Customer Segmentation and Fairness: A Queueing Perspective

Jian Liu[1,2], Yong-Pin Zhou[2]

[1]School of Economics and Management, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, China

[2]Foster School of Business, University of Washington, Seattle, Washington, 98195, USA

e-mail: jianlau@njust.edu.cn; yongpin@uw.edu

Abstract

In this paper we analyze how customer perception of fairness impacts a service provider who offers a priority service option. Customers differ in their waiting time costs, but that information is private to each customer. To maximize revenue, the service provider can offer a regular queue and a priority queue with additional charge, to induce customers to self-select into the two queues. We model customer fairness perception as a negative utility on the regular customers which is proportional to the waiting time difference between the regular and priority queues. Analyzing a stylized $M/M/1$ queue, we derive results that reaffirm existing research on the benefits of differentiated service and differentiated pricing in some situations, as well as results that challenge conventional wisdom in other situations. These results also lead to insights about how the service provider should position and promote the two queueing options.

Keywords: Priority Queues, Captive/Non-Captive Customer Segmentation, Fairness Perception, Differentiated Services, Differentiated Pricing

1. Introduction

Matching capacity with demand is a central theme in service operations management (Cachon and Terwiesch 2012). When capacity-demand mismatches occur, due to either long-term insufficient capacity or short-term fluctuations in service load, customer waiting results. Americans spend roughly 37 billion hours each year waiting in line (Morrow 1984, Stone 2012) in regard to services related to health care, call centers, banks, restaurants, transportation, and so forth. Customer sensitivity to waiting is a key factor in the service provider's capacity decisions as well as in the customer's queue-joining behavior. Such behavior is typically modeled as a waiting cost related to the wait time. Whereas for a customer (hereafter, he), the waiting cost includes the loss of time, emotional stress, and boredom, for a service provider (hereafter, she), the waiting cost could mean lost revenue or customer dissatisfaction (Stone 2012).

Service providers can take measures to mitigate the negative consequences of waiting, so as to reduce customer loss and maximize revenue. Besides the common approaches, such as adding capacity, limiting demand, and reducing variability from the system (Cachon and Terwiesch 2012), process design is an equally important, albeit subtler, approach. When there exists significant customer heterogeneity in waiting cost, service providers can use customer segmentation and prioritization as an effective tool.

For instance, when a service provider has identified certain characteristics of each arriving customer, she could give higher priority to customers who are more important, less patient, or require less service time, or a combination thereof. Examples include emergency room prioritization based on patient severity, or the "shortest processing time first" (Pinedo 2016) and "$c\mu$" (Cox and Smith 1961, Smith 1956) scheduling rules, which have been well studied (Van Mieghem 1995, Ward and Armony 2013).

Customer segmentation becomes more difficult when the differentiating information about each customer is private. The service provider may know the general distribution of such information across all of the customers but not individual ones. Thus, a "forced" segmentation and prioritization cannot be used. Instead, incentive mechanisms are needed to induce customers to differentiate themselves in regard to prioritization. This is often achieved by pricing, which not only helps to segment customers due to their heterogeneous price/cost sensitivity but also generates additional revenue for the service provider. For example, in Afèche and Pavlin (2016), a price/lead-time (wait time) menu is designed, along with a scheduling policy, to segment customers properly and maximize service provider revenue.

This pricing and prioritization combination is commonly used in practice. Various forms of "VIP lines" are illustrations of this concept: Bank branches often have VIP customer service lines, and airlines give priority to business passengers at check-in and boarding. Even airport security lines have such a feature. Universal Studio's Express Pass, which partly motivated this research, allows holders to bypass the regular waiting lines (which could be hours long during peak time), but can cost as much as a park ticket itself. This appears to be a rational way to fully utilize the park's limited resource (ride capacity), as the park can generate additional, substantial revenue from customers who are willing to pay more to wait less. In contrast, Disney theme parks do not charge for the FastPass, which is available to anyone with a park entrance ticket.[1] It would be interesting to determine the factors that can be used to explain such different approaches in practice.

[1]Disney is selling FastPass for an extra fee in its newest park in Shanghai. One of its executives, however, indicated that the purpose is to combat scalping.

Another example is the express check-in option provided by hotels such as Treasure Island in Las Vegas, where customers can pay a fee (in the range of $30–$40) ahead of time to stand in a shorter, priority check-in line upon arrival. Such features are especially useful during peak time for customers who want to reduce waiting but can be controversial because every customer who gets priority service simply adds to the waiting time of the non-priority customers. Even though it is the regular customers' choice not to pay the priority fee, there is still resentment due to perceived unfairness (which is studied in studied in Rafaeli *et al*. 2002, 2005, and also reflected in some negative reviews of Universal's Express Pass online). Such concerns and complaints about the priority fee are to be expected because inherently priority service results in a redistribution of waiting time from higher-priority customers to lower-priority customers.

Traditional queueing research focuses on the mathematical evaluation of various waiting measures (see, for example, Davis 1966, Kleinrock 1975, Sztrik 2016), but several recent papers incorporate the human perspective in waiting systems. Two excellent review books on this topic are by Hassin and Haviv (2003), and Hassin (2016). Additional work includes the study of waiting line design and information availability (Shunko *et al.* 2017) and human server social loafing (Wang and Zhou 2017). The subject of customer perception of fairness and how that affects both customer behavior and system performance has been understudied. To the best of our knowledge, this is the first paper that provides a systematic study of the impact of fairness perception in an analytical queueing framework. Using a stylized single-server queue, we incorporate fairness into the customers' queueing joining behavior and examine how this affects the service provider's determination of differentiated services and pricing as well as her revenue.

We separately examine a captive service system, whereby no customers balk but they can choose whether to pay to get priority service, and a non-captive service system, whereby customers can balk if they find the wait and price unacceptable. In particular, we address the following questions: How does fairness matter, and what are its implications for queue design (pricing and priority), customer choices, and service provider revenue?

The contribution of our paper is threefold. First, our research indicates that, for a captive system, the service provider should adopt priority segmentation to obtain the maximum revenue. When the perception of fairness is stronger, more customers will pay to join the priority queue, leading to higher revenue for the service provider. This result is consistent with the predominant view in the service operations literature that customer segmentation and differentiated pricing are an important tool for the service provider. In our model, this logic works in a captive system because the service provider does not need to worry about losing customers. Her objective is to create a way to extract the most revenue out of "moving" customers from the regular queue to the priority queue.

This logic does not necessarily work in a non-captive system, however, because customers always have the choice of leaving. Therefore, the service provider must delicately balance the trade-off between (1) making the regular queue less appealing so that more customers pay to join the priority queue, and (2) alienating customers such that they leave the system, taking away their service revenue. Our second contribution is to show that, in a non-captive system, when the customer fairness perception is not strong and the value of a regular customer is low, the service provider should continue to follow conventional wisdom to provide differentiated services and pricing. Otherwise, the service provider should focus on

offering one type of service so as not to alienate customers and cause high lost revenue. This observation is both sensible and novel in the service operations literature. We view it as a major contribution of our paper.

Our third contribution is to shed light on how, if possible, the service provider should try to influence customer fairness perception. Prominent placement and broad advertising of the priority service and its associated fee can serve to heighten customers' awareness of the differentiated service and increase their perception of service fairness. The results from our model regarding equilibrium sensitivity to the fairness perception parameter provide useful guidance on how to design, construct, and operate the waiting queues accordingly. It is intuitive that, in a captive service system, it is to the service provider's benefit to highlight the contrast between the two queues (for example, by putting the priority queue right next to the regular queue) to stimulate regular customers to "upgrade" to the priority service. In a non-captive system, when it is possible for the service provider to reduce the customers' unfairness perception, we show that she should do so. As a result, she can provide differentiated services and pricing and do better. (This is consistent with our previous result that it is optimal for the service provider to not differentiate customers and to focus on just one queueing option when the fairness perception is sufficiently high.) This can be achieved, for example, by moving the two queues far away from each other. During the World Expo 2010 in Shanghai China, waiting times at popular pavilions often exceeded three hours. Customers of expensive restaurants in some pavilions could get priority access, but their entrance was hidden from view by those in the regular queue. Some pavilions, such as the Singapore Pavilion, deliberately used the exit as the VIP queue entrance so as to minimize exposure to regular customers who were waiting long hours to get in.

2. Literature Review

Our paper is most related to two streams of research: classification service (customer segmentation and differentiated service) and fairness. Each of these streams is discussed below.

2.1. Classification service and the generalized $c\mu$ rule

In a service system in which service capacity is limited and customers are heterogeneous, it is often necessary to segment customers into different classes and to provide differentiated services based on this classification. If a *c* index is used to represent the unit waiting cost of each customer class, and a μ index is used for the service rate of each customer class, then their product $c\mu$ represents the rate at which the customer waiting cost can be removed from the system. It then makes sense to give priority to the customer class with the highest $c\mu$ index. Smith (1956) first suggested the optimality of this so-called $c\mu$ rule in a deterministic and static environment. Cox and Smith (1961) then extended the optimality of the $c\mu$ rule to a stochastic and dynamic environment with an arbitrary time horizon. Later, Pinedo (1983) proposed that, in stochastic and static settings, the $c\mu$ rule is also optimal. Van Mieghem (1995) developed an optimal $Gc\mu$ policy for the queueing system that operates in heavy traffic, with a non-decreasing convex waiting cost function. Atar *et al.* (2010) proved that under multi-server fluid scaling and overload conditions, a server-scheduling policy that assigns priority to classes based on $c\mu/\theta$, where θ is the abandonment rate of waiting customers, is asymptotically optimal for minimizing the overall long-run average holding cost when customers are heterogeneous but their classification is endogenous. Ward and Armony (2013) showed

that it is optimal to classify customers and give priority to those who may have a higher waiting cost, similar to the use of the $Gc\mu$ rule.

Whereas it is generally accepted that the service provider should classify the customers and employ some variant of the $c\mu$ rule, Afèche (2013) discussed situations in which the $c\mu$ rule is not optimal for revenue maximization. In our paper, we study customers who have the same μ but different c. The information is private to each customer, however, so the service provider cannot use a $c\mu$-type index to prioritize the services. Instead, she relies on differentiated prices to induce customers to self-segment. We find that, when customers' fairness perception is high, there exist situations in which the service provider should avoid differentiated services. Other recent papers on customer segmentation and service revenue maximization include Afèche and Pavlin (2016), who studied the segmentation of customers based on time sensitivity, and Gavirneni and Kulkarni (2016), who analyzed charging for priority service in a single-server queueing system with either captive or non-captive customers, whose waiting cost follows a Burr distribution.

There is also an active stream of research that focuses on customer queueing behavior. Bassamboo and Randhawa (2016) believed that customers become differentiated as time progresses and, thus, investigated scheduling polices that differentiate between customers to optimize system performance metrics. Choi *et al*. (2001) analyzed the *M/M/1* queues with two classes of customers, in which the priority-class customers have impatience of constant duration and low-priority customers are infinitely patient. Iravani and Balcıoğlu (2008) studied priority-class customers with impatience in *M*/*GI*/1+*M* queue. Wang *et al*. (2015) described a preemptive *M/M/c* queue with two priority classes that have different service

rates, in which the high-priority class is completely impatient. Kleinrock (1967) considered customers' inconvenience and impatience in a queue and suggested the need to determine a customer's relative position in the queue based on the size of his bribe, which is paid before the customer sees the queue length. Yang *et al.* (2017), in contrast, proposed a time-trading mechanism, in which customers who are privately informed about their waiting costs mutually agree on the ordering in the queue by trading positions.

2.2. Fairness

The aforementioned literature on customer queueing behavior focuses mainly on the psychological impatience of customers and how they respond to paying an extra fee for priority service, either by going to a different queue or moving ahead in the same queue. Not enough attention has been paid to the psychological responses of non-priority customers who may perceive unfairness when later-arriving, but higher-priority, customers overtake them in the queue. Yet, these fairness issues have been empirically shown to be very important to customers (Rafaeli *et al*. 2002, 2005), sometimes even more important than the waiting time itself. Waits deemed unreasonable were often those for which fairness was violated or for which the rules of behavior were not well stated.

First come, first served (FCFS) is commonly accepted and adopted as the fairest mechanism (Larson 1987). Norman (2009) noted that customers have negative feelings due to others' cutting in line or having to wait longer than expected. Avi-itzhak *et al*. (2008) acknowledged this but also pointed out that the amount of work should also matter; it could also seem unfair to serve a very long job before a very short job even though the former arrived first (a situation often encountered in supermarket checkout lines). Indeed, in many situations, the shortest-job-first (SJF) queue discipline minimizes the average wait for *all*

of the customers (Lariviere 2014) and can be considered by some to be fair. In regard to priority queueing systems, Rafaeli *et al.* (2002, 2005) revealed a connection between queue structure and fairness perceptions, confirming that seeing others wait in a shorter line is perceived as unfair even if these others paid for this right. Maister (1985) and Larson (1987) used slips and skips to measure social injustice in queues and pointed out that the classification services of customers violated the rule of FCFS. Nageswaran and Scheller-Wolf (2017) analyzed the fairness policy for a two-queue system that serves two classes of customers, one of which is redundant. They provided the analytical results regarding fairness and identified when redundancy harms or benefits non-redundant customers. They also found that joining the shortest queue is optimal only if the queue lengths are observable. In other words, in this case, redundancy is fair. If the queues are unobservable, however, there may be situations in which the non-redundant class is worse off under redundancy.

There is also a large literature on fairness, outside of the queueing setting, focused on more general social justice issues. The main fairness standards concern the following (Cui *et al.* 2007, Geng *et al.* 2015, Ho and Su 2009, Jin *et al.* 2014, Li and Jain 2016, Wu and Niederhoff 2014, Xia *et al.* 2004): (1) whether the two sides are identical, (2) the relationship between input and output for two sides, and (3) the outcome relative to an artificial psychological benchmark.

Rabin (1993) explained that individuals' notions of fairness are heavily influenced by the status quo and other reference points. Fehr and Schmidt (1999) presented an inequity aversion model based on situations, whereby, if a player is worse off in material terms than other players, then he or she feels inequity and suffers more from inequity to his or her material disadvantage than that to his or her material advantage.

Boiney (1995) extended the theories of ex-ante and ex-post equitable distributions from the social risk literature with modified envy-based fairness measures and developed a decision model with limited options under uncertainty, considering preferences and heterogeneity. Ho and Su (2009), Jin *et al.* (2014), and Li and Jain (2016) noted that fairness comes from customer perception and comparison of expenditure (e.g., price of similar products purchased in the past, price of similar products that other customers pay) or obtainment (e.g., outcome, utility, waiting time). Ward and Armony (2013) and Geng *et al*. (2015) studied fairness in service environments and focused mainly on unfairness caused by unbalanced workload among different servers.

In this paper, we study fairness in the queueing system by comparing the expected waiting time of customers in different queues. This is based on the observations in Maister (1985), Larson (1987), Norman (2009), and Lariviere (2014).

3. Model and Optimization

We model the service provider as a single-server queue. Customer arrivals follow a stationary Poisson Process with rate λ, and each customer service time has *i.i.d.* exponential distribution with rate μ. For every service received, we assume that the customer pays a base fee of c and obtains a fixed value of R'. Although customers value the service time, they dislike the waiting time. We assume that customers are identical except for their waiting cost. For every unit of time spent waiting in the queue, each customer incurs a cost of H, where H is a random variable with a *i.i.d.* distribution across the entire customer base.

