Submitted to *Operations Research*
manuscript

Pricing and Prioritizing Time-Sensitive Customers with Heterogeneous Demand Rates

Philipp Afèche, Opher Baron, Joseph Milner, Ricky Roet-Green
Rotman School of Management, University of Toronto, 105 St. George Street, Toronto, Ontario, M5S 3E6
afeche@rotman.utoronto.ca, opher.baron@rotman.utoronto.ca, jmilner@rotman.utoronto.ca,rgricky@gmail.com

We consider the pricing/lead-time menu design problem for a monopoly service where time-sensitive customers have demand on multiple occasions. Customers differ in their demand rates and valuations per use. We compare a model where the demand rate is the private information of the buyers to a model where the firm has full information. The model assumes that customers queue for a finite-capacity service under a general pricing structure. Customers choose a plan from the menu to maximize their expected utility. In contrast to previous work, we assume customers do *not* differ in their waiting cost. Yet we show that in the private information case prioritizing customers may be optimal as a result of demand rate heterogeneity. We provide necessary and sufficient conditions for this result. In particular, we show that for intermediate capacity, more frequent-use customers that hold a lower marginal value per use should be prioritized. Further, less frequent-use customers may receive a consumer surplus. We demonstrate the applicability of these results to relevant examples. The structure of the result implies that in some cases it may be beneficial for the firm to prioritize a customer class with a *lower* marginal cost of waiting.

Key words: Capacity Pricing, Heterogeneous Usage Rates, Priority queues

1. Introduction

Service firms sell memberships that lower the price paid by customers, yet raise the revenue received. And membership has its privileges. Season passes to leisure activities such as ski mountains and amusement parks are often accompanied by perks such as access to priority queues at theme parks (e.g., Universal Studios Express Pass) or early admission (e.g., Stratton Mountain Summit Pass). Memberships allow line-jumping for exhibit entrance at cultural institutions and early registration for classes at social organizations. Firms typically offer several different pricing plans, e.g., unlimited access (a season pass), limited access (a multiple use ticket), two-part tariffs (paid-for discount cards such as tastecard), and a pay-per-use price, and these often come with differing benefits. The customer's choice of which price to pay depends on the value expected to be derived from use and the total cost including the cost of waiting. And by inducing different customers to pay different prices, firms can increase their total revenue.

Consider, for example, the choice of whether to purchase a season pass at a ski resort allowing unlimited access. This may be of interest to skiers residing near the mountain (the 'locals'). They are likely to have a higher frequency of use than vacationers coming to the resort (the 'aways'). However, a local also may derive less enjoyment from any particular day of skiing than an away, as s/he may see multiple opportunities each season, reducing the marginal value. If the mountain's management does not offer a season pass, the locals may reduce their skiing. But if the season pass is priced too low, the aways may purchase them, reducing the mountain's revenue. Thus the mountain's management has the problem of pricing a season pass to attract locals, but not aways. And unless they require proving residency to purchase such a pass, it is difficult to distinguish locals from aways. But as noted above, the firm has another tool, the perks it offers along with the pass. In particular, priority services such as pass-holders' lift lines and early mountain access are of value when the system is congested. We show that these perks are not simply additional benefits of membership, but are necessary to maximize the mountain's revenue.

This paper considers the problem of designing price/lead-time menus and the corresponding scheduling policy for a profit-maximizing service provider serving customers with private information on their preferences. Customers are risk-neutral and maximize their expected utility by choosing whether to buy service, and if so, which service class (price-leadtime option on the menu). The key novelty is that the paper studies settings where customers have demand for multiple uses, and they are heterogeneous in these demand rates, as is the case for example in ski resorts or amusement parks. Most previous studies restrict attention to the case where customers have *unit* demand, that is, they have identical infinitesimal demand rates. The few papers that do consider heterogeneous demand rates typically restrict attention either to the case in which the provider observes customer preferences, or to undifferentiated First In First Out (FIFO) service which misses the value of differentiated service.

The paper deliberately focuses on the simplest model to understand the *minimal* conditions for differentiated service to be profit-maximizing when customers have heterogeneous demand rates. Specifically, we model the service facility as M/M/1 and consider customer types that differ only in their demand rates and their marginal (per-use) valuations; we do *not* assume the customers have differing marginal costs of waiting. Rather, we show that differences in valuation and demand rate are sufficient in some conditions to make prioritized service optimal for revenue maximization.

Customers repeatedly use the service (if they find it economical to do so) at a given rate that is inherent to their type. We assume the marginal value for both customer types is constant in usage, but apply a strict ordering on the marginal value for the types. We allow the marginal value of an additional usage for the more frequent-use type to be either higher than or lower than that of the less frequent-use type. Both cases are possible and represent alternate orderings of the

marginal rate of substitution between usage and price depending on the customer's type. (This is the constant sign assumption, cf. Fudenberg and Tirole (1991).) For example, skiing enthusiasts may find higher marginal value than a skiing novice at all frequencies of use. Alternatively, the marginal value of use of a visitor to a ski resort may be higher than that of a local for whom there are multiple opportunities to ski in a season. We investigate the firm's policy for both cases, comparing the firm's optimal policy under full and private information.

Hassin and Haviv (2003) provide a comprehensive literature review of research into the equilibrium behavior of customers and servers in queueing systems with pricing. The vast majority of pricing studies for queues restrict attention to the case where customers have unit demand, that is, they have identical infinitesimal demand rates. Naor (1969) and Mendelson (1985) consider first-in-first-out (FIFO) service for customers with homogeneous delay costs. Mendelson and Whang (1990), Hassin (1995), and Hsu et al. (2009) characterize the socially optimal price-delay menu and scheduling policy for heterogeneous customers. Some papers on the revenue-maximization problem for heterogeneous customers restrict the scheduling policy, customers' service class choices, or both (cf. Lederer and Li (1997), Boyaci and Ray (2003), Maglaras and Zeevi (2005), Allon and Federgruen (2009), Zhao et al. (2012), Afèche et al. (2013)). Afèche (2004) initiated a stream of revenue-maximization studies that design jointly optimal prices and scheduling policies in the presence of incentive-compatibility constraints (cf. Katta and Sethuraman (2005), Yahalom et al. (2006), Afèche (2013), Maglaras et al. (2014)). The conventional wisdom that emerges from all of these unit-demand studies is that offering priorities has positive value only if customers have heterogeneous delay costs. In contrast, only a few papers consider customers who have demand for multiple uses and who are heterogeneous in this attribute: some have high, others have low demand. Rao and Petersen (1998) and Van Mieghem (2000) consider the welfare-maximization problem. Rao and Petersen (1998) study a model with pre-specified priority delay functions, which eliminates the scheduling problem. Van Mieghem (2000) considers the menu design question jointly with the optimal scheduling problem under convex increasing waiting cost functions. Papers that consider the revenue-maximization problem under restriction to FIFO service establish the optimality of fixed-up-to tariffs (Masuda and Whang 2006) or compare the performance of simpler tariffs, namely, subscription-only versus pay-per-use pricing (Randhawa and Kumar (2008), and Cachon and Feldman (2011)). Finally, Plambeck and Wang (2013) consider revenue maximization with multiple-use customers whose service valuations are subject to hyperbolic discounting. This model captures the preference structure for unpleasant services. The optimal mechanisms they study are tailored to such settings, which are in marked contrast to the more pleasant services that fit our model.

This paper makes three contributions on the design of differentiated price-service mechanisms for queueing systems. First, we demonstrate the fundamental point that when customers differ in their demand rates, it may be optimal to offer delay-differentiated services (through priorities) rather than uniform service (e.g., FIFO), even though all customers are equally delay-sensitive. In fact, it follows from our derivation that a firm may prioritize customers that are *less* sensitive to delay. This result runs counter to the conventional wisdom given by the extensive literature on systems serving customers with equal demand rates that only prioritizing customers with higher delay costs has positive value. Second, we provide necessary and sufficient conditions, in terms of the demand and capacity characteristics, for priority service to be optimal. In brief, priority service is optimal only if customers with higher demand rates have lower marginal valuations than their low-demand counterparts, a plausible condition for several applications including entertainment parks. Under this condition, priority service is optimal if the aggregate willingness-to-pay of all potential high-demand users is sufficiently high, and there is sufficient, but not excessive, capacity. Furthermore, when priority service is optimal, the menu is designed such that high-demand/low-value customers buy the high-priority service for a subscription fee. The result implies that the use of priority queues seen in many environments such as amusement parks and ski resorts is not just a reward for loyal, season-ticket purchasing customers, but part of the mechanism design that allows the firm to differentiate between customer types. Third, we show that offering optimal delay-differentiated services can generate significant profit gains, compared to FIFO service, in many cases double-digit percentage gains.

2. Model

We consider a capacity-constrained monopoly firm that designs a menu of price-service plans for customers that differ based on their demand rate for the service and the value they derive from each usage. There are two customer types, indexed by $i = 1, 2$. The market for each type consists of a fixed, large number of potential customers, N_i. Type-i customers receive a (constant) value r_i for each service usage. Each type-i customer experiences a stream of service opportunities that arrive at rate γ_i, the expected number of service opportunities per year. This rate is fixed and inherent to the customer's type. Without loss of generality (w.l.o.g.), we assume $\gamma_1 > \gamma_2$.

Customers are delay-sensitive and prefer faster service. We assume that all customers have the *same* waiting cost, c, per unit time in the system (including service). This assumption eliminates waiting cost heterogeneity as the driver of delay differentiation, the focus of virtually the entire previous literature on priority pricing. Rather, we focus on identifying the conditions for optimal delay differentiation to arise as a result of demand rate heterogeneity.

The firm operates a service facility with fixed capacity, μ. (Our results characterize the optimal menu as a function of μ.) For simplicity we assume that the service operates as an $M/M/1$ queue. Let $\Lambda^{Max} = \gamma_1 N_1 + \gamma_2 N_2$ denote the maximum potential arrival rate. We assume that Λ^{Max} is on the order of μ, while $\gamma_i << \mu$. This implies that while customers may use the service multiple times during the year, each customer's usage of the capacity is relatively insignificant. This assumption is consistent with the types of service firms we are modeling (amusement parks, ski resorts, etc.).

The service provider first designs and announces a static menu of up to two service classes, indexed by $j = 1, 2$. Customers then choose from the menu the class of service to purchase as detailed below. The restriction to two service classes is w.l.o.g. in our model. (If the provider offers more than two plans and each type chooses the plan that maximizes its utility, then more than two plans would be used only if some customers are indifferent between two or more of the plans. However, the firm would only offer those plans that maximize its revenue, so there is no advantage derived from offering more plans than customer types.) The menu specifies for each class a usage rate-dependent tariff (or price function) and the expected waiting time a customer will encounter at each visit to the facility. To be clear, "class" refers to the attributes of a service option, "type" refers to those of a customer. We also refer to class-j service as plan j where it is natural to do so.

We assume that the firm knows the aggregate demand information (r_i, γ_i, and N_i for $i = 1, 2$, and c). With respect to customer-level demand information, we consider two settings. In the Full Information benchmark the firm can distinguish customer types. Our main results focus on the Private Information setting where the firm cannot distinguish customer types. We formalize these problems in Sections 2.1 and 2.2.

Let $P_j(\gamma)$ be the total annual revenue generated by a customer with usage rate γ who chooses class-j service, $j \in \{1, 2\}$. This form is general and can represent any pricing scheme including a service class with unlimited usage at a subscription price, a two-part tariff with or without a maximum usage rate, or a simple per use price. If, for example, $P_j(\gamma)$ were a two-part tariff with subscription fee F_j and price per use of p_j, then $P_j(\gamma) = F_j + p_j\gamma$.

Let W_j be the expected waiting time (or lead time, including service time) for class-j service. We require W_j's to be consistent with the average steady-state wait times that are realized given the provider's scheduling policy and the customers' purchase rates induced by the menu. This consistency requirement may be enforced by auditors or third party review sites. Practically, for the motivating examples, social media provides a means for customers to learn prior to purchase the expected wait times and determine if there are any inconsistencies with the posted times. See Afèche (2013) for further discussion.

We do not assume a specific scheduling policy but rather let the provider choose any non-anticipative and regenerative policy. This appears to be the most general, easily described restriction of admissible policies that guarantees the existence of long run waiting time averages. We allow

preemption, which simplifies the analysis without affecting the results (under priority scheduling, with preemption the waiting time of a given class does not depend on the arrival rate of the lower class).

Given the menu, customers decide whether to seek service, and if so, choose a plan to maximize their expected total (annual) utility. Customers are risk neutral. They do not observe the queue and base their decisions on the posted expected waiting times. This assumption is common in related papers. For the motivating applications, the notion is that the queue cannot be observed by the customer even at the time of purchase as it may be spatially or temporally removed from the ticket window.

