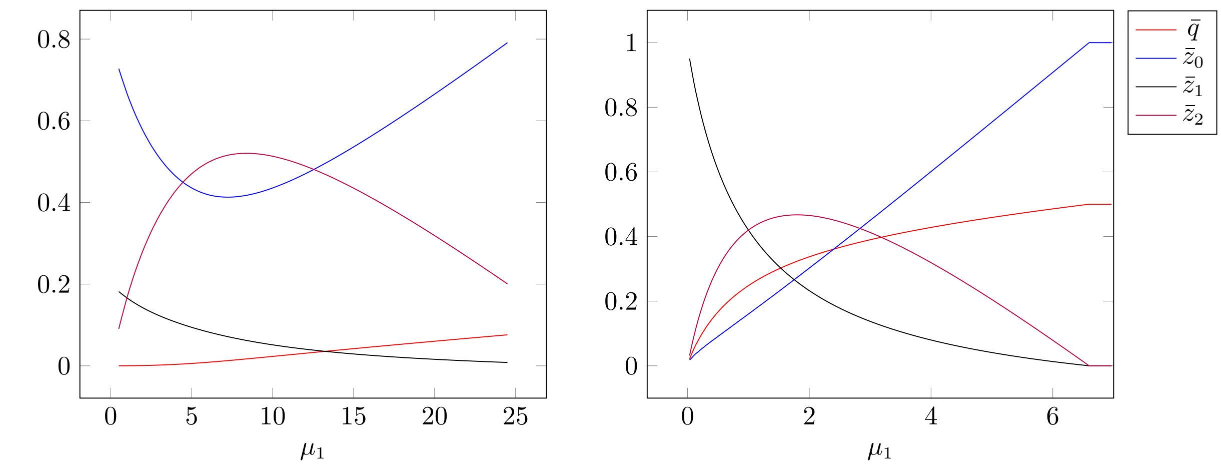
5. Prescriptive Analysis

Before getting into the optimization of the system, we first perform sensitivity analysis to provide insights into how the decision variable μ_1, intuitively interpreted as the reciprocal of the matching radius, impacts on different performance metrics such as the number of *requesting* passengers and the number of drivers in different states.

Lemma 1. *For any fluid model $\phi_f=(\lambda,\mu_1,\mu_2,\theta_0,\theta_1,C,\alpha_1,\alpha_2)$ that satisfies $\theta_1>\mu_2$, its equilibrium $\bar{x}=(\bar{q},\bar{z}_0,\bar{z}_1,\bar{z}_2)$ satisfies the following properties:*

1. *$\bar{q}$ is increasing in μ_1;*
2. *$\bar{z}_1$ is decreasing in μ_1;*
3. *$\bar{z}_2$ is quasi-concave in μ_1.*

Figure 4 plots how the equilibrium $\bar{x}$ changes as the decision variable μ_1 varies. As μ_1 increases, the matching radius is reduced, and it is harder for the platform to find feasible passenger-driver pairs. Consequently, the number of *assigned* drivers z_1 decreases. Lemma 1 also echoes the intuition that more passengers would stay in the waiting pool when a platform applies a shorter matching



(a) $\lambda = 1, C = 100, \alpha_1 = \alpha_2 = 0.5$

(b) $\lambda = 5, C = 10, \alpha_1 = 0.6, \alpha_2 = 0.8$

Figure 4 Fluid equilibrium $\bar{x}$ under different μ_1.

(Common parameters: $\theta_0 = 10, \theta_1 = 5, \mu_2 = 1$)

radius. In general, the monotonicity of $\bar{z}_0$ does not hold, as shown in Figure 4. For the parameter settings in Figure 4(a), $\bar{z}_0$ decreases first then increases in μ_1, while for the parameter settings in Figure 4(b) $\bar{z}_0$ is monotonically increasing in μ_1. The intuition behind the possible non-monotonicity is that a larger threshold μ_1 makes pick-ups more likely to be successful, and when trips take much longer time than pick-ups, more drivers may end up being *busy* due to the higher successful pick-up rate, leaving fewer drivers in state *idle*.

5.1. Optimization

Based on the fluid equilibrium introduced in Section 4.2, we formulate an optimization problem. Our focus is on the steady-state performance under the threshold policy, which is approximated by the fluid equilibrium (23)-(26). We obtain an elegant optimality condition that sheds light on the optimal control. Further analysis of the optimality condition also yields interesting managerial insights.

A well-designed matching policy improves the efficiency of the system and generally benefits all parties. When there are $\bar{z}_2$ *busy* drivers in the fluid equilibrium, trips are completed at rate $\bar{z}_2\mu_2$. Therefore, following (10), the total revenue is generated at rate $(p_f\mu_2 + p_s)\bar{z}_2$, where $p_f\mu_2 + p_s$ can be interpreted as the revenue generated by one unit of drivers during one unit of time if they are *busy*. We assume that the platform splits the revenue with the drivers according to a proportion specified in advance. The splitting proportion, p_f, μ_2 and p_s are all considered exogenous parameters. As a result, if we measure the system performance by revenue generated by the platform (or drivers), our objective reduces to one of maximizing the number of busy drivers $\bar{z}_2$ among all fluid equilibria. It is interesting to note that the probability of completing the trip for a new passenger is $z_2\mu_2/\lambda$, which is also proportional to z_2. In summary, the benefit of drivers, passengers, and the platform boils down to z_2 consistently, and we therefore formulate the optimization problem as

$$\begin{aligned} \max_{\mu_1,q,z_0,z_1,z_2} \quad & z_2, \\ \text{s.t.} \quad & (23)-(26), \\ & q, z_0, z_1, z_2 \geq 0. \end{aligned} \tag{27}$$

Intuitively, if the platform matches fewer passengers, the density of *requesting* passengers and *idle* drivers will be higher on the map, and it is therefore more likely for the platform to find matching pairs with shorter pick-up distances. To find the optimal threshold, the platform needs to strike a balance between matching more passengers (*quantity*) and shortening the pick-up distance (*quality*). Let $(\mu_1^*, q^*, z_0^*, z_1^*, z_2^*)$ be the optimal solution to (27). Theorem 2 below provides the condition under which this tradeoff is optimized.

Theorem 2 **(optimality condition)**. *The optimal solution to problem* (27) *satisfies*

$$\alpha_1 \frac{z_1^* \theta_1}{q^* \theta_0} + \alpha_2 \frac{z_1^*}{z_0^*} = 1. \tag{28}$$

The proof of Theorem 2 is postponed to Appendix D. We refer to (28) as the key optimality equation. This equation characterizes the optimality condition and also leads to an easy-to-implement optimal control policy, as will be shown in Section 6. Notably, larger z_0 and q enable the platform to match passengers and drivers at a higher pick-up rate, which measures the *quality* of service. On the other hand, z_1 is the total number of on-going pick-ups, which measures the *quantity* of matches. Then it is easy to see that the two fractions on the left-hand side of (28) measures the balance between quality and quantity of our matching policy. Our analysis reveals that, in order to maximize its revenue, the platform has to fine-tune the pick-up rate such that the linear combination of the two fractions, with coefficients α_1 and α_2, is equal to 1.

The key optimality equation provides an elegant way to determine whether a decision μ_1 is optimal. Inspired by (28), we define the *key matching index* ζ of an equilibrium as

$$\zeta = \alpha_1 \frac{z_1 \theta_1}{q \theta_0} + \alpha_2 \frac{z_1}{z_0}. \tag{29}$$

Theorem 2 states that, for different decisions μ_1, the *optimal* equilibrium is one such that the corresponding ζ equals 1. Note that quantities involved in computing ζ can be easily translated to observable metrics of the system, with $z_1\theta_1$ and $q\theta_0$ being the rate of cancellations and abandonments, and, z_1 and z_0 being the numbers of *assigned* and *idle* drivers. Therefore, an estimator of the *key matching index* is simply

$$\hat{\zeta} = \alpha_1 \frac{\#\text{ cancellations}}{\#\text{ abandonments}} + \alpha_2 \frac{\#\text{ assigned drivers}}{\#\text{ idle drivers}}. \tag{30}$$

Although ζ is derived from the fluid equilibrium, its estimator $\hat{\zeta}$ is easy to compute for a practical ride-hailing system, as the related quantities are usually observable to the platform. Corollary 1 below shows how the platform should adjust its decision based on the value of ζ.

Corollary 1. *To approach the optimal solution μ_1^*, whenever $\zeta > 1$, the system should increase μ_1; whenever $\zeta < 1$, the system should decrease μ_1.*

Corollary 1 follows from the quasi-concavity of $\bar{z}_2$ on μ_1. Combining Corollary 1 and the estimator $\hat{\zeta}$ leads to a simple, data-driven, and self-adaptive control policy to find the optimal matching radius for the system. The implementation of the control policy, together with the estimation of α_1 and α_2, will be presented with a concrete example in Section 6.

5.2. More on the Optimal Decision

In this section, we investigate the sensitivity of the optimal decision with respect to a few system inputs, and compare it with fixed-threshold policies to gain more insights into the optimal policy.

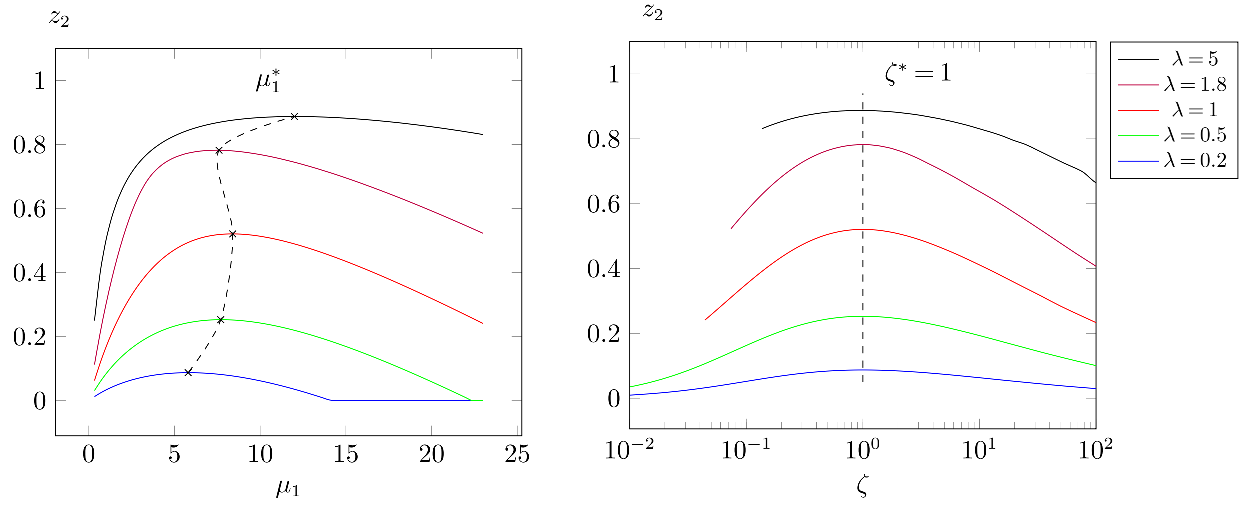
Passenger Arrival Rate The passenger arrival rate λ may be different depending on various factors such as working day v.s. holiday and rush hour v.s. non-rush hour. It is worth studying how the optimal matching rate and the platform's optimal revenue change with λ.