For simplicity, we follow Afèche and Pavlin (2016), Kulshreshtha (2003), Gavirneni and Kulkarni (2011), and Gavirneni and Kulkarni (2014) and let H be uniformly distributed on the interval $[0,1]$.

The service provider is a revenue maximizer. She receives a fixed c from each served customer. In addition, due to customer heterogeneity in waiting cost, she knows that those who are more sensitive to waiting (i.e., having a large H) may be willing to pay extra to reduce wait time. This means that there is an opportunity for the service provider to create priority service classes and charge extra fees for their access. Customers can self-select based on the priority and fee structure, allowing the service provider to make extra revenue from the priority fees. In the extreme situation, she should be able to create an infinite menu of priority-price combinations. In practice, service providers are more likely to provide a limited number of such choices (Nazerzadeh and Randhawa 2017). For simplicity, we focus on just two classes: a regular queue and a priority queue, which is most commonly studied in the literature (Cui *et al.* 2017, Gavirneni and Kulkarni 2014, Gavirneni and Kulkarni 2014, Lajos 1968, Wang *et al*. 2015) and observed in practice. In our model, the service provider must decide whether to offer a priority queue option and, if offered, how much she should charge customers to use it. We denote the priority fee by a fixed $K \geq 0$. Results from the two-class model should provide insights about complex systems with more than two priority queues.

This priority fee can be equivalently implemented in practice as a discount. That is, the service provider provides a normal queue at a charge of K, and customers who are willing to take a standby option (agreeing to be served at a lower priority) will receive a discount of K. The regular customers in this framework correspond to the priority customers in our current framework, and the standby customers

correspond to our regular customers. For ease of exposition, discussions in this paper will be based solely on our current "base fee + priority fee" framework.

Once the service provider announces the priority fee K, it is fixed. Thus, a customer's choice of queue depends on how much waiting cost he can save by joining the priority queue. For a customer with unit waiting cost of H, the utility that he will obtain by joining either queue or balking is given as follows:

Join the priority queue

$$U_1 = R' - c - K - HW_1. \quad (1)$$

Join the regular queue

$$U_2 = R' - c - HW_2 - \alpha(W_2 - W_1). \quad (2)$$

Balking (if allowed)

$$U_0 = 0. \quad (3)$$

Here, α is a parameter that represents the strength of customers' fairness perception. Because each customer makes his own queue-joining decision based on his H value, each queue-joining decision, in turn, affects the expected waiting time of both queues. Therefore, W_1 and W_2 in equations (1) and (2) represent the *equilibrium* expected waiting time in queue for the priority and regular customers, respectively.

If a customer joins the priority queue, his utility in (1) will be the fixed value R' minus the base fee c, the priority fee K, and the expected waiting cost. If a customer joins the regular queue, then he would save the priority fee but incur a higher waiting cost (presumably $W_2 > W_1$), plus the fairness disutility. Note that both (1) and (2) differ from the customer utility typically found in the literature: $R' - HW - c$ (e.g., Afèche and Pavlin 2016, Gavirneni and Kulkarni 2016), because, in our model, regular

customers incur an additional fairness disutility of $\alpha(W_2 - W_1)$, and the priority customers pay an extra fee of K.

When no customers balk, we call the system a captive one. This happens when the customers have no outside choices (e.g., they need account services) or when the value of service R' is much higher than the waiting cost. Each customer will pick the queue with a higher utility to join by comparing (1) and (2) based on his individual H. In other situations, customers may choose not to join either queue, and they can balk and receive a fixed utility U_0 from an outside option. We call this the non-captive service system. Without loss of generality, we can normalize U_0 to zero (for example, we can adjust R' to accommodate a positive U_0). A zero utility for the outside option also makes practical sense because the balking customers neither receive any service benefits nor incur any waiting cost.

It is important to note that, when formulating customer utilities, we used only expected waiting time information. There are two main reasons for this. First, for the problems that we are modeling, customers must make decisions about whether to join and whether to pay the priority fee before they see the actual queues. In the hotel express check-in example, customers can pay for the priority service ahead of time without knowing the exact waiting situation when they arrive. They must make such a decision based on their expectation of the waiting time. For Universal Studio (and other theme parks like it), customers pay for express passes before they enter the park. For a similar reason, we also model the fairness disutility term based on expected waiting times. In the Universal Studios example, although customers experience the fairness issue more acutely in person, that real-time perception happens after the customer already has decided whether to get an Express Pass. When he is deciding beforehand whether to get such a

pass, he does not have the real-time wait information and must rely on online reviews, friends' stories, or other sources. Such data points represent the realizations of a random outcome, but the aggregation of all this information gives a sense of the equilibrium average. As we described, such historical information can also pass on a sense of unfairness to the regular customers. Thus, it is more reasonable to use average waiting time in his decision.

Second, while it would be very interesting to model situations in which customers make utility comparison and purchase decisions based on real-time wait information, such a model would solve a different type of problem and require a different mathematical model. Thus it is beyond the scope of this paper.

The use of priority service is, in essence, a redistribution of waiting time from the priority customers to the regular customers. A distinguishing feature of our model is the inclusion of customer fairness perception. Although the priority and fee structures are transparent and known to all of the customers before they make decisions, customers in the regular queue may still feel a sense of unfairness when they know that priority customers get faster service and can overtake them in the service order, even though they themselves decided not to do so (Rafaeli *et al.* 2005). This perception of unfairness imposes a disutility on the regular customers, which we model as $\alpha(W_2 - W_1)$. It affects a customer's queue-joining decision (Sim 2010). The parameter $\alpha \geq 0$ measures the intensity of customers' preference for fairness (Bertsimas *et al.* 2012, Nicholson and Snyder 2011, Rafaeli *et al.* 2002, 2005).

Note that we define the disutility as a result of perceived unfairness by the regular customer based on the difference of expected waiting time between the two queues. It is also possible to define the disutility

on the difference between expected waiting time of the regular customers and of all the customers (i.e., the *extra* time the regular customers spend due to the priority customers). This alternative definition will certainly affect the boundary conditions of the optimal actions, but the qualitative insights remain the same.

Captive and non-captive systems differ in whether customers can balk. In the following two sections, we analyze them separately and show that the optimal system design and priority fee could differ significantly between the two types of systems.

3.1. Captive service systems

In a captive system, each customer decides whether to pay the base fee c to join the regular queue or to pay a fee of $c + K$ to join the priority queue. Each customer decides independently which queue to join by comparing (1) and (2). The resultant arrivals to the two queues form independent Poisson processes. We denote the rates to priority and regular queues by θ and ξ, and then we have $W_1 = \frac{\rho(\theta+\xi)}{\mu(1-\rho\theta)}$ and $W_2 = \frac{\rho(\theta+\xi)}{\mu(1-\rho(\theta+\xi))(1-\rho\theta)}$ (Gross and Harris 1998), where $\rho = \lambda/\mu$ is the system offered load.

Similarly, from (1) and (2), we see that customers pay to join the priority queue if and only if $H \geq \frac{K}{W_2 - W_1} - \alpha$. Therefore, we have

$$\theta = \left[1 - \frac{K}{W_2 - W_1} + \alpha\right]^+ \text{ and } \xi = 1 - \theta. \quad (4)$$

Together, from (1), (2), and (3), we can solve for the equilibrium customer behavior.

In a captive system, the service provider receives a base fee of c from every customer, but a priority fee only from a θ portion of the priority customers. We denote her total revenue rate by $R(K) =$

$c\lambda + K\theta\lambda$ for a given K. Her objective is to find the right priority fee K so as to maximize revenue: $\max_{K\geq 0} R(K) = c\lambda + \lambda \max_{K\geq 0} K\theta$. A higher K will increase the revenue per priority customer, but a lower K will increase the number of priority customers. The optimal K must seek a balance between this tradeoff. From a technical standpoint, this is equivalent to picking an optimal θ to maximize $K\theta$.

To simplify analysis, we fix c and vary K only. In practice, it is often more difficult to change the mandatory base fee, which is more scrutinized by the market than the optional priority fee. Further, in customer service contexts, the regular service is usually free, and a fee is charged only for priority service.

For any fixed value of $0 \leq \theta \leq 1$, we can solve (1), (2), and (4) to get:

$$\begin{cases} K(\theta)_{(\alpha)} = \dfrac{\rho^2}{\mu(1-\rho)(1-\theta\rho)}\left[(1-\theta)+\alpha\right] \\ R(\theta)_{(\alpha)} = \dfrac{\rho^3\theta}{(1-\rho)(1-\theta\rho)}\left[(1-\theta)+\alpha\right] + \lambda c \end{cases} . \tag{5}$$

Here we use subscript (α) to emphasize the dependence of the optimal solution on customer's fairness perception parameter. Later, when clear from the context, we will suppress the α subscript to simplify exposition.

Thus, the optimization problem can be simplified as $\max_{\theta\geq 0} R(\theta)_{(\alpha)}$, subject to $0 \leq \theta \leq 1$. The optimal solution is given in the following lemma. All of the proofs in this paper can be found in the appendix.

Lemma 1.

(1) For $0 \leq \alpha < 1-\rho$,

$$K^*_{(\alpha)} = \frac{\rho(1-\sqrt{1-\rho(1+\alpha)})}{\mu(1-\rho)}; \quad \theta^*_{(\alpha)} = \frac{1-\sqrt{1-\rho(1+\alpha)}}{\rho}; \quad R^*_{(\alpha)} = \frac{\rho(1-\sqrt{1-\rho(1+\alpha)})^2}{(1-\rho)} + \lambda c . \tag{6}$$

(2) For $\alpha \geq 1-\rho$,

$$K^*_{(\alpha)} = \frac{\rho^2}{\mu(1-\rho)^2}\alpha; \qquad \theta^*_{(\alpha)} = 1; \qquad R^*_{(\alpha)} = \frac{\rho^3}{(1-\rho)^2}\alpha + \lambda c \ . \tag{7}$$

In the first case, when $\alpha < 1-\rho$, we see that $K^*_{(\alpha)}$ is strictly positive and $\theta^*_{(\alpha)}$ is strictly between 0 and 1. This indicates that, when customer fairness perception is sufficiently small, it is optimal for the service provider to use priority service to segment customers and to generate extra revenue from the priority service.

In the second case when $\alpha \geq 1-\rho$, $K^*_{(\alpha)}$ is strictly positive and $\theta^*_{(\alpha)} = 1$. This indicates that, when customer fairness perception is sufficiently large, the service provider should charge a priority fee that induces all of the customers to join the priority queue. This will give the service provider the maximum possible revenue, but it is clearly a result based on the captive assumption. When the fairness perception is strong enough, the regular queue where customers experience fairness disutility becomes very unattractive, and all of the customers will move to the priority queue. In essence, the service provider takes advantage of the captiveness of the system and uses fairness disutility to extract maximum priority fee. We will see this result change in Section 3.2 when the system is non-captive.

The following proposition shows that the change from Case (1) to Case (2) in Lemma 1 actually happens continuously in a captive system: At any level of fairness perception (α), when the perception gets stronger, customers get more disutility being stuck in the regular queue. Hence, more customers are willing to pay the extra fee K to join the priority queue. This allows the service provider to charge a higher priority fee and collect higher revenue as a result.

Proposition 1: *When customers are captive, the equilibrium proportion of priority customers* $\theta^*_{(\alpha)}$*, the optimal priority fee* $K^*_{(\alpha)}$*, and the maximum revenue* $R^*_{(\alpha)}$ *are all increasing in the fairness parameter* α*. Moreover, the equilibrium expected waiting time* $W^*_{1\,(\alpha)}$ *and* $W^*_{2(\alpha)}$ *are both increasing in* α*.*

The results for the captive service system are summarized in Figure 1 below.

$0 \leq \alpha < 1 - \rho$	Two active queues, $K^* > 0$ **(Lemma 1)**
$\alpha \geq 1 - \rho$	Priority queue only, $K^* > 0$ **(Lemma 1)**

Figure 1: Service provider's optimal actions in the captive service system ($\theta + \xi = 1$)

Proposition 1 has an important implication for the service provider. When the customers are captive, the service provider should strive to foster a strong contrast between the two priority classes of services being offered and promote the priority service heavily to upgrade more customers from the regular queue to the priority queue. Through all of this, the service provider does not have to worry about losing customers because they are captive.

This insight can be used to partially explain the observation that some services do not make any serious attempt to "hide" the priority queue. This could be due to the fact that their customers are accustomed to paying different prices for different services and do not feel that there is anything unusual about it. Thus, regular customers do not derive a disutility from being treated differently.

3.2. Non-captive service systems

The insights presented in the last section depend heavily on the fact that customers are captive. When customers have outside options, they do not have to choose between just the priority and regular queues,

hence, $\theta + \xi \leq 1$. The remaining $1 - \theta - \xi$ portion of the customers will balk and not join the service system at all.

A customer's queue-joining decision is based on where $\max\{0, U_1, U_2\}$ is achieved, and the service provider will determine the system design and price accordingly. In the equilibrium, we note that there exist four possible decision outcomes by the service provider:

(A) Provide two queueing options to the customers, and price the priority queue such that both queues are actively used by customers (i.e., achieve true customer segmentation)

(B) Provide two queueing options to the customers, and price the priority queue such that customers enter only the priority queue (i.e., no customer segmentation)

(C) Provide two queueing options to the customers, and price the priority queue such that customers enter only the regular queue (i.e., no customer segmentation)

(D) Provide one queueing option to the customers

Outcome D is the standard single-queue, FCFS service system. In Outcome C, although the service provider may still charge a priority fee K, it does not matter to her revenue because nobody pays it. For all practical purposes, Outcomes C and D are interchangeable. In our analysis, we will focus on Outcomes A–C only; whenever the single-queue option comes into the discussion, we will reference Outcome C.

To achieve Outcomes A–C, the service provider offers both queue options but charges the priority fee K differently. If the service provider charges a sufficiently low K (specifically, when $K \leq \alpha(W_2 - W_1)$ holds), no customer will join the regular queue, even if it is offered. This leads to Outcome

B. If the service provider charges a sufficiently high K (specifically, when $K \geq \frac{(\alpha W_1 + R'')(W_2 - W_1)}{W_2}$ holds), no customer will join the priority queue, even if it is offered. This leads to Outcome C. For the intermediate range of K, when $\alpha(W_2 - W_1) < K < \frac{(\alpha W_1 + R'')(W_2 - W_1)}{W_2}$, customers will self-segment, and both queues are active. This leads to Outcome A.

In the following lemma, we derive a threshold on θ that corresponds to a limit on K beyond which no customers will join the regular queue. We define $R'' = R' - c$ and $\bar{\bar{\theta}} = \frac{1}{\rho} \frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda} + \sqrt{R''\lambda + 4}}$.

Lemma 2. *The following equation has a unique solution $\bar{\theta} \in (0, \bar{\bar{\theta}}]$:*

$$\alpha \bar{\theta}^2 \rho^2 = [R'' \mu (1 - \bar{\theta}\rho) - \bar{\theta}^2 \rho](1 - \bar{\theta}\rho). \tag{8}$$

Moreover, the following three regions of θ correspond precisely to Outcomes A–C:

(A) $0 < \theta < \bar{\theta}$. In this case, both queues are used by customers.

(B) $\bar{\theta} \leq \theta \leq \min\{\bar{\bar{\theta}}, 1\}$. In this case, $\xi(\theta) = 0$, and only the priority queue is used by customers.

(C) $\theta = 0$. In this case, $\xi(\theta) > 0$, and only the regular queue is used by customers.

