Submitted to
manuscript

Harnessing the Double-edged Sword via Routing: Information Provision on Ride-hailing Platforms

Leon Yang Chu
Department of Data Science and Operations, Marshall School of Business, University of Southern California leonyzhu@usc.edu

Zhixi Wan
Department of Operations and Business Analytics, Lundquist College of Business, University of Oregon zwan@uoregon.edu

Dongyuan Zhan
School of Management, University College London d.zhan@ucl.ac.uk

We consider a ride-hailing platform that provides free information to taxi drivers. Upon receiving a rider's request, the platform broadcasts the rider's origin and destination to idle drivers, who accept or ignore the request depending on the profitability considerations. We show that providing such information may reduce drivers' equilibrium profit. Hence information provision is a double-edged sword: the drivers may choose to take more profitable riders via "strategic idling". When multiple drivers compete for the same request, how the platform breaks the tie affects the incentives of the drivers. We propose a routing policy that can align the incentives and achieve the first-best outcome for large systems.

Key words: Ride-hailing Platform, Queueing Game, Information and Incentive, Routing, Taxi Industry

1. Introduction

It was summer 2017, and one couldn't find a better time than a Saturday night to appreciate Shanghai at the end of a scorching hot week. The city was vibrant with flashing neon lights and crowds of people. We walked out of a shopping center, looking for a taxi to escape the heat. Luckily, there was a taxi stand right in front of the shopping center entrance. About twenty people were waiting, and we joined the line.

The line hardly moved. Looking into the streets, empty taxis passed by our stand every minute, but they did not enter our taxi stand to pick up riders. It was a bizarre scene in a surreal city: anxious, would-be riders being ignored by unoccupied taxi drivers. The crowd started to experiment with other tactics: some moved away from the stand and tried to hail one, some stayed in the line and tried to use the ride-hailing app; however, all these efforts were in vain. The theoretical analysis we had previously undertaken—and recent news articles on taxi drivers' discriminatory behavior—came to mind. "This is for real!"

As compared with its Western counterparts, Chinese taxi fares are quite affordable for the mass consumer, which leads to a massive taxi industry. According to Ministry of Transport China (2017) estimates, more than 100 million riders are served daily nationwide in 2016. In treating taxi service as a quasi-public transportation service, each city's transportation administration council tightly regulates its local taxi market with regard to fare rates, total supply of licenses, and codes of service. Regulations generally enforce metering, discourage tipping, and—in particular—do not allow taxi drivers to refuse service based on the rider's destination.

Starting in 2012, smartphone-based ride-hailing applications (apps) emerged and quickly penetrated China's taxi market. These apps involved similar hailing processes. First, a rider requests a ride by submitting her origin–destination (OD) information to the app. Next, the app finds a cohort of nearby taxi drivers based on their real-time GPS location (likely along with other algorithmic rules) and then broadcasts the rider's request to the cohort's app terminals. To gain a foothold in the industry, the ride-hailing platforms routinely share riders' OD information with the drivers,[1] who indicate their acceptance of the rider's request by pressing a button on the app. Finally, the platforms choose a responding driver to match the rider. In case multiple drivers answer, the platforms typically use a "first-come, first-served" criterion to select a driver. In the event that no driver answers, a platform will usually broadcast the request again to a different cohort of drivers; this sequence continues until either the request is answered by at least one driver or is canceled by the prospective rider. The response rate of drivers can be quite low. Sometimes drivers are unaware of the request, especially when they are en route in the service of other riders. More common is a lack of interest in the rider's request—that is, given the supplied OD information.

At this point we should distinguish between these smartphone-based taxi-hailing platforms and platforms, such as Uber, for hailing private cars. Whereas broadcasting is the dominant matchmaking mechanism in smartphone-based taxi-hailing platforms, algorithmic car-dispatching mechanisms are used by platforms for hailing private cars. Unlike Uber and Lyft, which began as businesses for hailing private cars, the ride-sharing platform companies in China (e.g., Didi, Kuaidi) started out in the taxi-hailing app business before they, jointly with Uber, fostered the private-car–hailing market in 2014. Those origins doubtless reflect the enormous size of the traditional taxi market in all major Chinese cities.

To encourage both riders and drivers to participate in these platforms (and because of heavy competition), ride-hailing apps charge no commission for their information service; in fact, the platforms sometimes subsidize drivers and riders in order to accelerate adoption. This intervention

[1] When information is *not* shared, the platforms observe high cancellation rates—when drivers learn the rider information and then abandon the job even after accepting a request.

in the taxi industry by ride-hailing platforms has been remarkably successful. Back in 2012, fewer than 10% of taxi drivers owned a smartphone; by 2017, however, more than 80% of taxi drivers in China used ride-hailing apps every day.[2] Another change during this transition is that the drivers started to behave differently, becoming more selective in the rides they accepted despite regulations to the contrary.[3] One might expect there to be both winners and losers from this change: riders with "bad" trips (from a driver's perspective) would be the losers, and riders with "good" trips would be the winners. One might also expect that drivers would benefit from the introduction of ride-hailing apps, because they have more information and more decision rights.

In this paper, we examine the effects of OD information on taxi drivers' equilibrium behavior and equilibrium profit. More specifically, we model the queueing network and analyze the "fluid" game when the system is large. We then evaluate how different policies yield incentives and alter equilibrium performance. Our analysis reveals a number of important insights into how *information provision* and the *priority rule* (i.e., routing policy) affect both equilibrium behavior and equilibrium profit.[4]

First, we find that information provision can lead to a "lose–lose" outcome in which drivers make less profit and individual riders encounter (weakly) worse service availability (Proposition 5). This inferior outcome arises because the drivers compete for more profitable riders via "strategic idling". That phenomenon is most detrimental when demand is close to the system's capacity (Proposition 6)—resembling the situation in major cities in China, where the taxi industry had managed a delicate balance between supply and demand prior to the emergence of the ride-hailing platforms.

Second, we explore the effects of information provision by establishing two benchmarks: the no-information case (Proposition 1) and the first-best case (Propositions 2) and then deriving the steady state fluid profit rate of each driver under various routing policies with information provision (Propositions 3–4; Propositions 10–12). Toward that end, we propose novel models for the analysis of queueing systems under which servers may strategically refuse to provide service; we then derive the fluid equilibrium of these systems.

[2] "CEO Chen Wei: Didi started with 16 drivers and I hope every driver can earn big money." http://tech.sina.com.cn/i/2017-12-21/doc-ifypvuqe4979206.shtml (retrieved May 20, 2018).

[3] "Short-trip customers cannot find a taxi due to the discriminatory behavior of drivers through online hail platform." http://www.xinhuanet.com/auto/2017-03/24/c_1120681481.htm (retrieved May 20, 2018).

[4] We use the terms "priority rule" and "routing policy" interchangeably for the following reason. To implement the broadcast system, the platform may either (i) broadcast the rider request to all (nearby) idle drivers simultaneously and then break any "ties" using some priority rule or (ii) broadcast the request to drivers sequentially based on the priority rule, a procedure that resembles the action when jobs are routed to idle servers. The actual implementation proceeds in stages and its characterization lies somewhere *between* broadcasting to all potential servers simultaneously and notifying them one by one.

Third, we propose remedies that align incentives and thereby improve both driver profit and service availability (Propositions 7–10). We show in particular that, without changing the tightly regulated pricing scheme, revising only the routing policy is enough to attain the first-best benchmark. This outcome is achieved by routing more profitable requests to drivers according to the "shortest idle server first" policy while routing less profitable requests according to either the "random routing" policy or the "longest idle server first" policy. Taxi drivers under that system are incentivized to minimize their idle time, which increases utilization of the resource.

In the next section, we review the literature. Section 3 stipulates our notation and characterizes the no-information and first-best benchmarks. In Section 4, we analyze how providing information affects the equilibrium behavior and equilibrium profit of taxi drivers. Section 5 proposes alternative policies that can recover the first-best benchmark, and Section 6 demonstrates the robustness of our results by generalizing the setup to allow for *offline* hailing. We conclude in Section 7 by summarizing the results and indicating directions for future research. All proofs are given in Appendix.

2. Literature Review

Our research is related to three streams of literature: information disclosure in economics; queueing games; and sharing economics and queueing networks for the ride-sharing systems.

In the economics literature, information disclosure is typically between a sender and a (representative) receiver. In the "cheap talk" model of Crawford and Sobel (1982), the sender tries to bias the receiver and only partial information can be communicated in equilibrium. Milgrom and Roberts (1986) coin the term "persuasion games" to describe game-theoretic settings with verifiable information. The challenge in our context arises because the platform serves multiple drivers who compete for the more profitable riders. A related paper, Romanyuk (2017), shows that partial disclosure problem may restore full efficiency following the Bayesian persuasion framework of Kamenica and Gentzkow (2011). Concerned about driver reneging after finding out the rider types, we study how to align incentives using alternative routing rules under full information disclosure, in a flavor similar to Su and Zenios (2004), who study the strategic behavior of the patients using an M/M/1 model.

There is a large literature, pioneered by Naor (1969), on queueing games; two frequently cited survey books are Hassin and Haviv (2003) and Hassin (2016). Recently, Hassin and Roet-Green (2017) and Yang et al. (2018) study the impact of customer search cost on the queueing performance and service design. Much of that literature assumes fixed service rates and focuses on how customers who strategically decide whether or not to queue, and where to queue, affect system performance.

Some exceptions allow at most two strategic servers to determine the equilibrium service rates; examples include Kalai et al. (1992), Gilbert and Weng (1998), Cachon and Harker (2002), Cachon and Zhang (2007), and Debo et al. (2008). A few papers of more recent vintage study large systems with strategic servers (see e.g. Gopalakrishnan et al. 2016, Ibrahim 2017, Zhan and Ward 2017); these papers focus on servers' strategic behavior, such as choices of speed and shift. In our ride-hailing context, servers can strategically turn away riders based on the available OD information.

There is also a stream of papers, initiated by George and Xia (2011), that use queueing networks to model the ride-sharing systems. George and Xia (2011) study the optimal fleet size for a vehicle rental company by treating bicycles (or automobiles) as "jobs" in the system and treating the arriving customers as servers. Using a closed queueing network that incorporates both M/M/1 and M/M/∞ queues, they show that—when customer arrivals to different locations are independent—the queueing network is a Jackson network (Jackson 1963) with explicit product-form steady state probabilities. That type of queueing network is adapted by Ozkan and Ward (2017) to address dynamic matching decisions and by Braverman et al. (2017) to show how the empty vehicles should be rebalanced. In this paper, we do not consider multiple locations but do allow taxi drivers to be strategic, which results in a similar closed queueing network. However, due to the routing policy adopted by the platform and the resulting strategic behavior of the drivers, job assignments to the drivers are not Markovian, and therefore Jackson network does not apply to our model.

A number of papers in the operations management field study so-called sharing economics; examples include Gurvich et al. (2016), Benjaafar et al. (2017), Cachon et al. (2017), Hu and Zhou (2017), Taylor (2017), and Bai et al. (2018). These works focus on the equilibrium price or price and wage controls. Hu and Zhou (2016) examine the dynamic type matching policy; Braverman et al. (2017) and He et al. (2018) investigate the optimal repositioning policy. Feng et al. (2017) compare the efficiency of platform matching with that of street hailing. Guda and Subramanian (2018) show that strategic surge pricing is an efficient tool to incentivize drivers to costly move to the high-demand location. Afeche et al. (2018) consider both admission control and repositioning decisions, and find it could be optimal to strategically reject demand at the low-demand location, to induce repositioning to the high-demand location. Because our paper's setting features regulated prices and information on profitability, we focus on drivers' equilibrium behavior and equilibrium profit while exploring the operational incentives to address the inefficiency caused by the strategic behavior of individual servers. That is, we seek an instrument *unrelated* to the price that can improve the performance and recover the first-best benchmark.

The centralized control problem in our setting is to determine—given a limited number of drivers—the optimal admission policy for the low-profit and high-profit customers. This problem

was first studied by Miller (1969), under common service rates, via a Markovian decision process. Ross and Tsang (1989) introduce the more general "stochastic knapsack" problem in the telecommunications context. Savin et al. (2005) consider a similar problem in rental businesses, using a fluid model to characterize the optimal threshold control. We adopt their fluid model and address the decentralized control problem when strategic servers may reject jobs and thus render the system less efficient.