Figure 5(a) plots how the objective z_2 changes with respect to the decision variable μ_1 under different arrival rates. As we vary the matching decision μ_1, the *key matching index* ζ also changes accordingly. In Figure 5(b), we plot how the objective z_2 changes with respect to ζ for comparison. In both plots, we highlight the optimal solution in terms of μ_1 and ζ. Although under different λ the optimal decision variable μ_1^* for the problem (27) varies, the optimal ζ^* remains at 1. Therefore, ζ can be used to determine whether a decision is optimal without knowing the arrival rate λ.

Proposition 2. *The optimal objective value z_2^* and the optimal decision μ_1^* of optimization problem* (27) *satisfies the following:*

1. *The optimal objective value z_2^* is increasing in passenger arrival rate λ;*
2. *The optimal decision μ_1^* is increasing in λ if $q^* > \frac{\alpha_2\theta_1}{\alpha_1^2\theta_0}(\alpha_2 - (\alpha_1+\alpha_2)z_0^*)z_0^*$ and deceasing in λ if $q^* \le \frac{\alpha_2\theta_1}{\alpha_1^2\theta_0}(\alpha_2 - (\alpha_1+\alpha_2)z_0^*)z_0^*$.*

The non-monotonicity of μ_1^* in Proposition 2 is illustrated in Figure 5(a). Intuitively, when demand grows, i.e., the arrival rate increases, the platform would be more selective about pick-up time, i.e., to increase μ_1^*. However, this is in general not true as shown in Figure 5(a) when the arrival rate increase from $\lambda = 1$ to $\lambda = 1.8$. A similar observation of the non-monotonicity is made by Feng et al. (2017), who show that the pick-up time is not monotone in the passenger arrival rate under an on-demand non-idling policy and that the on-demand policy may result in inefficiency compared with a no-call policy (a policy that lets passengers and drivers meet randomly, analogous



(a) Threshold Value

(b) Key Matching Index

Figure 5 **Objective value $\bar{z}_2$ as a function of μ_1 and ζ.**

$C = 100$, $\theta_0 = 10, \theta_1 = 5$, $\mu_2 = 1$, $\alpha_1 = \alpha_2 = 0.5$.

to the traditional taxi model). Our analysis reveals that even if the platform always applies an optimal matching radius, the pick-up time is still not monotone in the passenger arrival rate.

On the passenger side, lowering the pick-up rate will result in lengthy pick-up times. However, passengers' overall experience may not necessarily be worse, since lowering the pick-up rate also helps to reduce the total waiting time. If the user experience, in particular, the waiting time for pick-up, is an important consideration (e.g., to enhance user loyalty), a platform can apply an upper bound on the matching radius to avoid long pick-up times.

Abandonments and Cancellations Let P_{ab} denote the probability that an arriving passenger abandons the platform while waiting to be matched, and P_c denote the probability that a passenger, who has been matched with a driver, cancels the service during the pick-up process. In the fluid equilibrium, the abandonment probability P_{ab} can be computed as $\bar{q}\theta_0/\lambda$ and the cancellation probability P_c as $\theta_1/(\theta_1 + \mu_1)$, both depending on the decision variable μ_1. It follows from the monotonicity of $\bar{q}$ that P_{ab} is increasing in μ_1, and it is clear that P_c is decreasing in μ_1. The solid line in Figure 6 depicts how both probabilities changes with μ_1. It follows from the fluid balance equation that the objective z_2 is linked to the probabilities of abandonment and cancellation according to the following equation

$$\frac{\mu_2 z_2}{\lambda} = (1 - P_{ab})(1 - P_c), \tag{31}$$

with both sides representing the probability that an arriving passenger completes a trip.

Equation (31) characterizes a unique feature of the on-demand ride-hailing market and offers another perspective from which to look at the tradeoff between the two probabilities. The dashed

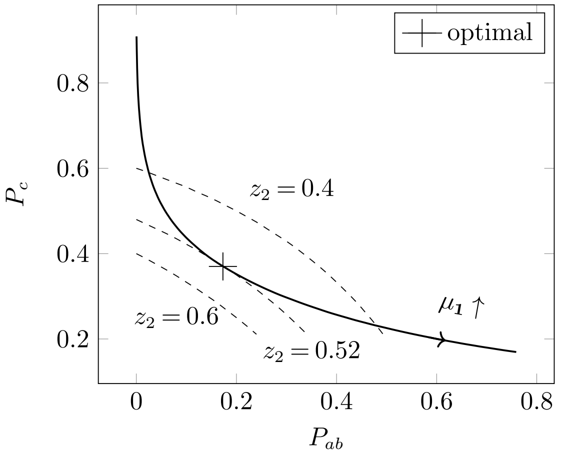


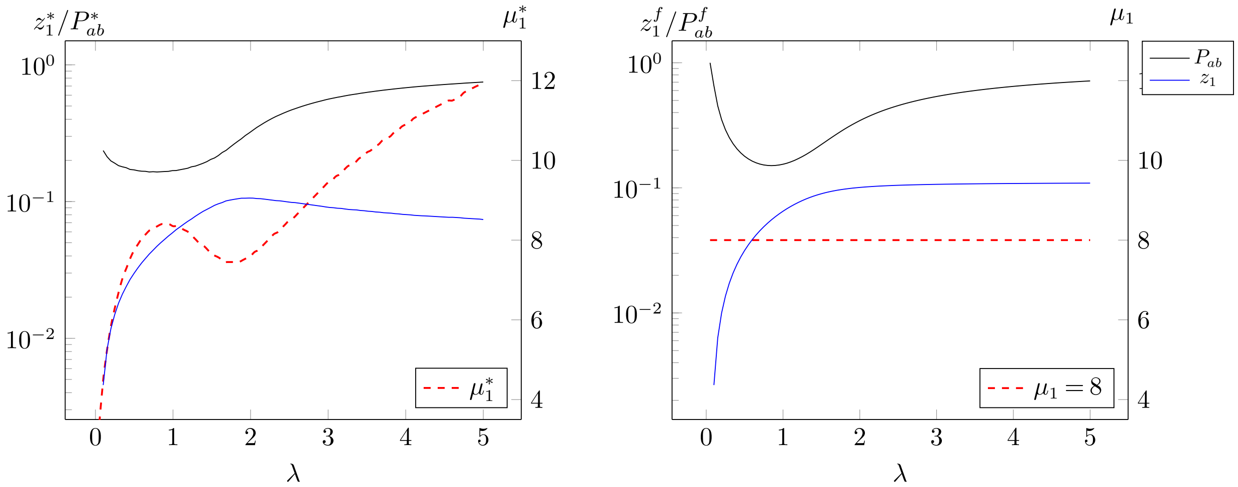
Figure 6 **Trade-off between abandonments and cancellations.**
$\lambda = 1, C = 100, \theta_0 = 10, \theta_1 = 5, \mu_2 = 1, \alpha_1 = \alpha_2 = 0.5.$

lines in Figure 6 are contours of P_{ab} and P_c at different values of the objective z_2. To strike a balance between these two probabilities, we need to find the contour of z_2 that is tangent to the trade-off curve (the solid line in the picture) based on (31). The tangent point is, in fact, the optimal point. The important message here is that abandonments and cancellations cannot be reduced at the same time. In order to completely avoid cancellations, the platform has to apply an extremely short matching radius, in which case the ride-hailing market reduces to a traditional taxi market that fails to utilize the spatial information. On the other hand, to reduce abandonments as much as possible, the platform must adopt a non-idling policy, which would cause a significant increase in pick-up time according to the spatial model.

Comparison with Fixed Threshold Policies To better understand the intuition behind the optimal decision, we plot how z_1 and P_{ab} change with the arrival rate λ under a fixed threshold in Figure 7(b) and under the optimal threshold in Figure 7(a). The fixed threshold in Figure 7(b) is chosen to be $\mu_1 = 8$, roughly the average of the optimal decisions as λ varies from 0 to 5.

The comparison in Figure 7 shows a significant difference between the two policies. Firstly, when λ is small, the abandonment probability under the optimal decision is significantly lower than that under a fixed threshold, which indicates that, when the passenger arrival rate is low, it is more profitable for the platform to increase the matching distance and try to pick up more passengers even if the expected pick-up time is long. Secondly, when λ is large, the proportion of *assigned* drivers decreases as λ increases under the optimal decision, meaning that the system can achieve a higher throughput with fewer *assigned* drivers. In contrast, the fixed matching threshold fails to take advantage of a high demand and ends up keeping more drivers in the *assigned* state, and

[2] z_1 and P_{ab} share the same axis as they both lie within [0,1]. Readers should note that the axis is interpreted differently for z_1 and P_{ab}.



(a) Optimal threshold μ_1^*

(b) Fixed Threshold $\mu_1 = 8$

Figure 7 Sensitivity analysis with respect to the arrival rate.[2]

$C = 100$, $\theta_0 = 10, \theta_1 = 5$, $\mu_2 = 1$, $\alpha_1 = \alpha_2 = 0.5$.

hence fewer drivers in the *busy* state compared to the optimal threshold. These observations are formalized in the Proposition 3 below, whose proof is postponed to Appendix D.

PROPOSITION 3. *Let P_{ab}^f and z_1^f be respectively the abandonment probability and number of assigned drivers in equilibrium under a given fixed threshold policy. Then*

$$\limsup_{\lambda\to 0} P_{ab}^* < 1 = \lim_{\lambda\to 0} P_{ab}^f, \quad \lim_{\lambda\to\infty} z_1^* = 0,$$

and z_1^f is increasing with λ.

6. Simulation & Implementation

In this section, we demonstrate the validity of our findings by implementing our matching policy in simulated environments. We will also discuss the implementation in more realistic settings with a road system, traffic information and real demand data.

Our numerical experiments are performed in two *virtual* cities. The first one (a *square* city) is a 100×100 square equipped with Euclidian distance. We assume that the pick-up location and the destination of each passenger are independently and uniformly distributed on the *square* map. The second one (a *grid* city), closer to reality, is a 100×100 grid, where drivers can only travel on the grid. To simplify the computation of routing and travel distance, on the *grid* map, we require the pick-up locations to be at a crossroads and the destinations to be on the grid (i.e. on the road). To be specific, when a passenger arrives, the pick-up location is first generated uniformly on the map and then rounded to the closest crossroads. The destinations are generated in the same way but rounded to the closest point on the grid. In both cities, we assume the driving speed is 1, and ignore all traffic conditions. Similar settings are adopted in the simulations of Besbes et al. (2018a) and Feng et al. (2017).

The initial locations of the drivers are assumed to be uniformly distributed on the map. For the *grid* city, these locations are rounded to the closest points on the grid. Passengers arrive according to a Poisson process with a specified rate, which is allowed to be time varying in Section 6.3. Upon the arrival of each passenger, a pick-up location and a destination are randomly generated on the map as discussed in the previous paragraph. As the system runs, the drivers' locations change as and when they are assigned to passengers. If a passenger cancels the trip during a pick-up, the driver would stop and wait for further assignment. For simplicity, we assume that idle drivers stay in the same location. The platform sets a matching radius and constantly checks for available passenger-driver pairs to match. The revenue from each trip is normalized to be the travel distance of the trip.

For demonstration purpose, we adopt the following exogenous parameters:

- Number of drivers: $N = 100$
- Passenger abandonment rate: $\theta_0 = 0.2$
- Passenger cancellation rate: $\theta_1 = 0.05$

Note that the platform is assumed to be unaware of any of the parameters above, nor the passenger arrival rate. The only parameters the platform needs to estimate are α_1 and α_2.