When the service provider offers both queues, her only remaining decision variable is how high a priority fee K to charge. This determines the values of θ and ξ. In the analysis below, we will follow a mathematically equivalent approach, as before, by using θ as the decision variable and expressing both ξ and K as functions of θ. This allows us to simplify the expressions.

Before finding the optimal equilibrium θ (hence, optimal K) across the three outcomes, we first analyze each outcome separately in the following analysis. Then, we compare the total revenue rate to get the optimal global solution.

Outcome A: Customers join both queues

For this outcome to happen, K must be in an intermediate range to achieve Outcome A. This translates to a constraint $0 < \theta < \bar{\theta}$, by Lemma 2. The following lemma provides the expressions of regular customer proportion, priority fee, and revenue rate for θ in this range.

Lemma 3. *For any given* $\theta \in (0, \bar{\theta})$,

$$\begin{cases} \xi(\theta) = \dfrac{1}{\rho}\dfrac{2\sqrt{R''\lambda(1-\theta\rho)}}{\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}} - \theta > 0 \\ K(\theta,\xi(\theta)) = \dfrac{(\xi(\theta)+\alpha)(\theta+\xi(\theta))^2\rho^2}{\mu(1-(\theta+\xi(\theta))\rho)(1-\theta\rho)} = R'' - \dfrac{(\theta+\xi(\theta))^2\rho}{\mu(1-\theta\rho)} = \dfrac{R''\rho(\xi(\theta)+\alpha)}{(\alpha\rho+(1-\theta\rho))} \\ R(\theta,\xi(\theta)) = \lambda\theta K(\theta,\xi(\theta)) + \lambda(\theta+\xi(\theta))c \end{cases} . \quad (9)$$

Hence, the total fraction of served customers is

$$\theta + \xi(\theta) = \frac{1}{\rho}\frac{2\sqrt{R''\lambda(1-\theta\rho)}}{\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}} .$$

Further,

1) $\frac{\partial\xi(\theta)}{\partial\theta} < 0$, $\frac{\partial^2\xi(\theta)}{\partial\theta^2} < 0$, *and* $\frac{\partial(\theta+\xi(\theta))}{\partial\theta} < 0$, $\frac{\partial^2(\theta+\xi(\theta))}{\partial\theta^2} < 0$;

2) $\frac{\partial\xi(\theta)}{\partial\alpha} < 0$, $\frac{\partial^2\xi(\theta)}{\partial\alpha^2} > 0$ *and* $\frac{\partial(\theta+\xi(\theta))}{\partial\alpha} < 0$, $\frac{\partial^2(\theta+\xi(\theta))}{\partial\alpha^2} > 0$.

Note that $\theta, \xi(\theta) > 0$ correspond to the outcome that customers will join both queues. The second part of Lemma 3 shows that any increase in the priority customer proportion is accompanied by a concave

decrease in regular customer proportion and an increase in customer balking. It is also not surprising to find that a stronger customer perception of fairness will (concavely) lead to greater customer loss.

The revenue rate function in (9) is difficult to optimize, but the following two lemmas provide the conditions under which it is monotone. This makes it easier to compare with the revenue rate functions in the other two outcomes (to be derived later) to find the global optimal solution.

$$\text{Define } \bar{\alpha} = \max\left\{\left(\frac{\lambda c\sqrt{\rho} + \sqrt{\lambda^2 c^2 \rho + 8\rho^2 R''\lambda(1-\rho) + 4\rho^2 R''\lambda\sqrt{R''\lambda(1-\rho)}}}{2\rho\sqrt{R''\lambda(1-\rho)}}\right)^2;\ \frac{1}{\rho}\right\},$$

$$\bar{c}^{(0)} = \left(\sqrt{R''\lambda(1-\rho)} - \frac{R''\lambda + 2\sqrt{R''\lambda(1-\rho)}}{\alpha}\right)\frac{\sqrt{\alpha\rho}}{\lambda}, \text{ and}$$

$$\bar{c}^{(1)} = \frac{(\alpha\rho+1)(1+\alpha)\sqrt{(R''\lambda)^2 + 4R''\lambda(\alpha\rho+1)}\cdot[\sqrt{R''\lambda} + \sqrt{R''\lambda + 4(\alpha\rho+1)}]^2}{4\lambda\alpha(\alpha\rho + (1-\rho))^2}.$$

Lemma 4. *For any given* $\theta \in (0, \bar{\theta})$*:*

1) *If* $c \geq \bar{c}^{(1)}$*, then* $\frac{\partial R(\theta, \xi(\theta))}{\partial \theta} < 0$ *and* $\frac{\partial^2 R(\theta, \xi(\theta))}{\partial \theta^2} < 0$*. That is, if* $c \geq \bar{c}^{(1)}$*, the revenue rate is a concave, decreasing function of* θ *on* $(0, \bar{\theta})$.

2) *If* $c \leq \bar{c}^{(0)}$ *and* $\alpha \geq \bar{\alpha}$*, then* $\frac{\partial R(\theta, \xi(\theta))}{\partial \theta} > 0$ *and* $\frac{\partial^2 R(\theta, \xi(\theta))}{\partial \theta^2} < 0$*. That is, if* $\alpha \geq \bar{\alpha}$ *and* $c \leq \bar{c}^{(0)}$*, the revenue rate is a concave, increasing function of* θ *on* $(0, \bar{\theta})$.

Lemma 4 gives us mathematical properties of the revenue rate function $R(\theta, \xi(\theta))$ for $\theta \in (0, \bar{\theta})$ that we need to find the optimum across all θ. The properties also imply that, when the system is such that customers join both queues, a strong customer perception of fairness (i.e., when $\alpha \geq \bar{\alpha}$) will make the regular queue less appealing (due to the fairness disutility). Thus, more customers will join the priority queue. There may be more lost customers, but this is outweighed by the increase in priority fees. On the

contrary, when regular customers are quite profitable to begin with (i.e., $c \geq \bar{c}$), the service provider should strive to reduce possible loss of customers and their base fee *c*. Although the discussion is for $\theta \in (0, \bar{\theta})$, we will show later that these qualitative insights hold in general.

Outcome B: Customers join only the priority queue

Compared with the priority customers, the regular customers suffer a longer wait and a disutility due to fairness perception. Therefore, when K is small, few customers will join the regular queue. This corresponds to a large θ value. In the following lemma, we will show that, when $\theta \geq \bar{\theta}$, we must have $\xi(\theta) = 0$. It is important to note that, even if $K = 0$, there will be lost customers (hence, θ is strictly less than 1) if the system is too congested. The upper bound on θ is shown to be $\bar{\bar{\theta}}$, as defined just before Lemma 2.

Lemma 5. *For any given* $\theta \in [\bar{\theta}, \bar{\bar{\theta}}]$,

$$\begin{cases} \xi(\theta) = 0 \\ K(\theta, 0) = R'' - \dfrac{\theta^2 \rho}{\mu(1-\theta\rho)} \\ R(\theta, 0) = \lambda\theta K(\theta, 0) + \lambda\theta c \end{cases} . \tag{10}$$

Moreover, the optimal θ^* *that maximizes the revenue rate expressed in (10) is a root to the following cubic equation:*

$$2\rho^2\theta^3 + \rho[\mu\rho(R''+c) - 3]\theta^2 - 2\mu\rho(R''+c)\theta + \mu(R''+c) = 0 . \tag{11}$$

The cubic equation (11) has a unique solution, but they are unwieldy to use. Fortunately, as with Outcome A, we can characterize the optimal solution under certain conditions. The monotonicity of the revenue rate in θ is also helpful in determining the global optimal solution across all three outcomes.

We define two additional threshold values of c as: $\bar{c}^{(2)} = R''\left(\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda+4}-\sqrt{R''\lambda}}+2\right)$ and $\bar{c}^{(3)} = \frac{\rho}{\mu}\frac{(3-2\bar{\theta}\rho)\bar{\theta}^2}{(1-\bar{\theta}\rho)^2} - R''$.

Lemma 6. *For any given* $\theta \in [\bar{\theta}, \bar{\bar{\theta}}]$*:*

(1) If $c \geq \bar{c}^{(2)}$*, then* $\frac{\partial R(\theta,0)}{\partial \theta} \geq 0$*, and the optimal solution satisfies* $\theta^* = \bar{\bar{\theta}}$ *and* $K^* = 0$.

(2) If $\bar{c}^{(3)} < c < \bar{c}^{(2)}$*, then the optimal solution satisfies* $\bar{\theta} < \theta^* < \bar{\bar{\theta}}$ *and* $K^* > 0$.

(3) If $c \leq \bar{c}^{(3)}$*, then the optimal solution satisfies* $\theta^* = \bar{\theta}$ *and* $K^* > 0$.

In all three cases, $R^* = \lambda\theta^* c$.

Again, we see in Outcome B that, when the base fee c is large (i.e., $c \geq \bar{c}^{(2)}$), the service provider is better off focusing on keeping the regular customers and their regular fee c. It is actually optimal for the service provider to set $K = 0$ and to completely forego the priority fee. As a result, no customers will join the regular queue.

Outcome C: Customers join only the regular queue

This outcome is the most straightforward to characterize.

Lemma 7. *For* $\theta = 0$, $\xi = \bar{\bar{\theta}}$ *and* $R = \lambda\bar{\bar{\theta}}c$.

Now that we have characterized the revenue rate function in the three possible outcomes (which correspond to different intervals of the θ value), we are able to study the global optimal solution across the three outcomes. It is important to observe that the revenue function is continuous across the three outcomes. That is, the revenue functions from Outcomes A and B agree on the boundary $\theta = \bar{\theta}$**,** and the revenue functions from Outcomes A and C agree on the boundary $\theta = 0$.

Lemma 8. The revenue function formed by all three outcomes is continuous on all $\theta \in [0, \bar{\bar{\theta}}]$.

Although we could not obtain a close-form expression for the global optimal θ (hence, K), we are able to provide asymptotic results and implications for how the service provider should design the system.

Proposition 2: *In a non-captive service system:*

(1) If $\alpha \geq \bar{\alpha}$ *and* $c \leq \min\{\bar{c}^{(0)}, \bar{c}^{(2)}\}$ *then* $\theta^* > 0$, $\xi^* = 0$, *and* $K^* > 0$.

(2) If $c \geq \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$, *then either* $\theta^* > 0$, $\xi^* = 0$, *and* $K^* = 0$ *or* $\theta^* = 0$, $\xi^* > 0$, *and* $K^* = 0$.

When the service provider maximizes her revenue, she must make the fundamental decision of whether the priority fee or the base fee is more important. There is an inherent tradeoff between the two: When incentivizing more customers to join the priority queue, she will alienate the remaining regular customers due to longer wait and higher fairness disutility. This leads to higher loss of base fee and vice versa. The two parts of Proposition 3 address this trade-off.

When customers have a strong perception of fairness, they find the regular queue less appealing and are more likely to either join the priority queue or leave the system. If, in addition, the base fee *c* is low, then the former effect dominates the latter, and the service provider should optimally set a positive *K* such that customers join only the priority queue or leave. The regular queue is an option for the customers, but they will not use it. This is Part (1) of the proposition. In practical terms, the service provider also could set the priority queue as the standard option and give a discount of *K* to the regular queue, or she can simply provide customers with just one queue (the priority queue in our model).

If, on the other hand, the base fee *c* is sufficiently high, then each lost customer means a significant loss of revenue. In such a case, the service provider should focus on retaining as many customers as possible.

She can do that by setting *K* to be zero, in which case no customers will join the regular queue, essentially offering only one queueing option to the customers and not differentiating them. This is Part (2) of the proposition. Practically, if a firm considers the long-term implication of a customer balking, then the value of *c* should include some form of the customer's lifetime value. Indeed, for firms that highly value customer satisfaction and long-term profit, we recommend that the service not be differentiated due to the effect of fairness perception and the higher loss of customers as a result.

Although these two conclusions appear similar, they are fundamentally different. In the first part, $K > 0$, so the purpose of moving customers to the priority queue is to maximize the priority-fee revenue. In the second part, $K = 0$, so the purpose of moving customers to the priority queue is to remove fairness disutility and maximize base fee revenue.

For intermediate levels of the c value, where $\min\{\bar{c}^{(0)}, \bar{c}^{(2)}\} < c < \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$, analytical results are difficult to derive. Intuitively, however, we believe that the optimal action for the service provider is an extension of the actions in the two extreme ranges depicted in Proposition 2; somewhere in the middle, there should be a monotone switching curve (on the α-c parameter space) to separate the two designs. In the next section, we will numerically investigate this.

In our analysis of the non-captive system so far, the customer fairness perception parameter α is treated as a fixed variable. For the most part, customers' fairness perception is an innate property of the customers and the cultural environment in which they live (Fehr and Schmidt 1999, Kahneman *et al.* 1986, Rabin 1993) and can be considered exogenous to our analysis. The service provider may still be able to exert some influence over α, however, through the design of the service process. For example, some

providers actively promote their VIP priority services, while others do it in a more muted fashion. When a service provider puts the priority entrance right next to the regular queue, she is promoting the difference between the two lines and increasing the value of α. This has the advantage of potentially motivating more regular customers to upgrade to the priority queue. Conversely, when a service provider uses the regular exit point as the queue entrance for the priority customers, she is seeking to minimize the contrast between the two services so as not to incite a negative fairness perception by the regular customers. This has the effect of reducing the value of α.

For the captive system, Proposition 1 demonstrates that it benefits the service provider to have a high value of α. The underlying intuition is built on the captiveness of the system and, hence, does not carry over to the non-captive system. In fact, one would expect the system performance not to be monotone in α. This is difficult to show, however, due to the complexity of expressions in the non-captive system. In what follows, we will analyze the special case of $\alpha = 0$.

3.3. The $\alpha = 0$ queues

The case of $\alpha = 0$ merits special discussion for two reasons. First, if the regular customers do not see the priority queue at all because the queues are virtual (e.g., call center queues, back-office work queues), or if the service provider has separated the two services and made the priority queue imperceptible to the regular customers, then the fairness concern vanishes. We can call this the $\alpha = 0$ case. It also can occur if customers have completely embraced the concept of a priority fee and do not feel any unfairness.

Second, because fairness does not play any role when $\alpha = 0$, this special case corresponds to existing research that does not model fairness. Hence, it serves as a good benchmark system to study the

impact of fairness perception. In this section, we will characterize the $\alpha = 0$ case and contrast it with $\alpha > 0$.

Lemma 1 already provides the expressions of performance as a function of α in a captive system. The next lemma provides the expression for $\alpha = 0$ in a non-captive system.

Lemma 9. *In a non-captive system, the service provider's optimal solution in the* $\alpha = 0$ *case is as follows:*

$$\begin{cases} \theta^* = \dfrac{1}{\rho}\left(1 - \dfrac{\sqrt{R''\lambda + 4} - \sqrt{R''\lambda}}{2}\right) \\ K^* = R''\left(1 - \dfrac{\sqrt{R''\lambda + 4} - \sqrt{R''\lambda}}{2}\right) \\ R^* = \lambda R'' \dfrac{1}{\rho}\left(1 - \dfrac{\sqrt{R''\lambda + 4} - \sqrt{R''\lambda}}{2}\right)^2 + \lambda \bar{\bar{\theta}} c \end{cases} . \quad (12)$$

Using these expressions, we are able to derive further important managerial implications for the service provider.