In queueing theory, an idling (or non-work conserving) policy refers to the routing that allows servers to be idle when customers are waiting or reneging. Rubinovitch (1985) formulates the "slow server problem" and suggests that intentionally idling a slow server may actually reduce the average waiting time. Afeche (2013) considers the service line design problem and finds that it could be optimal to intentionally idle customers (i.e., to provide damaged goods/services) to discourage high-type customers from joining a low-type queue. Maglaras et al. (2017) study the service line design in a large many-server system, and they also report that intentional idling could be optimal to maximize revenue. Intentional idling in our setting arises from servers' strategic behavior and hampers system performance; we shall propose ways to limit such strategic behavior.

3. Model and Benchmarks

We consider a model with homogeneous drivers and heterogeneous riders who demand trips within a city. Some riders are more profitable for a driver, and some are less profitable, depending on several factors (e.g., the traffic congestion along the route). If a rider hails a taxi on the street, the driver has only a prior belief about the rider's profitability; this setting is no different from the platform providing no information about customer destinations. If the platform does broadcast the rider's OD information, then drivers can ascertain the ride's profitability. We, therefore, use queueing models to analyze how a ride-hailing platform affects driver behavior and the resulting equilibrium outcomes.

3.1. Model and Notation

Let n denote the number of taxi drivers in the system. Riders arrive according to a Poisson process with rate λ. Each rider is either a high-type H (i.e., more profitable) or a low-type L (less profitable). A type-H rider arrives at the rate $\alpha\lambda$ and yields the driver a profit of P_H, whereas a type-L rider arrives at the rate $(1-\alpha)\lambda$ and yields a profit of P_L. We normalize the profit to zero when drivers are not serving any customers and assume that $P_H > P_L > 0$. To focus on the effect of trip information, we assume that the service time of each rider type is an exponentially distributed random variable with rate parameter μ and ignore the taxi's location both before and

after each ride. The system load is defined as $\rho \equiv \frac{\lambda}{n\mu}$, the relative profitability of rider types is $r \equiv \frac{P_L}{P_H} \in (0,1)$, and the ex ante profitability of a rider is given as $P_E \equiv \alpha P_H + (1-\alpha)P_L$. Given the common service time distribution, service availability is calculated as the number of matched riders divided by the total number of ride requests at equilibrium. As we will see, when comparing different information provision policies, higher equilibrium driver profit corresponds in large part to higher service availability. Therefore, we focus on equilibrium driver profit in our analysis. We shall adopt nomenclature from the queueing literature and use the terms "servers" and "drivers" interchangeably. The notation is summarized in Table 1.

Table 1 Notation

n	number of taxi drivers (servers)
μ	taxi driver service rate
λ	rider arrival rate
ρ	load factor, $\rho = \lambda/(n\mu)$
α	proportion of high-type riders
P_H (P_L)	profitability of high-type (low-type) riders
r	relative profitability, $r = P_L/P_H < 1$
P_E	ex ante profitability, $P_E = \alpha P_H + (1-\alpha)P_L$
π	equilibrium profit rate of a driver

3.2. No-information Benchmark

Before the emergence of the ride-hailing platforms, a server had no prior information about the profitability of riders and so could not discriminate against them on that basis. In other words, drivers maximized their profit by serving all encountered riders. Because we ignore the locations of the taxis in the model, an arriving rider would be randomly matched by one of the idle drivers. It turns out that in this no-information setting, we can use an M/M/n/n blocking queue to analyze the system. We could also use an equivalent closed queueing network in which drivers are the n jobs circulating within the system; see Figure 1.

We show that fluid approximation gives limit equilibrium server behavior in the M/M/n/n queue as the system becomes large. Consider a sequence of systems with an increasing number n of servers and increasing arrival rates $\lambda^n = \rho n\mu$ (while leaving unchanged the parameters ρ, μ, P_H, and P_L). Let π^{no} denote the profit rate of each server—in the no-information case—at the equilibrium of this fluid game. We can then use Whitt (2006) to establish the following proposition.

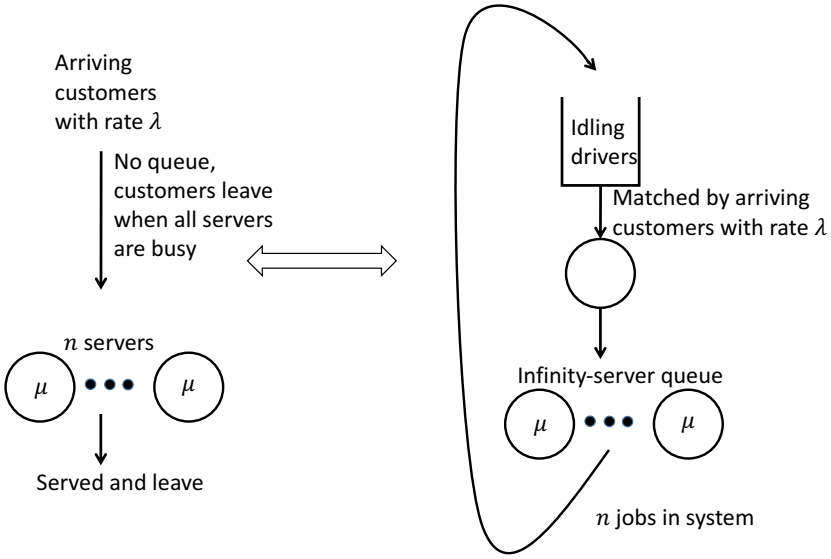


Figure 1 Equivalence between M/M/n/n Queue and Closed Queueing Network

PROPOSITION 1 **(No-information Benchmark)**. *Under any non-idling policy, the utilization rate of the servers converges to* $\min\{\rho, 1\}$ *as* $n \to \infty$. *The steady state fluid profit rate of each driver in the system is therefore*

$$\pi^{\text{no}}(\rho) = \begin{cases} \rho P_E \mu, & \text{if } \rho \leq 1, \\ P_E \mu, & \text{if } \rho > 1. \end{cases}$$

Proposition 1 characterizes the expected profit in the fluid limit, which serves as a benchmark profit when we consider the platform's effect. The *no-information* benchmark has two linear segments: the profit rate increases linearly from the origin when the load factor is less than 1, and is then capped when the load factor is greater than 1.

3.3. The First-best Admission Control

In a centralized control system, the platform routes customers to the servers, who cannot decline incoming requests. The platform's key task is to determine—as a function of profitability and the current system load—whether or not to admit an arriving ride request. Miller (1969) was the first to propose this problem and established that the optimal policy is (i) to admit all high-type riders but (ii) to admit low-type riders only when the system's total number of customers does not exceed some threshold.

Calculating the threshold is quite involved by using policy iteration of a Markovian decision process. To address this difficulty, Savin et al. (2005) propose a fluid approximation for finding the

proper threshold of a large system (high n) while maintaining a constant load factor ρ. Our next proposition holds by Theorem 5 in Savin et al. (2005).

PROPOSITION 2 **(First-best Benchmark)**. *Under the optimal control, the steady state fluid profit rate of each driver in the system is*

$$\pi^\star(\rho) = \begin{cases} \rho P_E \mu & \text{if } \rho \le 1, \\ (\rho\alpha P_H + (1-\rho\alpha)P_L)\,\mu & \text{if } 1 < \rho \le \frac{1}{\alpha}, \\ P_H \mu & \text{if } \rho > \frac{1}{\alpha}. \end{cases}$$

The *first-best* benchmark has three linear segments: the profit rate increases linearly from the origin when the load factor is less than 1; it then increases (but at a lower rate) when the load factor is greater than 1 and the system declines more and more low-type requests; and finally it is capped when the servers are fully utilized for high-type requests. Under the optimal fluid control, essentially all high-type riders are served, and all the remaining capacity is used to serve low-type riders. In contrast to Proposition 1, if $\rho > 1$ then the platform can obtain higher server profit via the admission control enabled by the profitability information.

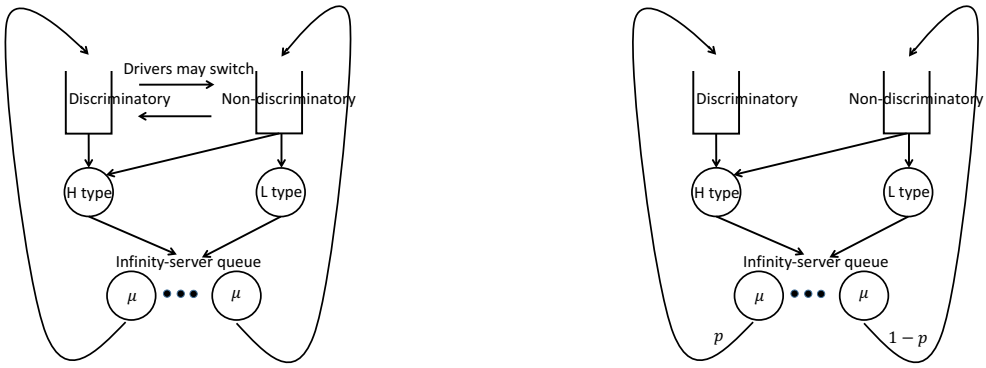
4. The Equilibrium Analysis under Full-Information

Although admission control and the resulting first-best outcome is clearly desirable, taxi drivers may choose to ignore incoming requests from the ride-hailing platform, which broadcasts a rider's request to all (or some) available drivers. Next, we show that the information on rider profitability leads to discriminatory behavior by taxi drivers.

In this section we assume that all riders book through the ride-hailing platform (this restriction is relaxed in Section 6); we then derive the servers' equilibrium behavior when they know the riders' profitability. Finally, we compare the equilibrium profit rate with both the no-information benchmark and the first-best benchmark. The platform must decide how to broadcast the ride request, as well as the associated priority rule (i.e., tie-breaking rule), should multiple drivers respond to the request. According to the broadcast policy, each server chooses between the discriminatory strategy that ignores the low-type requests and the non-discriminatory strategy that accepts all the requests. Given that the leading platform broadcasts ride information to nearby available drivers, of which the first to respond serves the rider, we start our analysis with the random routing assumption: all available drivers who are willing to accept the request have the same likelihood of being matched to that rider. Later, we show that our results are unaffected if instead the "longest idle server first" (LISF) rule is adopted. In line with the models from the related literature (Savin et al. 2005, George and Xia 2011, Braverman et al. 2017, Ozkan and Ward 2017), we assume that arriving riders who are not matched to an available driver use alternative transportation and therefore leave the system.

4.1. Full-information Equilibrium

Idle servers with OD information can choose between two strategies: discriminatory and non-discriminatory. All idle servers compete for high-type requests, whereas non-discriminatory servers respond also to low-type requests. Observe that servers need not declare their types and can freely switch, at any idle time, between the two strategies. These considerations lead to a closed queueing network that consists of two G/G/1 queues and one G/M/∞ queue, as shown in Figure 2(a). Whether and when a server switches his strategy depends on details of the priority rule/routing rule.



(a) Drivers switch (b) Drivers do not switch

Figure 2 Closed Queueing Network in Which Drivers May Be Discriminatory

If switching during waiting for customers does not occur, then, after each service, drivers must choose their respective discriminatory probabilities p; see Figure 2(b), where the n drivers correspond to jobs in the closed network. Here the G/M/∞ queue represents the busy drivers, with service rate μ, and the two G/G/1 queues correspond to the idle drivers who are (respectively) discriminatory and non-discriminatory. Note that this is not a generalized Jackson network because the customer assignments to the drivers is not Markovian; hence, the steady state does not have explicit product-form probabilities. It is difficult to characterize the equilibrium, so we resort to fluid approximation. There are n units of fluid in the system. Let $I_D(t), I_N(t)$ denote respectively the amount of idle fluid that is discriminatory and the amount of non-discriminatory fluid. The idle fluid is matched by arrival customers with rate λ according to the routing rule. The amount of fluid which is occupied by customers is $n - I_D(t) - I_N(t)$, which becomes idle with rate $\mu(n - I_D(t) - I_N(t))$. We first show there exists a symmetric mixed equilibrium with $p \in [0, 1]$ and derive the server's fluid equilibrium profit rate.