6.1. Estimation of α_1 and α_2

As mentioned in the Introduction, the implementation of our algorithm requires estimating the spatial parameters α_1 and α_2, for which we propose a simple statistical method. We assume that the platform has a collection of samples of form (Q, Z_0, T_{min}), where Q is the number of *requesting* passengers, Z_0 is the number of *idle* drivers and T_{min} is the average of the minimum pairwise pick-up time when there are Q passengers and Z_0 drivers. In our case, data is generated by random sampling in the above introduced two virtual cities introduced above. For each pair of (Q, Z_0) with $Q \in \{5, 10, 15, ..., 100\}$ and $Z_0 \in \{5, 10, 15, ..., 100\}$, we compute $\bar{T}_{min}$ as the average of T_{min} over 100 randomly generated samples. Because we assume the travel speed of drivers is 1, the pick-up time reduces to the pick-up distance, which is measured by the Euclidean distance for the *square* city and the Manhattan distance for the *grid* city. For the implementation in a more general setting, with a road system and traffic, one may replace the distance measures with the pick-up travel time estimated by the map service provider. The origin and destination may also be generated according to the distribution obtained from historical data.

The first step of our estimation is to apply transformation $V = \log(1/\bar{T}_{\min})$, $U = \log(Q)$ and $W = \log(Z_0)$. Next, we use the least squares method to fit the function:

$$V = \alpha_1 U + \alpha_2 W + \beta$$

and obtain $\hat{\alpha_1}$ and $\hat{\alpha_2}$ as the estimates of α_1 and α_2. The regression results for the *square* map and the *grid* map are summarized in Table 2. We observe that the results are significant with $R^2 > 0.99$ for both cities. The *grid* city has a slightly lower value of $\hat{\alpha}_1(\hat{\alpha}_2)$ than the *square* city. Due to high symmetry in our simulation setting, the estimation of α_1 and α_2 returns the similar values. In general, our method also applies to the case when α_1 and α_2 take significantly different values. In fact, as mentioned in the Introduction, α_1 and α_2 are expected to be different depending on the road and traffic condition in different areas.

	Pick-up Distribution	Destination Distribution	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{C}$	R^2
square	Uniform	Uniform	0.507	0.506	0.019	0.99
grid	Uniform, rounded to crossroads	Uniform, rounded to grid	0.525	0.526	0.015	0.99

Table 2 Regression results for estimating α_1 and α_2.

6.2. Constant Passenger Arrival

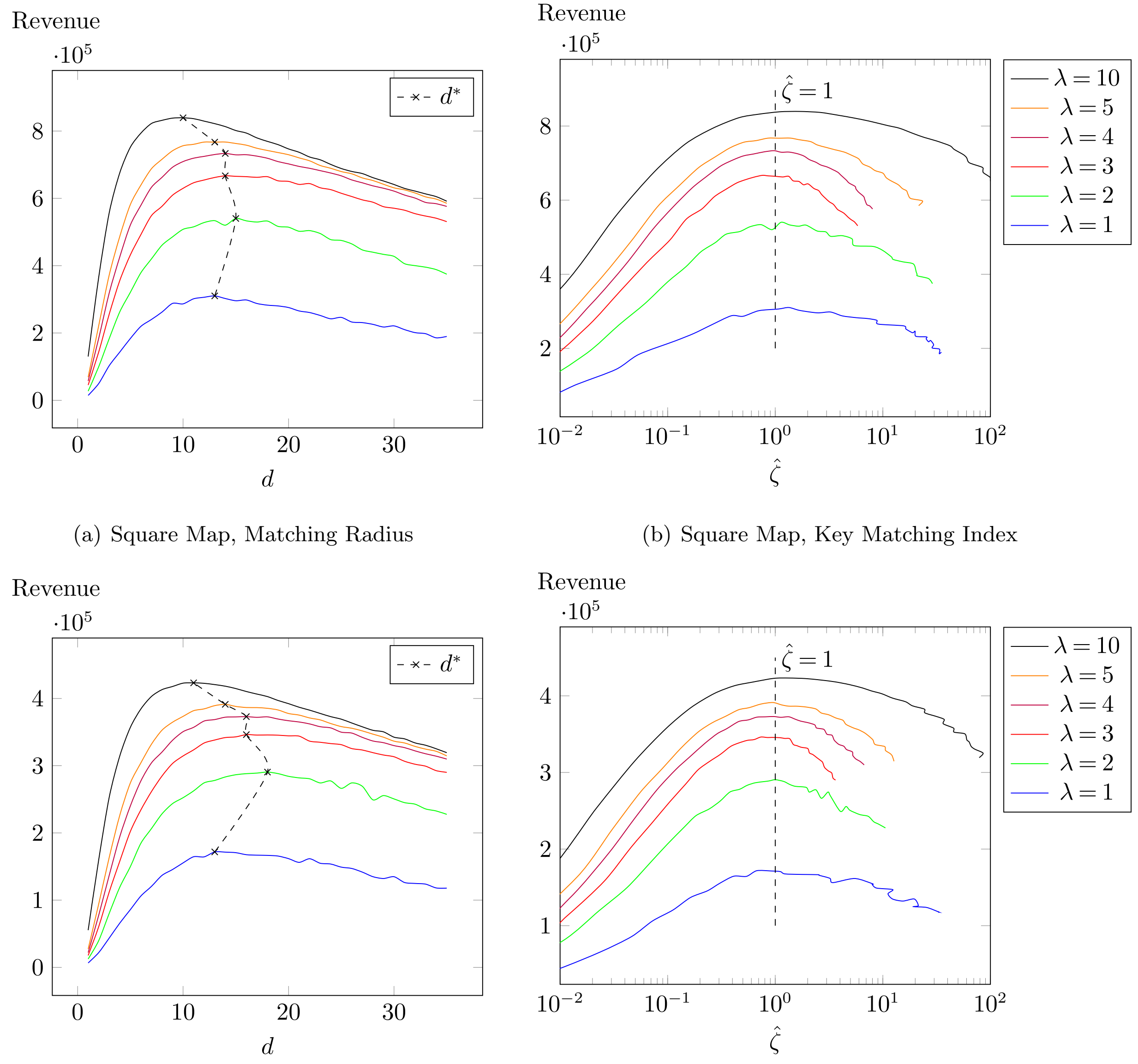
In this section, we assume a stationary environment where the passengers arrival rate is constant λ. For different λ, we tested different *matching radiuses* (d) and plotted the corresponding revenue over 10000 time units in Figure 8. The similarity between Figure 5 and Figure 8 also shows that our theoretical result is still valid in the more realistic setting as described in this section. Figure 8(a) and 8(c) show how revenue changes for different matching radiuses d. The dashed line marks the optimal radius d^*, which is unique, under different arrival rates λ. Note that the *key matching index* ζ is not rigorously defined in the simulation setting and we therefore use its estimator $\hat{\zeta}$ defined in (30). For different matching radiuses d, we plot how the revenue change with respect to $\hat{\zeta}$ in Figures 8(b) and 8(d) for comparison. We can see that for the optimal decision, $\hat{\zeta}$ stays quite close to 1, even though the optimal matching radius d^* varies for different λ. The observations hold for both *square* and *grid* cities, demonstrating the robustness of our results for roads of different shapes.

6.3. Self-adaptive Control Policy

Corollary 1 motivates us to propose the following simple, data-driven and self-adaptive policy in a changing environment. We assume that the platform can only collect data as the system is running but has no *prior* knowledge of any system parameters except for α_1 and α_2.

We divide the execution of the simulation into multiple epochs, each of the same length. In practice, the length of the epochs should be chosen such that it is long enough for the platform to obtain a statistical estimation and short enough to capture the dynamics of market demand. Let τ be the index of epochs. For each epoch, the platform has to choose a matching radius from a set of matching radiuses $S = \{d_0, d_1, ...\}$, where the number of elements in S can be finite or infinite.[3]

[3] S is chosen at the platform's discretion, and we do not address theoretical learning efficiency here. In practice, the density of elements in S should not be too high to ensure learning efficiency.



(a) Square Map, Matching Radius

(b) Square Map, Key Matching Index

(c) Grid Map, Matching Radius

(d) Grid Map, Key Matching Index

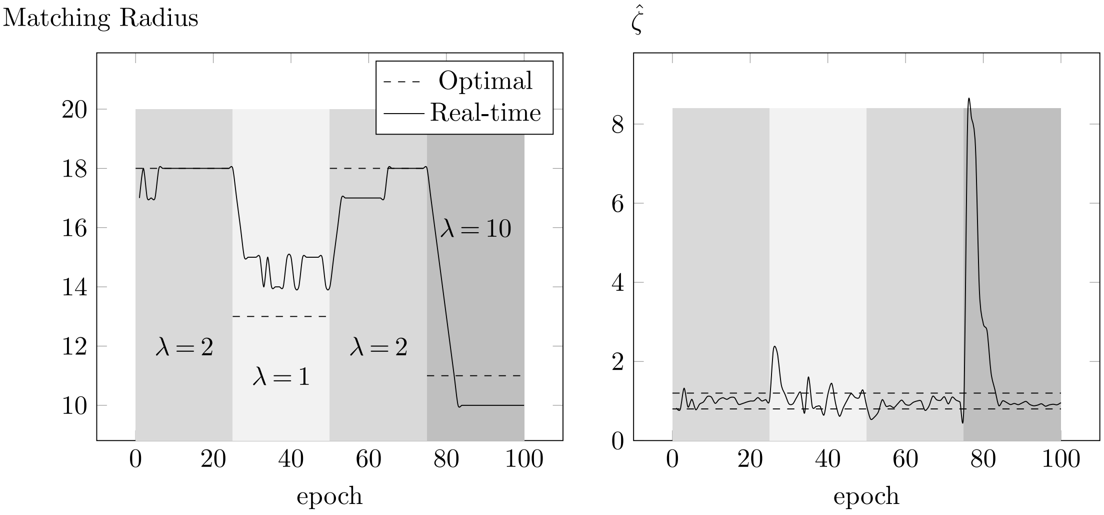
Figure 8 Revenue v.s. matching radius (left) and revenue v.s. ζ (right).

The self-adaptive control policy works as follows. Using the estimator $\hat{\zeta}$ defined in equation (30), at the beginning of epoch $\tau+1$, the platform estimates the *key matching index* for epoch τ as

$$\hat{\zeta}_\tau = \hat{\alpha}_1 \frac{\#\text{ cancellations in epoch } \tau}{\#\text{ abandonments in epoch } \tau} + \hat{\alpha}_2 \frac{\text{average } \#\text{ of assigned drivers in epoch } \tau}{\text{average } \#\text{ of idle drivers in epoch } \tau}.$$

If $\hat{\zeta}_\tau$ falls into a target neighborhood of 1, i.e., $\hat{\zeta}_\tau \in (1-\varepsilon_1, 1+\varepsilon_2)$ for some ϵ_1 and ϵ_2, the platform applies the same matching radius. Otherwise, the platform decrease the matching radius if $\hat{\zeta}_\tau > 1+\epsilon_2$ and increase the matching radius if $\hat{\zeta}_\tau < 1-\epsilon_1$. In other words, the platform aims to dynamically adjust the matching radius to keep the key matching index close to 1, as required by Theorem 2. The following two numerical experiments are conducted on the *grid* city. For demonstration purpose, we run the simulation for 100 epochs, each of 1000 time units. We also assume S to be the set of all positive integers that are smaller than 200, the largest possible distance between two points, and the target neighborhood is set to be $(0.8, 1.2)$.