Proposition 3:

(1) *In both the captive and non-captive service systems, when* $\alpha = 0$, *it is optimal for the service provider to charge a positive priority fee and to have customers join both queues (i.e.,* $K^* > 0$, $\theta^* > 0$, *and* $\xi^* > 0$*).*

(2) *In non-captive service system, when* $c \geq max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$, $R^*_{\alpha>0} < R^*_{\alpha=0}$. *That is, when the base fee c is high enough, the service provider derives higher revenue from the* $\alpha = 0$ *system than from any* $\alpha > 0$ *system.*

Part (1) of the proposition can be easily verified in the captive setting: Letting $\alpha = 0$ in Lemma 1, we see that θ^* and $\xi^* = 1 - \theta^*$ are both strictly positive, and so is K^*. Its proof for the non-captive

system is based on (12) and is a bit more involved (see the appendix). Whereas this result is consistent with the "conventional wisdom" in the literature (e.g., Gavirneni and Kulkarni 2011, Gavirneni and Kulkarni 2014), that segmented customer services plus differentiated pricing is an effective way to increase revenue for the service provider, it does not account for the fairness perception. As an important contrast, the results in Proposition 3 – that the service provider should focus on only one service offering under certain conditions – *do* account for the fairness perception. Such a contrast serves to highlight the importance of considering customers' fairness perception in a non-captive service system, which is the most distinguishing feature of our paper.

Finally, we ask the question: What if the service provider does not account for the fairness perception and sets the priority fee optimally based on $\alpha = 0$? The following proposition provides a glimpse of how customer queue-joining behavior will deviate:

Proposition 4: *In a non-captive service system, if the service provider adopts a fixed priority fee* $K^*_{\alpha=0}$, *then* θ *is increasing in* α *and* ξ *is decreasing in* α.

Proposition 4 is quite intuitive. It states that, for fixed pricing, customers are more likely to seek priority service if they feel more strongly about the fairness comparison between the queues and do not want to be stuck in the regular queue.

3.4. Numerical Studies for the Non-Captive Service System

Figure 1 provides a summary of the service provider's optimal actions in the captive service system. An analogous summary can be gathered from results in the non-captive service systems and is presented in Figure 2 below.

<table>
<tr><th></th><th>Small c
($c \le \min\{\bar{c}^{(0)}, \bar{c}^{(2)}\}$)</th><th>Intermediate c
($\min\{\bar{c}^{(0)}, \bar{c}^{(2)}\} < c < \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$)</th><th>Large c
($c \ge \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$)</th></tr>
<tr><td>$\alpha = 0$</td><td colspan="3">Two active queues, $K^* > 0$ (Proposition 3)</td></tr>
<tr><td>Small α</td><td>Two active queues, $K^* > 0$ (Conjecture confirmed by numerical tests)</td><td rowspan="2">Only one active priority queue, or two active queues, depending on the α and c values; the decisions are separated by a switching curve on the α-c parameter space (Conjecture confirmed by numerical tests)</td><td rowspan="2">Only one active queue (either regular queue only, $K^* = 0$ or priority queue only, $K^* > 0$) (Proposition 2 for $c \ge \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$)</td></tr>
<tr><td>Large α</td><td>Priority queue only, $K^* > 0$ (Proposition 2 for $\alpha \ge \bar{\alpha}$ and $c \le \min\{\bar{c}^{(0)}, \bar{c}^{(2)}\}$)</td></tr>
</table>

Figure 2: Service provider's optimal actions in the non-captive service systems ($\theta + \xi \le 1$)

We are able to analytically characterize the service provider's optimal action for most combinations of the α and c values in the non-captive service system, and the results provide managerial insights.

- When $\alpha = 0$, we show that the service provider should behave just as in the captive system: offer differentiated services and charge a positive fee for the priority option.
- As long as customers have non-zero fairness perception (i.e., $\alpha > 0$), the service provider needs to account for the negative impact of differentiated services: It may turn off some customers and cause higher loss. Therefore, if each lost customer carries a high value (i.e., when c is large), the service provider should design the system so that customers receive non-differentiated services. If c is small, however, we should expect the service provider to behave similarly to the case of captive systems.

Indeed, we are able to show that, when α is large, the service provider should have only priority customers. When α is small, the service provider should provide differentiated services and charge a positive fee for the priority service, and the intermediate range of c should see a mixture of the two extreme (large and small *c*) actions. Due to model complexity, we are able to give analytical results only for large *c*. We will numerically test the other cases below.

To test a wide range of parameter combinations we:

- normalize $\lambda = 1$ and $R'' = 1$ without loss of generality
- let $\mu =$ 1.1, 1.3, and 1.5 to represent various system loads
- let α vary between 0 and 1.2 to measure the intensity of the preference for fairness of the customer
- let c vary between 0 and 80 to represent the importance of the regular customer to the provider

For all of the parameter combinations, we numerically find the service provider's optimal service design decisions in the non-captive system and plot them in Figure 3.

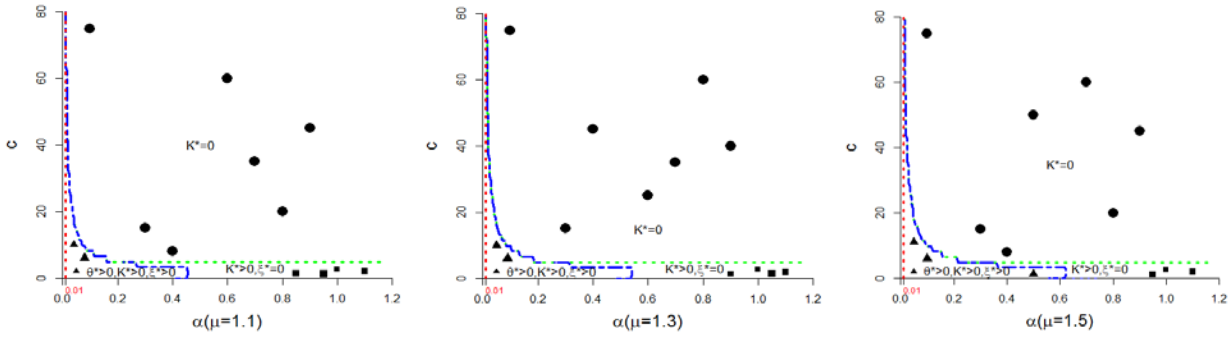

Figure 3: Service provider's optimal service design decisions in the non-captive system

The three panels of Figure 3 correspond to the cases of μ =1.1 (left), $\mu =$ 1.3 (middle) and $\mu =$ 1.5 (right), respectively. We are able to verity numerically that the service provider's optimal system design

can be any of the Outcomes A–C that we described earlier. In Figure 3, we mark the outcomes by triangles (Outcome A, both queues), squares (Outcome B, only priority queue), and circles (Outcome C, only regular queue). Further, the size of the markers represents the level of priority fee (the larger the marker, the higher the K^*). Of course, $K^* = 0$ in all Outcome C cases. Therefore, all of the circles are the same size.

The observations from Figure 3 confirm our analytical results and conjectures presented in Figure 2 for the non-captive system. For $\alpha = 0$, we already know that the service provider should always segment customers, which is verified in Figure 3 because the blue line never touches the vertical axis, meaning that the service provider should always adopt Outcome A when fairness does not matter ($\alpha = 0$). As soon as the fairness perception is positive (marked in Figure 3 by the $\alpha = 0.01$ vertical line), things change. In particular, when the base fee c is large (marked by green dash line), the service provider should choose to cancel the priority fee. The service provider collects no revenue from the priority queue but is better off because many fewer customers will balk, saving significant c amounts.

In contrast, when the base fee c is small, the service provider's optimal action switches from Outcome A to Outcome B when α increases, separated by the blue threshold curve: for small α, the service provider wants the customers to self-segment and charge for differentiated services. As α increases, regular customers will gradually decrease and even disappear. These analytical results from Proposition 2 are clearly confirmed in these numerical tests.

For the intermediate range of c ($\min\{\bar{c}^{(0)}, \bar{c}^{(2)}\} < c < \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$), we could not get close-form analytical results. Intuitively, we can conjecture that the service provider's optimal action should naturally extend the two extremes; numerically, we find this to be the case. When c is closer to the upper end of

$\max\{\bar{c}^{(1)},\bar{c}^{(2)}\}$, the service provider should focus on having just a regular queue and achieving Outcome C (marked by all of the circles). When *c* is closer to the lower end of $\min\{\bar{c}^{(0)},\bar{c}^{(2)}\}$, the service provider's optimal decision should change from Outcome A (triangles) to B (squares) as α increases (separated by the blue curve). The blue dash line marks the switch curve boundary between the two types of service provider actions. The curve boundaries in Figure 3 not only confirm our analytical results and conjectures, they are also quite novel and insightful, providing an additional layer of consideration for the service provider when designing their service process.

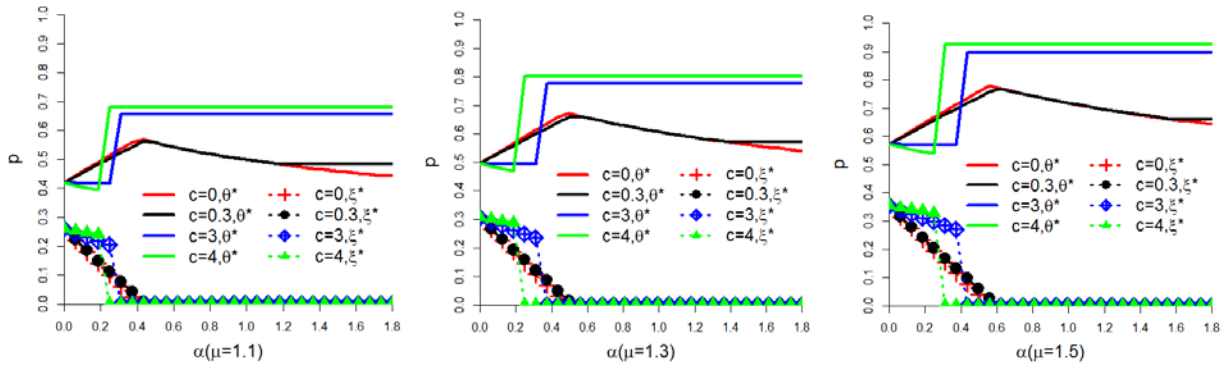

Figure 4: The optimal proportions of priority customers and regular customers

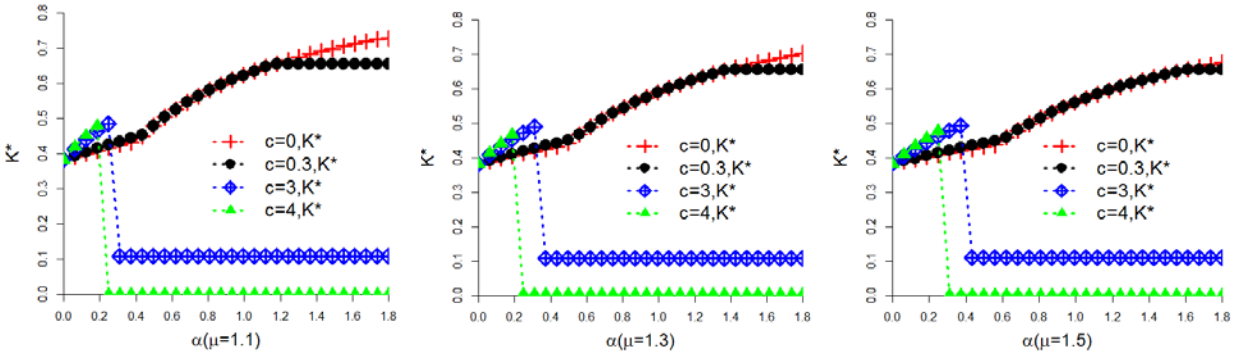

Figure 5: The optimal priority fee fees

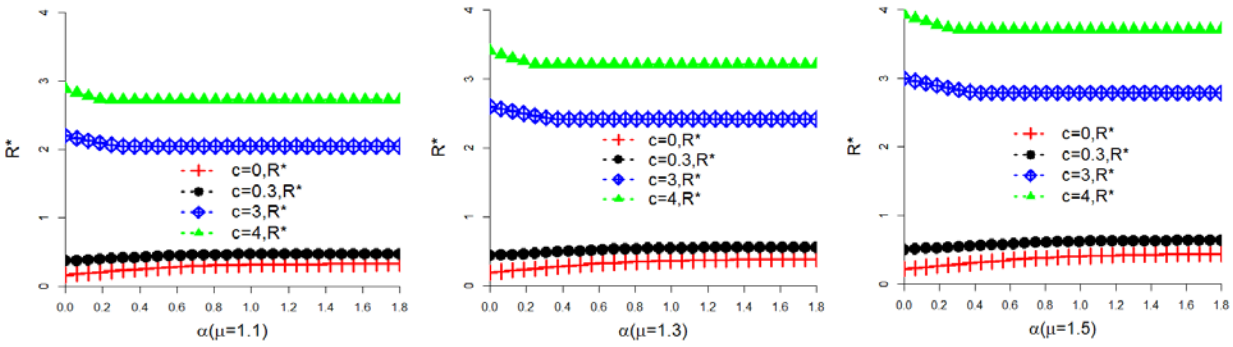

Figure 6: The maximization revenues

For Figures 4–6, we tested values of c from 0 to 5 in steps of 0.1 but show only two lower values of $c = 0, 0.3$ and two higher values of $c = 3, 4$ to illustrate the contrast. In addition, we tested values of μ from 1.1 to 2 in steps of 0.1, but show only three (1.1, 1.3, 1.5) here. The results presented here are robust and representative of the broader set of parameters that we have tested; they again confirm our analytical results and insights.

Figure 4 shows how the customer queue choices change with the values of α and c. Analytically, we already know that, for large c values, the system will be designed such that customers will choose only the regular queue (or priority queue with $K = 0$). Now, numerically, we can see from Figure 4 that, even for small values of c, the customers do not need to have a very high fairness perception value for customer segmentation to be suboptimal. In these cases, the service provider should induce all customers to choose the same queueing option.

Figure 5 not only reiterates the dichotomy on the c value (e.g., higher c value causes the service provider to focus on only one queue), but also clearly illustrates that the optimal priority access fee K^* is not necessarily a monotone function of α. A moderately higher α makes customers more willing to pay

extra K to avoid the fairness disutility experienced in the regular queue. There will be greater customer loss, but it is more than compensated by the extra fee from K. Therefore, the service provider focuses on the extra fee and increases K^* for higher α. For small c, customer loss is not costly, so the provider consistently raises K^*. For large c, however, customer loss becomes more important as α gets higher. After a certain point, the service provider should drop the optimal fee to $K^* = 0$ and focus on only one customer queue.

Figure 6 shows that the optimal revenue is decreasing in α when c is large, while it is increasing when c is small. This perfectly illustrates the interaction between α and c, just as we saw in Figure 5, but from a different perspective: A higher perception of unfairness α drives some customers away but, at the same time, also induces some regular customers to upgrade to priority. When c is large, the former effect dominates, so the service provider's revenue suffers when customers are more perceptive of fairness. Hence, she should try to minimize the contrast of the two queueing options as much as possible. On the contrary, when c is small, the latter effect dominates, and a higher α actually benefits the service provider, as it allows her to have a higher degree of differentiated pricing. Hence, she should seek to promote the two queue options and highlight the contrast between them. We find this to be an insightful, yet intuitive, observation.

4. Conclusions and Further Work

Our study is among the first to analyze the impact of fairness perception in service systems. A stylized, non-preemptive *M/M/1* queueing model is used to capture the essential system design decisions for the service

provider – whether to provide a priority queue and how much to charge for it – and queue-joining decisions for the customers that involve a fairness disutility due to the comparison of the two queues. After solving the optimal equilibrium actions by both the service provider and the customers, we are able to establish the benefit to the service provider by carefully chosen differentiated service and pricing in the captive system where customers do not balk. It is also true that the higher customer perception of fairness, the higher the revenue for the service provider, as more of the customers will pay to use the priority service.