PROPOSITION 3 **(Fluid Equilibrium under Random Routing)**. *Under full information and a random routing rule, the steady state fluid profit rate of each driver in the system is as follows:*

$$\pi^{\text{full}}(\rho) = \begin{cases} \rho P_E \mu & \text{if } \rho < \frac{P_L}{P_E}, \\ P_L \mu & \text{if } \frac{P_L}{P_E} \le \rho \le \frac{r}{\alpha}, \\ \rho \alpha P_H \mu & \text{if } \frac{r}{\alpha} < \rho \le \frac{1}{\alpha}, \\ P_H \mu & \text{if } \rho > \frac{1}{\alpha}. \end{cases}$$

The *full-information* equilibrium profit rate has four linear segments: the profit rate increases linearly when it is less than $P_L\mu$; it then plateaus at $P_L\mu$; it again increases linearly when all servers compete only for high-type requests; and last it is capped when the servers are fully utilized. In the first and second segments, both discriminatory and non-discriminatory drivers co-exist, whereas the system only has discriminatory drivers in the third and fourth segments. In the first segment, strategic idling occurs but the system still has enough capacity to serve all the riders. In the second segment, strategic idling occurs and some low-type riders are lost. Under this case, a non-discriminatory driver has zero waiting time, which results in a constant equilibrium profit rate of $P_L\mu$. In the third segment, the drivers are better off by focusing on the high-type riders exclusively and all low-type riders are lost. In the fourth segment, the high-type riders alone is sufficient to keep the drivers fully utilized.

It turns out that, even if we change the routing rule from random routing to LISF (which is implemented at popular origins, such as airports), the servers' fluid equilibrium profit rate as well as the bounds of the four cases remain the same, despite the servers being more inclined to wait for high-type riders as the increasing duration of their idleness increases their priority to obtain a high-type rider. Formally, we show there exists a symmetric mixed equilibrium with $p \in [0,1]$ and derive the server's fluid equilibrium profit rate.

PROPOSITION 4 **(Fluid Equilibrium under LISF)**. *When there is full information, the steady state fluid profit rate of each driver in the system under LISF is equal to the corresponding profit rate under a random routing rule.*

REMARK 1. While Propositions 3 and 4 are derived assuming that each driver chooses to be discriminatory with some probability after each service, these propositions and proof techniques still hold if each driver chooses to be discriminatory with some probability for the entire life.

REMARK 2. There exists other symmetric and asymmetric equilibria under both random routing and LISF. For example, under random routing, when the equilibrium discriminatory probability p is in between 0 and 1, after each service, the driver is indifferent between being discriminatory or not while waiting for a customer. Therefore, each idle server may switch between being discriminatory and not over time and one can create alternative equilibria as long as the overall expected number

of discriminatory servers stays the same. Similarly, under LISF, if the non-discriminatory driver is expected to wait for a positive time period W_N, there exist alternative equilibria under which each idle server switch between being discriminatory and not when waiting time is less than W_N. We have focused on the symmetric mixed equilibrium under the closed network illustrated by Figure 2(b) because it enables us to establish the existence result of the fluid equilibrium.

REMARK 3. For other potential equilibria under both random routing and LISF, Propositions 3 and 4 continue to characterize the correct fluid equilibrium profit rate for the servers. In Appendix, we establish this statement rigorously by considering three possible scenarios according to which the equilibrium profit rate is either larger than, equal to, or smaller than $P_L\mu$. These three scenarios correspond to the cases when, respectively, all low-type riders are forgone (and all drivers are always discriminatory), some low-type riders are forgone (and the non-discriminatory drivers are fully utilized), or all riders are served regardless of profitability.

4.2. The Impact of Information Provision

One feature of the full-information equilibrium is that strategic idling may arise, accompanied by some potential riders not being served. In other words, the servers may wait for high-type riders while low-type riders are forgone. Such strategic idling does not occur in the no-information equilibrium, under which either each server is fully utilized or all riders are served.

We can analyze how information provision affects servers' equilibrium profit rate by comparing its value to the no-information benchmark. Proposition 5 summarizes the result.

PROPOSITION 5 **(Fluid Comparison)**. *Under fluid approximation, as compared with the no-information benchmark, servers in the full-information equilibrium earn:*

- *Case 1—the same profit (i.e.,* $\pi^{\text{full}}(\rho) = \pi^{\text{no}}(\rho)$*) when* $\rho \le P_L/P_E$;
- *Case 2—lower profit (i.e.,* $\pi^{\text{full}}(\rho) < \pi^{\text{no}}(\rho)$*) when* $P_L/P_E < \rho < P_E/(\alpha P_H)$;
- *Case 3—higher profit (i.e.,* $\pi^{\text{full}}(\rho) \ge \pi^{\text{no}}(\rho)$*) when* $\rho \ge P_E/(\alpha P_H)$.

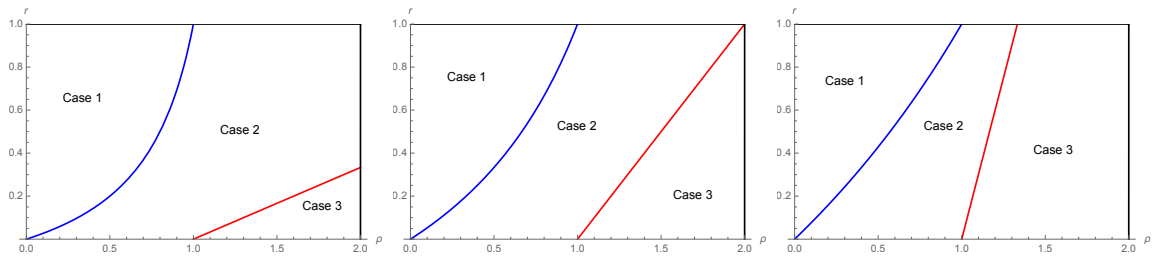
When the system load is small (i.e., Case 1), the drivers are not the system "bottlenecks" and they serve any rider notwithstanding their knowledge (if any) about rider profitability. Although having information on profitability results in a greater proportion of drivers remaining idle so they can to serve high-type riders, that idleness does not have a negative effect on the experience of low-type riders. In equilibrium, all drivers share an equal split of the maximum profit from the entire rider group—that is, the equilibrium profit rate equals the no-information benchmark.

When the system load is intermediate (Case 2), surprisingly, providing the information on profitability to the servers can actually *reduce* their profit. To see this, consider two separate regions:

in one region, ρ is between P_L/P_E and 1; in the other region, ρ is between 1 and $P_E/(\alpha P_H)$. In the first region, the drivers serve all demand under the no-information benchmark and thus earn the first-best profit rate from the riders. Yet if profitability information is provided, not enough drivers are willing to serve low-type riders despite zero waiting time. This dynamic reduces driver utilization and results in unserved riders. Hence the servers, as a group, fail to earn the maximum profit possible under full information. Note that, in this region, high-type riders receive the same service availability as before but both the low-type riders and the drivers are worse-off owing to strategic idling.

In the second region, where $\rho > 1$, not all riders can be served in the absence of information; it, therefore, makes sense for drivers, as a group, to cherry-pick the high-type riders. Thus, there is a competing factor that *favors* information provision. Despite this favorability, it turns out the servers still lose profits in the full-information equilibrium if some of them are non-discriminatory (and the servers may continue to earn a lower profit when all servers are discriminatory).

It is only in Case 3, when the system load is high enough that the profit rate from high-type riders alone exceeds the maximum profit rate under no information, that profitability information enables drivers to achieve a higher profit. In this case, all drivers are discriminatory and all low-type riders are excluded from the market.



(a) $\alpha = 1/4$ (b) $\alpha = 1/2$ (c) $\alpha = 3/4$

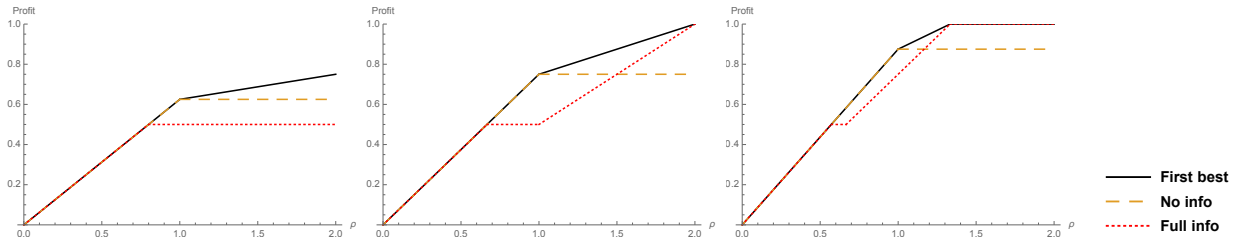
Figure 3 Case Boundaries by Proportion α of High-type Riders

Figure 3 illustrates the boundaries of our three cases for different values of α (1/4, 1/2, 3/4). In each graph, the x-axis is the load factor ρ and the y-axis is the relative profitability r of low-type riders. As we compare the equilibrium profit rate under full information to the no-information benchmark, the outcome moves from Case 1 (equal) to Case 2 (lower) to Case 3 (higher) when the load factor ρ increases and/or the relative profitability r decreases. That is: in the no-information case, drivers are better off when both the load factor and the relative profitability are at intermediate levels; under full information, they are better-off at the when the load factor is high and the relative profitability is low. We also observe that, as α increases (i.e., an increased proportion of

high-type riders), the area associated with Case 2 shrinks and the full-information equilibrium's relative performance improves.

4.3. Further Illustration and Performance Bound

Figure 4 illustrates how the servers' equilibrium profit rate varies with ρ for different α values (1/4, 1/2, 3/4) when $r = 1/2$, $P_H = 1$, $\mu = 1$. The dashed (resp. dotted) line plots the no-information (resp. full-information) case, and the solid line plots the first-best benchmark.



(a) $\alpha = 1/4$ (b) $\alpha = 1/2$ (c) $\alpha = 3/4$

Figure 4 Profit Comparison

Recall that the *no-information* benchmark has two linear segments and the *first-best* benchmark has three linear segments, whereas the *full-information* equilibrium fluid profit rate of each server has four linear segments. Moreover, the third segment of the first-best benchmark and the fourth segment of the full-information equilibrium correspond to the case that the drivers are fully utilized for high-type requests. When α is small, ρ must be large in order to reach this region. Hence that region appears in part (c) of Figure 4 but not in part (a) or part (b).

Figure 4 also demonstrates that drivers may experience a significant loss of profit at the full-information equilibrium as compared with the no-information benchmark. Specifically, we may consider a simple setting under which $\alpha = 1/2$, $\rho = 1$, and $r = 1/2$. When no information is provided, all riders are served and the drivers' expected profit rate is $(P_H + P_L)\mu/2$. At the full-information equilibrium, the drivers serve only high-type riders ($\beta = 1$); here the drivers' utilization rate is 50% and their expected profit rate is $P_H\mu/2$. Since $P_L/P_H = r = 1/2$ in this example, it follows that a third of the potential profit is lost. In general, the magnitude of this loss (compared with the no-information benchmark) is maximized at $\rho = 1$ as formalized by the following proposition.

PROPOSITION 6 **(Profit Ratio)**. *The ratio of the full-information equilibrium profit rate to the no-information benchmark, $\pi^{\text{full}}(\rho)/\pi^{\text{no}}(\rho)$, is minimized when $\rho = 1$; in particular, the minimum value is $\frac{\max\{\alpha,r\}}{\alpha+r-\alpha r}$.*

Proposition 6 shows that the detrimental effect of information provision is most prominent when the utilization rate is intermediary and close to 1, a region that regulators and taxi firms seek to occupy in order to maximize social welfare and profit. Working in this region requires a delicate balance between supply and demand, since one can expect that the strategic idling induced by information provision will severely compromise system efficiency.

Observe that the no-information case achieves the first-best benchmark when $\rho \leq 1$ and that the full-information case achieves the first-best benchmark when $\rho \geq 1/\alpha$. However, neither the no-information nor the full-information case achieves the first-best benchmark at intermediate levels of ρ. In the next section, we propose some remedies to improve both driver profits and service availability.

5. Engineering the Equilibrium via Additional Instruments

In this section, we discuss how it may be possible to achieve the first-best benchmark without subsidies from the platform. Our first two proposals involve a pricing adjustment and limiting the dispersion of information; although both methods can recover the first-best benchmark, each is severely constrained by regulations. Finally, we show that the first-best outcome can be achieved—without overhauling the system—by adopting an alternative priority rule (i.e., routing policy).

5.1. Alternative Schemes for Pricing and Information Provision

5.1.1. Alternative Pricing Scheme A natural way of improving resource allocation is to introduce a price instrument. For example, the ride-hailing platform could charge a service fee $s \in (0, P_H - P_L)$ for each successful high-type matching. Such a service fee would reduce the payoff for serving high-type riders from P_H to $P_H - s$ and thereby discourage drivers from self-induced idling in their pursuit of high-profit customers.

Note that the full-information equilibrium corresponds to the limit case $s \downarrow 0$ and the first-best benchmark can be induced by setting $s \uparrow (P_H - P_L)$. One important observation is that a service fee $s \in (0, P_H - P_L)$ monotonically improves resource allocation from a social welfare perspective. Proposition 7 formalizes this aspect and identifies the region in which an increased service fee may lead to a win–win outcome for riders and the platform without reducing driver profits.