We further assume that customers do not change their type based either on the menu of prices offered or their experience of the service. In our model a customer has no incentive to switch between service classes during the year. As such, the customers decide once, at the start of the year (or when their first service opportunity arises), which class of service to purchase (if any). Then customers who buy a plan have an incentive to join the facility at each service opportunity.

From these definitions, the total expected utility of a type-i customer for class-j service is

$$(r_i - cW_j)\gamma_i - P_j(\gamma_i),$$

where $r_i - cW_j$ is her expected net value from every service opportunity. As further discussed below, we can restrict attention w.l.o.g. to menus that sell class-i to type-i customers. We write u_i for the utility of a type-i customer for class-i.

Let n_i be the number of class-i plans sold. We assume that if customers of type-i find strictly positive utility from a plan, the firm must allow all customers of type-i to join, i.e. $n_i = N_i$. As such the firm does not discriminate between customers of the same type if the pricing and prioritization policy provide a positive surplus to the class. If the firm chooses to limit the number of customers from a type that has positive utility from joining, it could do better by raising the price for the corresponding class of service, until the point that they are indifferent between joining and balking. Thus, if it does not raise the price, it would not limit the number. However, if type-i customers are indifferent between joining and balking, the firm can restrict the number of class-i plans, i.e., it is possible for $n_i < N_i$. By these assumptions the firm can tailor the experience received by different types of customers by lowering the demand from some classes while raising their price. For simplicity, we treat n_i as a continuous variable throughout, rather than as an integer; given large N_i, this is a mild assumption.

2.1. Full Information Setting

In the Full Information (FI) setting the firm can distinguish between the customer types. It can therefore assign a price for each customer type and enforce the customers to pay that price if they use the service. We assume there are two service classes and that type-i customers are offered class-i service. In this case, the firm maximizes its profit by choosing the pricing, $P_i(\gamma_i)$, the participation, n_i, and the prioritization that subsequently defines the waiting time, W_i, for each class. The firm's policy is constrained by the need for customers to see non-negative utility in joining the service.

In the FI setting the problem is:

$$\Pi^{\text{FI}} = \max_{n_i,\, W_i,\, P_i(\gamma_i)} \quad \sum_i n_i P_i(\gamma_i) \tag{1a}$$

$$\text{subject to} \quad (r_i - cW_i)\gamma_i \geq P_i(\gamma_i) \quad \text{for } i = 1,2 \tag{1b}$$

$$((r_i - cW_i)\gamma_i - P_i(\gamma_i))(N_i - n_i) = 0 \quad \text{for } i = 1,2 \tag{1c}$$

$$W_i \geq \frac{1}{\mu - n_i\gamma_i} \quad \text{for } i = 1,2 \tag{1d}$$

$$\sum_i n_i\gamma_i W_i \geq \frac{\sum_i n_i\gamma_i}{\mu - \sum_i n_i\gamma_i} \tag{1e}$$

$$0 \leq n_i \leq N_i \quad \text{for } i = 1,2. \tag{1f}$$

The objective function gives the total revenue of the firm. Constraint (1b) is the individual rationality (IR) constraint. Constraint (1c) is a complementary slackness constraint that verifies that if type-i customers have positive surplus from joining, then all customers of that type join. Alternatively, if there is no surplus for type-i customers, any feasible number can be assigned. Constraint (1d) ensures for each class that the waiting time is bounded below by the minimum feasible waiting time for class-i in an M/M/1 queue. Constraint (1e) verifies that the (weighted) average wait time for both service classes is bounded below by the minimum achievable, non-idling waiting time. This defines the achievable region for the waiting time. Constraint (1f) enforces the non-negativity and market size bounds for each customer type.

We can simplify the problem by eliminating the pricing, $P_i(\gamma_i)$, and waiting times, W_i, as follows. Observe that in maximizing the objective function, the individual rationality constraints, (1b), are binding. Therefore letting $P_i^*(\gamma_i)$ be the optimal price for class-i service in the FI solution,

$$P_i^*(\gamma_i) = (r_i - cW_i)\gamma_i. \tag{2}$$

The objective function can then be written as

$$\max_{n_i, W_i} \sum_{i=1,2} n_i\gamma_i r_i - c\sum_{i=1,2} n_i\gamma_i W_i \tag{3}$$

From (1e) it is evident that for any fixed n_1 and n_2, every work-conserving policy is optimal. In particular, let W be the waiting time under FIFO service. That is,

$$W = \frac{1}{\mu - \sum_i \gamma_i n_i}$$

is optimal given n_i. Therefore, we can reduce the Full Information problem to

$$\text{(FI)} \qquad \Pi^{\text{FI}} = \max_{n_1,n_2} \quad \sum_i n_i \gamma_i \left(r_i - \frac{c}{\mu - n_1\gamma_1 - n_2\gamma_2} \right)$$
$$\text{subject to} \quad 0 \le n_i \le N_i \text{ for } i = 1, 2.$$

2.2. Private Information Setting

In the Private Information (PI) setting, we assume the firm cannot distinguish between customers' types. To be specific we assume the firm cannot determine the type of a customer prior to the season or based on their usage during the season. Alternatively, if the type may be identified during the season, the firm cannot take advantage of that information. First, customers' types may not be fixed year to year, and so the firm may not be able to identify customer types before the start of a season based on previous usage. Second, customers that purchase a season pass or otherwise fix their service class and payment at the start of the year may reveal their true type during the season, but that is immaterial as the firm has already been paid and differentiating service based on a customer's true type would be difficult as the service class has already specified the expected waiting time. Third, customers that pay per use may do so without identifying themselves so tracking their type may be difficult. If their type can be identified we assume the firm does not change their service menu during the year, i.e., as assumed the menu is static.

As before we assume customers choose from the menu to maximize their expected utility. Now, the firm can no longer designate a class of service to a particular type of customer without providing an incentive to ensure they choose one plan over another. We restrict attention w.l.o.g. to menus ensuring incentive compatibility (IC) that target class i to type i customers such that they weakly prefer class i or no service over service in class $j \neq i$. (Based on the revelation principle (e.g., Myerson (1997)), mechanism design problems restrict attention w.l.o.g. to IC direct revelation mechanisms in which each customer directly reveals her type. The mechanism described below, while strictly speaking an "indirect mechanism", is equivalent to a direct mechanism and more naturally describes the purchase process.)

The IC constraints ensure that the expected annual cost for a type-i customer to use class-i service is less than the cost to use class-k service, $k \neq i$.

$$P_i(\gamma_i) + cW_i\gamma_i \le P_k(\gamma_i) + cW_k\gamma_i \quad \text{for } i, k \in \{1, 2\}. \tag{4}$$

Further, we assume that the annual price paid is non-increasing in usage for any class of service to ensure one cannot misrepresent one's type by higher usage for some gain. That is we impose the monotonicity constraint:

$$P_i(\gamma_2) \le P_i(\gamma_1) \qquad \text{for } i=1,2. \tag{5}$$

Remark 1. If fractional service could be purchased, we may need to also eliminate the possibility of a high-use type-1 customer representing himself as a type-2 customer by purchasing multiple copies of class-2 service. In particular, a type-1 customer would require (γ_1/γ_2) copies of class-2 service. That is, we require the constraint

$$P_1(\gamma_1) + cW_1\gamma_1 \le \frac{\gamma_1}{\gamma_2}P_2(\gamma_2) + cW_2\gamma_1. \tag{6}$$

We can show that at optimality (6) is either redundant to the IC constraint (4) or to the IR constraint $(r_1 - cW_1)\gamma_1 \ge P_1(\gamma_1)$ or both. First observe that (6) is redundant to (4) if

$$\begin{aligned} P_2(\gamma_1) + cW_2\gamma_1 &\le \frac{\gamma_1}{\gamma_2}P_2(\gamma_2) + cW_2\gamma_1 \\ \text{or } \gamma_2 P_2(\gamma_1) &\le \gamma_1 P_2(\gamma_2). \end{aligned}$$

Then, if $P_2(\gamma) = b\gamma$ for some $b > 0$, condition (6) holds. We show in Section 3.1.2 that $P_2(\gamma)$ has this form if $r_1 > r_2$. If on the other hand $r_1 \le r_2$, we show in Section 3.2.2 it is possible for $P_2(\gamma)$ to be a two-part tariff, i.e., $P_2(\gamma) = b\gamma - a$ for some $a, b > 0$. In that case, we show below that (6) is redundant to the IR constraint. Thus, we ignore constraint (6) in our formulation.

Under these conditions we can restrict the solution to truthful revelation so that type-i customers only purchase class-i service. As before we assume the firm does not limit the number of customers that purchase class-i service if type-i customers see strictly positive utility, and only sets $n_i < N_i$ if customers are indifferent between purchasing or not. The firm's Private Information problem is

$$\begin{aligned} \Pi^{PI} = \max_{n_i, W_i, P_i(\gamma_i)} \quad & \sum_i n_i P_i(\gamma_i) \\ \text{subject to} \quad & (r_i - cW_i)\gamma_i \ge P_i(\gamma_i) \qquad \text{for } i=1,2 && \text{(7a)} \\ & P_1(\gamma_1) + cW_1\gamma_1 \le P_2(\gamma_1) + cW_2\gamma_1 && \text{(7b)} \\ & P_2(\gamma_2) + cW_2\gamma_2 \le P_1(\gamma_2) + cW_1\gamma_2 && \text{(7c)} \\ & P_i(\gamma_2) \le P_i(\gamma_1) \qquad \text{for } i=1,2 && \text{(7d)} \\ & (P_i(\gamma_i) - (r_i - cW_i)\gamma_i)(N_i - n_i) = 0 \qquad \text{for } i=1,2 && \text{(7e)} \\ & W_i \ge \frac{1}{\mu - n_i\gamma_i} \qquad \text{for } i=1,2 && \text{(7f)} \\ & \sum_i n_i\gamma_i W_i \ge \frac{\sum_i n_i\gamma_i}{\mu - \sum_i n_i\gamma_i} && \text{(7g)} \\ & 0 \le n_i \le N_i \qquad \text{for } i=1,2. && \text{(7h)} \end{aligned}$$

The problem formulation is almost identical to Problem (1), with two additions: Constraints (7b) and (7c) are the IC constraints that verify that the total cost for type-i customers from choosing plan i is always less than choosing plan k; and constraint (7d) is the monotonicity constraint that verifies that the total payment of the customer cannot be reduced by increasing her demand.

We can simplify the problem (7a)–(7h) as follows. As noted above, (7a) is always binding for $i=1$. That is, $P_1(\gamma_1)=(r_1-cW_1)\gamma_1$. Further, because $\gamma_1>\gamma_2$, (7b) and (7d) imply $P_2(\gamma_1)$ can be increased arbitrarily so that pricing alone is sufficient to deter type-1 customers from buying class-2 service. This implies that we can extract all type-1 utility and that we can drop the type-1 IC constraint (7b).

However, pricing alone may not suffice to deter type-2 customers from buying class-1 service. Let $u_2=(r_2-cW_2)\gamma_2-P_2(\gamma_2)$ denote the type-2 expected utility from class-2 service. The type-2 IC constraint (7c) is equivalent to

$$u_2 \geq (r_2-cW_1)\gamma_2-P_1(\gamma_2).$$

That is, the type-2 utility decreases in its payment for class-1 service. In other words, a given u_2 provides a lower bound on $P_1(\gamma_2)$. On the other hand, since tariffs must be increasing in usage, we get

$$(r_2-cW_1)\gamma_2-u_2 \leq P_1(\gamma_2) \leq P_1(\gamma_1)=(r_1-cW_1)\gamma_1.$$

It follows that if class-1 service is offered, type-2 can be deterred from buying it if and only if

$$(r_2-cW_1)\gamma_2-u_2 \leq (r_1-cW_1)\gamma_1.$$

or, equivalently,

$$W_1 \leq \frac{r_1\gamma_1-r_2\gamma_2}{c(\gamma_1-\gamma_2)}+\frac{u_2}{c(\gamma_1-\gamma_2)}. \tag{8}$$

Observe that if no type-1 customers are served ($n_1=0$), then (7c) can be satisfied by setting $P_1(\gamma_2)$ arbitrarily high. Moreover, maximizing the profit implies minimizing u_2. Therefore (8) need only hold when $n_1>0$. Combining these simplifications we can write the PI problem as:

$$\begin{aligned}
\text{(PI)} \qquad \Pi^{\text{PI}} = \max_{n_i, W_i} \quad & \sum_i (n_i\gamma_i(r_i-cW_i))-u_2n_2 \\
\text{subject to} \quad & n_1\left(W_1-\left(\frac{r_1\gamma_1-r_2\gamma_2}{c(\gamma_1-\gamma_2)}+\frac{u_2}{c(\gamma_1-\gamma_2)}\right)\right)\leq 0 && \text{(9a)}\\
& W_i \geq \frac{1}{\mu-n_i\gamma_i} \qquad \text{for } i=1,2 && \text{(9b)}\\
& \sum_i n_i\gamma_i W_i \geq \frac{\sum_i n_i\gamma_1}{\mu-\sum_i n_i\gamma_1}, && \text{(9c)}
\end{aligned}$$

$$u_2(N_2 - n_2) = 0 \tag{9d}$$

$$0 \le n_i \le N_i \qquad \text{for } i = 1, 2 \tag{9e}$$

$$u_2 \ge 0. \tag{9f}$$

Constraint (9a) expresses the conditional constraint bounding the waiting time for the type-1 customers when they are present and subsumes the incentive compatibility and monotonicity constraints. As noted, the individual rationality constraint is given by (9f).