The results for the non-captive systems in which customers can balk are not so straightforward. We find the regular customer base fee c and customer fairness perception α to play important roles in deciding how the service system is designed and charged. The most interesting result is that differentiated service may cause too many regular customers to balk when customers' fairness perception is high (large α). If, in addition, the regular customer value c is high, then such lost revenue cannot be compensated by the priority fee. In such cases, the service provider should actually forego customer segmentation and focus instead on providing one type of service to all customers. This challenges conventional wisdom but is very reasonable in the context of customer fairness perception.

To the extent that the service provider can influence the customer perception of fairness, for example, through placement of queues and promotion of queue choices, we also find conditions under which a heightened fairness perception could be beneficial or detrimental to the service provider. These new findings add to the collective knowledge on how to manage differentiated services and pricing, and constitute significant contribution to the research on this topic.

To establish a reasonable and tractable framework, and to derive insightful results, we have made simplifying assumptions about the service process and customer choices. Additional work can be done to relax these assumptions and further extend our understanding of this problem. For analytical tractability, we have used a uniform distribution for *H*. We conjecture that the main structural results in our paper will continue to hold for other types of distribution, but it would be nice to quantify what types of distribution.

In addition, it would be worthwhile to address the following: What if customers are heterogeneous on more than just the waiting time cost (e.g., they also differ on service value)? If customers can make upgrading decisions in real time after seeing the queues (if the service provider allows that), then their queue-joining decisions would certainly change. How will that affect the optimal system design? More interestingly, when more than two priority classes are possible, how do customers even perceive fairness in such a context (for example, when customers are overtaking some customers but being overtaken by others at the same time)? Field data and experimental work are needed to answer these questions before appropriate assumptions can be made in analytical models.

To keep the paper focused, we have studied only the service provider's objective to maximize revenue. It would be important to also understand the impact of differentiated service on each customer's waiting time, the variation of waiting time among all customers, customers' overall utility, and even the social welfare perspective of this issue.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 71671092; 71301075), the National Natural Science Foundation of Jiangsu Province, China (No. BK20130770), the International Postdoctoral Exchange Fellowship Program of China (No. 20140072), and the Postdoctoral Science Foundation funded project of Jiangsu Province, China (No. 1501040A).

References

Afèche P (2013) Incentive-Compatible Revenue Management in Queueing Systems: Optimal Strategic Delay. *Manufacturing & Service Operations Management*,15(3): 423-443.

Afèche P, Pavlin J M (2016) Optimal Price/Lead-Time Menus for Queues with Customer Choice: Segmentation, Pooling, and Strategic Delay. *Management Science*, 62(8): 2412-2436.

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ Rule for Many-Server Queues with Abandonment. *Operations Research*, 58(5):1427-1439.

Avi-itzhak B, Levy H, Raz D (2008) Quantifying fairness in queuing systems: Principles, approaches, and applicability. *Probability in the Engineering and Informational Sciences*, 22(4): 495-517.

Bassamboo A, Randhawa R S (2015) Scheduling Homogeneous Impatient Customers. *Management Science*, 62(7):2129-2147.

Bertsimas D, Farias V F, Nikolaos Trichakis N (2012) On the Efficiency-Fairness Trade-off. *Management Science*, 58(12):2234-2250.

Boiney L G (1995) When Efficient Is Insufficient: Fairness in Decisions Affecting a Group. *Management Science*, 41(9):1523-1537.

Cachon G, Terwiesch C (2012) Matching Supply with Demand: An Introduction to Operations Management, 3rd edition. McGraw-Hill Education, New York, USA

Choi B D, Kim B, Chung J M (2001) M/M/1Queue with Impatient Customers of Higher Priority. *Queueing Systems*, 38(1):49-66.

Cox D R, Smith W L (1961) Queues. Methuen, London, UK.

Cui S L, Wang J T, Wang Z B (2017) Equilibrium Strategies in M/M/1 Queues with Priorities. Working Paper, Georgetown University, Washington, D.C.

Cui T H, Raju J S, Zhang Z J (2007) Fairness and Channel Coordination. *Management Science*, 53(8): 1303-1314.

Davis R H (1966) Waiting-Time Distribution of a Multi-Server, Priority Queueing System. *Operations Research*, 14(1): 133-136.

Fehr E, Schmidt K M (1999) A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3): 817-868.

Gavirneni S, Kulkarni V G (2011) Concierge Option for a Service Offering: Design, Analysis, Impact, and Adoption. Working Paper, Cornell University, Ithaca, NY.

Gavirneni S, Kulkarni V G (2014) Concierge Medicine: Applying Rational Economics to Health Care Queuing. *Cornell Hospitality Quarterly*, 55(3): 314-325.

Gavirneni S, Kulkarni V G (2016) Self-Selecting Priority Queues with Burr Distributed Waiting Costs. *Production and Operations Management*, 25(6): 979-992.

Geng X, Huh W T, Nagarajan M (2015) Fairness Among Servers When Capacity Decisions Are Endogenous. *Production and Operations Management*, 24(6): 961-974.

Gross D, Harris C M (1998) Fundamentals of Queueing Theory, Third Edition. Wiley, NY.

Hassin R (2016) Rational Queueing. CRC Press, Taylor & Francis Group, Boca Raton, FL.

Hassin R, Haviv M (2003) To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems. Springer, Norwell, MA.

Ho T H, Su X M (2009) Peer-Induced Fairness in Games. *The American Economic Review*, 99(5): 2022-2049.

Iravani F, Balcıoğlu B (2008) On priority queues with impatient customers. *Queueing Systems*, 58(4): 239-260.

Jin L Y, He Y Q, Zhang Y (2014) How Power States Influence Consumers' Perceptions of Price Unfairness. *Journal of Consumer Research*, 40(5): 818-833.

Kahneman D, Knetsch J L, Thaler R H (1986) Fairness and the Assumptions of Economics. *The Journal of Business*, 59(4): Part2: The Behavioral Foundations of Economic Theory, S285-S300.

Kleinrock L (1967) Optimum Bribing for Queue Position. *Operations Research*, 15(2):304-318.

Kleinrock L (1975) Queueing Systems. Volume 1: Theory. Wiley-Interscience, NY.

Kulshreshtha P (2003) Rationing by Waiting, Opportunity Costs of Waiting and Bribery. *Indian Economic Review*, 38(1):59-75.

Lajos T (1968) Two Queues Attended by a Single Server. *Operations Research*, 16(3):639-650.

Lariviere M (2014) How should a supermarket organize its checkout lanes? *Kellogg insight presents* (T*he Operations Room*). October 17, 2014, https://operationsroom.wordpress.com/2014/10/17/how-should-a-supermarket-organize-its-checkout-lanes/

Larson R C, (1987) Perspectives on Queues: Social Justice and the Psychology of Queueing. *Operations Research*, 35(6):895-905.

Li K J, Jain S (2016) Behavior-Based Pricing: An Analysis of the Impact of Peer-Induced Fairness. *Management Science*, 62(9): 2705-2721.

Maister D (1985) The Psychology of Waiting Lines, In J. A. Czepiel, M. R. Solomon & C. F. Surprenant (Eds.), The Service encounter: managing employee/customer interaction in service businesses. Lexington, MA: D. C. Heath and Company, Lexington Books.

Morrow L (1984) Waiting as way of life. *Time*, July 23, 1984, P.65.

Nageswaran L, Scheller-Wolf A (2017) Queues with Redundancy: Is Waiting in Multiple Lines Fair?. Working Paper, Carnegie Mellon University, Pittsburgh, PA

Nazerzadeh H, Randhawa R S (2017) Near-Optimality of Coarse Service Grades for Customer Differentiation in Queueing Systems. *Production and Operations Management*, published online at http://onlinelibrary.wiley.com/doi/10.1111/poms.12818/full

Nicholson W, Snyder C M (2011) Microeconomic Theory: Basic Principles and Extension, 11th edition. South-Western College, Chula Vista, CA.

Norman D A (2009) Designing Waits That Work. *MIT Sloan Management Review*, 50(4): 23-28.

Pinedo M L (1983) Stochastic Scheduling with Release Dates and Due Dates. *Operations Research*, 31(3):559-572.

Pinedo, M L (2016) Scheduling: Theory, Algorithms, and Systems, 5th Edition. Springer, NY.

Rabin M (1993) Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, 83(5):1281-1302.

Rafaeli A, Barron G, and Haber K (2002) The Effects of Queue Structure on Attitudes. *Journal of Service Research*, 5(2):125-139.

Rafaeli A, Kedmi E, Vashdi D, Barron, G (2005) Queues and Fairness: A multiple Study Experimental Investigation. Technical report, Israel Institute of Technology, Haifa, Israel.

Shunko M, Niederhoff J, Rosokha, Y (2017) Humans Are Not Machines: The Behavioral Impact of Queueing Design on Service Time. *Management Science*, https://doi.org/10.1287/mnsc.2016.2610

Sim N (2010) Editorial: Express Pass to hell - why paid-for queue-jumping systems have got to go, Theme Park Tourist, March 25, 2010, http://www.themeparktourist.com/features/20100325/1314/editorial-express-pass-hell-why-paid-queue-jumping-systems-have-got-go

Smith W E (1956) Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 3(1-2):59-66.

Stone A (2012) Why Waiting is Torture. *The New York Times*, August 19, 2012, P. Sunday Review 12.

Sztrik J (2016) Basic Queueing Theory. GlobeEdit **S**aarbrucken, Germany.

Van Mieghem J A (1995) Dynamic Scheduling with Convex Delay Costs: The Generalized $c\mu$ Rule. *Annals of Applied Probability*, 5(3): 809-833.

Wang J, Zhou Y P (2017) Impact of Queue Configuration on Service Time: Evidence from a Supermarket. *Management Science,* http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2017.2781

Wang J F, Baron O, Scheller-Wolf A (2015) M/M/c Queue with Two Priority Classes. *Operations Research*, 63(3): 733-749.

Ward A R, Armony M (2013) Blind Fair Routing in Large-Scale Service Systems with Heterogeneous Customers and Servers. *Operations Research*, 61(1): 228-243.

Wu X L, Niederhoff J A (2014) Fairness in Selling to the Newsvendor. *Production and Operations Management*, 23(11): 2002-2022.

Xia L, Monroe K B, Cox J L (2004) The Price Is Unfair! A Conceptual Framework of Price Fairness Perceptions. *Journal of Marketing*, 68(4): 1-15.

Yang L Y, Debo L, Gupta V (2017) Trading Time in a Congested Environment. *Management Science*, 63(7):2377-2395.

Appendix

Proof of Lemma 1.

In a captive service system, a customer joins the priority queue if and only if $U_1 \geq U_2$. For a given value of θ (recall that using θ as the decision variable is equivalent to using K), the service provider's priority fee and revenue are found to be:

$$\begin{cases} K(\theta)_{(\alpha)} = (1-\theta+\alpha)(W_2 - W_1) = \dfrac{\rho^2}{\mu(1-\rho)(1-\theta\rho)}(1-\theta+\alpha) \\ R(\theta)_{(\alpha)} = \lambda\theta(K(\theta)_{(\alpha)} + c) + \lambda\xi c = \dfrac{\rho^3\theta}{(1-\rho)(1-\theta\rho)}(1-\theta+\alpha) + \lambda c \end{cases} \quad (0 \leq \theta \leq 1). \qquad \text{(A1)}$$

To optimize $R(\theta)_{(\alpha)}$ subject to $0 \leq \theta \leq 1$, we first find the first order derivative of the objective function:

$$\frac{\partial R(\theta)_{(\alpha)}}{\partial \theta} = \frac{\rho^3}{(1-\rho)}\left\{\frac{1}{(1-\theta\rho)^2}\left[(1-\theta)+\alpha\right] - \frac{\theta}{(1-\theta\rho)}\right\} = \frac{\rho^3(\theta^2\rho - 2\theta + 1 + \alpha)}{(1-\rho)(1-\theta\rho)^2}.$$

Then we find the second order derivative at any θ satisfying the first order condition $\dfrac{\partial R(\theta)_{(\alpha)}}{\partial \theta} = 0$ to be negative:

$$\frac{\partial^2 R(\theta)_{(\alpha)}}{\partial \theta^2} = -\frac{2\rho^3}{(1-\rho)}\frac{1}{(1-\theta\rho)} < 0.$$

Thus, we can find the optimal solutions via first order condition:

(1) If $0 \leq \alpha < 1-\rho$, (we use the superscript * notation to indicate optimal solutions):

$$\left\{\theta^*_{(\alpha)} = \frac{1-\sqrt{1-\rho(1+\alpha)}}{\rho};\quad K^*_{(\alpha)} = \frac{\rho(1-\sqrt{1-\rho(1+\alpha)})}{\mu(1-\rho)};\quad R^*_{(\alpha)} = \frac{\rho(1-\sqrt{1-\rho(1+\alpha)})^2}{(1-\rho)} + \lambda c. \right. \qquad \text{(A2)}$$

(2) If $\alpha \geq 1-\rho$,

$$\left\{\theta^*_{(\alpha)} = 1;\quad K^*_{(\alpha)} = K(1)_{(\alpha)} = \frac{\rho^2}{\mu(1-\rho)^2}\alpha;\quad R^*_{(\alpha)} = R(1)_{(\alpha)} = \frac{\rho^3}{(1-\rho)^2}\alpha + \lambda c. \right. \qquad \text{(A3)}$$

Proof of Proposition 1.

The monotonicity of $K^*_{(\alpha)}$, $\theta^*_{(\alpha)}$, and $R^*_{(\alpha)}$ in both cases is evident from equations (A2) and (A3). Moreover,

$$\frac{\partial W^*_{1\,(\alpha)}}{\partial \alpha} = \frac{\rho^2}{\mu(1-\theta^*_{(\alpha)}\rho)^2}\frac{\partial \theta^*_{(\alpha)}}{\partial \alpha} > 0 \quad \text{and} \quad \frac{\partial W^*_{2(\alpha)}}{\partial \alpha} = \frac{\rho^2}{\mu(1-\rho)(1-\theta^*_{(\alpha)}\rho)^2}\frac{\partial \theta^*_{(\alpha)}}{\partial \alpha} > 0. \tag{A4}$$

This means, both equilibrium expected waiting time $W^*_{1\,(\alpha)}$ and $W^*_{2(\alpha)}$ increase with α.

Proof of Lemma 2.

In a non-captive service system, the service provider provides two queueing options - priority and regular - to the customers and charges $K \geq 0$ to use the priority queue. Customers also have the option of balking and getting $U_0 = 0$. Comparing the expressions in (1) - (3), we can split customers into balking, regular, and priority proportions: $1-\theta-\xi$, ξ and θ. It is possible that the service provider prices its priority queue in such a way that customers end up choosing to use only one (or none) of them. Therefore, we discuss these scenarios separately.

Scenario 1) $\theta > 0$ and $\xi > 0$.

In this case, we have $W_1 > 0$ and $W_2 > 0$ and

$$\begin{aligned} \theta &= P\left(\frac{K}{W_2 - W_1} - \alpha < H < \frac{R'' - K}{W_1}\right) \\ \xi &= P\left(H \leq \min\left\{\frac{K}{W_2 - W_1} - \alpha; \frac{R'' - \alpha(W_2 - W_1)}{W_2}\right\}\right) \\ 1-\theta-\xi &= P\left(H \geq \max\left\{\frac{R'' - \alpha(W_2 - W_1)}{W_2}; \frac{R'' - K}{W_1}\right\}\right). \end{aligned} \tag{A5}$$

Clearly, we have $\lambda_0 = \lambda \cdot (1-\theta-\xi)$, $\lambda_1 = \lambda \cdot \theta$, $\lambda_2 = \lambda \cdot \xi$, and $\lambda_0 + \lambda_1 + \lambda_2 = \lambda$.