PROPOSITION 7 **(Fee Monotonicity)**. *The proportion of unserved low-type riders, the proportion of unserved high-type riders, and the equilibrium profit rate of servers are all (weakly) decreasing functions of s, while the platform's profit increases with s. Moreover, if $\rho \in \left[P_L/P_E, P_L/(\alpha P_H)\right]$ then, in equilibrium, the servers' profit rate is constant (at $P_L\mu$) when $s \leq (P_E - P_L/\rho)/\alpha$.*

So if $s \uparrow (P_H - P_L)$ then the platform would earn the maximum possible profit while the drivers would earn the minimum profit—that is, as if all riders were of the low-type. This outcome follows the classic economic insight that, if all services are priced such that servers are indifferent between customer types and thus avoid idleness, then the optimal allocation can arise as an equilibrium. In practice, effecting such "perfect" pricing requires the platform to have precise knowledge of profitability as a function of time and of each rider's origin and destination. Another challenge for the platform is to convince drivers that the price scheme has, in fact, been set up fairly. Charging a service fee is also likely to induce backlash from drivers and regulators both, especially when withholding OD information might have achieved the first-best outcome (i.e., when $\rho \leq 1$).

5.1.2. Limited Information Provision A ride-hailing platform can limit the information provision either by limiting the information's content or by limiting the proportion of servers who receive the information. The former case is studied by Romanyuk (2017), who shows that a (randomized) partial information provision policy can achieve the first-best benchmark when the drivers cannot renege. In the remainder of this subsection, we focus on the latter case.

Suppose that the platform discloses full information on ride types and gives choice priority to some proportion $\kappa \leq 1$ of the servers (i.e., the platform routes high-type riders to those servers first). Given Proposition 4's analysis of the LISF rule, Proposition 8 establishes that the platform can recover the first-best benchmark by limiting the number of informed servers because doing so, in turn, limits the proportion of discriminatory servers.

PROPOSITION 8 **(Limited Information Achieves First-best Outcome).** *The first-best outcome is achieved:*

- *by any* $\kappa \in [0,1]$ *when* $\rho \leq P_L/P_E$;
- *by any* $\kappa \in [0, 1-\rho(1-\alpha)]$ *when* $P_L/P_E < \rho \leq 1$;
- *by* $\kappa = \min\{\rho\alpha, 1\}$ *when* $\rho > 1$.

Observe that $\kappa = \min\{\rho\alpha, 1\}$ always induces the first-best benchmark for any value of ρ. For such κ, the platform essentially adjusts the proportion of the informed servers to match the demand of high-type customers. Of course, the platform cannot implement this scheme without a precise estimate of the high-type demand. Moreover, the load factor ρ and the high-type demand $\alpha\rho$ may vary depending on the time of the day and special events. Furthermore, informed servers may earn a higher profit than uninformed servers and so the ride-hailing platform would, in essence, be picking winners and losers; that outcome clearly raises fairness concerns, from which associated regulation and monitoring hurdles would follow.

5.2. Alternative Priority Rule

The platform must decide how to broadcast ride requests and the associated priority rule. While providing all drivers with equal access to the OD information, can the platform use non-pricing instruments to discourage strategic idling and thus improve system efficiency? Here we consider a simple priority rule under full information disclosure that achieves the first-best benchmark.[5]

Recall from our previous discussions that both random routing and LISF rule lead to reduced server equilibrium profit rate because of strategic idling. Now, toward the end of discouraging servers' discriminatory behavior, we consider the *shortest* idle server first (SISF) routing rule for high-type riders but use either the random routing rule or LISF routing rule for low-type riders.[6] Under such a routing rule, the longer a server is idle, the less incentive for him to be discriminatory because his priority for high-type requests decreases over time. Therefore, after each service, a server may be discriminatory for T time units but switch to be non-discriminatory if he has not been matched with a high-type rider after T time units. This scenario is illustrated in Figure 5.

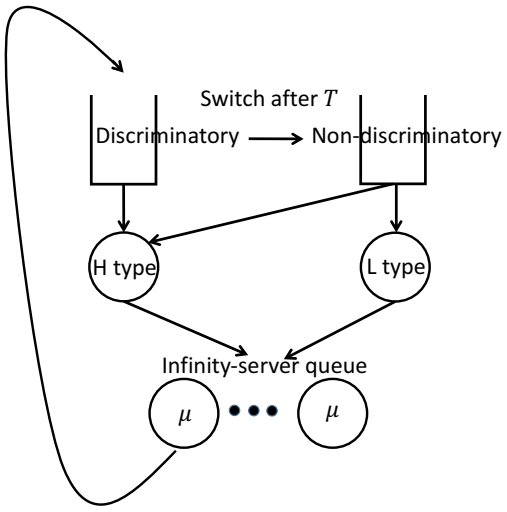


Figure 5 Closed Queueing Network where Drivers become Non-discriminatory after T

[5] Our assessment of the various Chinese ride-hailing platforms revealed several different priority rules (i.e., routing rules). One platform, for example, decided against "naive" adoption of LISF routing *at* the airport; instead, a driver joins the queue as soon as he picks up a rider *destined for* the airport.

[6] A careful reader may wonder why we do not use SISF routing rule for all the riders. This is because such a routing rule may create income disparity among the drivers. Specifically, when $\rho < 1$, some drivers would be idle all the time and obtain a steady state fluid profit rate of zero.

Observe that it is less efficient (than in the first-best benchmark) to let every server be discriminatory for some positive T time units. It would be ideal if all high-type customers were served and then the remaining capacity were all used to serve low-type customers—provided $\rho < 1/\alpha$ (i.e., total capacity exceeds high-type demand). In the fluid limit under our proposed routing, it follows that, if a newly idled driver is not initially matched with a high-type rider, he would not be so matched simply by idling longer and hence would immediately switch to being non-discriminatory. Therefore, the fluid limit achieves the ideal scenario: all high-type riders are served, and the remaining system capacity is all used to serve low-type riders. We formally verify that the first-best admission control identified by Theorem 5 of Savin et al. (2005) forms an equilibrium under the proposed routing rule (i.e., each individual server cannot obtain a higher fluid profit rate by deviating from the assignment of the first-best admission control, whether it is to accept a rider or reject a rider, when the rider requests are routed to individual servers according to the proposed routing rule) and all equilibria under the proposed routing rule achieve the first-best benchmark.

PROPOSITION 9 **(Fluid Equilibrium under SISF)**. *Suppose the SISF (resp. random or LISF) routing rule is used to match drivers with high-type (resp. low-type) riders. Under full information, the steady state fluid profit rate of each driver in the system achieves the first-best benchmark:*

$$\pi^{\mathrm{SISF}}(\rho) = \begin{cases} \rho P_E \mu & \text{if } \rho \le 1, \\ (\rho\alpha P_H + (1-\rho\alpha)P_L)\,\mu & \text{if } 1 < \rho \le \frac{1}{\alpha}, \\ P_H \mu & \text{if } \rho > \frac{1}{\alpha}. \end{cases}$$

Proposition 9 shows that under the proposed routing rule, inefficient strategic idling is negligible for a large system and the proof of Proposition 9 implies that drivers' switching time T is zero in the fluid equilibrium. This does not mean, however, that all idle servers are non-discriminatory and accept all incoming requests; if that were the case, we would have achieved only the no-information benchmark rather than the first-best benchmark. We highlight this difference by studying how fast drivers' equilibrium switching time T approaches zero as n goes to infinity via a limit diffusional process analysis assuming the time-based threshold equilibrium structure.

PROPOSITION 10 **(Order of Equilibrium Waiting Time)**. *Under the proposed routing rule summarized in Proposition 9, if $\rho < 1$, the drivers are discriminatory at first and wait for a positive amount $O(1/n)$ of time before starting to serve any low-type riders; if $1 \le \rho < 1/\alpha$, the drivers are discriminatory at first and wait for a positive amount $O(\log(n)/n)$ of time before starting to serve any low-type riders; and if $\rho \ge 1/\alpha$, the drivers take only high-type riders.*

By encouraging utilization, the proposed routing rule achieves the first-best benchmark and, for all potential load factors ρ, improves on both the no-information benchmark and the full-information equilibrium. While traditional market interactions are often constrained by fairness

concerns and the resulting queues are often regulated by rules like longest idle server first to ensure (ex post) fairness, platforms can leverage the digital technologies and create virtual queues that adopt more flexible routing rules that achieve (ex ante) fairness among the servers. The concept of (ex ante) fairness allows the regulators to enlarge the feasible policy space, while the (ex post) fairness concern is limited if the platforms interact with the service providers repeatedly. As such, the ride-hailing platform can discourage strategic idling and serve the society better.

6. Further Discussion: Equilibrium with Offline Hailing

So far, we have analyzed equilibrium performance while assuming that all riders submit their requests through the platform. Here, we suppose that some percentage γ of riders, independent of their types, hail taxis on the street instead of via the ride-hailing platform. Taxi regulations prohibit drivers who pick up such customers from refusing service in response to the rider's stated destination. From the perspective of drivers, then, there are three types of riders: high-types (H) and low-types (L), who can be distinguished based on the platform's supplied OD information; and riders of undisclosed type (M).

We first derive the equilibrium driver profit in the fluid system when the platform adopts the LISF routing rule for all rider types.

PROPOSITION 11 **(Fluid Equilibrium with Offline Hailing)**. *With offline hailing, the steady state fluid profit rate of each driver in the system is,*

$$\pi(\rho) = \begin{cases} \rho P_E \mu & \text{if } \rho \le \frac{P_L}{P_E}, \\ P_L \mu & \text{if } \frac{P_L}{P_E} < \rho \le \frac{P_L}{\alpha(1-\gamma)P_H + \gamma P_E}, \\ \rho(\alpha(1-\gamma)P_H + \gamma P_E)\mu & \text{if } \frac{P_L}{\alpha(1-\gamma)P_H + \gamma P_E} < \rho \le \frac{P_E}{\alpha(1-\gamma)P_H + \gamma P_E}, \\ P_E \mu & \text{if } \frac{P_E}{\alpha(1-\gamma)P_H + \gamma P_E} < \rho \le \frac{P_E}{\alpha(1-\gamma)P_H}, \\ \rho\alpha P_H \mu & \text{if } \frac{P_E}{\alpha(1-\gamma)P_H} < \rho \le \frac{1}{\alpha(1-\gamma)}, \\ P_H \mu & \text{if } \rho > \frac{1}{\alpha(1-\gamma)}. \end{cases}$$

In these six cases, if the load ρ is small then all riders are served. As the load ρ increases, the drivers become increasingly discriminatory and become more likely to discriminate by waiting for street hails or H-type customers; they begin declining service to low-type riders and, for sufficiently large ρ, eventually stop serving street hails as well. Although we derive Proposition 11 under the LISF routing rule, it is readily verified—by the arguments used when analyzing Propositions 3 and 4—that the equilibrium structure and the equilibrium profit rate are the same under random routing.

Next, we establish drivers' profit in the fluid system when the platform adopts the shortest idle server first (SISF) routing rule for high-type riders but uses either the random routing or the LISF routing rule for low-type riders.

PROPOSITION 12 **(Fluid Equilibrium with Offline Hailing under SISF)**. *Under the proposed routing rule summarized in Proposition 9, the steady state fluid profit rate of each driver in the system is*

$$\pi(\rho)=\begin{cases}\rho P_E\mu & \text{if } \rho\le\frac{P_L}{\alpha\gamma P_H+(1-\alpha\gamma)P_L},\\ (\rho\alpha(1-\gamma)P_H+(1-\rho\alpha(1-\gamma))P_L)\mu & \text{if } \frac{P_L}{\alpha\gamma P_H+(1-\alpha\gamma)P_L}<\rho\le\frac{P_L}{\alpha\gamma P_H+(\alpha+\gamma-2\alpha\gamma)P_L},\\ \rho(\alpha(1-\gamma)P_H+\gamma P_E)\mu & \text{if } \frac{P_L}{\alpha\gamma P_H+(\alpha+\gamma-2\alpha\gamma)P_L}<\rho\le\frac{1}{\gamma+\alpha(1-\gamma)},\\ (\rho\alpha(1-\gamma)P_H+(1-\rho\alpha(1-\gamma))P_E)\mu & \text{if } \frac{1}{\gamma+\alpha(1-\gamma)}<\rho\le\frac{1}{\alpha(1-\gamma)},\\ P_H\mu & \text{if } \rho>\frac{1}{\alpha(1-\gamma)}.\end{cases}$$

Note that the SISF routing rule summarized in Proposition 9 increases system efficiency but cannot achieve the first-best outcome because drivers may strategically reject low-type riders and wait for customers of undisclosed type (i.e., offline-hailing customers), over whom the platform has no control. It is worth noting also that, when either the random or the LISF routing rule is implemented, equilibrium driver profit and service availability may both decline with an increased number of riders seeking taxis via the platform (i.e., as γ decreases); yet when the SISF routing rule summarized in Proposition 9 is implemented, equilibrium driver profit and service availability are both increasing in the number of such online riders.