Remark 2 Let

$$\widetilde{W} = \frac{r_1\gamma_1 - r_2\gamma_2}{c(\gamma_1 - \gamma_2)}.$$

Here, $\widetilde{W}$ is a critical waiting time dependent only on the model parameters. Then (9a) implies that if type-1 customers are served, i.e., $n_1 > 0$, incentive compatibility requires that

$$W_1 \le \widetilde{W} + \frac{u_2}{c(\gamma_1 - \gamma_2)}. \tag{10}$$

Inequality (10) is the fundamental constraint governing the solution in the Private Information case. It implies that if, for a given capacity, the FIFO waiting time, W, exceeds $\widetilde{W}$, then it must be that either type-1 customers are not served, or if they are served, they are served with priority ($W_1 < W_2$) or type-2 customers receive some surplus utility ($u_2 > 0$), or both. Useful in determining what is the case is the reciprocal of $\widetilde{W}$, the critical capacity level

$$\widetilde{\mu} = \frac{1}{\widetilde{W}} = \frac{c(\gamma_1 - \gamma_2)}{r_1\gamma_1 - r_2\gamma_2}. \tag{11}$$

3. Optimal Price–Service Plans

In this section we develop the solutions for the Full Information and Private Information settings. Recall the two types are ordered by their inherent demand rates with $\gamma_1 > \gamma_2$. While the type-1 may use the service more frequently, it is not necessarily the case that the marginal value derived from a single usage by a type-1 customer, r_1, is greater than that of a type-2 customer, r_2. We consider two cases. In the first, referred to as the Increasing Ordering, we assume $r_1 \ge r_2$. We present the results for this case in Section 3.1. Here customers that have a higher valuation per usage use the service more. In this case we show that the firm can achieve the same profit in the FI and PI settings. This is the case investigated by Masuda and Whang (2006). As the FI setting solution provides no priority to one type of customer over another, the same holds for the PI setting. In the second case, referred to as the Decreasing Ordering, we assume $r_1 < r_2$. In Section 3.2, we show for this case it is possible that there is value to the information on the customer type and a prioritization policy may be optimal in the PI setting. In all cases, we determine the customer mix, service policy, and optimal pricing. These are dependent on the service capacity. Recall that

for the Full Information setting, all service is FIFO and the prices are given by (2). For the Private Information setting we provide the optimal service policy and prices. We summarize and discuss the theoretical results in Section 3.3. All proofs appear in Appendix A.

3.1. Increasing Ordering: Transaction Value Increases in Demand Rate

We first consider the full information setting, and subsequently the private information setting.

3.1.1. Increasing Ordering, Full Information. Let n_i^* be the optimal number of class-i customers that are served in the FI setting. The solution to (FI) for the increasing ordering is characterized by the following proposition:

Proposition 1 *For $r_1 \geq r_2$, there exist four thresholds over the capacity:*

$$\begin{aligned}
\mu_0 &= \frac{c}{r_1}, \\
\mu_1 &:= \arg\underset{\mu}{\{r_1 = \frac{c\mu}{(\mu - N_1\gamma_1)^2}\}}, \\
\mu_2 &:= \arg\underset{\mu}{\{r_2 = \frac{c\mu}{(\mu - N_1\gamma_1)^2}\}}, \\
\mu_3 &:= \arg\underset{\mu}{\{r_2 = \frac{c\mu}{(\mu - N_1\gamma_1 - N_2\gamma_2)^2}\}},
\end{aligned}$$

with $\mu_0 < \mu_1 \leq \mu_2 < \mu_3$ such that:

1. *For $\mu \leq \mu_0$, the provider does not serve any customers, $n_1^* = n_2^* = 0$.*
2. *For $\mu_0 < \mu < \mu_1$, the provider serves type-1 customers exclusively, but only partially, such that $0 < n_1^* < N_1, n_2^* = 0$.*
3. *For $\mu_1 \leq \mu \leq \mu_2$, the provider serves type-1 customers exclusively and fully, such that $n_1^* = N_1, n_2^* = 0$.*
4. *For $\mu_2 < \mu < \mu_3$, the provider serves type-1 customers fully, and type-2 customers partially, such that $n_1^* = N_1, 0 < n_2^* < N_2$.*
5. *For $\mu \geq \mu_3$, the provider serves type-1 and type-2 customers fully, such that $n_1^* = N_1, n_2^* = N_2$.*

Proposition 1 implies that as the capacity of the firm grows, first type-1 customers and subsequently type-2 customers are served as would be expected for the increasing ordering. In doing so, the firm engages in a revenue skimming policy for its capacity.

3.1.2. Increasing Ordering, Private Information. Under the increasing ordering, we observe that the firm can achieve the same revenue under the PI setting as in the FI setting, without offering any priorities. Observe that for FIFO waiting time W, $P_1(\gamma_1) = \gamma_1(r_1 - cW) > \gamma_2(r_2 - cW) = P_2(\gamma_2)$, so type-2 has no incentive to buy class-1 (we extract all type-1 utility, which is higher than for type-2 because of the increasing ranking) so long as we set $P_1(\gamma_2)$ high

enough. Similarly, type-1 has no incentive to buy class-2 because we can set $P_2(\gamma_1)$ to make the class prohibitively expensive. Formally we have:

Proposition 2 *When $r_1 \geq r_2$, the problem (PI) is maximized by offering FIFO service with $P_i(\gamma_i)$ set equal to the solution for the FI setting $P_i^*(\gamma_i)$, and achieves the same revenue.*

3.2. Decreasing Ordering: Transaction Value Decreases in Demand Rate

We next consider the decreasing ordering where $r_1 < r_2$, i.e., the customers with higher demand ($\gamma_1 > \gamma_2$) have lower marginal value per use. We show that in the FI setting, the solution is similar to that for the increasing ordering. That is, the firm should serve first the customers with the higher valuation for the service. However, the FI solution does not hold for the PI setting.

3.2.1. Decreasing Ordering, Full Information. The solution to Problem (FI) for the decreasing ordering is characterized by the following proposition:

Proposition 3 *For $r_1 < r_2$, there exist four thresholds over the capacity:*

$$\begin{aligned}
\mu_0 &= \frac{c}{r_2}, \\
\mu_1 &:= \arg\underset{\mu}{\{r_2 = \frac{c\mu}{(\mu - N_2\gamma_2)^2}\}}, \\
\mu_2 &:= \arg\underset{\mu}{\{r_1 = \frac{c\mu}{(\mu - N_2\gamma_2)^2}\}}, \\
\mu_3 &:= \arg\underset{\mu}{\{r_1 = \frac{c\mu}{(\mu - N_1\gamma_1 - N_2\gamma_2)^2}\}},
\end{aligned}$$

with $\mu_0 < \mu_1 \leq \mu_2 < \mu_3$ such that:

1. *For $\mu \leq \mu_0$, the provider does not serve any customers, $n_1^* = n_2^* = 0$.*
2. *For $\mu_0 < \mu < \mu_1$, the provider serves type-2 customers exclusively, but only partially, such that $n_1^* = 0$, $0 < n_2^* < N_2$.*
3. *For $\mu_1 \leq \mu \leq \mu_2$, the provider serves type-2 customers exclusively and fully, such that $n_1^* = 0$, $n_2^* = N_2$.*
4. *For $\mu_2 < \mu < \mu_3$, the provider serves type-2 customers fully, and type-1 customers partially, such that $0 < n_1^* < N_1$, $n_2^* = N_2$.*
5. *For $\mu \geq \mu_3$, the provider serves type-1 and type-2 customers fully, such that $n_1^* = N_1$, $n_2^* = N_2$.*

For the FI setting for the decreasing ordering, the solution again is given by a price skimming policy. In this case as the capacity increases, the type-2 customers are allocated capacity initially, and type-1 customers are served only if there is sufficient capacity. We now turn to the case with private information.

3.2.2. Decreasing Ordering, Private Information. Under the decreasing ordering we find that the firm may not achieve the same revenue in the PI setting as in the FI setting. As before, the firm's price/lead-time menu depends on the capacity. Here it also depends on the aggregate or total valuation of the service for the year for each customer. We consider two sub-cases:

- the total valuation of a type-1 customer is less than that of a type-2 customer ($r_1\gamma_1 \leq r_2\gamma_2$).
- the total valuation of a type-1 customer is higher than that of a type-2 customer ($r_1\gamma_1 > r_2\gamma_2$).

We refer to these as the "low total valuation" and "high total valuation" sub-cases, respectively. In the low total valuation case, type-1 customers are not particularly attractive customers for the firm. In contrast, the high total valuation sub-case provides an opportunity for the firm to set prices and service priorities so as to capture the type-1 customers' value while ensuring the type-2 customers identify themselves as such and extracts a higher marginal revenue from them.

Our main result is that for both sub-cases, we find that there may be a range of capacity where to maximize its revenue, the firm may need to prioritize the type-1 customers ($W_1 < W$) and/or provide positive consumer surplus to the type-2 customers.

Low total valuation sub-case. With both low valuation for each usage ($r_1 < r_2$) and low total valuation ($r_1\gamma_1 \leq r_2\gamma_2$), the firm would need both sufficient capacity and a sufficient number of type-1 customers for it to find value in serving these customers. The following proposition identifies these conditions:

Proposition 4 *Suppose $r_1 < r_2$ and $r_1\gamma_1 \leq r_2\gamma_2$. Let μ_2 be defined as in Proposition 3.*

1. *If $\mu \leq \mu_2$, the FI solution solves Problem (PI): $\Pi^{PI} = \Pi^{FI}$, $n_1 = n_1^* = 0$ and type-2 customers are served under FIFO service with $P_i(\gamma_i) = P_i^*(\gamma_i)$.*
2. *If $\mu > \mu_2$, $\Pi^{PI} < \Pi^{FI}$, type-2 customers are served fully and there exists $\overline{\mu} > \mu_2$ such that type-1 customers are served if and only if $\mu > \overline{\mu}$ and*

$$\frac{N_1}{N_2} > \frac{r_2\gamma_2 - r_1\gamma_1}{r_1\gamma_1}.$$

In this case type-1 get strict priority and zero utility, whereas type-2 customers receive positive utility.

In Proposition 4 the condition in part 2.,

$$\frac{N_1}{N_2} > \frac{r_2\gamma_2 - r_1\gamma_1}{r_1\gamma_1} \text{ or, equivalently, } (N_1 + N_2)(r_1\gamma_1) > N_2(r_2\gamma_2),$$

holds for sufficiently large N_1. In this case the total potential value of all customers at the lower total value per customer $r_1\gamma_1$ exceeds the value of the type-2 customers alone. That is, if all customers were to pretend to be type-1, the firm could extract higher revenue than if they served

only the type-2 customers. This is the only circumstance in the low total valuation sub-case that there is value to be extracted from the type-1 customers. In this case a limited number, say $\overline{n}_1$, of the type-1 customers will be served. However, these customers are served with priority and charged a premium price, $\overline{P}_1(\gamma_1) > P_1^*(\gamma_1)$ in order to deter type-2 customers from buying class-1 service. This results in a delay for the type-2 customers, but they are compensated by receiving a discount so their price $\overline{P}_2(\gamma)$ is lower than $P_2^*(\gamma_2)$. The discount provides consumer surplus to the type-2 customers and so the firm does not receive the same revenue as in the FI setting. That is, in Lemma 3 (given in the proof) we show for $\mu > \mu_2$, the prices as functions of γ are:

$$\overline{P}_1(\gamma) = (r_1 - cW_1)\gamma_1 > P_1^*(\gamma_1), \tag{12}$$

$$\overline{P}_2(\gamma) = (r_2 - cW_2)\gamma - \frac{c(\gamma_1 - \gamma_2)}{\mu - n_1\gamma_1} + (r_1\gamma_1 - r_2\gamma_2) < P_2^*(\gamma_2), \tag{13}$$

and $\Pi^{PI} < \Pi^{FI}$. Observe that the price for class-1 service is independent of γ, i.e., a subscription price, whereas that for class-2 service is a two-part tariff.