Note that if $\frac{R''-K}{W_1} \geq 1$ then we have $\lambda_0 = 0$. This is a captive system. If $\frac{R''-K}{W_1} \leq 0$ then $\theta = 0$. Both violate the condition for Scenario 1). Thus, we will focus on $0 < \frac{R''-K}{W_1} < 1$.

For $\theta > 0$, we must have $\frac{K}{W_2 - W_1} - \alpha < \frac{R''-K}{W_1}$. This means $K < \frac{(\alpha W_1 + R'')(W_2 - W_1)}{W_2}$, from which we can directly obtain the following two inequalities after simple algebraic manipulation:

$$\frac{R''-K}{W_1} > \frac{R''-\alpha(W_2 - W_1)}{W_2} \quad \text{and} \quad \frac{R''-\alpha(W_2 - W_1)}{W_2} > \frac{K}{W_2 - W_1} - \alpha .$$

Then, (A5) can be simplified to

$$\begin{aligned} \theta &= P\left(\frac{K}{W_2 - W_1} - \alpha < H < \frac{R''-K}{W_1} \right) \\ \xi &= P\left(H \leq \frac{K}{W_2 - W_1} - \alpha \right) \\ 1 - \theta - \xi &= P\left(H \geq \frac{R''-K}{W_1} \right). \end{aligned} \tag{A6}$$

The second one gives us $K = (\xi + \alpha)(W_2 - W_1)$, and the last inequality gives us $K = R'' - (\theta + \xi)W_1$.

Plugging in W_1 and W_2 as functions of θ and ξ, we get the following two equations:

$$K = \frac{(\xi + \alpha)(\theta + \xi)^2 \rho^2}{\mu(1 - (\theta + \xi)\rho)(1 - \theta\rho)} = R'' - \frac{(\theta + \xi)^2 \rho}{\mu(1 - \theta\rho)}, \tag{A7}$$

Using θ as the free variable, we can solve ξ as its function, and rewrite (A7) as $(\theta + \xi)^2 \rho = \frac{R''\mu(1 - (\theta + \xi)\rho)(1 - \theta\rho)}{\alpha\rho + (1 - \theta\rho)}$. Since $\theta + \xi \geq 0$, we get a unique solution

$$(\theta + \xi) = \frac{-R''\mu(1 - \theta\rho)\rho + \sqrt{\left(R''\mu(1 - \theta\rho)\rho\right)^2 + 4\rho\left(\alpha\rho + (1 - \theta\rho)\right)R''\mu(1 - \theta\rho)}}{2\rho\left(\alpha\rho + (1 - \theta\rho)\right)}.$$

Hence

$$\xi(\theta) = \frac{1}{\rho} \frac{2\sqrt{R''\lambda(1 - \theta\rho)}}{\sqrt{R''\lambda(1 - \theta\rho)} + \sqrt{R''\lambda(1 - \theta\rho) + 4(\alpha\rho + (1 - \theta\rho))}} - \theta . \tag{A8}$$

By examining the first order derivative of $\xi(\theta)$, we show that it is a decreasing function of θ:

$$\frac{\partial \xi(\theta)}{\partial \theta} = -\frac{\dfrac{4R''\lambda\alpha\rho}{\sqrt{R''\lambda(1-\theta\rho)}\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}}{\left(\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}\right)^2} - 1 < 0. \quad \text{(A9)}$$

With this monotonicity, the maximum value of θ that satisfies $\xi(\theta) > 0$ must be the one where $\xi(\theta) = 0$ in (A8). Denote it by $\bar{\theta}$. Then $\bar{\theta}$ satisfies

$$\alpha\bar{\theta}^2\rho^2 = [R''\mu(1-\bar{\theta}\rho) - \bar{\theta}^2\rho](1-\bar{\theta}\rho). \quad \text{(A10)}$$

Next, we show a bound on $\bar{\theta}$. Based on (A10), we get

$$[\bar{\theta}^2 + \alpha 2\bar{\theta}\frac{\partial\bar{\theta}}{\partial\alpha}]\rho^2 = (R''\mu(-\frac{\partial\bar{\theta}}{\partial\alpha}\rho) - 2\bar{\theta}\frac{\partial\bar{\theta}}{\partial\alpha}\rho)(1-\bar{\theta}\rho) + (R''\mu(1-\bar{\theta}\rho) - \bar{\theta}^2\rho)(-\frac{\partial\bar{\theta}}{\partial\alpha}\rho).$$

So,

$$\frac{\partial\bar{\theta}}{\partial\alpha} = -\frac{\bar{\theta}^2}{\alpha 2\bar{\theta} + \frac{1}{\rho}(R''\mu + 2\bar{\theta})(1-\bar{\theta}\rho) + \frac{1}{\rho}(R''\mu(1-\bar{\theta}\rho) - \bar{\theta}^2\rho)} < 0.$$

The largest value $\bar{\theta}$ can achieve is at $\alpha = 0$. When we plug $\alpha = 0$ into (A10), we get $0 = [R''\mu(1-\bar{\theta}\rho) - \bar{\theta}^2\rho](1-\bar{\theta}\rho)$. Since $\bar{\theta}$ cannot be negative and $1-\bar{\theta}\rho > 0$, we must have the following equation when $\alpha = 0$:

$$0 = [R''\mu(1-\bar{\theta}\rho) - \bar{\theta}^2\rho] = R''\mu - R''\mu\rho\bar{\theta} - \bar{\theta}^2\rho.$$

So,

$$\bar{\theta}\,|_{any\ \alpha} \le \bar{\theta}\,|_{\alpha=0} = \frac{-R''\mu\rho + \sqrt{(R''\mu\rho)^2 + 4R''\mu\rho}}{2\rho} = \frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda} + \sqrt{R''\lambda + 4}} = \bar{\bar{\theta}}. \quad \text{(A11)}$$

To simplify exposition, we define an auxiliary function by using Eq.(A10):

$$f(\bar{\theta}) = [R''\mu(1-\bar{\theta}\rho) - \bar{\theta}^2\rho](1-\bar{\theta}\rho) - \alpha\bar{\theta}^2\rho^2 = R''\mu - 2R''\mu\bar{\theta}\rho + R''\mu\bar{\theta}^2\rho^2 - \bar{\theta}^2\rho + \bar{\theta}^3\rho^2 - \alpha\bar{\theta}^2\rho^2.$$

We can find the derivative of $f(\bar{\theta})$ as follows:

$$\frac{\partial f(\bar{\theta})}{\partial \bar{\theta}} = -2R''\mu\rho + 2R''\mu\bar{\theta}\rho^2 - 2\bar{\theta}\rho + 3\bar{\theta}^2\rho^2 - 2\alpha\bar{\theta}\rho^2$$

Based on Eq. (A11) we must have $\bar{\theta} \le \bar{\bar{\theta}}$ and $0 = R''\mu - R''\mu\rho\bar{\bar{\theta}} - \bar{\bar{\theta}}^2\rho$. Thus, we will get $R''\mu - R''\mu\rho\bar{\theta} - \bar{\theta}^2\rho > 0$ when $\alpha > 0$. So, $-2R''\mu\rho + 2R''\mu\bar{\theta}\rho^2 + 2\bar{\theta}^2\rho^2 < 0$. We can rewrite the $\frac{\partial f(\bar{\theta})}{\partial \bar{\theta}} = (-2R''\mu\rho + 2R''\mu\bar{\theta}\rho^2 + 2\bar{\theta}^2\rho^2) - 2\bar{\theta}\rho(1 - \frac{1}{2}\bar{\theta}\rho) - 2\alpha\bar{\theta}\rho^2 < 0$. With this monotonicity, the equation (A10) has a unique solution for $\bar{\theta} \in (0, \bar{\bar{\theta}}]$.

So far, we have shown that if in Scenario 1), we must have $0 < \theta < \bar{\theta}$. Conversely, any $0 < \theta < \bar{\theta}$ and $\xi(\theta)$ as in (A8) will lead to Scenario 1. This proves part A) of Lemma 2. That is, Outcome A happens if and only if $0 < \theta < \bar{\theta}$.

Scenario 2) $\xi = 0$ and $\theta > 0$.

This means customers only join the priority queue or balk. Please note that the system still provides the two queue options; it's just that customers do not find it beneficial to join the regular queue. Therefore, as each customer considers whether to join the regular queue, the formulas $W_1 = \frac{\rho(\theta+\xi)}{\mu(1-\rho\theta)}$ and $W_2 = \frac{\rho(\theta+\xi)}{\mu(1-\rho(\theta+\xi))(1-\rho\theta)}$ remain valid with $\xi = 0$.

So, we have $W_1 = \frac{\rho\theta}{\mu(1-\rho\theta)}$ and $W_2 = \frac{\rho\theta}{\mu(1-\rho\theta)^2} = \frac{W_1}{1-\rho\theta}$. Moreover, the U_2 term should still contain the fairness disutility, $\alpha(W_2 - W_1) = \frac{\alpha\rho\theta W_1}{1-\rho\theta}$.

Just as in Scenario 1), $\theta > 0$ means $\frac{R'' - K}{W_1} > \frac{R'' - \alpha(W_2 - W_1)}{W_2} > \frac{K}{W_2 - W_1} - \alpha$.

Then, $\xi = P\left(H \le \min\left\{\frac{K}{W_2 - W_1} - \alpha; \frac{R'' - \alpha(W_2 - W_1)}{W_2}\right\}\right) = P\left(H \le \frac{K}{W_2 - W_1} - \alpha\right)$. The fact that $\xi = 0$ means $\frac{K}{W_2 - W_1} - \alpha \le 0$. There are two immediate implications:

1. $$K \le \alpha(W_2 - W_1) = \alpha \frac{\theta\rho W_1}{1-\theta\rho} = \alpha \frac{\theta^2\rho^2}{\mu(1-\theta\rho)^2}. \qquad \text{(A12)}$$

2. $\theta = P\left(\frac{K}{W_2 - W_1} - \alpha \le H \le \frac{R'' - K}{W_1} \right) = P\left(0 \le H \le \frac{R'' - K}{W_1} \right) = \frac{R'' - K}{W_1}$. From this we get

$$K = R'' - \theta W_1 = R'' - \frac{\theta^2\rho}{\mu(1-\theta\rho)}. \qquad \text{(A13)}$$

Combining (A12) and (A13), we get

$$R'' - \frac{\theta^2\rho}{\mu(1-\theta\rho)} \le \frac{\alpha\theta^2\rho^2}{\mu(1-\theta\rho)^2}.$$

The left hand side starts at R'' and is decreasing in θ; the right hand side starts at 0 but is increasing in θ. For the inequality to hold, we must have θ greater than or equal to the intersection of the two sides. The intersection can be found as:

$$R'' - \frac{\rho\theta^2}{\mu(1-\rho\theta)} = \frac{\alpha\rho^2\theta^2}{\mu(1-\rho\theta)^2} \qquad \Rightarrow \qquad \alpha\rho^2\theta^2 = \left[R''\mu(1-\theta\rho) - \theta^2\rho \right](1-\theta\rho).$$

This is exactly equation (A10), which defines $\overline{\theta}$, so $\theta \ge \overline{\theta}$. It is easy to see that θ can be no more than $\overline{\overline{\theta}}$. Therefore, $\overline{\theta} \le \theta \le \min\{\overline{\overline{\theta}}, 1\}$.

What we have shown is that in Scenario 2, we must have $\overline{\theta} \le \theta \le \min\{\overline{\overline{\theta}}, 1\}$. Recall also that $\theta \ge \overline{\theta}$ and $\xi > 0$ is impossible (at the end of the proof of Scenario 1). Therefore, we have shown that Outcome B (part B of Lemma 2) happens if and only if $\overline{\theta} \le \theta \le \min\{\overline{\overline{\theta}}, 1\}$.

Scenario 3) $\xi > 0$ and $\theta = 0$.

This case corresponds to Outcome C precisely. That is, only the regular queue is used by customers. This proves part C) of Lemma 2.

Proof of Lemma 3.

For any given $0 < \theta < \overline{\theta}$, customers join both two queues in the equilibrium. We will get the following functions for priority fee and revenue by using Eqs. (A6) - (A8).

$$\begin{cases} \xi(\theta) = \dfrac{1}{\rho}\dfrac{2\sqrt{R"\lambda(1-\theta\rho)}}{\sqrt{R"\lambda(1-\theta\rho)}+\sqrt{R"\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}} - \theta > 0 \\ K(\theta,\xi(\theta)) = \dfrac{(\xi(\theta)+\alpha)(\theta+\xi(\theta))^2\rho^2}{\mu(1-(\theta+\xi(\theta))\rho)(1-\theta\rho)} = R" - \dfrac{(\theta+\xi(\theta))^2\rho}{\mu(1-\theta\rho)} = \dfrac{R"\rho(\xi(\theta)+\alpha)}{(\alpha\rho+(1-\theta\rho))} \\ R(\theta,\xi(\theta)) = \lambda\theta K(\theta,\xi(\theta)) + \lambda(\theta+\xi(\theta))c \end{cases} \quad \text{(A14)}$$

The first order derivative of $\theta + \xi(\theta)$ is shown based on Eq. (A8).

$$\frac{\partial(\theta+\xi(\theta))}{\partial\theta} = -\frac{\dfrac{4R"\lambda\alpha\rho}{\sqrt{R"\lambda(1-\theta\rho)}\sqrt{R"\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}}{\left(\sqrt{R"\lambda(1-\theta\rho)}+\sqrt{R"\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}\right)^2} < 0. \quad \text{(A15)}$$

To simplify exposition, we define two auxiliary functions:

$$\begin{cases} g(\theta) = \dfrac{4R"\lambda\alpha\rho}{\sqrt{R"\lambda(1-\theta\rho)}\sqrt{R"\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}} \\ f(\theta) = \left(\sqrt{R"\lambda(1-\theta\rho)}+\sqrt{R"\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}\right)^2 \end{cases}.$$

Then, $\dfrac{\partial(\theta+\xi(\theta))}{\partial\theta} = -\dfrac{g(\theta)}{f(\theta)}$. We can find the derivatives of $f(\theta)$ and $g(\theta)$ as follows:

$$\begin{cases} \dfrac{\partial g(\theta)}{\partial\theta} = 4R"\lambda\alpha\rho(-\frac{1}{2})\dfrac{2(R"\lambda)^2(1-\theta\rho)(-\rho)+4R"\lambda(-\rho)(\alpha\rho+(1-\theta\rho))+4R"\lambda(1-\theta\rho)(-\rho)}{2\sqrt{(R"\lambda)^2(1-\theta\rho)^2+4R"\lambda(1-\theta\rho)(\alpha\rho+(1-\theta\rho))}} > 0 \\ \dfrac{\partial f(\theta)}{\partial\theta} = 2[\sqrt{R"\lambda(1-\theta\rho)}+\sqrt{R"\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}][\dfrac{-R"\lambda\rho}{2\sqrt{R"\lambda(1-\theta\rho)}}+\dfrac{-R"\lambda\rho+4(-\rho)}{2\sqrt{R"\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}} < 0 \end{cases}.$$

Therefore, we get the second order derivative of $\theta + \xi(\theta)$ as follows:

$$\frac{\partial^2(\theta+\xi(\theta))}{\partial\theta^2} = -\frac{\dfrac{\partial g(\theta)}{\partial\theta}f(\theta) - g(\theta)\dfrac{\partial f(\theta)}{\partial\theta}}{f(\theta)^2} < 0 \quad \text{(A16)}$$

Thus, more customers joining the priority queue will lead to more loss and $\theta + \xi(\theta)$ is concave decreasing in θ.