Stepwise Demand Function. We first verify that our policy adapts to different passenger arrival rates. The 100 epochs are further divided into four periods of equal length, the first and the third of which have passenger arrival rate 2, indicated by the different shades of gray in Figure 9. In the second period, we set the passenger arrival rate to 1 to simulate a drop in demand. In the last period, we set the passenger arrival rate to 10 to simulate an spike in demand. In Figure 9(a), we compare the real-time matching radius to the optimal radius given by the simulation in the Section 6.2. The real-time key matching index is plotted in Figure 9(b). One can see that our policy is self-adaptive, i.e., when the demand rate switches across periods, the key matching index first bounces out of the target neighborhood, but then falls back within the neighborhood quickly.

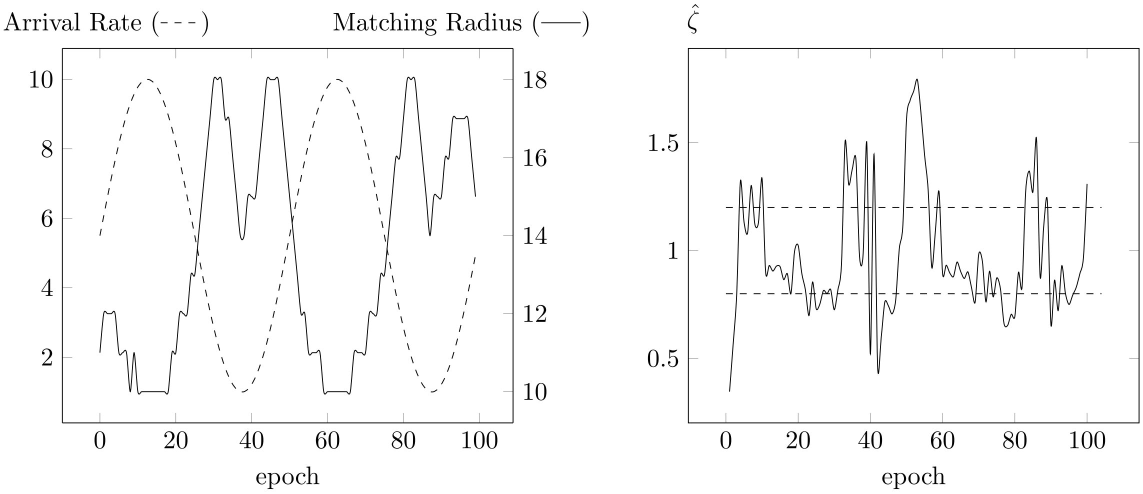


(a) Matching Radius, d_τ

(b) Key Matching Index, $\hat{\zeta}_\tau$

Figure 9 Experiment: step-wise demand.

Continuously Varying Demand Function. In our final experiment, we assume that the passenger arrival rate is sinusoidal as shown in Figure 10(a). As demand changes, the matching radius is adjusted automatically, so that the key matching index is kept near the target neighborhood, as indicated in Figure 10(b). As a comparison, we consider a static policy where the platform uses a fixed matching radius that cannot adjust itself in the changing environment. Table 3 compares the adaptive policy to ones with ones having a fixed matching radius. The result shows that the adaptive policy outperforms any static policy, and the improvement ranges from 0.5% to 37% depending on the choice of the fixed matching radius. In fact, sourcing a satisfactory fixed matching radius, even with the assistance of our results, requires demand forecasting and intense simulation. In that sense, the adaptive policy not only delivers superior performance but is also more robust to demand fluctuation.



(a) Continuously Varying Arrival Rate

(b) Key Matching Index

Figure 10 Experiment: continuously varying demand.

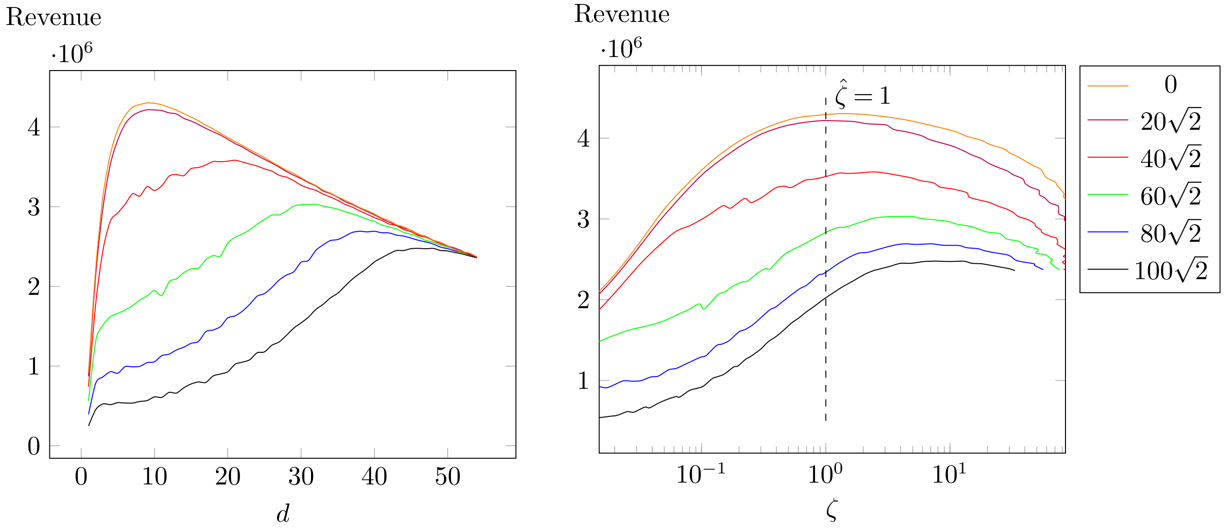
Matching Policy	Revenue(10^6)	Average Pool Size	Average Number of *assigned* Drivers	Revenue Increase
self-adaptive	6.42	18.23	9.9	-
Radius= 5	4.67	21.36	2.47	37.3%
Radius=10	6.22	19.06	7.15	3.3%
Radius=11	6.30	18.73	8.07	1.9%
Radius=12	6.38	18.45	9.04	0.6%
Radius=13	6.39	18.22	9.92	0.5%
Radius=14	6.38	17.95	10.85	0.6%
Radius=15	6.37	17.74	11.79	0.8%
Radius=16	6.34	17.48	12.68	1.2%
Radius=17	6.33	17.29	13.63	1.4%
Radius=18	6.28	17.11	14.48	2.2%
Radius=19	6.22	17.02	15.40	3.2%
Radius=20	6.15	16.82	16.25	4.4%
Radius=25	5.79	16.01	20.35	10.9%

Table 3 Comparison of self-adaptive policy and static polices with different matching radiuses

7. Discussion & Conclusions

7.1. Spatial Imbalance

In our numerical experiments, both the locations of requesting passengers and idle drivers are uniformly sampled. However, the real world may be more complicated. For example, during the morning rush hour, more passengers will want to travel from residential areas to business districts, creating an imbalance between demand and supply in the spatial model. As an extension, we provide a pilot experiment and a brief discussion on *spatial imbalance*. We sample pick-up locations and destinations (hence the locations of idle drivers) using different distributions. Specifically, we use a bivariate normal distribution with zero covariance and truncate the distribution so that the support is within the map. We separate the means of the two distributions along a diagonal of the map, while keeping the standard deviation fixed, to obtain different levels of imbalance.



(a) Matching Radius

(b) Key Matching Index

Figure 11 Revenue v.s. matching radius (left) and Revenue v.s. ζ (right) for imbalanced case

In Figure 11, we plot how revenue changes for different matching radiuses d and the indexes $\hat{\zeta}$ under different level of spatial imbalance. The figure legend shows the distance between the centers of the demand and supply distributions, with a larger difference indicating a larger imbalance level. Our optimality condition works reasonably well for lower levels of imbalance since the optimal revenue occurs at somewhere close to $\hat{\zeta} = 1$ for the curves corresponding to imbalance levels 0, $20\sqrt{2}$ and $40\sqrt{2}$ in Figure 11(b). However, as the imbalance level becomes higher (see curves corresponding to $60\sqrt{2}$, $80\sqrt{2}$ and $100\sqrt{2}$) the optimal revenue occurs when $\hat{\zeta}$ is larger than 1. As shown in Figure 11(a), the platform needs to apply a larger matching radius to get the optimal revenue as the imbalance become more severe. One possible explanation is that, in the absence of an empty-car routing policy, the platform must balance the supply and demand by matching drivers and passengers with long distances. Even if the passenger cancels their request during pick up, the driver would still have moved closer to the high demand area. The above discussion suggests a possible extension of our work, which is to combine the matching policy with empty-car routing (e.g., the one in Braverman et al. (2019)) in the case where there is a spatial imbalance between demand and supply.

7.2. Conclusion

The pick-up process is an important feature of the ride-hailing system, in the sense that it adds to the passengers' waiting time, wastes drivers time, and also wastes the capacity of the platform. Although nobody likes the pick-up process because of additional passenger waiting time, possible misuse of driver time and wasted capacity of the platform, it is inevitable due to the spatial nature of the system. In this paper, we studied the problem where a ride-hailing platform tries to find the optimal matching radius. The matching radius allows the platform to balance the trade-off between the quantity and quality of service. For tractability, we modeled the pick-up rate as a Cobb-Douglas product function of the number of requesting passengers and the number of idle drivers. Using fluid approximation, we obtain an optimization problem and identify an elegant optimality condition featuring the key matching index ζ, defined in (29). Based on our theoretical findings, we also proposed a self-adaptive that approaches the optimal decision, even without knowing the system parameters such as the demand rate and the passenger patience time. Simulations on a square map and a grid map suggest that our research can potentially be applied in real-world operations.

Some of the settings in our work can be relaxed or extended for future work. The first is *spatial imbalance* which can potentially be alleviated by empty-car routing, as discussed in Section 7.1. The second is pricing. In our setting, we focused on the matching policy and considered pricing policy as exogenously given. However, as Castillo et al. (2017) argue, pricing can also be used to solve the Wild Goose Chase problem. This emphasizes the possibility of studying the joint optimization problem of pricing and matching. Another possible direction is the customized matching decision. In practice, passengers use ride-hailing services for different purposes and their patience times will be different. Furthermore, different passengers (or drivers) may have different preferences for driver ratings and vehicle types (or pick-up distances and travel destinations). Optimization of the system efficiency with consideration of individual preferences remains an open question.

References

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

Akbarpour, M., S. Li, and S. Oveis Gharan (2017). Thickness and information in dynamic matching markets. *Working paper*.

Baccara, M., S. Lee, and L. Yariv (2018). Optimal dynamic matching. *Working paper*.

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

Banerjee, S., D. Freund, and T. Lykouris (2016). Pricing and optimization in shared vehicle systems: An approximation framework. *arXiv preprint arXiv:1608.06819*.

Banerjee, S., Y. Kanoria, and P. Qian (2018). The value of state dependent control in ridesharing systems. *arXiv preprint arXiv:1803.04959*.

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

Benjaafar, S., J.-Y. Ding, G. Kong, and T. Taylor (2018). Labor welfare in on-demand service platforms. *Available at SSRN 3102736*.

Besbes, O., F. Castro, and I. Lobel (2018a). Spatial capacity planning. *Available at SSRN*.

Besbes, O., F. Castro, and I. Lobel (2018b). Surge pricing and its spatial supply response. *Columbia Business School Research Paper* (18-25).

Bimpikis, K., O. Candogan, and D. Saban (2019). Spatial pricing in ride-sharing networks. *Oper. Res.*.