Observe that as $\mu \to \infty$, $W_1, W_2 \to 0$. Therefore at the limit $\overline{P}_1(\gamma_1) = P_1^*(\gamma_1) = r_1\gamma_1$ from (12). But for type-2 customers, from (13) $\overline{P}_2(\gamma_2) = r_1\gamma_1 < P_2^*(\gamma_2) = r_2\gamma_2$. The price that is paid by type-2 customers is reduced even when the capacity is large. In order to serve type-1, one cannot charge more than $r_1\gamma_1$ when $W_1 = 0$, but then one cannot charge type-2 more than $r_1\gamma_1$ as well because otherwise they would represent themselves as type-1. As a result, $\Pi^{PI} < \Pi^{FI}$.

As we noted in Remark 1, if the price for class-2 service is given by a two-part tariff of the form $P_2(\gamma) = b\gamma - a$, for $a, b > 0$, a type-1 customer may prefer to purchase multiple services of class-2 service. Incentive compatibility requires (6) to hold:

$$P_1(\gamma_1) + cW_1\gamma_1 \leq \frac{\gamma_1}{\gamma_2}P_2(\gamma_2) + cW_2\gamma_1.$$

Substituting $P_1(\gamma_1)$ and $P_2(\gamma_2)$ given by (12) and (13) into the above and simplifying, implies (6) holds if

$$\frac{c}{\mu - n_1\gamma_1} \leq r_1$$

or, equivalently,

$$r_1 - cW_1 \geq 0,$$

as (17) in the proof of Proposition 4 shows that $W_1 = 1/(\mu - n_1\gamma_1)$. But this is the IR constraint for type-1 customers and therefore type-1 customers would not purchase multiple copies of class-2 service and (6) is redundant to the formulation.

High total valuation sub-case. We now consider the sub-case where $r_1\gamma_1 > r_2\gamma_2$. Here, the type-1 customers are very attractive if one considers the total revenue they could provide. (They

Form	Description	n_1	Class-1 Priority	Class-2 Price Form	Class-2 Price Function
(i)	PI equal to FI solution $\Pi^{PI}(\mu)=\Pi^{FI}(\mu)$	n_1^*	No	Fixed	$P_2(\gamma)=(r_2-cW)\gamma_2$
(ii)	Class-2 only served $\Pi^{PI}(\mu)<\Pi^{FI}(\mu)$	$n_1=0$	NA	Pay-per-use	$P_2(\gamma)=(r_2-cW)\gamma$
(iii)	Class-1 priority, Class-2 surplus $\Pi^{PI}(\mu)<\Pi^{FI}(\mu)$	$n_1<n_1^*$	Yes	Two-part tariff	$P_2(\gamma)=(r_2-cW_2)\gamma$ $-\frac{c(\gamma_1-\gamma_2)}{\mu-n_1\gamma_1}+(r_1\gamma_1-r_2\gamma_2)$
(iv)	Class-1 priority Class-2 no surplus $\Pi^{PI}(\mu)<\Pi^{FI}(\mu)$	$n_1<n_1^*$	Yes	Pay per use	$P_2(\gamma)=(r_2-cW_2)\gamma$
(v)	Class-1 priority Class-2 no surplus $\Pi^{PI}(\mu)=\Pi^{FI}(\mu)$	$n_1=n_1^*$	Yes	Pay-per-use	$P_2(\gamma)=(r_2-cW_2)\gamma$

Table 1: Solution forms for high total valuation sub-case.

are still less attractive than type-2 who have a higher marginal value and so will be completely served as capacity expands, i.e., $n_2=N_2$, before any type-1 customers are served.)

The solution to the PI case can be classified as being in one of five forms. Each solution is defined by three elements: n_1, the number of type-1 customers served, W_1, their waiting time, and u_2, the consumer surplus of the type-2 customers. (In all cases $n_2=n_2^*$.) These also define W_2, the class-2 waiting time, $P_2(\gamma)$, the class-2 price function, and Π^{PI}, the revenue. In particular, we show that again there is a range of capacity where the firm will serve the type-1 customers with priority while providing a positive surplus to the type-2 customers. The five forms are summarized in Table 1; the details for all the solutions are given in Appendix B. As before W is the waiting time under FIFO service. Fixed prices imply a subscription price for unlimited usage; per use prices are just that. (Recall in all cases $P_1(\gamma)=(r_1-cW_1)\gamma_1$, a fixed price.) To emphasize the dependency of these cases on the capacity, we make explicit the dependency of the revenues as functions of μ, $\Pi^{PI}(\mu)$ and $\Pi^{FI}(\mu)$.

The solution form that holds depends on the capacity in the system. Recall from Proposition 4 that $\overline{\mu}$ is the lowest capacity for which the firm would serve type-1 customers and must be at least μ_2. For $\mu<\overline{\mu}$, either solution (i) or (ii) holds. Also recall IC requires

$$W_1 \le \widetilde{W} + \frac{u_2}{c(\gamma_1-\gamma_2)}.$$

The waiting time for a class-1 customer, if given priority, must be at least the service time, $1/\overline{\mu}$. Thus if $1/\overline{\mu} > \widetilde{W}$ or equivalently from (11), $\overline{\mu} < \widetilde{\mu}$, IC also requires that $u_2>0$ which is solution (iii). In this case, as capacity expands, at some point, $W_1=\widetilde{W}$. At this point, say $\widehat{\mu}$, the firm does not need to give additional incentive to type-2 customers, and for larger capacities, $u_2=0$ and solution (iv) holds. If $\overline{\mu}>\widetilde{\mu}$, then for all capacities greater than $\overline{\mu}$, $u_2=0$. Let $K_1=r_1(\widetilde{\mu}-N_2\gamma_2)^2/c$

and $K_2 = r_1\widetilde{\mu}^2/c$. These are the critical capacities that determine what happens when $n_1 = n_1^*$, i.e., cases (i) and (v) above. Proposition 5 summarizes all of the possibilities.

Proposition 5 *Suppose $r_1 < r_2$ and $r_1\gamma_1 > r_2\gamma_2$. Let μ_2 be defined as in Proposition 3.*

1. *If $\mu \le \mu_2$ or $\mu \ge K_2$, solution (i) holds; the PI solution = FI solution. For $\mu \le \mu_2$, $n_1 = n_1^* = 0$; for $\mu \ge K_2$, $n_1 = n_1^* > 0$.*
2. *If $\mu_2 > \widetilde{\mu}$, then for $\mu_2 < \mu < K_2$ solution (v) holds.*
3. *If $\mu_2 \le \widetilde{\mu}$, there exists $\overline{\mu}$ such that*
   a. *if $\overline{\mu} < \widetilde{\mu}$, there exists $\widehat{\mu}$ such that for $\mu < \overline{\mu}$ solution (ii) holds; for $\overline{\mu} \le \mu < \widehat{\mu}$ solution (iii) holds; for $\widehat{\mu} \le \mu < K_1$ solution (iv) holds; and for $K_1 \le \mu < K_2$ solution (v) holds.*
   b. *if $\overline{\mu} \ge \widetilde{\mu}$, then for $\mu < \widetilde{\mu}$ solution (ii) holds; and for $\widetilde{\mu} \le \mu < K_1$ solution (iv) holds; and for $K_1 \le \mu < K_2$ solution (v) holds.*

In Proposition 5, the PI and FI solutions are equivalent, for both very low ($\mu \le \mu_2$) and high ($\mu > K_2$) capacity, in contrast to Proposition 4. In the intervening range which case holds depends on the various model parameters. Case 3a illustrates the full range of solutions. We highlight that for $\mu < \overline{\mu}$, we do not serve type-1 customers – solution (ii), but for slightly higher capacity ($\overline{\mu} \le \mu \le \widehat{\mu}$), not only do we serve them, but we give them priority – solution (iii). As in Proposition 4, prioritization of the high-demand rate customers will be joined with discounting for low-demand rate customers. By prioritizing the class-1 service the firm can raise the total revenue it receives from these customers. However, the waiting time for the class-2 service will necessarily increase. To compensate, and continue to attract them, their price will decrease. Further, the class-2 price is given as a two-part tariff where the fixed part of the tariff represents the surplus utility the type-2 customers receive for this service. For higher capacity ($\widehat{\mu} < \mu < K_1$), class-1 service is still prioritized, but no surplus is given to the type-2 customers as with sufficient priority $W_1 \le \widetilde{W}$ – solution (iv). Finally for $K_1 \le \mu < K_2$ all of the customers served in the FI solution can be served, but prioritization is still required to ensure $W_1 \le \widetilde{W}$ – solution (v).

3.3. Summary of Results

To summarize, under the increasing ordering ($r_1 \ge r_2$), the FI and PI solutions are identical. The firm can skim the price as one would expect, allocating capacity to the higher value customers before serving lower value ones. However in the decreasing ordering ($r_1 < r_2$) this may not be the case. When the frequent customers do not value the service highly (the low total valuation subcase, $r_1\gamma_1 \le r_2\gamma_2$), the FI solution value dominates the PI solution value when there is sufficient capacity to serve both types of customers. Some n_1 of these customers will be prioritized. The type-2 customers' price will be lower because of their lower priority. Further, we show the price is even

lower than required for the individual rationality constraints to hold as the incentive compatibility constraints force the firm to provide a consumer surplus.

A similar result holds for the high total valuation sub-case ($r_1\gamma_1 > r_2\gamma_2$). Here, because of the high long-term value of the type-1 customers, there are several possibilities, depending on the capacity. In particular, the firm may choose not to serve type-1; serve a limited number of them with priority and provide a consumer surplus to type-2; serve a limited number and provide no consumer surplus to type-2; serve both types as in the FI setting though prioritize type-1; or simply use the FI solution.

The main point here is that for the decreasing ordering where high frequency customers value each interaction lower than the lower frequency customers, there are solutions where incentive compatibility requires prioritization even when all customers value waiting equally. Firms should be especially careful in their pricing when a class of customers would choose to frequent the service, while deriving smaller marginal benefit from each use. As long as their marginal value is not too low, the firm can benefit by serving them, possibly with priority. N.B., while our analysis assumes identical sensitivity to waiting for both types, it is clear that the firm can benefit from prioritizing type-1 customers even if their sensitivity to waiting is *less* than that of the type-2 customers.

4. The Value of Priority vs. FIFO Service

We compare the revenue received and the customer served under the Full Information, and Private Information solutions. We focus on the decreasing ordering case ($r_1 < r_2$). Recall that in this case when there is sufficient capacity ($\mu > \mu_2$), all N_2 type-2 customers and some of the type-1 customers are served. Recall also that the PI solution requires adherence to the incentive compatibility constraints, (7b) and (7c). We have shown that if the pricing menu determined in the FI solution is applied to the PI case, these constraints will not be observed and priority service may improve the revenue. To evaluate the benefit of prioritization, we also compare the PI solution to the best policy under restriction to FIFO service in the private information case. The firm has a single price and offers only FIFO service, but optimizes over the total number of customers. Let $n_1^{FI} = n_1^*$ be the optimal number served under the FI solution and n_1^{PI} be the optimal number served under the PI solution. Let n_1^{Sub} be the optimal number of type-1 customers served under the suboptimal policy. Let Π^{Sub} be the suboptimal revenue.

We compare the percentage difference between the FI solution and the PI solution to measure the value of information, and the difference between the PI solution and the suboptimal solution to measure the benefit of prioritization. Let

$$\Delta_{PI}^{FI} = \frac{\Pi^{FI} - \Pi^{PI}}{\Pi^{FI}} \times 100\% \quad \text{and} \quad \Delta_{Sub}^{PI} = \frac{\Pi^{PI} - \Pi^{Sub}}{\Pi^{PI}} \times 100\%.$$

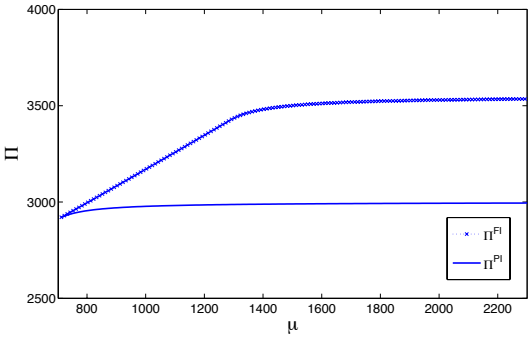

Figure 1: Example 1 – Decreasing order, low total valuation, with fewer type-1 customers $N_1 = 50$, $N_2 = 150$. Here $r_1 = 1$, $r_2 = 5$, $\gamma_1 = 11$, $\gamma_2 = 4$, and $c = 15$.