COROLLARY 1 **(Offline-Hailing Improvement)**. *In the fluid setting with offline hailing, the SISF routing rule summarized in Proposition 9 yields a Pareto improvement of the equilibrium driver profit and of the equilibrium service availability. Furthermore, both the equilibrium driver profit and service availability decrease (weakly) under the proposed SISF routing as γ increases.*

7. Conclusion

The context for this paper is ride-hailing apps, and we study the effects of information provision and illustrate the importance of routing rule. We show that providing taxi drivers both rider information and decision rights can reduce their equilibrium profits because of competition among drivers and the consequent strategic idling.

In seeking to improve on this sub-optimal outcome, we study an alternative pricing instrument and an alternative level of information provision. We also demonstrate that an alternative routing rule can better align incentives and enhance social welfare. Namely, routing profitable riders to drivers who have shorter idle times should increase server utilization and reduce strategic idling.

The analysis presented here abstracts from location considerations. Yet, our model could easily be extended to a setting with multiple locations, and our derived insights regarding information as a double-edged sword would be unaffected. Future research could profitably examine routing rules for queueing games, especially with respect to online platforms; this is a very promising area of study because such platforms can adopt various routing rules through virtual queues.

References

Afeche, P. 2013. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing Service Oper. Management* **15**(3) 423–443.

Afeche, P., Z. Liu, C. Maglaras. 2018. Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. Working Paper.

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

Benjaafar, S., G. Kong, X. Li, C. Courcoubetis. 2017. Peer-to-peer product sharing: Implications for ownership, usage, and social welfare in the sharing economy. To appear in *Management Sci.*

Braverman, A., J.G. Dai, X. Liu, L. Yin. 2017. Empty-car routing in ridesharing systems. Working Paper.

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

Cachon, G.P., P.T Harker. 2002. Competition and outsourcing with scale economies. *Management Sci.* **48**(10) 1314–1333.

Cachon, G.P., F. Zhang. 2007. Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Sci.* **53**(3) 408–420.

Crawford, Sobel. 1982. Strategic information transmission. *Econometrica* **50**(6) 1431–1451.

Debo, L.G, L.B. Toktay, L.N. Van Wassenhove. 2008. Queuing for expert services. *Management Sci.* **54**(8) 1497–1512.

Feng, G., G. Kong, Z. Wang. 2017. We are on the way: Analysis of on-demand ride-hailing systems. Working Paper.

George, D.K., C.H. Xia. 2011. Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research* **211**(1) 198–207.

Gilbert, S.M., Z.K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Sci.* **44**(12) 1662–1669.

Gopalakrishnan, R., S. Doroudi, A.R. Ward, A. Wierman. 2016. Routing and staffing when servers are strategic. *Oper. Res.* **64**(4) 1033–1050.

Guda, H., U. Subramanian. 2018. Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication and worker incentives. To appear in *Management Sci.*

Gurvich, I., M. Lariviere, A. Moreno-Garcia. 2016. Operations in the on-demand economy: Staffing services with self-scheduling capacity. Working Paper.

Hassin, R. 2016. *Rational Queueing*. CRC Press, Boca Raton, FL.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.

Hassin, R., R. Roet-Green. 2017. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Oper. Res.* **65**(3) 804–820.

He, L., Z. Hu, M. Zhang. 2018. Robust repositioning for vehicle sharing. To appear in *Manufacturing Service Oper. Management*.

Hu, M., Y. Zhou. 2016. Dynamic type matching. Working Paper.

Hu, M., Y. Zhou. 2017. Price, wage and fixed commission in on-demand matching. Working Paper.

Ibrahim, R. 2017. Managing queueing systems where capacity is random and customers are impatient. Forthcoming in *Prod. and Oper. Management*.

Jackson, J.R. 1963. Jobshop-like queueing systems. *Management Sci.* **10**(1) 131–142.

Kalai, E., M. I. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Sci.* **38**(8) 1154–1163.

Kamenica, E., M. Gentzkow. 2011. Bayesian persuasion. *Amer. Econom. Rev.* **101**(6) 2590–2615.

Maglaras, C., J. Yao, A. Zeevi. 2017. Optimal price and delay differentiation in large-scale queueing systems. Forthcoming in *Management Sci.*

Milgrom, P., J. Roberts. 1986. Relying on the information of interested parties. *RAND J. Econom.* **17**(1) 18–32.

Miller, B. 1969. A queueing reward system with several customer classes. *Management Sci.* **16**(3) 234–245.

Ministry of Transport China. 2017. National report on urban passenger transport development. China Communications Press.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Ozkan, E., A.R. Ward. 2017. Dynamic matching for real-time ridesharing. Working Paper.

Romanyuk, G. 2017. Ignorance is strength: Improving the performance of matching markets by limiting information. Working Paper.

Ross, K.W., D.H.K. Tsang. 1989. The stochastic knapsack problem. IEEE *Transaction on Communications* **37**(7) 740–747.

Rubinovitch, M. 1985. The slow server problem. *Journal of Applied Probab.* **22**(1) 205–213.

Savin, S., M.A. Cohen, N. Gans, Z. Katalan. 2005. Capacity management in rental businesses with two customer bases. *Oper. Res.* **53**(4) 617–631.

Su, X., S. Zenios. 2004. Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing Service Oper. Management* **6**(4) 280–301.

Taylor, T. 2017. On-demand service platforms. To appear in *Manufacturing Service Oper. Management*.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Oper. Res.* **54**(1) 37–54.

Yang, L., L. Debo, V. Gupta. 2018. Search among queues under quality differentiation. To appear in *Management Sci.*

Zhan, D., A.R. Ward. 2017. Incentive based service system design: Staffing and compensation to trade off speed and quality. Working Paper.

Appendix. Proofs

Proof of Proposition 1 Each server can finish μ services per unit time and the expected profit of each service is P_E. When servers are idle, the profit rate is 0. From Whitt (2006), the utilization rate converges to $\min\{\rho,1\}$. Therefore, the expected profit rate is $\min\{\rho,1\}P_E\mu$. $\square$

Proof of Proposition 2 For our setting, the first-best fluid control result (Theorem 5) in Savin et al. (2005) can be reduced to: *In the fluid centralized control problem, the optimal policy is to admit all high-type riders, and the admission threshold for the low-type riders $K(n)$ is non-increasing in ρ:*

$$K(n)=\begin{cases}0 & \text{for } \rho>\frac{1}{\alpha(1-r)},\\ n\left(\rho\alpha-\frac{\rho\alpha-1}{r}\right) & \text{for } \frac{1}{\alpha}\le\rho\le\frac{1}{\alpha(1-r)},\\ n^- & \text{for } 1\le\rho<\frac{1}{\alpha}\\ n & \text{for } \rho<1\end{cases},$$

where threshold n^- means when the busy driver number is n, the policy limits but does not eliminate the low-type riders into the system.

In steady state, when $\rho\alpha\ge 1$, the high-type customers occupy all servers, and the threshold $K(n)$ does not play a role. The fluid profit is $P_H\mu$. If $\rho<1$, then $K(n)=n$. The system admits all the customers and the fluid profit is $P_E\mu$. If $1\le\rho\le\frac{1}{\alpha}$, then $K(n)=n^-$. The system admits all the high-type riders $\alpha\lambda$ and uses all the remaining capacity $n-\alpha\lambda$ to serve the low-type riders. The fluid profit of each server is

$$\frac{\alpha\lambda P_H\mu+(n-\alpha\lambda)P_L\mu}{n}=(\rho\alpha P_H+(1-\rho\alpha)P_L)\mu. \quad \square$$

Proof of Proposition 3 Under a fluid equilibrium, all the server units have the same equilibrium profit rate. We consider the following three mutually exclusive and jointly exhaustive equilibrium cases.

1. All servers are discriminatory ($\beta=1$). Under this case, the resulting profit rate per unit server is $\frac{\min\{\alpha\lambda,n\mu\}P_H}{n}=\min\{\rho\alpha,1\}P_H\mu$. An infinitesimal amount of servers who deviates to be non-discriminatory does not need to wait and serves low-type riders immediately, and the resulting profit rate per unit server is $P_L\mu$. To sustain this equilibrium, we need $\rho\alpha P_H\mu\ge P_L\mu$, or $\rho\ge r/\alpha$.

2. Servers choose to be discriminatory with probability $\beta<1$ and $I_N=0$. Under this case, all high-type customers are served by discriminatory servers under the random routing rule. The flow balance functions give $\alpha\lambda=\beta\Lambda$ and $\Lambda=(n-I_D)\mu$, where Λ is the service completion rate. Therefore, $\Lambda=\frac{\alpha\lambda}{\beta}$ and $I_D=n-\frac{\alpha\lambda}{\beta\mu}$. According to Little's Law, the expected waiting time of servers in the discriminatory queue is $\frac{I_D}{\beta\Lambda}$. The profit rate of discriminatory servers is $\frac{P_H}{\frac{1}{\mu}+\frac{I_D}{\beta\Lambda}}$. Observe that non-discriminatory servers do not need to wait and serve low-type riders immediately, and the resulting profit rate of non-discriminatory servers is $P_L\mu$. To form this mixed equilibrium, the two profit rates must be the same: $\frac{P_H}{\frac{1}{\mu}+\frac{I_D}{\beta\Lambda}}=P_L\mu$. Solving out the equations

gives $\beta = \frac{\rho\alpha r}{r-\rho\alpha(1-r)}$. To sustain this equilibrium, we need $\rho < \frac{r}{\alpha}$ due to $\beta < 1$. Moreover, $I_N = 0$ implies that $(1-\alpha)\lambda \geq (1-\beta)\Lambda$, which is equivalent to $\rho \geq \frac{r}{\alpha+(1-\alpha)r} = \frac{P_L}{P_E}$. Notice that under this case, the servers' profit rate is $P_L\mu$.

3. Servers choose to be discriminatory with probability $\beta < 1$ and $I_N > 0$. Under this case, all customers are served and the resulting profit rate per unit server is $\rho P_E\mu$. The flow balance functions give $\Lambda = (n - I_D - I_N)\mu = \lambda$. According to Little's Law, the expected waiting time of servers in the discriminatory queue is $W_D = \frac{I_D}{\beta\lambda}$ and that in the non-discriminatory queue is $W_N = \frac{I_N}{(1-\beta)\lambda}$. Because some high-type riders may be matched with non-discriminatory servers, we must have $\beta \leq \alpha$. The expected profit of a customer served by the non-discriminatory servers is $P_N = \frac{(1-\alpha)P_L+(\alpha-\beta)P_H}{1-\beta}$. The two queues have the same profit rate per unit server, which equals the profit rate averaged by all servers, i.e., $\frac{P_H}{1/\mu+W_D} = \frac{P_N}{1/\mu+W_N} = \frac{P_E\lambda}{n} = \rho P_E\mu$. Therefore, $I_D = \beta n\left(\frac{P_H}{P_E} - \rho\right)$ and $I_N = n(1-\rho) + \beta n\left(\rho - \frac{P_H}{P_E}\right)$. Moreover, under random routing, $\frac{I_D}{I_D+I_N}\alpha = \beta$. When $\rho < \frac{P_L}{P_E}$, the solution is $\beta = 0$. All drivers are non-discriminatory. Therefore, the profit rate per unit server is $\frac{P_E\lambda}{n} = \rho P_E\mu$ and $I_N = (1-\rho)n$. If an infinitesimal amount of drivers becomes discriminatory, the total idle server is still $(1-\rho)n$. According to Little's Law, $W_D = \frac{I_D}{\alpha\lambda\cdot\frac{I_D}{(1-\rho)n}} = \frac{1-\rho}{\alpha\rho\mu}$, and the resulting profit rate per unit server is $\frac{P_H}{1/\mu+W_D} = \frac{\alpha P_H\rho\mu}{1-(1-\alpha)\rho}$ To sustain this equilibrium, this expected profit per unit server should be less than $\rho P_E\mu$, which is equivalent to $\rho < \frac{P_L}{P_E}$.