Braverman, A., J. G. Dai, X. Liu, and L. Ying (2019). Empty-car routing in ridesharing systems. *Oper. Res.*.

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

Chan, C. W., G. Yom-Tov, and G. Escobar (2014). When to use speedup: An examination of service systems with returns. *Oper. Res. 62*(2), 462–482.

Cheng, S.-F., M. Hu, and J. Keppo (2019). Tragedy of the ride-hailing. *Fifth Marketplace Innovation Workshop*.

Coma, C. W. and P. H. Douglas (1928). A theory of production. In *Proceedings of the Fortieth Annual Meeting of the American Economic Association*, Volume 139, pp. 165.

DiDi (2019). Company information: https://www.didiglobal.com/law. [Online; accessed 16-May-2019].

Dong, J., P. Feldman, and G. B. Yom-Tov (2015). Service systems with slowdowns: Potential failures and proposed solutions. *Oper. Res. 63*(2), 305–324.

Feng, G., G. Kong, and Z. Wang (2017). We are on the way: Analysis of on-demand ride-hailing systems. *Working paper*.

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

Gurvich, I. and A. Ward (2014). On the dynamic control of matching queues. *Stochastic Systems 4*(2), 479–523.

Hu, M. and Y. Zhou (2018). Dynamic type matching. *Rotman School of Management Working Paper* (2592622).

Hu, M. and Y. Zhou (2019). Price, wage and fixed commission in on-demand matching. *Available at SSRN 2949513*.

Iglesias, R., F. Rossi, R. Zhang, and M. Pavone (2016). A bcmp network approach to modeling and controlling autonomous mobility-on-demand systems. *arXiv preprint arXiv:1607.04357*.

Korolko, N., D. Woodard, C. Yan, and H. Zhu (2018). Dynamic pricing and matching in ride-hailing platforms. *Available at SSRN*.

Luxen, D. and C. Vetter (2011). Real-time routing with openstreetmap data. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '11, New York, NY, USA, pp. 513–516. ACM.

Mandelbaum, A. and G. Pats (1995). State-dependent queues: approximations and applications. *Stochastic networks 71*, 239–282.

Mandelbaum, A., G. Pats, et al. (1998). State-dependent stochastic networks. part i. approximations and applications with continuous diffusion limits. *The Annals of Applied Probability 8*(2), 569–646.

Ozkan, E. (2018). Joint pricing and matching in ridesharing systems.

Ozkan, E. and A. R. Ward (2016). Dynamic matching for real-time ridesharing. *Working paper*.

Powell, S. G. and K. L. Schultz (2004). Throughput in serial lines with state-dependent behavior. *Management Science 50*(8), 1095–1105.

Rochet, J.-C. and J. Tirole (2003). Platform competition in two-sided markets. *Journal of the european economic association 1*(4), 990–1029.

Taylor, T. A. (2018). On-demand service platforms. *Manufacturing & Service Operations Management*.

Teschl, G. (2009). *Ordinary Differential Equations and Dynamical Systems*. Universiät Wien.

Uber (2018a). Cancelling an uber ride: https://help.uber.com/h/56270015-1d1d-4c08-a460-3b94a090de23.

Uber (2018b). Company information: https://www.uber.com/newsroom/company-info/. [Online; accessed 15-May-2019].

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

Xu, Z., Y. Yin, and J. Ye (2019). On the supply curve of ride-hailing systems. *Transportation Research Part B: Methodological*.

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

Appendix A: Detailed Description of the Stochastic Processes

The number of abandonments and completed trips by time t can be written as

$$R_0(t) = W_1\Big(\theta_0 \int_0^t Q(s)ds\Big), \tag{32}$$

and

$$D_2(t) = W_2\Big(\mu_2 \int_0^t Z_2(s)ds\Big), \tag{33}$$

where $W_1(\cdot)$ and $W_2(\cdot)$ are independent Poisson processes with unit rate.

Processes $R_1(t)$ and $D_1(t)$ are related to initial state of the system. At time 0, there are $Z_1(0)$ matched passengers and driver pairs. Let $Y_1, ..., Y_{Z_1(0)} \sim \exp(\mu_1(0))$ be the remaining pick-up time and $V_1, ..., V_{Z_1(0)} \sim \exp(\theta_1)$ be the remaining patience of these passengers. At time $t > 0$, system has matched $M_{\mu_1}(t)$ passengers. Let $\tau_{Z_1(0)+1}, ..., \tau_{Z_1(0)+M_{\mu_1}(t)}$ be the arrival time, $Y_{Z_1(0)+1}, ..., Y_{Z_1(0)+M_{\mu_1}(t)}$ be the remaining pick-up time and $V_{Z_1(0)+1}, ..., V_{Z_1(0)+M_{\mu_1}(t)}$ be the remaining patience of these passengers in arbitrary order, where $Y_i \sim \exp(\mu_1(\tau_i))$ and $V_i \sim \exp(\theta_1)$. We then have:

$$R_1(t) = \sum_{i=1}^{Z_1(0)} \mathbf{1}_{\{V_i < \min\{Y_i, t\}\}} + \sum_{i=Z_1(0)+1}^{Z_1(0)+M_{\mu_1}(t)} \mathbf{1}_{\{V_i < \min\{Y_i, t-\tau_i\}\}},$$
$$D_1(t) = \sum_{i=1}^{Z_1(0)} \mathbf{1}_{\{Y_i < \min\{V_i, t\}\}} + \sum_{i=Z_1(0)+1}^{Z_1(0)+M_{\mu_1}(t)} \mathbf{1}_{\{Y_i < \min\{V_i, t-\tau_i\}\}}.$$

Note that the pool of requesting passengers is not necessarily first come first served and that the order in which passengers are served depends on the spatial distributions of passengers and drivers. The stochastic model cannot characterize such a spatial relationship, and therefore we utilize the memoryless property of the exponential distribution and model the waiting pool as a pure counting process, without specifying the serving order.

Appendix B: Numerical Justification of the Spatial Model

To justify the usage of the Cobb-Douglas type function as a spatial model, we adopt the following regression analysis approach. The analysis follows a similar procedure to the one in Section 6.1 but is made for the purpose of providing justification instead of estimating parameters.

1. Stylized Maps. Our first set of experiments is conducted on a set of stylized maps, including

(a) A straight line of length 100, where drivers can travel in both directions,

(b) A square map, where drivers can travel anywhere within the map, and

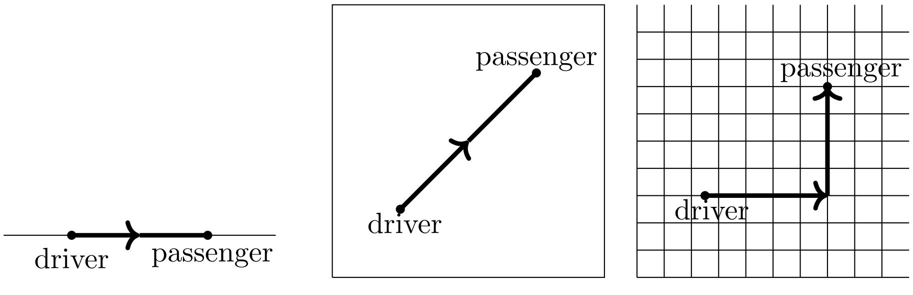
(c) A square map with grid roads, where drivers can only travel on the grids.

While maps a and b are highly representative of one and two-dimensional maps, map c lies somewhere in between. We randomly sample m passengers requesting for service and l idle drivers and record the minimum pairwise distance $d_{\min}$. On map c, as drivers can only travel on the grids, the Manhattan distance is used to measure distance. For each pair of values m and l, we take 100 samples and take the average over all $d_{\min}$ to obtain $E(d_{\min})$. Then we conduct the following regression analysis

$$\log(E(d_{\min})) = \beta + \alpha_1 \log m + \alpha_2 \log l. \tag{34}$$

	Stochastic Model	Fluid Model
Spatial Model	$\mu_1(t) = \tilde{C}(Q(t))^{\alpha_1}(Z_0(t))^{\alpha_2}$	$\mu_1(t) = C(q(t))^{\alpha_1}(z_0(t))^{\alpha_2}$
State Quantities	$Q(t) = Q(0) + A(t) - R_0(t) - M(t)$	$q(t) = q(0) + \lambda t - \theta_0 \int_0^t q(s)ds - m(t)$
	$Z_0(t) = Z_0(0) + R_1(t) + D_2(t) - M(t)$	$z_0(t) = z_0(0) + r_1(t) + \mu_2 \int_0^t z_2(s)ds - m(t)$
	$Z_1(t) = Z_1(0) + M(t) - D_1(t) - R_1(t)$	$z_1(t) = z_1(0) + m(t) - d_1(t) - r_1(t)ds$
	$Z_2(t) = Z_2(0) + D_1(t) - D_2(t)$	$z_2(t) = z_2(0) + d_1(t) - \mu_2 \int_0^t z_2(s)ds$
Process Quantities	$R_0(t) = W_1\left(\theta_0 \int_0^t Q(s)ds\right)$	$r_0(t) = \theta_0 \int_0^t q(s)ds$
	$R_1(t) = \sum_{i=1}^{Z_1(0)} \mathbf{1}_{\{V_i < \min\{Y_i, t\}\}} +$	$r_1(t) = \frac{\theta_1}{\mu_1+\theta_1} z_1(0)(1 - e^{-(\mu_1+\theta_1)t}) +$
	$\sum_{i=Z_1(0)+1}^{Z_1(0)+M_{\mu_1}(t)} \mathbf{1}_{\{V_i < \min\{Y_i, t-\tau_i\}\}}$	$\int_0^t \frac{\theta_1}{\mu_1(s)+\theta_1}(1 - e^{-(\mu_1(s)+\theta_1)(t-s)})dm(s)$
	$D_1(t) = \sum_{i=1}^{Z_1(0)} \mathbf{1}_{\{Y_i < \min\{V_i, t\}\}}$	$d_1(t) = \frac{\mu_1}{\mu_1+\theta_1} z_1(0)(1 - e^{-(\mu_1+\theta_1)t})$
	$\sum_{i=Z_1(0)+1}^{Z_1(0)+M_{\mu_1}(t)} \mathbf{1}_{\{Y_i < \min\{V_i, t-\tau_i\}\}}$	$\int_0^t \frac{\mu_1(s)}{\mu_1(s)+\theta_1}(1 - e^{-(\mu_1(s)+\theta_1)(t-s)})dm(s)$
	$D_2(t) = W_2\left(\mu_2 \int_0^t Z_2(s)ds\right)$	$d_2(t) = \mu_2 \int_0^t z_2(s)ds$
Matching Process	$M_{\mu_1}(t) = \int_0^t \mathbf{1}_{\{\mu_1(t) \geq \mu_1, Z_0(s-)>0\}} dA(s) +$	$\mu_1(t) \leq \mu_1$
	$\int_0^t \mathbf{1}_{\{\mu_1(t) \geq \mu_1, Q(s-)>0\}} d(R_1(s) + D_2(s))$	$\int_0^t \mathbf{1}_{\{\mu_1(s) < \mu_1\}} dm(s) = 0$
Initial Conditions	$Z_0(0) + Z_1(0) + Z_2(0) = K$	$z_0(t) + z_1(0) + z_2(0) = 1$
	$\mu_0(t) \leq \mu_1$	$\mu_0(t) \leq \mu_1$

Table 4 **Stochastic model and fluid model.**



(a)

(b)

(c)

Figure 12 **City configurations**

Results of the regression analysis are summarized in Table 5. Regression analysis shows that the spatial model (1) can be applied to all three scenarios, with different parameter configurations. Another important observation is that the values of the parameters α_1 and α_2 changes across different maps. In practice, the

Map a. Line (goodness of fit: $R^2 = 0.99$)

coef	est.	std. err	T-value	p-value	95% CI
β	3.9473	0.034	116.603	0	[3.881, 4.014]
α_1	-1.0067	0.006	-158.466	0	[-1.019, -0.994]
α_2	-1.0022	0.006	-157.759	0	[-1.015, -0.990]

Map b. Square (goodness of fit: $R^2 = 0.99$)

coef	est.	std. err	T-value	p-value	95% CI
β	4.1950	0.020	209.399	0.000	[4.156, 4.234]
α_1	-0.5246	0.004	-139.533	0.000	[-0.532, -0.517]
α_2	-0.5260	0.004	-139.917	0.000	[-0.533, -0.519]

Map c. Grid (goodness of fit: $R^2 = 0.99$)

coef	est.	std. err	T-value	p-value	95% CI
β	4.2083	0.007	573.496	0	[4.194, 4.223]
α_1	-0.5274	0.001	-382.973	0	[-0.530, -0.525]
α_2	-0.527	0.001	-382.722	0	[-0.530, -0.524]

Table 5 Regression analysis - stylized maps

$m, l \in \{5, 10, 15, ..., 100\}$**, 100 samples at each** (m, l)

road system, traffic conditions, and the distributions of pick-up locations and drivers' locations may also affect these parameters.