μ	Δ_{PI}^{FI}	n_1^{FI}	ρ^{FI}	ρ^{PI}
703	0.0	0.02	0.854	0.854
900	3.65	16.7	0.871	0.667
1100	8.5	33.8	0.883	0.545
1300	13	50	0.885	0.462
1500	14.6	50	0.767	0.4

Table 2: Example 1 – Percentage difference in profit, type-1 customers served, and utilization at various capacity levels.

Let ρ^{FI}, ρ^{PI}, and ρ^{Sub} the utilization under the FI, the PI and the suboptimal solutions, respectively. For all of the figures and tables, we present results for $\mu > \mu_2$, noting μ_2, as given by Proposition 3, depends on r_1, N_2, γ_2 and c.

4.1. The Low Total Valuation Sub-case

In this case, the high frequency of the type-1 customers does not result in higher total value from the type-1 customers, i.e., $r_1\gamma_1 < r_2\gamma_2$. We consider two examples. For both we let $r_1 = 1$, $r_2 = 5$, $\gamma_1 = 11$, $\gamma_2 = 4$, and $c = 15$.

Example 1. In the first example we let $N_1 = 50$ and $N_2 = 150$ so that the total value from type-2 only dominates the value if all customers purchase class-1: $N_2 r_2 \gamma_2 > (N_1 + N_2) r_1 \gamma_1$ or

$$\frac{N_1}{N_2} = \frac{1}{3} < \frac{9}{11} = \frac{r_2\gamma_2 - r_1\gamma_1}{r_1\gamma_1}.$$

In this case, by Proposition 4, $n_1^{PI} = 0$ so $\Pi^{Sub} = \Pi^{PI}$. The revenue provided by the few type-1 customers does not justify serving them even if they can be prioritized. In the FI solution type-1 customers are valuable for $\mu > \mu_2 = 703$. Figure 1 presents the revenues as a function of μ. Table 2 details the profit reduction percentage, the number of type-1 customers served under FI, and the utilization at various capacity levels. The PI optimal two-part tariff, attractive to type-2 only, provides significantly more revenue compared with providing a subscription price that would be attractive to type-1 as well. In this example the revenue and utilization under the FI case are substantially higher than in the PI case.

Example 2. We now let $N_1 = 150$ and $N_2 = 50$, reversing their values from Example 1. Now there are sufficient type-1 customers to make selling them a subscription while prioritizing them attractive. That is, because $N_1/N_2 = 3 > 9/11$, Proposition 4 implies $n_1^{PI} > 0$ when there is sufficient

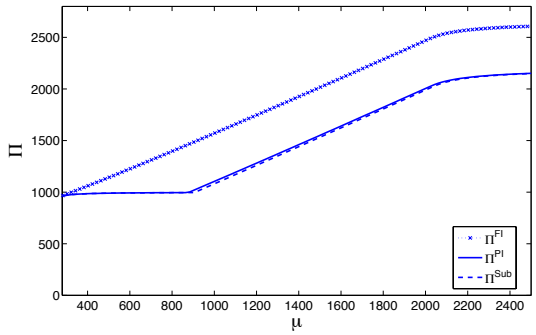

Figure 2: Example 2 – Decreasing order, high total valuation, with many type-1 customers $N_1 = 150$, $N_2 = 50$. Here $r_1 = 1$, $r_2 = 5$, $\gamma_1 = 11$, $\gamma_2 = 4$, and $c = 15$.

μ	Δ_{Sub}^{PI}	Δ_{PI}^{FI}	n_1^{FI}	n_1^{PI}	n_1^{Sub}	ρ^{FI}	ρ^{PI}	ρ^{Sub}
500	0	13.3	19.4	0	0	0.827	0.4	0.4
1000	2.1	29.7	61.6	61.3	59.8	0.877	0.874	0.858
1500	1.2	23.1	104.5	104.2	103	0.9	0.898	0.889
2000	0.75	18.8	147.9	147.6	146.6	0.913	0.912	0.906
2500	0	17.5	150	150	150	0.74	0.74	0.74

Table 3: Example 2 – Percentage difference in profits, type-1 customers served, and utilization at various capacity levels.

capacity. When capacity exceeds $\overline{\mu} \approx 870$, $n_1^{PI} > 0$. The type-1 customers are prioritized and the type-2 customers receive a positive consumer surplus. With sufficient capacity the waiting time tends to zero and by Lemma 3, $P_1(\gamma) = P_2(\gamma) = r_1\gamma_1$. That is, the type-2 customers receive a substantial discount of $r_2\gamma_2 - r_1\gamma_1 = 20 - 11 = 9$ or a 45% discount. This translates into the 13%–30% difference between the FI and PI solution – see Table 3. On the other hand, there is little difference between the profit of the PI solution and that of the Suboptimal solution that uses FIFO (1%–2%). Prioritizing the type-1 customers is of little benefit. By comparing the number of type-1 customers served and the system utilizations in each of the solutions (FI, PI, and Sub), we observe that almost all of the difference is attributable to the discount given to the type-2 customers in the PI and Sub cases.

4.2. The High Total Valuation Sub-case

In this case $r_1 < r_2$ and $r_1\gamma_1 > r_2\gamma_2$. That is, the type-1 customers are very attractive and the firm would want to serve as many as possible. However, the profitability from doing so depends on the cost of waiting. In Example 3 we assume a low cost of waiting; in Example 4, a high cost. For both examples we let $r_1 = 1$, $r_2 = 5$, $\gamma_1 = 21$, $\gamma_2 = 4$, $N_1 = N_2 = 100$.

Example 3. Here we let $c = 25$. Then from (11),

$$\widetilde{\mu} = \frac{1}{\widetilde{W}} = \frac{c(\gamma_1 - \gamma_2)}{r_1\gamma_1 - r_2\gamma_2} = \frac{25(21-4)}{21-20} = 425 < \mu_2 = 513.$$

μ	Δ_{Sub}^{PI}	n_1^{FI}	n_1^{Sub}	ρ^{FI}	ρ^{Sub}
600	1.5	3.7	0	0.796	0.667
800	7.8	12.3	7.14	0.823	0.687
1000	11	21	16.2	0.842	0.74
1200	9.2	29.8	25.3	0.856	0.776
1800	5.8	56.6	52.6	0.882	0.836
2400	4	83.6	80	0.898	0.867
3000	0	100	100	0.833	0.833

Table 4: Example 3 – Low waiting cost, $c = 25$. Type-1 receives priority and is served fully. Type-2 does not receive a discount. Here $r_1 = 1$, $r_2 = 5$, $\gamma_1 = 21$, $\gamma_2 = 4$, $N_1 = N_2 = 100$.

μ	Δ_{Sub}^{PI}	Δ_{FI}^{PI}	n_1^{FI}	n_1^{PI}	n_1^{Sub}	ρ^{FI}	ρ^{PI}	ρ^{Sub}
800	3	1.89	9.5	8.3	2.2	0.75	0.717	0.558
1200	14	1.2	26.4	25	20	0.796	0.773	0.683
1600	10.5	0.8	43.7	42.4	37.8	0.823	0.806	0.746
2000	8.3	0.5	59.8	61.1	55.7	0.842	0.828	0.785
2400	6.7	0.35	78.7	77.4	73.7	0.856	0.844	0.811
3000	1.82	0	100	100	100	0.833	0.833	0.833
3500	0	0	100	100	100	0.714	0.714	0.714

Table 5: Example 4 – High waiting cost, $c = 50$. Type-1 receives priority and is served partially. Type-2 receives a discount. Here $r_1 = 1$, $r_2 = 5$, $\gamma_1 = 21$, $\gamma_2 = 4$, $N_1 = N_2 = 100$.

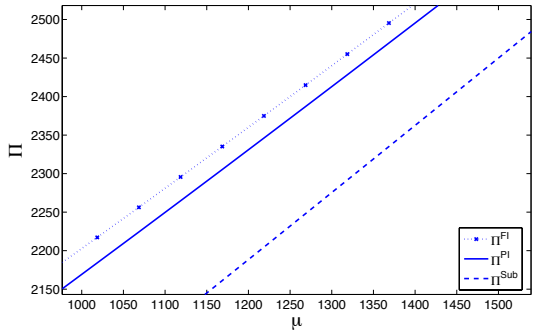

Figure 3: Example 4 – Decreasing order, high total valuation, with high waiting cost $c = 50$.

In this case, from Proposition 5.2, there is no difference between the number of customers served in the FI and PI solutions and $\Delta_{PI}^{FI} = 0$. However, in the PI solution priority service is given to the type-1 customers. This raises the price they pay. But the corresponding gain in revenue is offset by the reduction in the price paid by the type-2 customers. The priority differentiation only serves to discriminate between the two classes. Without prioritizing, all customers would pay the lower price resulting in the same waiting cost, but lowering the total revenue. This is expressed in that $\Delta_{Sub}^{PI} > 0$.

Example 4. Next we let $c = 50$. Then from (11),

$$\widetilde{\mu} = \frac{1}{\widetilde{W}} = \frac{c(\gamma_1 - \gamma_2)}{r_1\gamma_1 - r_2\gamma_2} = \frac{50(21-4)}{21-20} = 850 > \mu_2 = 569.$$

In this case, for $\mu > \mu_2$, case 3a of Proposition 5 holds. Because of the higher cost of waiting, in order to encourage type-1 customers to join, they need to be given an incentive in the form of priority service. As depicted in Table 5, however, the number is limited so that fewer are served than under the FI-case, ensuring the effectiveness of the waiting time reduction. Similarly, the type-2 customers would prefer such a priority, eliminating its effectiveness. To discourage them from switching, they receive a discount, at least while $\mu < \hat{\mu}$. In Table 5, this is demonstrated by the small percentage loss, given by Δ_{PI}^{FI}. Under the restriction to FIFO service, the suboptimal solution (Sub) would hold and fewer type-1 customers would join, lowering the revenue significantly, relative to the PI solution. Figure 3 illustrates the small loss vs. the FI solution and the large gain over the suboptimal FIFO service. (From Table 5 $\Delta_{Sub}^{PI} = 14\%$ at $\mu = 1200$ and stays above 6% even as capacity doubles.) Because the type-1 customers have a high total valuation, ensuring their full participation is of value. By prioritizing them the firm can significantly increase profits over FIFO service. We want to highlight that though some type-1 customers are excluded in the PI solution compared with the FI solution, the capacity utilization in the example is maintained between 70%-85%. This range may be considered moderate indicating that our example does not result from an extreme choice of parameters.

5. Discussion

Priority queues have long been studied as a means of discriminating between customers that differ in their cost of waiting. In this paper, we assume a common cost of waiting, and still demonstrate that priority queues may be revenue maximizing. Our insight is that if customers differ in their demand rate and marginal valuation of a service, prioritization provides a means of encouraging high-frequency, low marginal value customers to pay a little more for the service while allowing the firm to reduce the price for lower frequency customers. What may be thought of as a benefit or privilege provided to loyal, frequent-use customers, is demonstrated to be a tool that allows the firm to improve profitability. Moreover, our analysis implies that prioritizing customers with higher demand rate and lower marginal value may also be optimal if they are *more* patient than their low-frequency counterparts.

Using our ski resort example, the season pass that is sold to the locals is accompanied by admission to priority queues at the lifts and early access to the mountain, not just as a perk, but as a means of raising the price of the pass and the revenue of the firm. The higher price discourages the aways from purchasing the season pass. Further, the price of the daily pass sold to the aways is discounted so that they see a consumer surplus. Effectively, the locals subsidize the price paid by the aways, and the total revenue generated grows. Of course, this result depends on the demand and capacity of the firm. With too few locals or too little capacity, only the high-marginal value

aways are served. Offering a season pass does not generate sufficient revenue compared with the additional cost of waiting incurred by the aways. If the capacity is very high, all customers are served FIFO, and a single price is charged to all. But, importantly, at moderate levels of capacity, priority queues are optimal.

One concern may be that the policy hinges on the assumption that the firm would restrict the number of season passes sold. As demonstrated in our example, there may be some restriction in the number, in order to achieve the desired waiting time for the priority customers. However, the restricted number may only be slightly lower than the number of class-1 customers that would be served in the full information case. If one accepts the premise of revenue maximization by limiting the total waiting time as in Cachon and Feldman (2011), then the assumption the firm would limit the number of priority customers served should be acceptable.