Next, we analyze the derivatives with respect to α.

$$\frac{\partial(\theta+\xi(\theta))}{\partial\alpha}=-\frac{\frac{4\sqrt{R''\lambda(1-\theta\rho)}}{\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}}{\left(\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}\right)^2}<0. \tag{A17}$$

Redefine functions *f* and *g* as follows:

$$\begin{cases} g(\alpha)=\dfrac{4\sqrt{R''\lambda(1-\theta\rho)}}{\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}} \\ f(\alpha)=\left(\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}\right)^2 \end{cases}.$$

Then, $\dfrac{\partial(\theta+\xi(\theta))}{\partial\alpha}=-\dfrac{g(\alpha)}{f(\alpha)}$. Moreover, we get the following relations

$$\begin{cases} \dfrac{\partial g(\alpha)}{\partial\alpha}=4\sqrt{R''\lambda(1-\theta\rho)}(-\dfrac{1}{2})\dfrac{4\rho}{[2\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}]^2}<0 \\ \dfrac{\partial f(\alpha)}{\partial\alpha}=\dfrac{4\rho[\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}]}{\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}>0 \end{cases}$$

Thus,

$$\frac{\partial^2(\theta+\xi(\theta))}{\partial\alpha^2}=-\frac{\frac{\partial g(\alpha)}{\partial\alpha}f(\alpha)-\frac{\partial f(\alpha)}{\partial\alpha}g(\alpha)}{f(\alpha)^2}>0. \tag{A18}$$

Hence, stronger fairness will lead to more customer loss, and $\theta+\xi(\theta)$ is convex decreasing in α.

Finally, the derivatives $\xi(\theta)$ of can be easily obtained from those of $\theta+\xi(\theta)$:

$$\frac{\partial\xi(\theta)}{\partial\theta}=\frac{\partial(\theta+\xi(\theta))}{\partial\theta}-1<0,\qquad \frac{\partial\xi(\theta)}{\partial\theta^2}=\frac{\partial(\theta+\xi(\theta))}{\partial\theta^2}<0.$$

$$\frac{\partial\xi(\theta)}{\partial\alpha}=\frac{\partial(\theta+\xi(\theta))}{\partial\alpha}-\frac{\partial\theta}{\partial\alpha}=\frac{\partial(\theta+\xi(\theta))}{\partial\alpha}<0,\qquad \frac{\partial\xi(\theta)}{\partial\alpha^2}=\frac{\partial(\theta+\xi(\theta))}{\partial\alpha^2}>0.$$

Proof of Lemma 4.

For any given $0<\theta<\overline{\theta}$, the revenue function is shown in Eq. (A14). We can obtain the first order derivative of the objective function $R(\theta,\xi(\theta))$ as follows.

$$\frac{\partial R(\theta,\xi(\theta))}{\partial\theta}=\lambda\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+(1-\theta\rho))^2}(\xi(\theta)+\alpha)+\lambda\frac{\partial(\theta+\xi(\theta))}{\partial\theta}\left(\frac{\theta R''\rho}{(\alpha\rho+(1-\theta\rho)}+c\right)-\lambda\frac{\theta R''\rho}{\alpha\rho+(1-\theta\rho)}$$

$$=\lambda\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+1-\theta\rho)^2}\left(\frac{1}{\rho}\frac{2\sqrt{R''\lambda(1-\theta\rho)}}{\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}-\theta+\alpha\right) \tag{A19}$$

$$-\lambda\frac{4R''\lambda\alpha\rho}{\left(\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}\right)^2}\left(\frac{\theta R''\rho}{\alpha\rho+(1-\theta\rho)}+c\right)-\lambda\frac{\theta R''\rho}{\alpha\rho+(1-\theta\rho)}.$$

Then we get the second order derivative of the objective function as follows:

$$\frac{\partial^2 R(\theta,\xi(\theta))}{\partial\theta^2}=-\lambda\frac{2R''\rho(\alpha\rho+1)\rho(\xi(\theta)+\alpha)}{(\alpha\rho+(1-\theta\rho))^3}+2\lambda\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+(1-\theta\rho))^2}\frac{\partial\xi(\theta)}{\partial\theta}+\lambda\frac{\partial(\theta+\xi(\theta))}{\partial\theta^2}\left(\frac{\theta R''\rho}{\alpha\rho+(1-\theta\rho)}+c\right).$$

Recall from the proof of Lemma 3, we get $\frac{\partial\xi(\theta)}{\partial\theta}<0$ and $\frac{\partial(\theta+\xi(\theta))}{\partial\theta^2}<0$.

Obviously, $\frac{2R''\rho(\alpha\rho+1)\rho(\xi(\theta)+\alpha)}{(\alpha\rho+(1-\theta\rho))^3}>0$, $\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+(1-\theta\rho))^2}>0$ and $\frac{\theta R''\rho}{\alpha\rho+(1-\theta\rho)}+c>0$. Hence,

$$\frac{\partial^2 R(\theta,\xi(\theta))}{\partial\theta^2}<0. \tag{A20}$$

Because $R(\theta,\xi(\theta))$ is concave in θ for $0<\theta<\overline{\theta}$, it has a unique maximum in this range of θ.

We are not able to get an explicit expression of the optimal θ^*, but we can study the asymptotic behavior of θ^* when the base fee c and the fairness parameter α are sufficiently large.

(1) The base fee c is sufficiently large

Define $\overline{c}^{(1)}>0$ to satisfy the following equality:

$$\lambda\frac{4R''\lambda\alpha\rho}{\sqrt{(R''\lambda)^2+4R''\lambda(\alpha\rho+1)}\cdot\left(\sqrt{R''\lambda}+\sqrt{R''\lambda+4(\alpha\rho+1)}\right)^2}\overline{c}^{(1)}=\lambda\frac{R''\rho(\alpha\rho+1)(1+\alpha)}{(\alpha\rho+(1-\rho))^2}.$$

Hence, $\overline{c}^{(1)}=\frac{(\alpha\rho+1)(1+\alpha)\sqrt{(R''\lambda)^2+4R''\lambda(\alpha\rho+1)}\cdot\left(\sqrt{R''\lambda}+\sqrt{R''\lambda+4(\alpha\rho+1)}\right)^2}{4\lambda\alpha(\alpha\rho+(1-\rho))^2}$.

For $c\geq\overline{c}^{(1)}$, we see that the following three inequalities are quite straightforward:

(a) $\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+(1-\theta\rho))^2}\left(\frac{1}{\rho}\frac{2\sqrt{R''\lambda(1-\theta\rho)}}{\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}-\theta+\alpha\right)<\frac{R''\rho(\alpha\rho+1)(1+\alpha)}{(\alpha\rho+(1-\rho))^2}$

(b) $\dfrac{\theta R''\rho}{\alpha\rho+(1-\theta\rho)}>0$

(c) $\dfrac{\dfrac{4R''\lambda\alpha\rho}{\sqrt{R''\lambda(1-\theta\rho)}\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}}{\left(\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}\right)^2}\left(\dfrac{R'\rho\theta}{(\alpha\rho+(1-\theta\rho))}+c\right)$

$$>\frac{4R''\lambda\alpha\rho}{\sqrt{(R''\lambda)^2+4R''\lambda(\alpha\rho+1)}\cdot\left(\sqrt{R''\lambda}+\sqrt{R''\lambda+4(\alpha\rho+1)}\right)^2}c.$$

Plugging these into (A19), we see that when $c\geq\overline{c}^{(1)}$,

$$\frac{\partial R(\theta,\xi(\theta))}{\partial\theta}<0. \tag{A21}$$

Therefore, $R(\theta,\xi(\theta))$ decreases with θ when $c\geq\overline{c}^{(1)}$, and achieves the maximum value at the left boundary of θ. Note that the range of θ in Outcome A is open at the left boundary (i.e. approaching 0), but we also know that the revenue function is continuous at $\theta=0$ (see Lemma 8). Thus, we get that for $c\geq\overline{c}^{(1)}$, the optimal point is achieved at $\theta^*=0$, $\xi^*=\xi(\theta^*)=\dfrac{1}{\rho}\dfrac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}}$, and $K^*=0$.

(2) The fairness parameter α is sufficiently large

Define $\overline{\alpha}>0$ to satisfy $\dfrac{R''\lambda}{2\sqrt{R''\lambda(1-\rho)}}\dfrac{R''}{\overline{\alpha}}+\dfrac{R''\lambda}{2\sqrt{R''\lambda(1-\rho)}\sqrt{\overline{\alpha}\rho}}c+\dfrac{R''}{\overline{\alpha}}=\dfrac{R''}{2}$. Then, we get

$$\overline{\alpha}=\max\left\{\left(\frac{\lambda c\sqrt{\rho}+\sqrt{\lambda^2c^2\rho+8\rho^2R''\lambda(1-\rho)+4\rho^2R''\lambda\sqrt{R''\lambda(1-\rho)}}}{2\rho\sqrt{R''\lambda(1-\rho)}}\right)^2;\frac{1}{\rho}\right\}.$$

Also, define

$$\overline{c}^{(0)}=\left(\sqrt{R''\lambda(1-\rho)}-\frac{R''\lambda+2\sqrt{R''\lambda(1-\rho)}}{\alpha}\right)\cdot\frac{\sqrt{\alpha\rho}}{\lambda}=\left(\frac{(\alpha-2)\sqrt{R''\lambda(1-\rho)}-R''\lambda}{\alpha}\right)\cdot\frac{\sqrt{\alpha\rho}}{\lambda}.$$

For $c\leq\overline{c}^{(0)}$ and $\alpha\geq\overline{\alpha}$, we can obtain the following three inequalities:

(d) $\dfrac{\theta R''\rho}{\alpha\rho+(1-\theta\rho)}<\dfrac{R''\rho}{\alpha\rho+(1-\rho)}<\dfrac{R''}{\alpha}$

(e)
$$\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+(1-\theta\rho))^2}\left(\frac{1}{\rho}\frac{2\sqrt{R''\lambda(1-\theta\rho)}}{\sqrt{R''\lambda(1-\theta\rho)}+\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}-\theta+\alpha\right)$$
$$>\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+1)^2}\left(\xi(\theta)+\alpha\right)>\frac{R''\rho(\alpha\rho+1)}{(\alpha\rho+1)^2}\alpha=\frac{R''\rho}{\alpha\rho+1}\alpha.$$

(f)
$$\frac{4R''\lambda\alpha\rho}{\sqrt{R''\lambda(1-\theta\rho)}\sqrt{R''\lambda(1-\theta\rho)+4(\alpha\rho+(1-\theta\rho))}}\left(\frac{R''\rho\theta}{(\alpha\rho+(1-\theta\rho))}+c\right)$$
$$<\frac{4R''\lambda\alpha\rho}{\left(\sqrt{R''\lambda(1-\rho)}+\sqrt{R''\lambda(1-\rho)+4(\alpha\rho+(1-\rho))}\right)^2}\left(\frac{R''\rho}{(\alpha\rho+(1-\rho))}+c\right)$$
$$<\frac{4R''\lambda\alpha\rho}{\sqrt{R''\lambda(1-\rho)}\sqrt{4\alpha\rho}}\left(\frac{R''\rho}{\alpha\rho}+c\right)=\frac{R''\lambda}{\sqrt{R''\lambda(1-\rho)}\sqrt{4\alpha\rho}}\left(\frac{R''}{\alpha}+c\right).$$

In inequalities (e) and (f), if $\alpha\geq\frac{1}{\rho}$ holds, we also can get the following two relations

(e.2) $$\frac{R''\rho}{\alpha\rho+1}\alpha\geq\frac{R''\rho}{2\alpha\rho}\alpha=\frac{R''}{2}$$

(f.2) $$\frac{R''\lambda}{\sqrt{R''\lambda(1-\rho)}\sqrt{4\alpha\rho}}\left(\frac{R''}{\alpha}+c\right)\leq\frac{R''\lambda}{2\sqrt{R''\lambda(1-\rho)}}\frac{R''}{\alpha}+\frac{R''\lambda}{2\sqrt{R''\lambda(1-\rho)}\sqrt{\alpha\rho}}c.$$

Plugging these inequalities into (A19), we get that when $c\leq\overline{c}^{(0)}$ and $\alpha\geq\overline{\alpha}$,

$$\frac{\partial R(\theta,\xi(\theta))}{\partial\theta}>0\,. \tag{A22}$$

Thus, when $c\leq\overline{c}^{(0)}$ and $\alpha\geq\overline{\alpha}$, $R(\theta,\xi(\theta))$ is increasing in θ, and achieves the maximum value at the right boundary. Similarly, although the range of θ in Outcome A is open at the right boundary (i.e. approaching $\overline{\theta}$), but we also know that the revenue function is continuous at $\theta=\overline{\theta}$ (see Lemma 8). Thus, we get that $\theta^*=\overline{\theta}$, $\xi^*=\xi(\overline{\theta})=0$ and $K^*=K(\overline{\theta})>0$.

Proof of Lemma 5.

Recall that in the proof of Lemma 2, in Outcome B, we have $\underline{\theta} \le \theta \le \bar{\theta} = \min\left\{\frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}},1\right\}$.

If $\bar{\theta} \ge 1$, no customers balk, and the system becomes captive. Therefore, we focus on $\bar{\theta} < 1$ (i.e. $\min\{\bar{\theta},1\} = \bar{\theta}$). Therefore, for any $\underline{\theta} \le \theta \le \bar{\theta}$, we get

$$\begin{cases} \xi(\theta) = 0 \\ K(\theta,0) = R'' - \dfrac{\theta^2\rho}{\mu(1-\theta\rho)} \\ R(\theta,0) = \lambda\theta(K(\theta,0)+c) = \lambda\theta\left(R'' - \dfrac{\theta^2\rho}{\mu(1-\theta\rho)}\right) + \lambda\theta c \end{cases} . \tag{A23}$$

We can obtain the first order derivative of the objective function $R(\theta,0)$ from Equation (A23):

$$\begin{aligned} \frac{\partial R(\theta,0)}{\partial\theta} &= \lambda\left(R'' - \frac{\theta^2\rho}{\mu(1-\theta\rho)} + c\right) + \lambda\theta\frac{\rho}{\mu}\left(-\frac{2\theta(1-\theta\rho)-\theta^2(-\rho)}{(1-\theta\rho)^2}\right) = \lambda R'' - \lambda\frac{\rho}{\mu}\theta^2\frac{3-2\theta\rho}{(1-\theta\rho)^2} + \lambda c \\ &= \lambda\frac{(R''+c)\mu - 2\rho(R''+c)\mu\theta + \rho\left[(R''+c)\mu\rho - 3\right]\theta^2 + 2\rho^2\theta^3}{\mu(1-\theta\rho)^2}. \end{aligned} \tag{A24}$$

We also obtain the second derivative of the objective function $R(\theta,0)$ from Equation (A24):

$$\frac{\partial R(\theta,0)}{\partial\theta^2} = -\lambda\frac{\rho}{\mu}\frac{(6\theta-6\theta^2\rho)(1-\theta\rho)+(3\theta^2-2\theta^3\rho)2(\rho)}{(1-\theta\rho)^3} < 0$$

Thus, we will get the cubic equation (11) and the unique optimal θ^* based on Eq. (24).