Moreover, under the first case, no idle server would switch from discriminatory to non-discriminatory at any time; under the second and the third cases, an idle server is indifferent from being discriminatory or non-discriminatory under the random routing rule and has no incentive to switch at any time. Therefore, the no-switching assumption is satisfied.

Notice that the regions of ρ are mutually exclusive and jointly exhaustive; therefore, we have established that there exists a unique symmetric equilibrium under which the servers do not switch between strategies while idling. Furthermore, the equilibrium profit rate of the servers is

$$\pi^{\text{full}}(\rho) = \begin{cases} \rho P_E\mu, & \text{if } \rho < \frac{P_L}{P_E}, \\ P_L\mu, & \text{if } \frac{P_L}{P_E} \leq \rho \leq \frac{r}{\alpha}, \\ \rho\alpha P_H\mu & \text{if } \frac{r}{\alpha} < \rho \leq \frac{1}{\alpha}, \\ P_H\mu & \text{if } \rho > \frac{1}{\alpha}. \end{cases} \quad \square$$

Proof of Proposition 4 Under a fluid equilibrium, all the server units have the same equilibrium profit rate. We consider the following three mutually exclusive and jointly exhaustive equilibrium cases.

1. All servers are discriminatory ($\beta = 1$). The resulting profit rate per unit server is $\min\{\rho\alpha, 1\}P_H\mu$. Under this case, the waiting time of the non-discriminatory servers is $W_N = 0$ and we have a single queue for the discriminatory servers with waiting time $W_D = \left[\frac{\beta N}{\alpha\lambda} - \frac{1}{\mu}\right]^+$. The profit rate of a discriminatory server is $\frac{P_H}{W_D+1/\mu}$, and the profit rate of a non-discriminatory server is $\frac{P_L}{1/\mu}$. To sustain the equilibrium, we need $\frac{P_H}{W_D+1/\mu} \geq \frac{P_L}{1/\mu} \Leftrightarrow \rho \geq \frac{P_L}{\alpha P_H} = \frac{r}{\alpha}$.

2. Servers choose to be discriminatory with probability $\beta < 1$ and $W_N = 0$. Under this case, we have a single queue for the discriminatory servers with waiting time $W_D = \left[\frac{\beta N}{\alpha\lambda} - \frac{1}{\mu}\right]^+$ and the waiting time of the non-discriminatory servers is $W_N = \left[\frac{(1-\beta)N}{(1-\alpha)\lambda} - \frac{1}{\mu}\right]^+ = 0$. The profit rate of a discriminatory server is $\frac{P_H}{W_D+1/\mu}$, and the profit rate of a non-discriminatory server is $\frac{P_L}{1/\mu}$. To form this mixed equilibrium, the two profit rates

must be the same: $\frac{P_H}{W_D+1/\mu} = \frac{P_L}{1/\mu}$. This equality implies that $W_D > 0$; thus, $W_D = \frac{\beta N}{\alpha\lambda} - \frac{1}{\mu}$. Plugging W_D into the equation, we have $\beta = \frac{\rho\alpha P_H}{P_L}$. To sustain this equilibrium, we need $\rho < \frac{r}{\alpha}$ due to $\beta < 1$. Moreover, $W_N = \left[\frac{(1-\beta)N}{(1-\alpha)\lambda} - \frac{1}{\mu}\right]^+ = 0$ implies that $\rho \geq \frac{1-\beta}{1-\alpha} \Leftrightarrow \rho \geq \frac{P_L}{P_E}$. Notice that under this case, the servers' profit rate is $P_L\mu$.

3. Servers choose to be discriminatory with probability $\beta < 1$ and $W_N > 0$. Under this case, all customers are served and the resulting profit rate per unit server is $\rho P_E\mu$. We can also view the system as two separate queues with waiting time $W_D = \left[\frac{\beta N}{\alpha\lambda} - \frac{1}{\mu}\right]^+$ and $W_N = \left[\frac{(1-\beta)N}{(1-\alpha)\lambda} - \frac{1}{\mu}\right]^+$. The profit rate of discriminatory servers is $\frac{P_H}{W_D+1/\mu}$, and the profit rate of non-discriminatory servers is $\frac{P_L}{W_N+1/\mu}$, while $W_D > W_N > 0$. To form this mixed equilibrium, $\frac{P_H}{W_D+1/\mu} = \frac{P_L}{W_N+1/\mu} \Leftrightarrow \beta = \frac{\alpha P_H}{P_E}$. To sustain the equilibrium, $W_N = \left[\frac{(1-\beta)N}{(1-\alpha)\lambda} - \frac{1}{\mu}\right]^+ > 0$ implies that $\rho < \frac{1-\beta}{1-\alpha} \Leftrightarrow \rho < \frac{P_L}{P_E}$.

Moreover, under the first case, no idle server would switch from discriminatory to non-discriminatory at any time; under the second and the third cases, an idle server is indifferent from being discriminatory or non-discriminatory before receiving a low-type request and prefers to be discriminatory after rejecting a low-type request under the LISF routing rule and has no incentive to switch at any time. Therefore, the no-switching assumption is satisfied.

Notice that the regions of ρ are mutually exclusive and jointly exhaustive; therefore, we have established that there exists a unique symmetric equilibrium under which the servers do not switch between strategies while idling. Furthermore, the equilibrium profit rate of the servers is

$$\pi^{\text{full}}(\rho) = \begin{cases} \rho P_E\mu, & \text{if } \rho < \frac{P_L}{P_E}, \\ P_L\mu, & \text{if } \frac{P_L}{P_E} \leq \rho \leq \frac{r}{\alpha}, \\ \rho\alpha P_H\mu & \text{if } \frac{r}{\alpha} < \rho \leq \frac{1}{\alpha}, \\ P_H\mu & \text{if } \rho > \frac{1}{\alpha}. \end{cases} \quad \square$$

Proof of Remark 1 Though the above results were derived while assuming that each driver chooses (after each service) to be discriminatory with some probability, the same profit cases still hold if servers choose with some probability to discriminate for all of their respective lives. In that case, we have two pools of drivers and it is an N-model and the cross-link (non-discriminatory drivers serving the high-type customers) may be used. Under both random routing and LISF routing, the fluid analysis is exactly the same as the above and the profit rate is the also the same. $\square$

Proof of Remark 3 Under a fluid equilibrium, all the servers have the same profit rate. We consider the following three mutually exclusive and jointly exhaustive equilibrium cases.

1. The equilibrium profit rate is smaller than $P_L\mu$. Under this case, if a server chooses to be non-discriminatory, it must be the case that he needs to wait for customers because otherwise by choosing to be non-discriminatory, he would earn a profit rate higher than the equilibrium profit rate. Therefore, all customers are served. The resulting equilibrium profit rate is $\rho P_E\mu$, and this case can happen only if $\rho < \frac{P_L}{P_E}$. This is true for both random routing and LISF routing.

2. The equilibrium profit rate is greater than $P_L\mu$. Under this case, if a server chooses to be non-discriminatory and serves the low-type riders, he earns a profit rate no more than the equilibrium profit rate while waiting and serving the low-type riders. Given either the random routing or the LISF routing rule, he

is strictly better off to be discriminatory and remain discriminatory during waiting after each service. As such, all low-type riders are lost. The resulting equilibrium profit rate is $\min\{\rho\alpha, 1\}P_H\mu$, and this case can happen only if $\rho > \frac{P_L}{\alpha P_H} = \frac{r}{\alpha}$.

3. The equilibrium profit rate equals $P_L\mu$. Under this case, either all servers are always discriminatory and $\rho = \frac{r}{\alpha}$ (due to $\min\{\rho\alpha, 1\}P_H\mu = P_L\mu$), or the servers may switching between discriminatory and non-discriminatory after each service. To form a mixed equilibrium, it must be the case that the equilibrium profit rates of both discriminatory servers and non-discriminatory servers are $P_L\mu$. This implies waiting for the discriminatory servers and all high-type riders are served, while only part of the low-type riders is served. To support the equilibrium profit rate $P_L\mu$, this case can happen only if $\rho geq \frac{P_L}{P_E}$ and $\rho \leq \frac{r}{\alpha}$.

Notice that the regions of ρ are mutually exclusive and jointly exhaustive. Therefore, if $\rho < \frac{P_L}{P_E}$, the equilibrium profit rate is $\rho P_E\mu$; if $\rho \in [\frac{P_L}{P_E}, \frac{r}{\alpha}]$, the equilibrium profit rate is $P_L\mu$; and if $\rho > \frac{r}{\alpha}$, the equilibrium profit rate is $\min\{\rho\alpha, 1\}P_H\mu$. □

Proof of Proposition 5 By Propositions 1 and 3, when $\rho \leq \frac{P_L}{P_E}$, which is less than 1, the servers' profit rate is $\rho P_E\mu$ under both equilibria; when $\frac{P_L}{P_E} < \rho \leq \min\left\{1, \frac{r}{\alpha}\right\}$, the servers' profit rate is $\rho P_E\mu$ under the no-information benchmark, which is greater than $P_L\mu$, the servers' profit rate under the full-information equilibrium.

Now, we consider the case that $\min\left\{1, \frac{r}{\alpha}\right\} < \rho < \max\left\{1, \frac{r}{\alpha}\right\}$. Suppose that $\frac{r}{\alpha} > 1$, the servers' profit rate is $P_E\mu$ under the no-information benchmark, which is greater than $P_L\mu$, the servers' profit rate under the full-information equilibrium. Suppose that $\frac{r}{\alpha} \leq 1$, the servers' profit rate is $\rho P_E\mu$ under the no-information benchmark, which is greater than $\rho\alpha P_H\mu$, the servers' profit rate under the full-information equilibrium.

When $\rho \geq \max\left\{1, \frac{r}{\alpha}\right\}$, the servers' profit rate is $P_E\mu$ under the no-information benchmark, while the servers' profit rate is $\min\{\rho\alpha, 1\}P_H\mu$ under the full-information equilibrium. The servers earn (weakly) higher profit if and only if $\rho \geq P_E/(\alpha P_H)$. □

Proof of Proposition 6 By Proposition 1, $\pi^{\text{no}}(\rho)/\rho$ is a constant when $\rho \leq 1$ and $\pi^{\text{no}}(\rho)$ is a constant when $\rho \geq 1$. It suffices to show that $\pi^{\text{full}}(\rho)/\rho$ is a weakly decreasing function of ρ when $\rho \leq 1$, and $\pi^{\text{full}}(\rho)$ is a weakly increasing function of ρ when $\rho \geq 1$.

By Proposition 3, when $\rho \in P_L/P_E$, $\pi^{\text{full}}(\rho)/\rho$ is a constant and $\pi^{\text{full}}(\rho)$ is an increasing function of ρ; when $P_L/P_E \leq \rho \leq r/\alpha$, $\pi^{\text{full}}(\rho)$ is a constant and $\pi^{\text{full}}(\rho)/\rho$ is a decreasing function of ρ; when $r/\alpha \leq \rho \leq 1/\alpha$, $\pi^{\text{full}}(\rho)/\rho$ is a constant and $\pi^{\text{full}}(\rho)$ is an increasing function of ρ; and when $1/\alpha \leq \rho$, π^{full} is a constant and $\pi^{\text{full}}(\rho)/\rho$ is a decreasing function of ρ. Therefore, $\pi^{\text{full}}(\rho)/\rho$ is a weakly decreasing function of ρ, and $\pi^{\text{full}}(\rho)$ is a weakly increasing function of ρ. This completes the proof. □

Proof of Proposition 7 By replacing P_H with $(P_H - s)$ in Propositions 3 and 4, the steady state fluid profit rate of each driver in the system is,

$$\pi^{\text{full}}(\rho) = \begin{cases} \rho(P_E - \alpha s)\mu, & \text{if } \rho \leq \frac{P_L}{P_E - \alpha s} \\ P_L\mu, & \text{if } \frac{P_L}{P_E - \alpha s} < \rho \leq \frac{P_L}{\alpha P_H - \alpha s} \\ \rho\alpha(P_H - s)\mu, & \text{if } \frac{P_L}{\alpha P_H - \alpha s} < \rho \leq \frac{1}{\alpha} \\ (P_H - s)\mu, & \text{if } \rho > \frac{1}{\alpha} \end{cases}.$$

Moreover, the portion of unserved low-type riders, the portion of unserved high-type riders, and the equilibrium servers' profit are all continuous (weakly) decreasing function of s for given ρ. Because both the

portion of unserved low-type riders and the portion of unserved high-type riders are (weakly) decreasing functions of s, the realized social welfare is a (weakly) increasing function of s, and the platform's profit is increasing in s due to that the equilibrium drivers' profit is (weakly) decreasing in s.