2. Map of New York City In order to find out whether the spatial model (1) can be applied to real-world situations, we use New York City as an example. We use point of interest(POI)[4] and roadbed[5] data hosted by the NYC Open Data[6]. Both datasets are provided by the Department of Information Technology & Telecommunications(DoITT). The POI dataset contains a list of about 20,000 common places and points of interest with in New York City. The roadbed dataset provides the roadbed in divided pieces, each having with a polygon shape. These polygons resemble the real road system. We assume that passengers are picked up at random POIs and idle drivers are randomly located on the roadbed. Figure 13(a) shows the distribution of the POIs in part of lower Manhattan, and Figure 13(b) shows the layout of the roadbed in the same region. It can be seen that the POIs cover a wide range of places where the need for ride-hailing services may emerge, and the roadbed is of a similar shape to the actual road system. The time it takes for a driver to reach a passenger is computed using Open Source Routing Machine(OSRM[7], Luxen and Vetter (2011)), which provides detailed driving routes between different places. Due to the high computational complexity, we reduce the sample size from 100 to 20. Regression results are shown in Table 6.

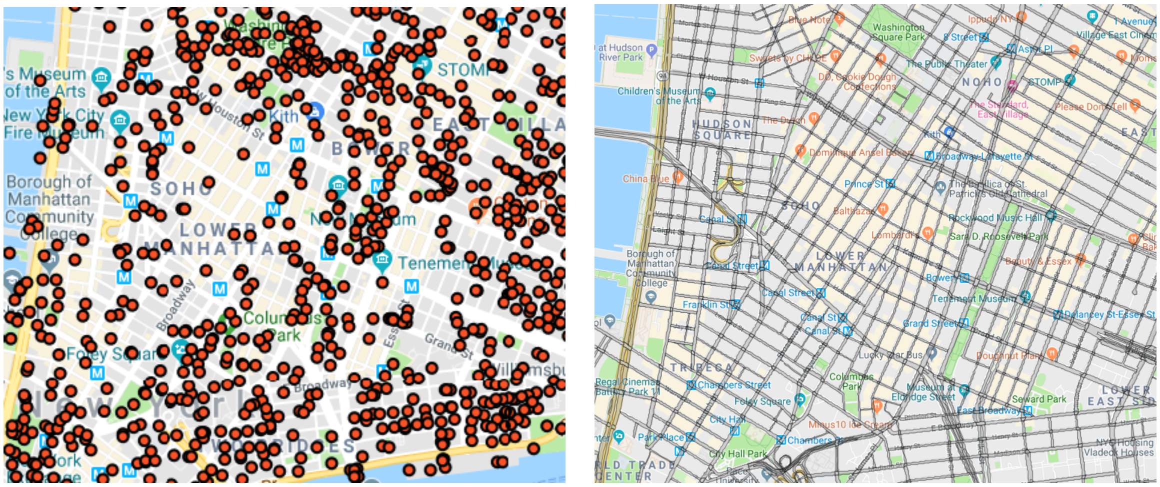
With a moderate sampling assumption, the regression results justify the potential use of the Cobb-Douglas function in real cities. Notably, the estimation of α_1 and α_2 in New York City returns values that are significantly different from 0.5, which indicates that α_1 and α_2 have to be chosen or estimated specifically for each scenario.

[4] `https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj`

[5] `https://data.cityofnewyork.us/City-Government/Roadbed/xgwd-7vhd`

[6] `https://opendata.cityofnewyork.us/`

[7] `http://project-osrm.org/`



(a) POIs: Lower Manhattan

(b) Roadbed: Lower Manhattan

Figure 13 POIs and roadbed in Manhattan

Map of New York City (goodness of fit: $R^2 = 0.93$)

coef	est.	std. err	T-value	p-value	95% CI
β	7.2849	0.067	108.152	0.000	7.151, 7.419
α_1	-0.3877	0.015	-25.712	0.000	-0.418, -0.358
α_2	-0.4077	0.015	-27.032	0.000	-0.438, -0.378

Table 6 Regression analysis - map of New York City

$m, l \in \{5, 10, 15, ..., 50\}$**, 20 samples at each** (m, l)

Appendix C: Approximation-Rounding Loss

As shown in Table 1, our approximation works reasonably well, even at a scale of $N = 100$. But we also observe that the accuracy of our approximation decreases as $E(Q)$ or $E(Z_0)$ becomes small primarily due to the subsequent *rounding loss*. In the original stochastic process, $\mu_1(t) = \tilde{C}(Q(t))^{\alpha_1} Z_0(t))^{\alpha_2}$ takes discrete values only. The jump size of the pick-up rate $\mu_1(t)$ is significant if one of $Q(t)$ and $Z_0(t)$ is small, and the other is large. For example, when $\tilde{\lambda} = 0.5N$ with $N = 100$, the average pool size is only 0.01 at the fluid scale while the amount of idle drivers is quite large at 0.73. Each time the pool size in the stochastic model changes by 1, the pick-up rate varies dramatically. Therefore, the pick-up rate in the stochastic system will be significantly larger than μ_1, the pick-up rate in equilibrium in the fluid model. Without rounding loss, one would expect $E(Z_2)/E(Z_1) \approx \mu_1/\mu_2$ because of (24). But in the case where $\tilde{\lambda} = 0.5N$ and $N = 100$, the ratio $E(Z_2)/E(Z_1) \approx 12.6$, significantly larger than $\mu_1/\mu_2 = 10$. Note that the rounding loss is mitigated as N grows large, as observed in Table 1 that the approximation accuracy increases as the system scales up. Therefore, the rounding loss is of less concern since a real ride-hailing system is usually large in scale, and the average pool size or the number of idle drivers is rarely too small. Actually, the optimality condition we derived in Section 5 does not require an accurate estimation of the pick-up rate.

Appendix D: Proofs

Proof of Proposition 1: For a given threshold μ_1 and initial condition $\mu_1(0) < \mu_1$, we show that the system must first go through the stage where $\mu_1(t) < \mu_1$ and then the stage where $\mu_1(t) = \mu_1$.

The system starts with first stage as $\mu_1(0) < \mu_1$. Since $m(t)$ does not increase at this stage (see (22)), the evolution equations (14)-(17) become extremely simple as follows

$$q(t) = q(0) + \lambda t - \theta_0 \int_0^t q(s)ds, \tag{35}$$

$$z_0(t) = z_0(0) + \int_0^t (z_2(s)\mu_2 + z_1(s)\theta_1)ds, \tag{36}$$

$$z_1(t) = z_1(0)e^{-(\mu_1(0)+\theta_1)t}, \tag{37}$$

$$z_2(t) = z_2(0) + \frac{\mu_1(0)}{\mu_1(0)+\theta_1} z_1(0)(1 - e^{-(\mu_1(0)+\theta_1)t}) - \mu_2 \int_0^t z_2(s)ds. \tag{38}$$

Evidently the derivatives of the terms (with respect to t) on the right hand side in (35)-(38) are Lipschitz continuous, so the existence and uniqueness of the process follows from the Picard-Lindelof theorem (Theorem 2.2 of Teschl (2009)). Note that $q(t)$ and $z_0(t)$ are increasing function of t (as long as $q(t) \le \lambda/\theta_0$, which can be easily shown to be true for any $t \ge 0$). We show by contradiction that the system must enter the second stage at some time point t_1. Suppose that the system is at the first stage for all $t \ge 0$. From (35)-(38) it can be easily shown that $q(t) \to \lambda/\theta_0$ and $z_0(t) \to 1$ when $t \to \infty$. Since $\mu_1 < C(\frac{\lambda}{\theta_0})^{\alpha_1}$, it follows from (20) that the system must enter the second stage when t is large enough, a contradiction.

We show that the pick-up rate $\mu_1(t)$ stays at μ_1 as the system enters the second stage. Suppose otherwise there exists $t_1 < t_2$ such that $\mu_1(t_1) = \mu_1 > \mu_1(t)$ when $t_1 < t \le t_2$. According to the analysis for the first stage, $\mu_1(t)$ is increasing on the interval $(t_1, t_2]$, contradicting the right continuity of $\mu_1(t)$ at $t = t_1$. What remains to be shown is the existence and uniqueness of the process at the second stage.

Note that $m(t)$ is monotone so $\pi(t) \triangleq m'(t)$ exists almost everywhere. Replace $\mu_1(t)$ by μ_1 in (14)-(17), rearrange and we obtain the following differential equations

$$q'(t) = \lambda - \theta_0 q(t) - \pi(t), \tag{39}$$

$$z_0'(t) = z_2(t)\mu_2 + z_1(t)\theta_1 - \pi(t), \tag{40}$$

$$z_1'(t) = -(\mu_1(0)+\theta_1)z_1(0)e^{-(\mu_1(0)+\theta_1)t} + \pi(t) - (\theta_1+\mu_1)z_1(t). \tag{41}$$

Since $z_0(t) + z_1(t) + z_2(t) = 1$, the expression of $z_2'(t)$ is omitted. Let $G(q,z) = Cq^{\alpha_1}z^{\alpha_2}$. For simplicity, let $\alpha_1 = \alpha_2 = \alpha$ (the case of $\alpha_1 \neq \alpha_2$ can be proven in a similar way). Since $\frac{d}{dt}G(q(t), z_0(t)) = 0$ at the second stage, i.e., $\frac{q'(t)}{q(t)} + \frac{z_0'(t)}{z_0(t)} = 0$, it is easy to derive

$$\pi(t) = \frac{z_0(t)}{z_0(t)+q(t)}(\lambda - \theta_0 q(t)) + \frac{q(t)}{z_0(t)+q(t)}(z_2(t)\mu_2 + z_1(t)\theta_1). \tag{42}$$

Plug (42) into (39)-(41). Note that $z_0(t) + q(t) \ge 2\sqrt{z_0(t)q(t)} = (\mu_1/C)^{1/(2\alpha)}$ which is bounded away from 0. It is easy to show that the right-hand sides of (39)-(41) are Lipschitz continuous and again, the existence and uniqueness of the fluid process at the second stage follows from the Picard-Lindelof theorem, completing the proof. □

Proof of Theorem 1: First we provide a roadmap for the proof. As shown in the proof of Proposition 1, the fluid process will go through two stages as time evolves: in the first stage where $\mu_1(t) < \mu_1$, when $\pi(t) = 0$, the pool size $q(t)$ and the number of idle drivers $z_0(t)$ will accumulate, until at some time point $\mu_1(t)$ reaches μ_1, and remain there afterwards, referred to as the second stage. Since it takes finite time for the system to evolve from the first stage to the second stage, we only need to focus on the second stage. From (39), (41) and (42), we can see that if $(q(t), z_0(t), z_1(t), z_2(t))$ converges at all, it must converge to $(\bar{q}, \bar{z}_0, \bar{z}_1, \bar{z}_2)$. To show the convergence of $(q(t), z_0(t), z_1(t), z_2(t))$, noting that each of them is bounded and differentiable, it is sufficient to show the monotonicity of each process. The proof shown below relies on a careful analysis of the derivative of each term. For simplicity assume $z_1(0) = 0$.