Proof of Lemma 6.

Recall from the proof of Lemma 5, let's define an auxiliary function $h_1(\theta) = \lambda\frac{\rho}{\mu}\theta^2\frac{(3-2\theta\rho)}{(1-\theta\rho)^2}$,

then $\max_{\theta} h_1(\theta)=R''\lambda\left(\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda+4}-\sqrt{R''\lambda}}+3\right)$. The following three cases correspond to the three cases in Lemma 6:

(1) If $c \geq \overline{c}^{(2)} = R''\left(\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda+4}-\sqrt{R''\lambda}}+2\right)$, then $\frac{\partial R(\theta,0)}{\partial \theta} \geq 0$. We get the following optimal solution:

$$\theta^* = \frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}} = \overline{\overline{\theta}}; \qquad K(\theta,0)^* = K(\theta^*,0) = 0; \qquad R(\theta,0)^* = \lambda\overline{\overline{\theta}}c.$$

(2) If $\overline{c}^{(3)} < c < \overline{c}^{(2)}$, we obtain $\overline{\theta} < \theta^* < \frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}} = \overline{\overline{\theta}}$ and $K(\theta,0)^* > 0$.

(3) If $c \leq \frac{\rho}{\mu}\frac{(3-2\overline{\theta}\rho)\overline{\theta}^2}{(1-\overline{\theta}\rho)^2} - R'' = \overline{c}^{(3)}$, we obtain $\theta^* = \overline{\theta}$, and $K(\theta,0)^* = K(\overline{\theta},0) > 0$.

Proof of Lemma 7.

For $\theta = 0$, W_2 simplifies to $W_2 = \frac{\xi\rho}{\mu(1-\xi\rho)}$. Moreover, $\xi = P\left(H \leq \frac{R''}{W_2}\right) = \frac{R''}{W_2}$, and $\xi = \frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}} = \xi^* = \overline{\overline{\theta}}$. Therefore, the maximum revenue of the service provider can be calculated as follows:

$$R(0,\xi)^* = \lambda\xi^* c = \lambda\xi c = \lambda\frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}}c = \lambda\overline{\overline{\theta}}c \qquad \text{(A25)}$$

Proof of Lemma 8.

Recall from the proof of Lemma 2 that, in Outcome A, we have

$\alpha(W_2 - W_1) < K < \dfrac{(\alpha W_1 + R'')(W_2 - W_1)}{W_2}$. We now consider its extreme points (boundaries of with Outcomes B and C) of Outcome A.

(1) If $K = \alpha(W_2 - W_1)$ (i.e. $\dfrac{K}{W_2 - W_1} - \alpha = 0$), we get the following based on Eq. (A6):

$$\xi = P\left(H \le \frac{K}{W_2 - W_1} - \alpha \right) = P(H \le 0) = 0 .$$

Then, we will get $\theta = \bar{\theta}$ and $\xi(\bar{\theta}) = 0$ based on Equations (A8), (A10), (A12) and (A13). Recall from the proofs of Lemma 3 and Lemma 5 that the revenue function $R(\theta, \xi(\theta))$ can be calculated as

$$R(\theta, \xi(\theta)) = \lambda\theta \frac{R''\rho(\xi + \alpha)}{\alpha\rho + (1 - \theta\rho)} + \lambda(\theta + \xi)c = \lambda\theta\left(R'' - \frac{(\theta + \xi)^2 \rho}{\mu(1 - \theta\rho)} \right) + \lambda(\theta + \xi)c = \lambda\theta\left(R'' - \frac{\theta^2 \rho}{\mu(1 - \theta\rho)} \right) + \lambda\theta c$$

$$= R(\theta, 0). \qquad \text{(A26)}$$

Thus, the revenue functions of Outcome A and Outcome B have the same value on the boundary point of $\theta = \bar{\theta}$ (i.e., $K = \alpha(W_2 - W_1)$).

(2) If $K = \dfrac{(\alpha W_1 + R'')(W_2 - W_1)}{W_2}$ (i.e. $\dfrac{K}{W_2 - W_1} - \alpha = \dfrac{R'' - K}{W_1}$), we get the following based on Eq. (A6):

$$\theta = P\left(\frac{K}{W_2 - W_1} - \alpha \le H \le \frac{R'' - K}{W_1} \right) = 0 .$$

When $\theta = 0$, recall also from the proofs of Lemma 3 and Lemma 7 that the revenue function $R(\theta, \xi(\theta))$ can be calculated as:

$$R(\theta, \xi(\theta)) = \lambda\theta \frac{R''\rho(\xi + \alpha)}{\alpha\rho + (1 - \theta\rho)} + \lambda(\theta + \xi)c = \lambda\theta\left(R'' - \frac{(\theta + \xi)^2 \rho}{\mu(1 - \theta\rho)} \right) + \lambda(\theta + \xi)c = \lambda\xi c = R(0, \xi). \qquad \text{(A27)}$$

Thus, the revenue functions of Outcome A and Outcome C have the same value on the boundary point of $\theta = 0$ (i.e., $K = \dfrac{(\alpha W_1 + R'')(W_2 - W_1)}{W_2}$).

Since the revenue functions are obviously continuous on the interior of the θ range of each

Outcome, and we just proved that they agree on the boundary points as well, we have now proved that the overall revenue function is continuous on the entire range of $\theta \in [0, \overline{\overline{\theta}}]$.

Proof of Proposition 2.

The revenue function on the range of θ for each individual Outcome is summarized below:

$$\begin{cases} R(\theta, \xi(\theta)) = \lambda\theta \dfrac{R''\rho(\xi+\alpha)}{(\alpha\rho + (1-\theta\rho))} + \lambda(\theta+\xi)c = \lambda\theta(R'' - \dfrac{(\theta+\xi)^2\rho}{\mu(1-\theta\rho)}) + \lambda(\theta+\xi)c \quad (0 \le \theta \le \overline{\theta}) \\ R(\theta, 0) = \lambda\theta(R'' - \dfrac{\theta^2\rho}{\mu(1-\theta\rho)}) + \lambda\theta c \quad (\overline{\theta} \le \theta \le \overline{\overline{\theta}}) \\ R(0, \xi) = \lambda\xi c \quad (\theta = 0) \end{cases} . \qquad \text{(A28)}$$

Recall from the proof of Lemma 8 that the overall revenue function in (A28) is a continuous function in the range of $0 \le \theta \le \overline{\overline{\theta}}$.

In summary, we get the following results:

- Based on the previous proof of Lemma 4, we get the optimal result $\theta^* = \overline{\theta}$, $\xi^* = \xi(\overline{\theta}) = 0$ and $K^* = K(\overline{\theta}) > 0$ when $c \le \overline{c}^{(0)}$ and $\alpha \ge \overline{\alpha}$.
- Based on the previous proof of Lemma 6, we get $0 < \overline{\theta} \le \theta^* < \overline{\overline{\theta}}$, $\xi^* = \xi(\theta^*) = 0$ and $K(\theta, 0)^* > 0$ if $c < \overline{c}^{(2)}$.
- Based on the previous proof of Lemma 4, we get that for $c \ge \overline{c}^{(1)}$, the optimal point is achieved at $\theta^* = 0$, $\xi^* = \xi(\theta^*) = \dfrac{1}{\rho} \dfrac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda} + \sqrt{R''\lambda + 4}} > 0$ and $K^* = 0$.
- Based on the previous proof of Lemma 6, we get $\theta^* = \overline{\overline{\theta}} > 0$, $\xi^* = 0$ and $K(\theta, 0)^* = 0$ if $c \ge \overline{c}^{(2)}$.

These lead to the following two results:

1) If $\alpha \geq \bar{\alpha}$ *and* $c \leq \min\{\bar{c}^{(0)}, \bar{c}^{(2)}\}$, *then* $\theta^* > 0$, $\xi^* = 0$, *and* $K^* > 0$.

2) If $c \geq \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$, *then either* $\theta^* > 0$, $\xi^* = 0$, *and* $K^* = 0$ *or* $\theta^* = 0$, $\xi^* > 0$, *and* $K^* = 0$.

Proof of Lemma 9.

Recall from the proof of Lemma 2 that when $\alpha = 0$, we have $\bar{\theta} = \bar{\bar{\theta}}$. When we plug α=0 into (A8), we get $\theta + \xi = \bar{\bar{\theta}}$.

Recall also from the proof of Proposition 2 that the revenue function is continuous in range of $\theta \in [0, \bar{\bar{\theta}}]$ for any α. It must be continuous on $\theta \in [0, \bar{\bar{\theta}}]$ for the special case $\alpha = 0$:

$$\begin{cases} R(\theta, \xi(\theta)) = \lambda\theta \dfrac{R''\rho\xi}{1-\theta\rho} + \lambda(\theta+\xi)c = \lambda\theta\left(R'' - \dfrac{(\theta+\xi)^2\rho}{\mu(1-\theta\rho)}\right) + \lambda(\theta+\xi)c \quad (0 \leq \theta \leq \bar{\theta}) \\ R(\theta, 0) = \lambda\theta\left(R'' - \dfrac{\theta^2\rho}{\mu(1-\theta\rho)}\right) + \lambda\theta c \quad (\bar{\theta} \leq \theta \leq \bar{\bar{\theta}}) \\ R(0, \xi) = \lambda\xi c \quad (\theta = 0) \end{cases} . \qquad \text{(A29)}$$

To optimize the revenue function on $0 \leq \theta \leq \bar{\bar{\theta}}$, we first take the first order derivative:

$$\frac{\partial R(\theta, \xi(\theta))}{\partial \theta} = \lambda R''\rho \frac{\frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}} - \theta}{1-\theta\rho} + \lambda R'\rho\theta \frac{-(1-\theta\rho) - \left(\frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}} - \theta\right)(-\rho)}{(1-\theta\rho)^2}$$

$$= \frac{\lambda R''\rho}{(1-\theta\rho)^2}\left(\frac{1}{\rho}\frac{2\sqrt{R''\lambda}}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}} - 2\theta + \theta^2\rho\right).$$

Let θ^* satisfy the first order condition (that is, $\frac{\partial R(\theta^*, \xi(\theta^*))}{\partial \theta} = 0$), We can show that the second order derivative at θ^* is negative:

$$\frac{\partial^2 R(\theta^*, \xi(\theta^*))}{\partial \theta^2} = -2\frac{\lambda R''\rho}{(1-\theta^*\rho)} < 0.$$

Thus, θ^* is an optimal solution, maximizing the objective function. We can use the first order condition to find an express form of θ^* and then K^* and R^* as well.

$$\begin{cases} \theta^*_{\alpha=0} = \dfrac{1}{\rho}\left(1-\dfrac{\sqrt{R''\lambda+4}-\sqrt{R''\lambda}}{2}\right) = \dfrac{1}{\rho}\left(1-\dfrac{2}{\sqrt{R''\lambda+4}+\sqrt{R''\lambda}}\right) \\ K^*_{\alpha=0} = K(\theta^*_{\alpha=0},\xi^*_{\alpha=0}) = R''\rho\theta^*_{\alpha=0} = R''\left(1-\dfrac{2}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}}\right) \\ R^*_{\alpha=0} = R(\theta^*_{\alpha=0},\xi^*_{\alpha=0}) = \lambda R''\dfrac{1}{\rho}\left(1-\dfrac{2}{\sqrt{R''\lambda+4}+\sqrt{R''\lambda}}\right)^2 + \lambda\bar{\bar{\theta}}c \end{cases} . \quad \text{(A30)}$$

Thus, the provider should induce the customers to self-segment and join both queues.

Proof of Proposition 3.

(1)Recall from the proof of Lemma 1 the following optimal results in the captive system:

When $\alpha = 0$, the optimal solution is shown:

$$\theta^*_{\alpha=0} = \frac{1-\sqrt{1-\rho}}{\rho};\quad K^*_{\alpha=0} = \frac{\rho(1-\sqrt{1-\rho})}{\mu(1-\rho)};\quad R^*_{\alpha=0} = \frac{\rho(1-\sqrt{1-\rho})^2}{(1-\rho)} + \lambda c . \quad \text{(A31)}$$

Based on the proof of Lemma 9, the optimal solutions in a non-captive system is shown in Eq.(A30) when $\alpha = 0$. Thus, in both service systems, it's optimal for the service provider to charge a positive priority fee, and have customers join both queues.

(2) Based on the proofs of Proposition 2, Lemmas 4, 6 and 7, the optimal maximum revenues satisfying the following:

When $c \geq \max\{\bar{c}^{(1)}, \bar{c}^{(2)}\}$, we have

$$\lambda\bar{\bar{\theta}}c = R^*_{\alpha>0} < R^*_{\alpha=0} = \lambda R''\frac{1}{\rho}\left(1-\frac{2}{\sqrt{R''\lambda+4}+\sqrt{R''\lambda}}\right)^2 + \lambda\bar{\bar{\theta}}c . \quad \text{(A32)}$$

That is, when the base fee *c* is high enough, the service provider obtains higher revenue at $\alpha = 0$.

Proof of Proposition 4.

For any given *K*, Equation (A14) in the proof of Lemma 3 gives us the resulting θ and ξ in a non-captive system. Therefore, we can plug the fixed value of $K^*_{\alpha=0}$ (A30) into (A14) to get:

$$K^*_{\alpha=0}=R''\left(1-\frac{2}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}}\right)=\frac{R''\rho(\xi(\theta)+\alpha)}{\alpha\rho+(1-\theta\rho)}=R''-\frac{(\theta+\xi(\theta))^2\rho}{\mu(1-\theta\rho)}. \quad \text{(A33)}$$

We can rewrite (A33) as $K^*_{\alpha=0}(\alpha\rho+(1-\theta\rho))=R''\rho(\xi(\theta)+\alpha)$ and $(\theta+\xi(\theta))^2\rho=(R''-K^*_{\alpha=0})\mu(1-\theta\rho)$.

At the same time, we also obtain the following two first order derivative functions:

$$K^*_{\alpha=0}\left(\rho-\frac{\partial\theta}{\partial\alpha}\rho\right)=R''\rho\left(\frac{\partial\xi(\theta)}{\partial\alpha}+1\right) \text{ and } 2(\theta+\xi)\frac{\partial(\theta+\xi(\theta))}{\partial\alpha}\rho=(R''-K^*_{\alpha=0})\mu\left(-\frac{\partial\theta}{\partial\alpha}\rho\right).$$

Thus, we get the following two inequalities, which hold because $\frac{\partial(\theta+\xi(\theta))}{\partial\alpha}<0$ and $R''-K^*_{\alpha=0}=R''\frac{2}{\sqrt{R''\lambda}+\sqrt{R''\lambda+4}}>0$:

$$\frac{\partial\theta}{\partial\alpha}=-\frac{2(\theta+\xi(\theta))}{(R''-K^*_{\alpha=0})\mu}\frac{\partial(\theta+\xi(\theta))}{\partial\alpha}>0 \quad \text{(A34)}$$

$$\frac{\partial\xi(\theta)}{\partial\alpha}=-\frac{K^*_{\alpha=0}}{R''}\frac{\partial\theta}{\partial\alpha}-\frac{R''-K^*_{\alpha=0}}{R''}<0. \quad \text{(A35)}$$

Therefore, when the service provider charges a fixed fee $K^*_{\alpha=0}$ based on the wrong perception that customers do not care about fairness, then the stronger the real customer fairness perception, the more of them will join the priority queue and the fewer the regular queue.