Furthermore, when $\rho \in [\frac{P_L}{P_E}, \frac{P_L}{\alpha P_H}]$, the equilibrium servers' profit rate is a constant at $P_L\mu$ when $s \leq (P_E - P_L/\rho)/\alpha$. $\square$

Proof of Proposition 8 As the probability of being discriminatory increases, the waiting time W_D weakly increases and results in weakly lower profit for discriminatory servers, and the waiting time W_N weakly decreases and results in weakly higher profit for discriminatory servers. The proof of Proposition 4 shows that there is a unique probability $p(<1)$ such that both types of servers earn the same profit when $\rho < r/\alpha$, and all servers become discriminatory (i.e., $\beta = 1$) when $\rho \geq r/\alpha$.

Therefore, when $\rho < r/\alpha$ and the portion of discriminatory servers at the full-information equilibrium is less than κ, the equilibrium portion of discriminatory servers remains the same; otherwise, the equilibrium portion of discriminatory servers is κ. When $\rho \geq r/\alpha$, the equilibrium portion of discriminatory servers is κ. As such, at the steady state fluid equilibrium, the portion of discriminatory servers is

$$\beta = \begin{cases} \min\{\kappa, \alpha P_H/P_E\}, & \text{if } \rho \leq P_L/P_E \\ \min\{\kappa, \rho\alpha P_H/P_L\}, & \text{if } P_L/P_E < \rho \leq P_L/(\alpha P_H) = r/\alpha \\ \kappa, & \text{if } \rho > r/\alpha \end{cases}.$$

When $\rho \leq P_L/P_E$, we show that all the riders are served and maximum social welfare is obtained independent of κ. Suppose that $\kappa \geq \alpha P_H/P_E$, the resulting equilibrium is identical to the full-information equilibrium, under which all the riders are served. Suppose that $\kappa < \alpha P_H/P_E$, at equilibrium, $\beta = \kappa$. We further consider the following two sub-cases: $\kappa \geq \alpha$ and $\kappa \leq \alpha$. In the former case, the N-model can be viewed as two separate queues because $\beta = \kappa \geq \alpha$ implies that $W_D \geq W_N$. Furthermore, because $\rho \leq P_L/P_E < 1$ and $(1-\alpha)\lambda \leq (1-\kappa)N\mu \Leftrightarrow \rho(1-\alpha) \leq (1-\kappa)$, which holds when $\rho \leq P_L/P_E$ and $\kappa < \alpha P_H/P_E$, all the riders are served. In the latter case, the N-model can be viewed as a pooled queue from rider's point of view and because $\rho \leq P_L/P_E < 1$, all the riders are served.

When $\rho \in (P_L/P_E, 1]$, we show that all the riders are served and maximum social welfare is obtained if and only if $\kappa \leq \min\{\rho\alpha/r, 1-\rho(1-\alpha)\} = 1-\rho(1-\alpha)$. First notice that maximum social welfare is not obtained if $\kappa \geq \rho\alpha/r$ because when $\kappa \geq \rho\alpha/r$, the resulting equilibrium is identical to the full-information equilibrium, under which some low-type riders are not served by Proposition 3. When $\kappa < \rho\alpha/r$, at equilibrium, $\beta = \kappa$. We further consider the following two sub-cases: $\kappa \geq \alpha$ and $\kappa \leq \alpha$. In the former case, the N-model can be viewed as two separate queues because $\beta = \kappa \geq \alpha$ implies that $W_D \geq W_N$. Furthermore, because $\rho < 1$, all the riders are served if and only $(1-\alpha)\lambda \leq (1-\kappa)N\mu \Leftrightarrow \rho(1-\alpha) \leq (1-\kappa) \Leftrightarrow \kappa < 1-\rho(1-\alpha)$. In the latter case, the N-model can be viewed as a pooled queue from rider's point of view and because $\rho \leq 1$, all the riders are served. Notice that $\alpha \leq 1-\rho(1-\alpha)$, all the riders are served and maximum social welfare is obtained if and only if $\kappa \leq 1-\rho(1-\alpha)$.

When $\rho > 1$, it is best to serve as many as high-type riders first before serving the low-type riders. This can be achieved by setting $\kappa = \min\{\rho\alpha, 1\}$, under which it is the best interest of the informed servers to be discriminatory and earn the highest possible profit rate $P_H\mu$. Notice that setting κ below $\min\{\rho\alpha, 1\}$ implies

that while some low-type riders are served, some high-type riders are lost because they need to be served by the non-discriminatory servers and the load factor $\rho > 1$, while setting κ above $\min\{\rho\alpha, 1\}$ implies that the portion of discriminatory server would exceed $\min\{\rho\alpha, 1\}$ as the informed servers compete and strategically idle themselves; thus, the resulting equilibrium would not maximize social welfare. □

Proof of Proposition 9 We first verify that the first-best admission control identified by Theorem 5 of Savin et al. (2005) forms an equilibrium. We consider the three cases.

If $\rho \geq \frac{1}{\alpha}$, each server earns a profit rate of $P_H\mu$ under the first best control identified by Theorem 5 of Savin et al. (2005). That is, each server is discriminatory and fully utilized by the high-type riders. This is the maximum profit rate a server can earn and the first best control identified forms an equilibrium.

If $1 \leq \rho < \frac{1}{\alpha}$, each server earns a profit rate of $(\rho\alpha P_H + (1-\rho\alpha)P_L)\mu$ under the first best control identified by Theorem 5 of Savin et al. (2005). That is, each server is fully utilized and the probability of serving a high-type rider after each service is $\rho\alpha$. To show that the first best control identified forms an equilibrium, it suffices to show that if all other servers follow the first best control, the probability of serving a high-type rider after each service is bounded by $\rho\alpha$. According to the threshold n^- in Savin et al. (2005), it suffices to show that staying discriminatory and waiting longer would not help improve the probability if there are newly arrived idle servers. Given the high-type riders are routed according to SISF, this probability is indeed zero because the arrival rate of the servers exceeds the rate of the high-type riders. Moreover, matching with a high-type rider can only happen at the moment when the server completes the previous job in the fluid model.

If $\rho < 1$, each server earns a profit rate of $\rho P_E\mu$ under the first best control identified by Theorem 5 of Savin et al. (2005). That is, servers split the entire demand and the probability of serving a high-type rider after each service is $\rho\alpha$. Moreover, all the servers accept all the requests under the first best control. To show that the first best control identified forms an equilibrium, it suffices to show that if all other servers follow the first best control, i) the probability of serving a high-type rider after each service is bounded by $\rho\alpha$, and matching with a high-type rider can only happen at the time when the server completes the previous job; and ii) the probability of a server receives a low-type rider request is 0 at the time when the server completes the previous job. The first statement is true by the analysis performed for the case $1 \leq \rho < \frac{1}{\alpha}$ because staying discriminatory and waiting longer would not improve the probability if the server is not matched with a high-type rider initially. The second statement is true under both the random routing and the LISF routing rule because there are positive idle servers in the system when $\rho < 1$.

Therefore, the first-best admission control identified by Theorem 5 of Savin et al. (2005) forms an equilibrium. Observe that under a fluid equilibrium, all the server units have the same equilibrium profit rate. Hence, the profit rate at any equilibrium is no more than the first-best benchmark. Now, we show that all equilibria under the proposed routing rule achieve the first-best benchmark. We consider the three cases. If $\rho \geq \frac{1}{\alpha}$, by choosing to be discriminatory, the server matches with a high-type rider in no time under the SISF routing rule. Therefore, the equilibrium profit rate must be $P_H\mu$, which is the first-best benchmark. If $1 \leq \rho < \frac{1}{\alpha}$, by choosing to switch from discriminatory to non-discriminatory when there are newly arrived idle servers, the focal server can guarantee a probability of serving a high-type rider after each service to be at

least $\rho\alpha$. This is because the high-type matching probability is minimized when the idle servers are discriminatory upon arrival as characterized by the first best control in Savin et al. (2005). Moreover, either the focal server is matched with a high-type rider in no time or he can be matched with a low-type rider immediately due to $\rho \geq 1$. As such, the focal server can guarantee a profit rate of at least $(\rho\alpha P_H + (1-\rho\alpha)P_L)\mu$, which is the first-best benchmark. Therefore, the equilibrium profit rate must be the first-best benchmark. If $\rho < 1$, there are positive idle servers in the system. Because matching with a high-type rider can only happen at the time when the server completes the previous job due to the SISF routing rule, all the demand would be served by all servers when the low-type riders are routed using either random routing or the LISF routing rule. Therefore, the equilibrium profit rate must be the first-best benchmark. □

Proof of Proposition 10 In this proof, we focus on the position threshold of a server, which would give the order of waiting time. For a tagged server, we define $X(t)$, the position of an idle server who becomes idle at time t, to be the number of idle servers who have shorter idle time (higher priority for high-type customers). $X(t)$ increases by 1 when a server becomes idle with rate $n\mu$ (almost all servers are busy), and decreases by 1 when a high-type customer arrives with rate $\alpha\lambda$. If the position reaches -1, that means the tagged server is matched with a high-type customer. Note that the embedded Markov chain is an asymmetric random walk. Each step the position goes up with probability $p = \frac{n\mu}{n\mu+\alpha\lambda} = \frac{1}{1+\rho\alpha} > 1/2$, goes down with probability $1-p$. Therefore $p(1-p) < 1/4$. Denote by S the steps when the random walk first hits -1. We have

$$\Pr(S = 2k+1) = \frac{1}{k+1}\binom{2k}{k}p^k(1-p)^{k+1}.$$

According to Stirling approximation, we have

$$\Pr(S = 2k+1) \sim \frac{1-p}{(k+1)\sqrt{\pi k}}(4p(1-p))^k.$$

Note in average each step takes time $\frac{1}{n\mu+\alpha\lambda}$. At step $2k-1$, if the server has not been matched with the high-type customers, the expected waiting cost is

$$\bar{P}\frac{2}{n\mu+\alpha\lambda} = \frac{2(P_H\rho\alpha + P_L(1-\rho\alpha))}{n\mu(1+\rho\alpha)}$$

The expected benefit is

$$\Pr(S = 2k+1)(P_H - P_L) = \frac{1}{k+1}\binom{2k}{k}p^k(1-p)^{k+1}(P_H - P_L),$$

which is decreasing in k. The server stops waiting when they are equal

$$\frac{1-p}{(k+1)\sqrt{\pi k}}(4p(1-p))^k(P_H - P_L) = \frac{2(P_H\rho\alpha + P_L(1-\rho\alpha))}{n\mu(1+\rho\alpha)}$$

When n is large, to match the two, we need

$$k = O(\log(n)).$$

Therefore, the waiting time

$$T = O\left(\frac{\log(n)}{n}\right).$$

If $\rho < 1$, then all customers will be served. Each server expects to wait $\frac{1}{\rho} - \frac{1}{\mu}$ between services. Similarly, we track the position $X(t)$ of a server after he becomes idle at $t = 0$. Note that the arrival rate of the idle

servers is exact λ. Therefore each step, the position increases by 1 with rate $\frac{1}{1+\alpha}$, the position decreases by 1 with rate $\frac{\alpha}{1+\alpha}$. Next, we find out the waiting time of those matched with a low-type rider. In fluid, there are $n-\lambda/\mu = n(1-\rho)$ idle servers. Due to the random routing or the LISF routing rule of the low-type rider, the probability of waiting time larger than t (the server misses all the $(1-\lambda)t$ low-type riders) is

$$\left(1-\frac{1}{n(1-\rho)}\right)^{(1-\alpha)\lambda t} \to \exp\left(-\frac{(1-\alpha)\rho\mu t}{1-\rho}\right) \text{ as } n\to\infty.$$

Therefore, as the system is large, the waiting time for a low-type rider is exponential distribution. No matter how long he has waited, the expected waiting is always $\frac{1-\rho}{(1-\alpha)\rho\mu}$. Assume step $2K-1$ is the indifferent point between being discriminatory and non-discriminatory. That means at step $2K-1$ if the server is routing with a low-type ride, he accepts it and obtains P_L. He rejects it and the expected extra profit from high-type riders depend on the probability of getting a high-type rider before the next low-type ride. Define $q := 4\frac{1}{1+\alpha}\frac{\alpha}{1+\alpha} = \frac{4\alpha}{(1+\alpha)^2} < 1$. The probability of getting a high-type rider after $2K-1$ step has an upper bound

$$\sum_{k=K}^{\infty}\Pr(S=2k+1) \sim \int_K^{\infty}\frac{\alpha}{1+\alpha}\frac{q^k}{(k+1)\sqrt{\pi k}}dk < \int_K^{\infty}\frac{\alpha}{(1+\alpha)\sqrt{\pi}}q^k dk = \frac{-\alpha}{(1+\alpha)\sqrt{\pi}\log(q)}q^K.$$