In this work we have assumed that the customer classes have inherent demand rates. This would seem to hold for the aways (the low-rate users) in our example. They have limited number of days they can ski, and changes in pricing would likely not change their usage, but rather whether they purchase at all (i.e., not go to the resort). For the locals (the high-rate users), this assumption may be more one of convenience than actuality. Certainly, if a season pass is purchased, then the assumption of a fixed expected number of ski-days is reasonable. But in our model, the high-rate users do not modify their usage if a season pass is not sold. Rather, some purchase and others do not. However, the fundamental point of this paper will continue to hold in a model that considers usage rate decisions. That is, offering priority service may increase revenues as a result of demand rate heterogeneity, even if all customers share the same delay cost.

Appendix A: Proofs

Proof of Proposition 1 For the FI setting, the solution is found by taking the derivatives of the objective function of (FI) with respect to n_i given as

$$\frac{\partial \Pi(n_1,n_2)}{\partial n_i} = \gamma_i \left(r_i - \frac{c\mu}{(\mu - n_1\gamma_1 - n_2\gamma_2)^2} \right), \quad i = 1,2. \tag{14}$$

(Below we denote $\Pi_{n_i} \equiv \partial \Pi / \partial n_i$.)

Because $r_1 \geq r_2$ and $\gamma_1 > \gamma_2$, it follows from (14) that

$$\Pi_{n_1} > \Pi_{n_2}. \tag{15}$$

Then we have the following after some algebra:

1. When $\mu \leq \mu_0$, for any $n_1, n_2 \geq 0$, Π_{n_1} is non-positive ($\Pi_{n_1} = 0$ only at $\mu = \mu_0$), and Π_{n_2} is strictly negative. Therefore the provider does not gain any positive revenue from serving any customers.

2. When $\mu_0 \le \mu < \mu_1$, for any $0 \le n_1 \le N_1$ and $n_2 \ge 0$, Π_{n_2} is strictly negative, and therefore the provider does not serve class-2 customers. But $\Pi_{n_1}(0,0) > 0 > \Pi_{n_1}(N_1,0)$, and since
$$\frac{\partial^2 \Pi}{\partial n_1^2}(n_1,n_2) = -\frac{c\mu\gamma_1}{(\mu - n_1\gamma_1 - n_2\gamma_2)^3} < 0, \tag{16}$$
then for every $n_2 \ge 0$, Π is concave in n_1 and there exists a unique n_1^* between $0, N_1$ that maximizes the revenue.
3. When $\mu_1 \le \mu \le \mu_2$, $\Pi_{n_1}(N_1,0) > 0$ and there exists enough capacity such that the provider exhausts all type-1 customers. Yet, for every $n_2 \ge 0$, $\Pi_{n_2} < 0$, and therefore it is not profitable to serve any type-2 customers.
4. When $\mu_2 < \mu < \mu_3$, $\Pi_{n_2}(N_1,0) > 0 > \Pi_{n_2}(N_1,N_2)$. And since similar to (16) Π is concave in n_2 for any $n_1 \ge 0$, there exists a unique n_2^* between 0 and N_2 that maximizes the revenue.
5. When $\mu \ge \mu_3$, as in the previous case, it is profitable to serve both classes, and since capacity is high the provider serves all customers from both classes. ■

Proof of Proposition 2 Let Λ^* be the total arrival rate of customers under the FI setting solution, and let $W^* = W(\Lambda^*) = 1/(\mu - \Lambda^*)$ be the correspondence waiting time under FIFO.

By Proposition 1, if $\mu \le \mu_0 = c/r_1$, the trivial solution to serve no customers holds for both FI and PI settings. For $\mu > \mu_0$, it is optimal serve some type-1 customers in the FI setting. Therefore, we need (9a) to hold in the PI setting, to achieve the FI solution. But from the IR constraint (1b), $r_1 - cW(\Lambda^*) > 0$, which implies that $\mu \ge \Lambda^* + \mu_0$. But observe
$$\mu_0 = \frac{c}{r_1} = \frac{c(\gamma_1 - \gamma_2)}{r_1(\gamma_1 - \gamma_2)} > \frac{c(\gamma_1 - \gamma_2)}{r_1\gamma_1 - r_2\gamma_2} = \widetilde{\mu},$$
therefore $\mu \ge \Lambda^* + \widetilde{\mu}$ or equivalently $W(\Lambda^*) \le \widetilde{W}$. Thus the solution to the FI setting is feasible for the PI setting, and since it is an upper bound on the objective value for the PI setting, it is optimal. ■

Proof of Proposition 3 We establish below that over each relevant capacity range we have $\Pi_{n_1} < \Pi_{n_2}$. Then all steps in proof of Proposition 1 hold, after switching the subscripts "1" and "2". We distinguish between two cases:

1. If $r_1\gamma_1 \le r_2\gamma_2$, the total valuation of the type-1 customer is less than the type-2 customer. In this case, from (14) we have:
$$\begin{aligned}\Pi_{n_1} &= r_1\gamma_1 - \gamma_1 \frac{c\mu}{(\mu - n_1\gamma_1 - n_2\gamma_2)^2} \\ &< r_2\gamma_2 - \gamma_2 \frac{c\mu}{(\mu - n_1\gamma_1 - n_2\gamma_2)^2} = \Pi_{n_2},\end{aligned}$$
2. If $r_1\gamma_1 > r_2\gamma_2$, the total valuation of the type-1 customer is higher than the type-2 customer. Then:

(a) If $\mu < \widetilde{\mu} = c(\gamma_1 - \gamma_2)/(r_1\gamma_1 - r_2\gamma_2)$, we show that $\Pi_{n_1} < \Pi_{n_2}$:

Let $\Lambda = n_1\gamma_1 + n_2\gamma_2$. For $0 < \Lambda < \mu$, $\frac{(\mu-\Lambda)^2}{\mu} < \mu$. Therefore

$$\begin{aligned}
&\frac{(\mu-\Lambda)^2}{\mu} < \frac{c(\gamma_1-\gamma_2)}{r_1\gamma_1 - r_2\gamma_2} \\
\Rightarrow \quad &\frac{r_1\gamma_1 - r_2\gamma_2}{c(\gamma_1-\gamma_2)} > \frac{\mu}{(\mu-\Lambda)^2} \\
\Rightarrow \quad &r_1\gamma_1 - \gamma_1\frac{c\mu}{(\mu-\Lambda)^2} < r_2\gamma_2 - \gamma_2\frac{c\mu}{(\mu-\Lambda)^2} \\
\Rightarrow \quad &\Pi_{n_1} < \Pi_{n_2}.
\end{aligned}$$

(b) If $\mu \geq \widetilde{\mu}$, observe that

$$\begin{aligned}
&\Lambda > \mu - \sqrt{\mu\widetilde{\mu}} \\
\Rightarrow \quad &\frac{(\mu-\Lambda)^2}{\mu} < \widetilde{\mu} = \frac{c(\gamma_1-\gamma_2)}{r_1\gamma_1 - r_2\gamma_2} \\
\Rightarrow \quad &\frac{(\mu-\Lambda)^2}{\mu} < \widetilde{\mu} = \frac{c(\gamma_1-\gamma_2)}{r_1\gamma_1 - r_2\gamma_2} \\
\Rightarrow \quad &r_1\gamma_1 - \gamma_1\frac{c\mu}{(\mu-\Lambda)^2} < r_2\gamma_2 - \gamma_2\frac{c\mu}{(\mu-\Lambda)^2} \\
\Rightarrow \quad &\Pi_{n_1} < \Pi_{n_2}.
\end{aligned}$$

We show that this condition is either satisfied or redundant for any capacity level. Observe that for any $\mu_0 < \mu < \mu_1$, Λ^* satisfies $r_2 = c\mu/(\mu-\Lambda^*)^2$, or

$$\Lambda^* = \mu - \sqrt{\frac{c\mu}{r_2}} = \mu - \sqrt{\mu\mu_0} > \mu - \sqrt{\mu\widetilde{\mu}}.$$

Therefore $\Pi_{n_1} < \Pi_{n_2}$ for $\mu < \mu_1$. For $\mu > \mu_1$, $n_2^* = N_2$. To increase Λ one needs to serve type-1 customers as well, and the provider does so only when $\Pi_{n_1} > 0$, which is when $\mu > \mu_2$. For $\mu_2 < \mu < \mu_3$, Λ^* satisfies $r_1 = c\mu/(\mu-\Lambda^*)^2$, or equivalently $\Lambda^* = \mu - \sqrt{\frac{c\mu}{r_1}} > \mu - \sqrt{\mu\widetilde{\mu}}$. For $\mu \geq \mu_3$, both types are fully served and therefore the condition is redundant.

■

Proof of Proposition 4 We consider two cases:

1. When $\mu \leq \mu_2$, the solution given in Proposition 3 implies that $n_1^* = 0$. If $n_1^* = 0$ for the PI setting, then constraint (9a) is not active. The remaining problem is identical to problem (FI). Therefore the solution to the FI setting is feasible and since the PI setting value is bounded by the FI solution, it is optimal.

2. If $\mu > \mu_2$, for the FI setting $n_1^* > 0$. Let $\overline{n}_1$ be the number of type-1 customers that are served in the PI setting. If $\overline{n}_1 > 0$, (9a) is active. Since $r_1\gamma_1 \leq r_2\gamma_2$, $\widetilde{W} \leq 0$ so for any class-1 customers to be served (and $W_1 > 0$), by (9a) we required $u_2 > 0$.

Observe further that any increase in u_1 would require an additional increase in u_2, both lowering the objective value of (PI). Therefore under any optimal solution $u_1 = 0$. Further, in any optimal solution (9b) will be tight for class-1 customers, i.e.,

$$W_1 = \frac{1}{\mu - \overline{n}_1\gamma_1}, \tag{17}$$

as increasing W_1 only increases u_2, lowering the objective function value. So, u_2 is given by solving (9a) at equality, i.e.,

$$u_2 = \frac{c(\gamma_1 - \gamma_2)}{\mu - \overline{n}_1\gamma_1} - (r_1\gamma_1 - r_2\gamma_2). \tag{18}$$

Note that $u_2 > 0$ if $\overline{n}_1 > 0$. Because $\mu > \mu_2$, the argument given in (15) (reversing subscripts "1" and "2") holds here, so $\Pi_{n_1} < \Pi_{n_2}$. Since $u_2 > 0$ implies that $n_2 = N_2$, we have $\Lambda = \overline{n}_1\gamma_1 + N_2\gamma_2$. In addition, because class-1 customers receive priority, there is no value in choosing a larger delay for class-2. Therefore constraint (9c) holds at equality, i.e.,

$$\sum_i n_i\gamma_i W_i = \frac{\sum_i n_i\gamma_1}{\mu - \sum_i n_i\gamma_1},$$

which implies that

$$W_2 = \frac{\mu}{(\mu - \overline{n}_1\gamma_1)(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)}.$$

So if $\overline{n}_1 > 0$, the objective value for (PI-2), say Π_1, is

$$\Pi_1 = N_2\gamma_2((r_2 - cW_2) - u_2) + \overline{n}_1\gamma_1(r_1 - cW_1).$$

If $n_1 = 0$, the objective value, say Π_0, is

$$\Pi_0 = N_2\gamma_2(r_2 - cW),$$

where

$$W = \frac{1}{\mu - N_2\gamma_2}.$$

As Π_0 and Π_1 together depend on μ, we would like to find the range of μ such that $\Pi_1 > \Pi_0$. We claim that there exists $\overline{\mu} > \mu_2$ such that $\mu < \overline{\mu}$ implies $\Pi_1 < \Pi_0$ and $\mu \geq \overline{\mu}$ implies $\Pi_1 \geq \Pi_0$. Let

$$\Delta\Pi \equiv \Pi_1 - \Pi_0 = N_2\gamma_2((r_2 - cW_2) - u_2) + \overline{n}_1\gamma_1(r_1 - cW_1) - N_2\gamma_2(r_2 - cW).$$

In Lemma 1 we find the value of $\overline{n}_1$ and show that $\overline{n}_1 \leq n_1^*$. In Lemma 2 we find that when substituting $\overline{n}_1, u_2, W_1, W_2$ and W as given above, for $\mu > \mu_2$, $\Delta\Pi$ is increasing with μ. Moreover, $\Delta\Pi < 0$ as $\mu \to \mu_2$. We also show that if $N_1/N_2 > (r_2\gamma_2 - r_1\gamma_1)/r_1\gamma_1$, $\Delta\Pi > 0$ when $\mu \to \infty$. The intermediate value theorem implies that in this case, there exists $\overline{\mu} > \mu_2$ such that for $\mu > \overline{\mu}$, $\Delta\Pi > 0$ and for $\mu < \overline{\mu}$, $\Delta\Pi < 0$. Otherwise, if $N_1/N_2 \leq (r_2\gamma_2 - r_1\gamma_1)/r_1\gamma_1$, $\Delta\Pi \leq 0$ for $\mu > \mu_2$.