- **Existence and uniqueness of the equilibrium** From equations (23)-(26) it is easy to derive that

  $$z_1 = \frac{(\lambda - q\theta_0) - \mu_2(1 - z_0)}{\theta_1 - \mu_2},$$
  $$z_2 = \frac{\mu_1}{\mu_2}\frac{(\lambda - q\theta_0) - \mu_2(1 - z_0)}{\theta_1 - \mu_2}.$$

  For a fixed μ_1, let $z_0 = 1$ and $q = (\frac{\mu_1}{C_0})^{1/\alpha_1}$. Then we gradually decrease z_0 and increase q such that (26) holds, and check if (25) also holds. If yes, then a feasible solution is found. Since z_1 and z_2 are both increasing in z_0 and decreasing in q, the uniqueness of the solution to (23)-(26) is proven. Moreover, when z_0 approaches 0, q approaches $+\infty$, and the right-hand side of (25) is less than 1, so there must exist a pair (z_0, q) such that (25) holds by the intermediate value theorem. The existence is thus proven.

- **Convergence to the equilibrium** As in the proof of Proposition 1, we assume $\alpha_1 = \alpha_2 = \alpha$ for the sake of simplicity. Since $\mu_1(t) = \mu_1$ at the second stage, $\mu_1'(t) = (q(t)z_0(t))' = 0$. Recall that $\pi(t) \triangleq m'(t)$, we have

  $$q'(t)z_0(t) + q(t)z_0'(t) = (A(t) - \pi(t))z_0(t) + q(t)(B(t) - \pi(t)) = 0,$$

  where

  $$A(t) = \lambda - \theta_0 q(t), \;\; B(t) = z_1(t)\theta_1 + z_2(t)\mu_2.$$

  It follows that $\pi(t)$ lies between the value of $A(t)$ and $B(t)$. We consider the following two cases:

  1. The process $q(t)$ is monotone, i.e., $A(t) \geq B(t)$ or $A(t) \leq B(t)$ holds for all $t \geq 0$.

     First consider the case where $A(t) \geq B(t)$ for all $t \geq 0$. Then $q(t)$ is an increasing function by (39), and $q(t) \leq \lambda/\theta_0$, so it has limit q^*. It follows that $z_0(t)$ also has limit, denoted as z_0^*. Observe the evolution of $z_2(t)$. If $z_2'(t) = z_1(t)\mu_1 - z_2(t)\mu_2 \geq 0$ or ≤ 0 for a large enough t, then $z_2(t)$ is monotone and has limit z_2^*. The proof is complete. Otherwise, by the continuity of $z_1'(t)$, there exists a t such that $z_2'(t) = z_1(t)\mu_1 - z_2(t)\mu_2 = 0$. By (41), $z_1'(t) = \pi(t) - (\mu_1 + \theta_1)z_1(t) > B(t) - (\mu_1 + \theta_1)z_1(t) = 0$ (we assume $A(t) > B(t)$, otherwise equilibrium is already reached). Hence $z_2''(t) = z_1'(t)\mu_1 > 0$, which implies that for a small enough $\epsilon > 0$, $z_2'(t + \epsilon) > 0$. In summary, we have shown that if $z_2'(t) = 0$ for some t, then it will stay positive within a small interval. By the continuity of $z_2'(t)$, it must lead to $z_2'(t) \geq 0$ for all $t \geq 0$, and hence a limit exists. The case of $A(t) \leq B(t)$ follows a similar analysis and thus is omitted.

  2. $A(t) - B(t)$ changes its sign infinitely many times and reaches 0 at some time point.

     Suppose $A(t_1) = B(t_1)$ for some $t_1 > 0$. If $z_2'(t_1) = z_1(t_1)\mu_1 - z_2(t_1)\mu_2 = 0$, then the equilibrium is reached; now consider the case where $z_2'(t_1) > 0$ (analysis of the other case is similar and thus omitted). We can directly verify that $A'(t_1) = 0$ and $B'(t_1) = z_2'(t_1)\mu_2 + z_1'(t_1)\theta_1 = (\mu_2 - \theta_1)z_2'(t_1) < 0$, implying that $A(t_i + \epsilon) > B(t_i + \epsilon)$ for a small enough $\epsilon > 0$. We claim that $A(t) - B(t)$ and $z_2'(t)$ are both nonnegative for $t \geq t_1$, hence both $q(t)$ and $z_2(t)$ are monotone and have limits, completing the proof. By the continuity of both functions, it only needs to be shown that whenever $A(t) = B(t)$ and $z_2'(t) > 0$, then $(A(t) - B(t))' > 0$ and whenever $z_2'(t) = 0$ and $A(t) > B(t)$, then $z_2''(t) > 0$. The first case was just proven above. As for the second case, it is easy to verify that $z_2''(t) = z_1'(t)\mu_1 - z_2'(t)\mu_2 = z_1'(t)\mu_1 = -z_0'(t)\mu_1 > 0$ (since $q'(t) > 0$ and $q(t)z_0(t)$ is constant, then $z_0'(t) < 0$).

□

Proof of Lemma 1: We first prove that $\bar{q}$ is increasing in μ_1. From (23)-(26), we have:

$$(\frac{\mu_1}{C})^{1/\alpha_2} = \bar{q}^{\alpha_1/\alpha_2}(1 - \frac{(\lambda - \bar{q}\theta_0)(\mu_1 + \mu_2)}{\mu_2(\theta_1 + \mu_1)})$$

Take derivative over μ_1, and let $t = \alpha_1/\alpha_2$. We have

$$\begin{aligned}&\frac{\mu_2}{C^{1/\alpha_2}}((1/\alpha_2 + 1)\mu_1^{1/\alpha_2} + \theta_1/\alpha_2\mu_1^{1/\alpha_2 - 1})\\ &= (t+1)\theta_0\bar{q}^t\frac{d\bar{q}}{d\mu_1}(\mu_1 + \mu_2) + \bar{q}^{1+t}\theta_0 + (\mu_2 - \lambda)\bar{q}^t + t(\mu_2(\theta_1 + \mu_1) - \lambda(\mu_1 + \mu_2))\bar{q}^{t-1}\frac{d\bar{q}}{d\mu_1}.\end{aligned}$$

Rearranging yields

$$\frac{d\bar{q}}{d\mu_1} = \frac{\frac{\mu_2}{C_0^{1/\alpha_2}}((1/\alpha_2 + 1)\mu_1^{1/\alpha_2} + \theta_1/\alpha_2\mu_1^{1/\alpha - 1}) - (\bar{q}^{1+t}\theta_0 + (\mu_2 - \lambda)\bar{q}^t)}{(t+1)\theta_0\bar{q}^t(\mu_1 + \mu_2) + t(\mu_2(\theta_1 + \mu_1) - \lambda(\mu_1 + \mu_2))\bar{q}^{t-1}}.$$

The denominator is clearly positive. Plug $\mu_1 = C\bar{q}^{\alpha_1}\bar{z}_0^{\alpha_2}$ into the above equation. The numerator becomes

$$\mu_2((1/\alpha_2 + 1)\bar{q}^t\bar{z}_0 + 1/\alpha_2\frac{\theta_1}{\mu_1}\bar{q}^t\bar{z}_0) - \bar{q}^t(\mu_2 - \lambda + \bar{q}\theta_0).$$

Apply the fact that $\lambda - \bar{q}\theta_0 = \bar{z}_1(\mu_1 + \theta_1)$ and $\bar{z}_1\mu_1 = \bar{z}_2\mu_2$. The numerator further reduces to

$$\mu_2\bar{q}^t(-1 + \bar{z}_0 + \bar{z}_2 + \bar{z}_1\frac{\theta_1}{\mu_2} + \bar{z}_0/\alpha_2 + \frac{\theta_1}{\alpha_2\mu_1}\bar{z}_0),$$

which is clearly positive given that $\theta_1 > \mu_2$.

Next we prove that $\bar{z}_1$ is decreasing in μ_1. Again we utilize the equality $\lambda - \bar{q}\theta_0 = \bar{z}_1(\mu_1 + \theta_1)$. Since the left hand side of the above equality is decreasing in μ_1, it must lead to $\frac{d\bar{z}_1}{d\mu_1} < 0$.

Since in the equilibrium $\bar{z}_0$ is not necessarily monotone in μ_1, we have to consider several cases to prove the quasi-concavity of $\bar{z}_2$. First note that by (24), the ratio between *busy* and *assigned* drivers $\bar{z}_2/\bar{z}_1 = \mu_1/\mu_2$ increases with μ_1. For the range of μ_1 such that $\bar{z}_0$ is decreasing in μ_1, it follows directly from (25) that $(\bar{z}_1 + \bar{z}_2)$ is increasing. Thus, whenever the number of *idle* drivers $\bar{z}_0$ is decreasing in μ_1, it is always better for the platform to increase μ_1. As μ_1 increases, once $\bar{z}_0$ starts to decrease, we show that it will continue to decrease as μ_1 increases further. Finally we prove that $\bar{z}_2$ is quasi-concave when $\bar{z}_0$ is increasing in μ_1. That is, the derivative of $\bar{z}_2$ with respect to μ_1 changes its sign for at most once. To show the quasi-concavity of $\bar{z}_2$ with respect to μ_1, first we take logarithm on both sides of (26) and then take derivative over μ_1,

$$\alpha_1\frac{d\bar{q}}{d\mu_1}/\bar{q} + \alpha_2\frac{d\bar{z}_0}{d\mu_1}/\bar{z}_0 = \mu_1. \tag{43}$$

Taking derivative over μ_1 on both sides of (23) and (24) gives

$$\frac{d\bar{q}}{d\mu_1} = (-\theta_1 \frac{d\bar{z}_1}{d\mu_1} - \mu_2 \frac{d\bar{z}_2}{d\mu_1})/\theta_0, \quad \frac{d\bar{z}_0}{d\mu_1} = -\frac{d\bar{z}_1}{d\mu_1} - \frac{d\bar{z}_2}{d\mu_1}.$$

Put back into (43), rearrange and we have

$$\frac{d\bar{z}_2}{d\mu_1} = [\alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0 - 1]/C_1,$$

where $C_1 = (\theta_1 + \mu_1\mu_2)/(\theta_0\bar{q}) + (1+\mu_1)/\bar{z}_0 > 0$.