We denote the upper bound by p_U. Therefore, when he rejects a low-type rider at step K, the upper bound of expected extra profit is

$$\begin{aligned}&\int_0^{\infty}\left(p_U(P_H-P_L)-(1-p_U)\bar{P}t\right)\left(\frac{(1-\alpha)\rho\mu}{1-\rho}\right)\exp\left(-\frac{(1-\alpha)\rho\mu t}{1-\rho}\right)dt\\ &= p_U(p_H-p_L)-(1-p_U)\frac{(1-\rho)\bar{P}}{(1-\alpha)\rho\mu}.\end{aligned}$$

To make the upper bound of expected profit zero, we have

$$K = O(1).$$

Note that each step takes $1/(\lambda+\alpha\lambda)$, the waiting time

$$T = O\left(\frac{1}{n}\right). \quad \square$$

Proof of Proposition 11 Suppose that the limit equilibrium waiting time for the discriminatory, moderately discriminatory, and non-discriminatory servers are W_D, W_M, and W_N, respectively. At equilibrium, we can view the system as three separate queues. Therefore, $W_D = \left[\frac{\beta N}{\alpha(1-\gamma)\lambda}-\frac{1}{\mu}\right]^+$, $W_M = \left[\frac{\beta' N}{\gamma\lambda}-\frac{1}{\mu}\right]^+$, and $W_N = \left[\frac{(1-\beta-\beta')N}{(1-\alpha)(1-\gamma)\lambda}-\frac{1}{\mu}\right]^+$. Furthermore, $W_D \geq W_M \geq W_N$ at equilibrium. We consider the following five mutually exclusive and jointly exhaustive equilibrium cases.

- At equilibrium, all servers are discriminatory. Under this case, $W_M = 0$. The profit rate of a discriminatory server is $\frac{P_H}{W_D+1/\mu}$, and the profit rate of a moderately discriminatory server is $\frac{P_E}{1/\mu}$. To sustain the equilibrium, we need to have

$$\frac{P_H}{W_D+1/\mu} \geq \frac{P_E}{1/\mu} \Leftrightarrow \rho \geq \frac{\beta P_E}{\alpha(1-\gamma)P_H}.$$

Notice that under this case, the servers' profit rate is $\min\{\rho\alpha(1-\gamma),1\}P_H\mu$. All the low-type and offline hailing riders are lost, and if $\rho\alpha(1-\gamma)\leq 1$, all the high-type online hailing riders are served, otherwise $\frac{\rho\alpha(1-\gamma)-1}{\rho\alpha(1-\gamma)}$ portion of them is lost.

- At equilibrium, servers are either discriminatory or moderately discriminatory with $W_M = 0$. Under this case, the profit rate of a discriminatory server is $\frac{P_H}{W_D+1/\mu}$, and the profit rate of a moderately discriminatory server is $\frac{P_E}{1/\mu}$. To sustain the equilibrium, we need to have

$$\frac{P_H}{W_D+1/\mu} = \frac{P_E}{1/\mu} \Leftrightarrow \beta = \frac{\rho\alpha(1-\gamma)P_H}{P_E}$$

Because $\beta < 1$, $\rho < \frac{P_E}{\alpha(1-\gamma)P_H}$. Furthermore, $W_M = \left[\frac{(1-\beta)N}{\gamma\lambda} - \frac{1}{\mu}\right]^+ = 0$ implies that $\rho \geq \frac{1-\beta}{\gamma} \Leftrightarrow \rho \geq \frac{P_E}{\alpha(1-\gamma)P_H+\gamma P_E}$. Notice that under this case, the servers' profit rate is $P_E\mu$. All the high-type riders are served and a portion of the low-type riders is lost.

- At equilibrium, servers are either discriminatory or moderately discriminatory with $W_D > 0$. Under this case, the profit rate of a discriminatory server is $\frac{P_H}{W_D+1/\mu}$, the profit rate of a moderately discriminatory server is $\frac{P_E}{W_M+1/\mu}$, and the profit rate of a non-discriminatory servers' profit rate is $\frac{P_L}{1/\mu}$, while $W_D > W_M > 0$. To sustain the equilibrium, we need to have

$$\frac{P_H}{W_D+1/\mu} = \frac{P_E}{W_M+1/\mu} \geq \frac{P_L}{1/\mu} \Leftrightarrow \beta = \frac{\alpha(1-\gamma)P_H}{\alpha(1-\gamma)P_H+\gamma P_E}, \rho \geq \frac{P_L}{\alpha(1-\gamma)P_H+\gamma P_E}$$

Moreover, $W_M = \left[\frac{(1-\beta)N}{\gamma\lambda} - \frac{1}{\mu}\right]^+ > 0$ implies that $\rho < \frac{1-\beta}{\gamma} \Leftrightarrow \rho < \frac{P_E}{\alpha(1-\gamma)P_H+\gamma P_E}$. Notice that under this case, the servers' profit rate is $\rho(\alpha(1-\gamma)P_H + \gamma P_E)\mu$. All the high-type riders are served and a portion of the low-type riders is lost.

- At equilibrium, some servers are non-discriminatory with $W_N = 0$. Under this case, the profit rate of a discriminatory server is $\frac{P_H}{W_D+1/\mu}$, the profit rate of a moderately discriminatory server is $\frac{P_E}{W_M+1/\mu}$, and the profit rate of a non-discriminatory servers' profit rate is $\frac{P_L}{1/\mu}$, while $W_D > W_M > W_N = 0$. To sustain the equilibrium, we need to have

$$\frac{P_H}{W_D+1/\mu} = \frac{P_E}{W_M+1/\mu} = \frac{P_L}{1/\mu} \Leftrightarrow \beta = \frac{\rho\alpha(1-\gamma)P_H}{P_L}, \beta' = \frac{\rho\gamma P_E}{P_L}$$

Because $\beta + \beta' < 1$, $\rho < \frac{P_L}{\alpha(1-\gamma)P_H+\gamma P_E}$. Furthermore, $W_N = \left[\frac{(1-\beta-\beta')N}{(1-\alpha)(1-\gamma)\lambda} - \frac{1}{\mu}\right]^+ = 0$ implies that $\rho \geq \frac{1-\beta-\beta'}{(1-\alpha)(1-\gamma)} \Leftrightarrow \rho \geq \frac{P_L}{P_E}$. Notice that under this case, the servers' profit rate is $P_L\mu$.

- At equilibrium, some servers are non-discriminatory with $W_N > 0$. Under this case, the profit rate of a discriminatory server is $\frac{P_H}{W_D+1/\mu}$, the profit rate of a moderately discriminatory server is $\frac{P_E}{W_M+1/\mu}$, and the profit rate of a non-discriminatory server is $\frac{P_L}{W_N+1/\mu}$, while $W_D > W_M > W_N > 0$. To sustain the equilibrium, we need to have

$$\frac{P_H}{W_D+1/\mu} = \frac{P_E}{W_M+1/\mu} = \frac{P_L}{W_N+1/\mu} \Leftrightarrow \beta = \frac{\alpha(1-\gamma)P_H}{P_E}, \beta' = \gamma$$

Furthermore, $W_N = \left[\frac{(1-\beta-\beta')N}{(1-\alpha)(1-\gamma)\lambda} - \frac{1}{\mu}\right]^+ > 0$ implies that $\rho < \frac{1-\beta-\beta'}{(1-\alpha)(1-\gamma)} \Leftrightarrow \rho < \frac{P_L}{P_E}$. Notice that under this case, all the riders are served and the servers' profit rate is $\rho P_E\mu$.

Notice that the regions of ρ are mutually exclusive and jointly exhaustive, therefore, we always have a single equilibrium for all potential parameter configurations and this completes the proof. □

Proof of Proposition 12 When $\rho > \frac{1}{\alpha(1-\gamma)}$, all the server capacity would be devoted to the high-type riders and the profit rate is $P_H\mu$. When $\rho \leq \frac{1}{\alpha(1-\gamma)}$, given the SISF routing rule, the servers capacity first cover all the high-type riders without idling and then the remaining capacity is used to cover the low-type

riders and the non-disclosed type riders by the analysis of Proposition 10. That is, the normalized remaining server capacity for these two types is $1-\rho\alpha(1-\gamma)$, while the loads for the non-disclosed type riders and the low-type riders are $\gamma\rho$ and $(1-\alpha)(1-\gamma)\rho$. Given the nature of offline hailing, it is best to model the routing rule for the non-disclosed type riders as random routing. As such, the steady state fluid profit rate of each driver derived from the remaining server capacity can be calculated using Proposition 4 by setting the effective load $\hat{\rho}=\frac{\gamma\rho+(1-\alpha)(1-\gamma)\rho}{1-\rho\alpha(1-\gamma)}$, the effective portion of high-type riders $\hat{\alpha}=\frac{\gamma}{\gamma+(1-\alpha)(1-\gamma)}$, the effective profitability of high-type riders $\hat{P}_H=P_E$, the relative profitability $\hat{r}=\frac{P_L}{P_E}$, and the ex ante profitability $E[\hat{P}]=\frac{\gamma P_E+(1-\alpha)(1-\gamma)P_L}{\gamma+(1-\alpha)(1-\gamma)}$. That is, the profit rate of the remaining capacity is,

$$\hat{\pi}(\rho)=\begin{cases}\hat{\rho}E[\hat{P}]\mu, & \text{if } \hat{\rho}\le\frac{P_L}{E[\hat{P}]}\Leftrightarrow\rho\le\frac{P_L}{\alpha\gamma P_H+(1-\alpha\gamma)P_L}\\ P_L\mu, & \text{if } \frac{P_L}{E[\hat{P}]}<\hat{\rho}\le\frac{\hat{r}}{\hat{\alpha}}\Leftrightarrow\frac{P_L}{\alpha\gamma P_H+(1-\alpha\gamma)P_L}<\rho\le\frac{P_L}{\alpha\gamma P_H+(\alpha+\gamma-2\alpha\gamma)P_L}\\ \hat{\rho}\hat{\alpha}\hat{P}_H\mu, & \text{if } \frac{\hat{r}}{\hat{\alpha}}<\hat{\rho}\le\frac{1}{\hat{\alpha}}\Leftrightarrow\frac{P_L}{\alpha\gamma P_H+(\alpha+\gamma-2\alpha\gamma)P_L}<\rho\le\frac{1}{\gamma+\alpha(1-\gamma)}\\ \hat{P}_H\mu, & \text{if } \hat{\rho}>\frac{1}{\hat{\alpha}}\Leftrightarrow\rho\ge\frac{1}{\gamma+\alpha(1-\gamma)}\end{cases}.$$

Given that the profit rate of the capacity devoted to high-type riders is $P_H\mu$ and the normalized remaining server capacity for these two types is $1-\rho\alpha(1-\gamma)$, the steady state fluid profit rate of each driver in the system is $\rho\alpha(1-\gamma)P_H\mu+(1-\rho\alpha(1-\gamma))\hat{\pi}(\rho)$. Plugging in the numbers, we have,

$$\pi(\rho)=\begin{cases}\rho P_E\mu, & \text{if } \rho\le\frac{P_L}{\alpha\gamma P_H+(1-\alpha\gamma)P_L}\\ (\rho\alpha(1-\gamma)P_H+(1-\rho\alpha(1-\gamma))P_L)\mu, & \text{if } \frac{P_L}{\alpha\gamma P_H+(1-\alpha\gamma)P_L}<\rho\le\frac{P_L}{\alpha\gamma P_H+(\alpha+\gamma-2\alpha\gamma)P_L}\\ \rho(\alpha(1-\gamma)P_H+\gamma P_E)\mu, & \text{if } \frac{P_L}{\alpha\gamma P_H+(\alpha+\gamma-2\alpha\gamma)P_L}<\rho\le\frac{1}{\gamma+\alpha(1-\gamma)}\\ (\rho\alpha(1-\gamma)P_H+(1-\rho\alpha(1-\gamma))P_E)\mu, & \text{if } \frac{1}{\gamma+\alpha(1-\gamma)}<\rho\le\frac{1}{\alpha(1-\gamma)}\\ P_H\mu, & \text{if } \rho>\frac{1}{\alpha(1-\gamma)}\end{cases}. \quad \square$$