Lemma 1 *If $r_1 < r_2$, $r_1\gamma_1 \le r_2\gamma_2$, and $\mu > \overline{\mu}$, the provider serves $\overline{n}_1$ class-1 customers, such that:*

$$\overline{n}_1 : \quad \min\left(n_1 : r_1 - \frac{c\mu}{(\mu - n_1\gamma_1 - N_2\gamma_2)^2} - \frac{N_2 c(\gamma_1-\gamma_2)}{(\mu - n_1\gamma_1)^2} = 0,\ N_1\right),$$

where

1. *$\overline{n}_1$ maximizes Π_1,*
2. *$\overline{n}_1 < n_1^*$ for $n_1^* < N_1$, and $\overline{n}_1 = n_1^*$ only when $\overline{n}_1 = N_1$.*

Proof of Lemma 1

(a) Observe

$$\begin{aligned}\Pi_1 &= N_2\gamma_2((r_2 - cW_2) - u_2) + n_1\gamma_1(r_1 - cW_1)\\ &= N_2(r_2\gamma_2 - cW_2\gamma_2 - c(\gamma_1-\gamma_2)W_1 + r_1\gamma_1 - r_2\gamma_2) + \overline{n}_1 r_1\gamma_1 - c\overline{n}_1 W_1\gamma_1\\ &= (n_1+N_2)r_1\gamma_1 - \frac{c(\overline{n}_1\gamma_1 + N_2\gamma_2)}{\mu - \overline{n}_1\gamma_1 - N_2\gamma_2} - \frac{cN_2(\gamma_1-\gamma_2)}{\mu - \overline{n}_1\gamma_1} \qquad (19)\end{aligned}$$

To find the n_1 that maximizes Π_1, we take the derivative of Π_1 with respect to n_1 and compare it to 0.

$$\frac{\partial \Pi_1}{\partial n_1} = \gamma_1\left(r_1 - \frac{c\mu}{(\mu - n_1\gamma_1 - N_2\gamma_2)^2} - \frac{N_2 c(\gamma_1-\gamma_2)}{(\mu - n_1\gamma_1)^2}\right) = 0. \qquad (20)$$

Let $\overline{n}_1$ be the solution of (20). Then $\overline{n}_1$ maximizes Π_1 if the second derivative of Π_1 with respect to n_1 at $\overline{n}_1$ is strictly negative. Observe

$$\frac{\partial^2 \Pi_1}{\partial {n_1}^2}|_{n_1=\overline{n}_1} = -\gamma_1^2\left(\frac{2c\mu}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^3} + \frac{2N_2 c(\gamma_1-\gamma_2)}{(\mu - \overline{n}_1\gamma_1)^3}\right) < 0. \qquad (21)$$

Inequality (21) because for every $\mu > \mu_2$, the FI solution implies that $\mu > n_1^*\gamma_1 + N_2\gamma_2$. If $\overline{n}_1 \le n_1^*$, then $\mu > \overline{n}_1\gamma_1 + N_2\gamma_2$ and the inequality holds. What remains to prove is that $\overline{n}_1 \le n_1^*$. We show this below.

(b) For simplicity, define

$$f(n_1) = r_1 - \frac{c\mu}{(\mu - n_1\gamma_1 - N_2\gamma_2)^2} - \frac{N_2 c(\gamma_1-\gamma_2)}{(\mu - n_1\gamma_1)^2}.$$

As we show above, $\overline{n}_1$ is the solution of $f(n_1) = 0$. Also define

$$\begin{aligned} g(n_1) &= r_1 - \frac{c\mu}{(\mu - n_1\gamma_1 - N_2\gamma_2)^2},\\ h(n_1) &= \frac{N_2 c(\gamma_1-\gamma_2)}{(\mu - n_1\gamma_1)^2}.\end{aligned}$$

Then $f(n_1) = g(n_1) - h(n_1)$. For every $n_1 \le n_1^*$, the FI solution implies that $\mu > n_1\gamma_1 + N_2\gamma_2$. Therefore, $h(n_1)$ is strictly positive for $n_1 \le n_1^*$. From Proposition 3, n_1^* is the solution of $g(n_1) = 0$ if $n_1^* < N_1$. Therefore, when substituting n_1^* into $f(n_1)$, we have

$$f(n_1^*) = -h(n_1^*) < 0$$

As we proved above, $f(n_1)$, which is the first derivative of Π_1 with respect to n_1, is strictly decreasing in n_1 (see inequality (21)). Therefore, the solution of $f(n_1)=0$ has to satisfy $\overline{n}_1 < n_1^*$. The equality $\overline{n}_1 = n_1^*$ is only possible when $\overline{n}_1 = N_1$. □

Lemma 2

1. *For $\mu > \mu_2$, $\Delta\Pi$ is increasing with μ.*
2. $\lim_{\mu\to\mu_2^+} \Delta\Pi < 0$.
3. *If $N_1/N_2 > (r_2\gamma_2 - r_1\gamma_1)/r_1\gamma_1$ then $\lim_{\mu\to\infty} \Delta\Pi > 0$. Otherwise, $\lim_{\mu\to\infty} \Delta\Pi \le 0$.*

Proof of Lemma 2

(a)

$$\begin{aligned}\Delta\Pi = \Pi_1 - \Pi_0 &= N_2((r_2 - cW_2)\gamma_2 - u_2) + \overline{n}_1\gamma_1(r_1 - cW_1) - N_2\gamma_2(r_2 - cW)\\ &= N_2(r_2\gamma_2 - cW_2\gamma_2 - c(\gamma_1-\gamma_2)W_1 + r_1\gamma_1 - r_2\gamma_2) + \overline{n}_1 r_1\gamma_1 - c\overline{n}_1 W_1\gamma_1 - N_2 r_2\gamma_2 + cN_2 W\gamma_2\\ &= (\overline{n}_1 + N_2)r_1\gamma_1 - N_2 r_2\gamma_2 - c(\overline{n}_1\gamma_1 W_1 + N_2\gamma_2 W_2) - cN_2(\gamma_1-\gamma_2)W_1 + cN_2\gamma_2 W. \qquad (22)\end{aligned}$$

Recall that the definitions of W_1 and W_2 imply that constraint (9c) is binding, meaning:

$$\overline{n}_1\gamma_1 W_1 + N_2\gamma_2 W_2 = \frac{\overline{n}_1\gamma_1 + N_2\gamma_2}{\mu - \overline{n}_1\gamma_1 - N_2\gamma_2}. \qquad (23)$$

Substituting (23) and W, W_1, W_2 into (22), we have:

$$\begin{aligned}\Delta\Pi &= (\overline{n}_1 + N_2)r_1\gamma_1 - N_2 r_2\gamma_2 - \frac{c(\overline{n}_1\gamma_1 + N_2\gamma_2)}{\mu - \overline{n}_1\gamma_1 - N_2\gamma_2} - \frac{cN_2(\gamma_1-\gamma_2)}{\mu - \overline{n}_1\gamma_1} + \frac{cN_2\gamma_2}{\mu - N_2\gamma_2}\\ &= (\overline{n}_1 + N_2)r_1\gamma_1 - N_2 r_2\gamma_2 - \frac{c\mu\overline{n}_1\gamma_1}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)(\mu - N_2\gamma_2)} - \frac{cN_2(\gamma_1-\gamma_2)}{\mu - \overline{n}_1\gamma_1}.\end{aligned}$$

The derivative of $\Delta\Pi$ with respect to μ is:

$$\begin{aligned}\frac{\mathrm{d}\Delta\Pi}{\mathrm{d}\mu} &= \frac{c\overline{n}_1\gamma_1(\mu^2 - N_2\gamma_2(\overline{n}_1\gamma_1 + N_2\gamma_2))}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^2(\mu - N_2\gamma_2)^2} + \frac{cN_2(\gamma_1-\gamma_2)}{(\mu - \overline{n}_1\gamma_1)^2}\\ &\quad + \gamma_1\left(r_1 - \frac{c\mu}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^2} - \frac{N_2 c(\gamma_1-\gamma_2)}{(\mu - \overline{n}_1\gamma_1)^2}\right)\frac{\partial\overline{n}_1}{\partial\mu} \qquad (24)\end{aligned}$$

The third term in (24) is the derivative of $\Delta\Pi$ with respect to n_1 at the point where $n_1 = \overline{n}_1$, times the derivative of $\overline{n}_1$ with respect to μ. Since $\overline{n}_1$ is defined in Lemma 1 such that $\frac{\partial\Delta\Pi}{\partial n_1}|_{n_1=\overline{n}_1} = 0$, this term vanishes. Therefore,

$$\frac{\mathrm{d}\Delta\Pi}{\mathrm{d}\mu} = \frac{c\overline{n}_1\gamma_1(\mu^2 - N_2\gamma_2(\overline{n}_1\gamma_1 + N_2\gamma_2))}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^2(\mu - N_2\gamma_2)^2} + \frac{cN_2(\gamma_1-\gamma_2)}{(\mu - \overline{n}_1\gamma_1)^2} \qquad (25)$$

All the elements in the second term of (25) are positive. Observe for every $\mu > \mu_2$, the FI solution implies that the provider has enough capacity to serve n_1^* class-1 customers, in addition to all

N_2 class-2 customers, or alternatively, $\mu > n_1^*\gamma_1 + N_2\gamma_2$. By Lemma 1, $\overline{n}_1 \le n_1^*$ implying that $\mu > \overline{n}_1\gamma_1 + N_2\gamma_2$. Then the numerator of the first term in (25)

$$\begin{aligned} c\overline{n}_1\gamma_1[\mu^2 - N_2\gamma_2(\overline{n}_1\gamma_1 + N_2\gamma_2)] &> c\overline{n}_1\gamma_1[\mu^2 - (\overline{n}_1\gamma_1 + N_2\gamma_2)^2] \\ &= c\overline{n}_1\gamma_1[\mu - (\overline{n}_1\gamma_1 + N_2\gamma_2)][\mu + (\overline{n}_1\gamma_1 + N_2\gamma_2)] > 0 \end{aligned}$$

which implies that $\mathrm{d}\Delta\Pi/\,\mathrm{d}\mu > 0$.

(b) The FI solution implies that when $\mu \to \mu_2^+$, $n_1^* \to 0^+$. As we showed in Lemma 1, $\overline{n}_1 \le n_1^*$ and therefore $\overline{n}_1 \to 0^+$. Therefore, $W_2 \to W$. But since $\overline{n}_1 \to 0^+$, class-1 customers are not being served but all class-2 customers get a consumer surplus. As a result, $\Pi_1 < \Pi_0$ since

$$\begin{aligned} \lim_{\mu\to\mu_2^+} \Delta\Pi &= \lim_{\mu\to\mu_2^+} (N_2((r_2 - cW_2)\gamma_2 - u_2) + \overline{n}_1(r_1 - cW_1)\gamma_1 - N_2[r_2 - cW]\gamma_2]) \\ &= -N_2 \lim_{\mu\to\mu_2^+} u_2 = -N_2 \lim_{\mu\to\mu_2^+} \left(\frac{c(\gamma_1 - \gamma_2)}{\mu - \overline{n}_1\gamma_1} - (r_1\gamma_1 - r_2\gamma_2)\right) \\ &= -N_2 \left(\frac{c(\gamma_1 - \gamma_2)}{\mu} - (r_1\gamma_1 - r_2\gamma_2)\right) < 0 \end{aligned}$$

(c) When $\mu \to \infty$, the waiting times vanish, meaning $W, W_1, W_2 \to 0$. Therefore,

$$\begin{aligned} \lim_{\mu\to\infty} \Delta\Pi &= \lim_{\mu\to\infty} \Big(N_2((r_2 - cW_2)\gamma_2 - c(\gamma_1 - \gamma_2)W_1 + r_1\gamma_1 - r_2\gamma_2) \\ &\qquad + \overline{n}_1(r_1 - cW_1)\gamma_1 - N_2(r_2 - cW]\gamma_2) \Big) \\ &= (\overline{n}_1 + N_2)r_1\gamma_1 - N_2 r_2\gamma_2. \end{aligned}$$

So, if $N_1/N_2 > (r_2\gamma_2 - r_1\gamma_1)/r_1\gamma_1$ then $\lim_{\mu\to\infty} \Delta\Pi > 0$. Otherwise $\lim_{\mu\to\infty} \Delta\Pi \le 0$. □

Lemma 3 *For $\mu > \mu_2$, the price functions are:*

$$\begin{aligned} \overline{P}_1(\gamma) &= (r_1 - cW_1)\gamma_1 > P_1^*(\gamma_1), \\ \overline{P}_2(\gamma) &= (r_2 - cW_2)\gamma - \frac{c(\gamma_1 - \gamma_2)}{\mu - n_1\gamma_1} + (r_1\gamma_1 - r_2\gamma_2) < P_2^*(\gamma_2). \end{aligned}$$

and $\Pi^{PI} < \Pi^{FI}$.