We discuss the sign of $\frac{d\bar{z}_0}{d\mu_1}$ when $\mu_1 \in (0, C(\lambda/\theta_0)^{\alpha_1}]$. If $\frac{d\bar{z}_0}{d\mu_1} > 0$ for a certain range of μ_1, combining that with Lemma 1, it is easy to see that $\alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0$ is decreasing in μ_1. On the other hand, if $\frac{d\bar{z}_0}{d\mu_1} \le 0$, combining that with $\frac{d\bar{z}_1}{d\mu_1} < 0$ it must lead to $\frac{d\bar{z}_2}{d\mu_1} = -\frac{d\bar{z}_0}{d\mu_1} - \frac{d\bar{z}_1}{d\mu_1} > 0$, i.e., $\alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0 > 1/$. Let $S_1 = \{\mu_1 : \alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0 \ge 1\}$ and $S_2 = \{\mu_1 : \alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0 \le 1\}$. It is easy to verify that $\mu_1 \in S_2$ when $\mu_1 \to C(\lambda/\theta_0)^{\alpha_1}$. We now argue that $\mu_1 \in S_1$ when μ_1 approaches 0. Since $\mu_1 = C(\bar{q})^{\alpha_1}(\bar{z}_0)^{\alpha_2}$, at least one of $\bar{q}$ and $\bar{z}_0$ approaches 0 when μ_1 approaches 0. In addition, as $\bar{z}_1$ is decreasing in μ_1, it follows that $\alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0$ goes to infinity when μ_1 approaches 0, hence the desired conclusion holds. Note that when $\mu_1 \in S_2$, it must hold that $\frac{d\bar{z}_0}{d\mu_1} > 0$ and hence $\alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0$ is decreasing in μ_1. So S_2 is an absorbing area, i.e., as μ_1 increases, once it enters S_2, it will always stay within S_2. It follows that S_1 and S_2 are intervals that complement each other, i.e., there exists a threshold value ξ such that $S_1 = (0, \xi]$ and $S_2 = [\xi, C(\lambda/\theta_0)^{\alpha_1}]$, and the quasi-concavity of $\bar{z}_2$ follows. □

Proof of Theorem 2: Recall in the proof of Lemma 1 that $S_1 = \{\mu_1 : \alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0 \ge 1\}$ and $S_2 = \{\mu_1 : \alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0 \le 1/\alpha\}$. Recall that $\mu_1 \in S_1$ when μ_1 approaches 0. First we show that there exist a unique μ_1 such that $\alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0 = 1$. By continuity, there must exists $\hat{\mu}_1 \in S_1 \cap S_2$, i.e., existence holds. At the point $\mu_1 = \hat{\mu}_1$, $\bar{z}_0$ is increasing in μ_1 and $\bar{z}_1$ is decreasing in μ_1, hence $\alpha_1\theta_1\bar{z}_1/(\theta_0\bar{q}) + \alpha_2\bar{z}_1/\bar{z}_0$ is strictly decreasing in μ_1, implying the uniqueness of the solution.

Our objective is to maximize the system output rate: $R = z_2\mu_2 = z_1\mu_1$. Use q and z_0 as the decision variable. By the balancing equations (23)-(26), we have:

$$z_1 = \frac{(\lambda - q\theta_0) - \mu_2(1 - z_0)}{\theta_1 - \mu_2}$$

Recall that $G(q, z_0) = C(q)^{\alpha_1}(z_0)^{\alpha_2}$. Our optimization problem can then be reformulated as:

$$\max_{q,z_0} \frac{(\lambda - q\theta_0) - \mu_2(1 - z_0)}{\theta_1 - \mu_2} G(q, z_0)$$

$$\text{s.t. } (G(q, z_0) + \mu_2)\lambda = (G(q, z_0) + \mu_2)q\theta_0 + \mu_2(1 - z_0)(\theta_1 + G(q, z_0)),$$

$$q, z_0 \ge 0,$$

where $G(q, z_0) = C(qz_0)^{\alpha}$.

The solution to the system takes one of the two forms:

1). The system runs at full capacity: either $q = 0$ or $z_0 = 0$, or both.

2). Both q and z_0 are positive.

It is easy to see that the optimal solution would take the second form. We now look into what properties the optimal solution should satisfy. The Lagrangian multiplier is:

$$L(q, z_0, \mu) = ((\lambda - q\theta_0) - \mu_2(1 - z_0))G(q, z_0) - \mu((G(q, z_0) + \mu_2)\lambda - (G(q, z_0) + \mu_2)q\theta_0 - \mu_2(1 - z_0)(\theta_1 + G(q, z_0)))$$

By the first order condition, we have:

$$\frac{\partial L}{\partial q} = -\theta_0 G(q,z_0) + A\frac{\partial G}{\partial q} - \mu(A\frac{\partial G}{\partial q} - \theta_0(G(q,z_0)+\mu_2)) = 0$$

$$\frac{\partial L}{\partial z_0} = \mu_2 G(q,z_0) + A\frac{\partial G}{\partial z_0} - \mu(A\frac{\partial G}{\partial z_0} + \mu_2(G(q,z_0)+\theta_1)) = 0,$$

where $A = (\lambda - q\theta_0) - \mu_2(1-z_0)$.

$$\frac{A\frac{\partial G}{\partial z_0} + \mu_2 G}{-A\frac{\partial G}{\partial q} + \theta_0 G} = \frac{\theta_1}{\theta_0}$$

$$\frac{A\frac{\partial log(G)}{\partial z_0} + \mu_2}{-A\frac{\partial \log(G)}{\partial q} + \theta_0} = \frac{\theta_1}{\theta_0}$$

$$A(\theta_0\frac{\partial \log(G)}{\partial z_0} + \theta_1\frac{\partial \log(G)}{\partial q}) = (-\mu_2 + \theta_1)\theta_0$$

Note that $A = (\lambda - q\theta_0) - \mu_2(1-z_0) = (\theta_1 - \mu_2)z_1$. We have the following:

$$z_1(\frac{\partial \log(G)}{\partial z} + \frac{\theta_1}{\theta_0}\frac{\partial \log(G)}{\partial q}) = 1 \qquad (44)$$

Since $G = Cq^{\alpha_1}z_0^{\alpha_2}$, it follows from (44) that

$$\alpha_1\frac{\theta_1}{\theta_0}\frac{z_1}{q} + \alpha_2\frac{z_1}{z_0} = 1$$

holds for the equilibrium solution $(\bar{q}, \bar{z}_0, \bar{z}_1, \bar{z}_2)$. □

Proof of Proposition 2: First we show that the optimal objective value z_2^* is increasing in λ. This is easily shown by finding a feasible solution with a larger objective value when λ increases by an amount of Δ: simply let $z_1' = z_1^* + \frac{\Delta}{\theta_1+\mu_2}, z_2' = z_2^* + \frac{\Delta}{\theta_1+\mu_2}$ while fixing all the other variables.

Secondly, we show that when $\lambda \to 0$ or $\lambda \to \infty$, the optimal decision μ_1^* is both increasing in λ. For a fixed λ, assume $z_0 = z_0^*, z_1 = z_1^*, z_2 = z_2^*$ and $q = q^*$ are optimal. Now increase λ to $\lambda + \Delta$ while keeping $\mu_1 = \mu_1^*$ fixed. So $(\bar{q}, \bar{z}_0, \bar{z}_1, \bar{z}_2)$ which solve (23)-(26) are functions of λ.

We are interested in the sign of $\frac{d}{d\lambda}(\alpha_1 z_1/z_0 + (\alpha_2 z_1\theta_1)/(q\theta_0))$. Since $\frac{\mu_1+\mu_2}{\mu_2}z_1 + z_0 = 1$, by replacing z_1 by the expression of z_0 and canceling constants we obtain

$$\frac{d}{d\lambda}(\alpha_1/z_0 + \alpha_2\theta_1(1-z_0)/(\theta_0 q)) = -\alpha_1\frac{dz_0}{d\lambda}/z_0^2 + \alpha_2\theta_1/\theta_0\cdot(-\frac{dz_0}{d\lambda}q - (1-z_0)\frac{dq}{d\lambda})/q^2. \qquad (45)$$

Since $G(z_0,q) = \mu_1^*$ is fixed, $q^{\alpha_1}z_0^{\alpha_2}$ is constant, implying that $\alpha_1\frac{dq}{d\lambda}/q + \alpha_2\frac{dz_0}{d\lambda}/z_0 = 0$. We show that $\frac{dq}{d\lambda} \geq 0$. Suppose otherwise $\frac{dq}{d\lambda} < 0$ and $\frac{dz_0}{d\lambda} > 0$, which according to (24) and (25) must lead to $\frac{dz_1}{d\lambda} < 0, \frac{dz_2}{d\lambda} < 0$, which contradicts (23).

Now we let $\frac{dq}{d\lambda} = \alpha_2 qw$, $\frac{dz_0}{d\lambda} = -\alpha_1 z_0 w$, where $w > 0$. Putting that back into the right-hand side of (45) yields

$$\alpha_1^2 w/z_0 - \alpha_2^2\theta_1/\theta_0\cdot(1-(1+\alpha_1/\alpha_2)z_0)w/q. \qquad (46)$$

It is clear that when $\lambda \to 0$, $q \to 0$ and $z_0 \to 1$, so (46) is positive; when $\lambda \to \infty$, $q \to \infty$ and $z_0 \to 0$, so (46) is also positive. By Corollary 1, when $\lambda \to 0$ or $\lambda \to \infty$, the optimal decision μ_1^* should be increasing in λ in order for (28) to hold. (46) also provides a criterion for the monotonicity of μ_1^* with respect to λ. Specifically, μ_1^* is increasing in λ if $q^* > \frac{\alpha_2\theta_1}{\alpha_1^2\theta_0}(\alpha_2 - (\alpha_1+\alpha_2)z_0^*)z_0^*$, and deceasing in λ otherwise. □

Proof of Proposition 3: We start by analyzing P^*_{ab} and z^*_1. Suppose otherwise $\limsup_{\lambda\to 0} P^*_{ab} = 1$. By taking subsequences, we can can assume w.l.o.g. that $P^*_{ab} = \frac{\theta_0 q^*}{\lambda} = \frac{\theta_0 q^*}{\theta_0 q^* + \theta_1 z^*_1 + \mu_2 z^*_2} \to 1$ when $\lambda \to \infty$, i.e., z^*_1/q^* and z^*_2/q^* both converge to 0. However, by (27) it follows that $z^*_1/z^*_0 \to 1/\alpha_2$, hence $z^*_0/q^* \to \infty$, which contradicts the fact that $z^*_0 + z^*_1 + z^*_2 = 1$ and $q^* \to 0$ when $\lambda \to \infty$. $\lim_{\lambda\to\infty} z^*_1 = 0$ follows from the fact that $z^*_2 \to 1$ when $\lambda \to \infty$.

Next we analyze P^f_{ab} and z^f_1. Under any fixed matching radius $\bar{\mu}_1$, when λ is small enough, the threshold $\bar{\mu}_1$ can never be achieved and the system makes no matches, resulting in $\lim_{\lambda\to 0} z^f_1 = \lim_{\lambda\to 0} z^f_1 = 0$ and hence $\lim_{\lambda\to 0} P^f_{ab} = 1$. It remains to show that z^f_1 is increasing with λ. Suppose z^f_1 decreases with λ at some point. By (24) z^f_2 also decreases with λ, which leads to z^f_0 increasing with λ in view of $z^f_0 + z^f_1 + z^f_2 = 1$. On the other hand, by (23) q^f must also increases with λ, contradicting (26) since the matching radius is fixed. □