Proof of Lemma 3 The proof follows directly from the definitions of W_1 in (17) and u_2 in (18). Since $u_2 > 0$, the reduction in the price of class-2 customers is strictly larger than the increase in the price for class-1 customers. Given that $\overline{n}_1 < n_1^*$ as we proved in Lemma 1, the total revenue $\Pi^{PI} = \overline{n}_1\overline{P}_1(\gamma_1) + N_2\overline{P}_2(\gamma_2)$ is strictly less then $\Pi^{FI} = n_1P_1^*(\gamma_1) + N_2P_2^*(\gamma_2)$. □

By proving that $\Delta\Pi$ is increasing with μ, we can claim the following: Let $\overline{\mu}$ be the largest root of $\Delta\Pi$. If $\overline{\mu} \le \mu_2$, then for all $\mu > \mu_2$, $\Delta\Pi > 0$, and the provider serves $\Lambda = \overline{n}_1\gamma_1 + N_2\gamma_2$. Otherwise, if $\overline{\mu} > \mu_2$, $\Delta\Pi < 0$ for all $\mu_2 < \mu < \overline{\mu}$ and $\Delta\Pi > 0$ for all $\mu > \overline{\mu}$. In the first case, the provider serves class-2 customers exclusively: $\Lambda = N_2\gamma_2$, while in the second case the provider serves both classes: $\Lambda = \overline{n}_1\gamma_1 + N_2\gamma_2$. ■

Proof of Proposition 5 We prove the proposition in several steps. Step 1. shows the result for $\mu \le \mu_2$ – part of case 1 in the proposition. Step 2. considers the case where $\mu_2 < \mu < K_2$ and $\mu_2 \le \widetilde{\mu}$ and establishes case 3a. and 3b. of the proposition. Step 3. considers the case $\mu_2 < \mu < K_2$ and $\widetilde{\mu} < \mu_2$ establishing case 2 of the proposition. Finally Step 4. considers the case $\mu \ge K_2$ which establishes the remainder of case 1.

Step 1. For $\mu \le \mu_2$ the proof is exactly the same as in Proposition 4, case 1.

Step 2. If $\mu > \mu_2$, the provider serves class-2 customers fully, $n_2^* = N_2$. The provider considers whether or not to also serve some class-1 customers. Recall constraint (9a):

$$n_1 > 0 \quad \Rightarrow \quad W_1 \le \frac{r_1\gamma_1 - r_2\gamma_2}{c(\gamma_1-\gamma_2)} - \frac{u_1-u_2}{c(\gamma_1-\gamma_2)} = \widetilde{W} + \frac{u_2}{c(\gamma_1-\gamma_2)},$$

that is, if $n_1 > 0$, the inequality must be satisfied.

We now consider cases 3a. and 3b. at the same time. We determine value of $\overline{\mu}$ and verify that 3a. holds for $\overline{\mu} < \widetilde{\mu}$ and 3b. holds if $\overline{\mu} \ge \widetilde{\mu}$.

Recall that $\widetilde{\mu}$ is the minimum capacity that is required in order to serve class-1 customers with $u_2 = 0$. For $\mu < \widetilde{\mu}$, if $n_1 > 0$, the minimum waiting time for class-1 customers does not satisfy constraint (9a):

$$W_1 = \frac{1}{\mu - n_1\gamma_1} > \frac{1}{\mu} > \frac{1}{\widetilde{\mu}} = \widetilde{W}.$$

Therefore, we require that class-2 customers receive consumer surplus, i.e., $u_2 > 0$. Recall $\overline{n}_1$ is the optimal number of customers that are served when $u_2 > 0$. We argue as before that under maximization, W_1 is tight, i.e.,

$$W_1 = \frac{1}{\mu - \overline{n}_1\gamma_1}.$$

Solving (9a) at equality we find

$$\begin{aligned} u_2 &= \frac{c(\gamma_1-\gamma_2)}{\mu-\overline{n}_1\gamma_1} - (r_1\gamma_1 - r_2\gamma_2) = \left(\frac{r_1\gamma_1 - r_2\gamma_2}{\mu-\overline{n}_1\gamma_1}\right)\left[\frac{c(\gamma_1-\gamma_2)}{r_1\gamma_1 - r_2\gamma_2} - (\mu - \overline{n}_1\gamma_1)\right] \\ &= \left(\frac{r_1\gamma_1 - r_2\gamma_2}{\mu-\overline{n}_1\gamma_1}\right)(\widetilde{\mu} - \mu + \overline{n}_1\gamma_1) > 0. \end{aligned}$$

Therefore, when $\mu < \widetilde{\mu}$, the provider compares the expected revenue from serving class-2 customers exclusively and fully (as in solution ii) and the expected revenue from serving $\overline{n}_1$ class-1 customers also (as in solution iii). We define $\Delta\Pi = \overline{\Pi}_1 - \Pi_0$. In Lemma 4 we prove that there exists a threshold $\overline{\mu}$ such that for $\mu < \overline{\mu}$, $\Delta\Pi < 0$, and for $\mu > \overline{\mu}$, $\Delta\Pi > 0$.

Although $\overline{\mu}$ always exists, it is not always a threshold between two different policies. When $\mu \ge \widetilde{\mu}$, the provider can satisfy constraint (9a) by serving class-1 customers with full priority and $u_2 = 0$. Therefore, for $\mu \ge \widetilde{\mu}$ we compare the revenue when $u_2 > 0$ (as in solution iii, when $\Pi = \overline{\Pi}_1$) and when $u_2 = 0$ (as in solution iv, when $\Pi = \widehat{\Pi}_1$).

We distinguish between two cases:

a. When $\overline{\mu} < \widetilde{\mu}$: we show that there exist a threshold $\widehat{\mu} > \widetilde{\mu}$ such that for $\mu < \widehat{\mu}$, $\overline{\Pi}_1 > \widehat{\Pi}_1$ and for $\widehat{\mu} < \mu < K_1$, $\overline{\Pi}_1 < \widehat{\Pi}_1$. Therefore, for $\mu < \overline{\mu}$ solution (ii) holds; for $\overline{\mu} \le \mu < \widehat{\mu}$ solution (iii) holds; and for $\widehat{\mu} \le \mu < K_1$ solution (iv) holds.
b. When $\overline{\mu} \ge \widetilde{\mu}$: we show that for $\widetilde{\mu} \le \mu < K_1$, $\overline{\Pi}_1 < \widehat{\Pi}_1$. Therefore, for $\mu < \widetilde{\mu}$ solution (ii) holds; and for $\widetilde{\mu} \le \mu < K_1$ solution (iv) holds.

In Lemma 5 we find the value of $\widehat{n}_1$, which is defined in solution (iv). In Lemma 6 we show that both $\overline{n}_1$ and $\widehat{n}_1$ increase with μ, although $\widehat{n}_1$ increases faster with μ. We also show that it is not feasible to serve $\overline{n}_1$ if $\overline{n}_1 < \widehat{n}_1$.

We use Lemma 6 to prove Lemma 7. Lemma 7 states that if $\overline{\mu} < \widetilde{\mu}$ there exists $\widehat{\mu}$ such that for $\mu < \widehat{\mu}$ the optimal solution is to serve $\overline{n}_1$ with $u_2 > 0$, and for $\mu > \widehat{\mu}$ the optimal solution is to serve $\widehat{n}_1$ with $u_2 = 0$. Otherwise, if $\overline{\mu} \ge \widetilde{\mu}$, then for $\mu > \widehat{\mu}$ the optimal solution is to serve $\widehat{n}_1$ with $u_2 = 0$.

It is left to show that in both cases, the revenue is less then the revenue of the FI solution. From Lemma 3, $\overline{\Pi} < \Pi^*$. In Lemma 8 we show that $\widehat{\Pi} < \Pi^*$.

Lemma 4 *Let $\Delta\Pi = \overline{\Pi}_1 - \Pi_0$. There exists $\overline{\mu}$ such that for $\mu < \overline{\mu}$, $\Delta\Pi = \overline{\Pi}_1 - \Pi_0 < 0$, and for $\mu \ge \overline{\mu}$, $\Delta\Pi \ge 0$, where the inequality is strict for $\mu > \overline{\mu}$.*

Proof of Lemma 4 The proof that such $\overline{\mu}$ exists is similar to the proof of Proposition 4 with the exception that $\overline{\mu}$ always exists in the current case. To be precise, the difference is in statement iii of Lemma 2: In the current case, $r_2\gamma_2 - r_1\gamma_1 < 0$. Therefore $\frac{N_1}{N_2} > 0 > \frac{r_2\gamma_2 - r_1\gamma_1}{r_1\gamma_1}$, and follows from that $\lim_{\mu\to\infty} \Delta\Pi > 0$ always. Therefore, by the intermediate value theorem, $\overline{\mu} > \mu_2$ is well defined. □

Lemma 5 *$\widehat{\Pi}_1$ is increasing with $\widehat{n}_1$ and is maximized at*

$$\widehat{n}_1 = \frac{\mu - \widetilde{\mu}}{\gamma_1}. \tag{26}$$

Proof of Lemma 5 From (9a) with $u_1 = u_2 = 0$:

$$W_1 = \frac{1}{\mu - \widehat{n}_1\gamma_1} \le \frac{1}{\widetilde{\mu}} = \widetilde{W},$$

which implies that

$$\widehat{n}_1\gamma_1 \le \mu - \widetilde{\mu}. \tag{27}$$

Let $\widehat{\Pi}_1$ be the revenue from serving $\widehat{n}_1$ with $u_2 = 0$. Then

$$\widehat{\Pi}_1 = N_2\gamma_2(r_2 - cW_2) + \widehat{n}_1\gamma_1(r_1 - cW_1), \tag{28}$$

where

$$W_1 = \frac{1}{\mu - \widehat{n}_1\gamma_1},$$

and

$$W_2 = \frac{\mu}{(\mu - \widehat{n}_1\gamma_1)(\mu - \widehat{n}_1\gamma_1 - N_2\gamma_2)}.$$

As we argued above,

$$\widehat{n}_1\gamma_1 W_1 + N_2\gamma_2 W_2 = (\widehat{n}_1\gamma_1 + N_2\gamma_2)W = \frac{\widehat{n}_1\gamma_1 + N_2\gamma_2}{\mu - \widehat{n}_1\gamma_1 - N_2\gamma_2}.$$

Substituting it into (28), we get:

$$\widehat{\Pi}_1 = N_2\gamma_2 r_2 + \widehat{n}_1\gamma_1 r_1 - \frac{c(\widehat{n}_1\gamma_1 + N_2\gamma_2)}{\mu - \widehat{n}_1\gamma_1 - N_2\gamma_2} \tag{29}$$

Taking the derivative of $\widehat{\Pi}_1$ with respect to $\widehat{n}_1$

$$\frac{\partial\widehat{\Pi}_1}{\partial\widehat{n}_1} = \gamma_1\left[r_1 - \frac{c\mu}{(\mu - \widehat{n}_1\gamma_1 - N_2\gamma_2)^2}\right] \tag{30}$$

From Proposition 4, for $\mu > \mu_2$,

$$r_1 = \frac{c\mu}{(\mu - n_1^*\gamma_1 - N_2\gamma_2)^2}. \tag{31}$$

To show that the derivative given in (30) is positive, it is enough to show $n_1^* > \widehat{n}_1$. From equation (31) we get:

$$n_1^*\gamma_1 = \mu - N_2\gamma_2 - \sqrt{\frac{c\mu}{r_1}}. \tag{32}$$

From (27), $\widehat{n}_1\gamma_1 \le \mu - \widetilde{\mu}$. Therefore, to show that $n_1^* > \widehat{n}_1$, we require

$$N_2\gamma_2 + \sqrt{\frac{c\mu}{r_1}} < \widetilde{\mu}. \tag{33}$$

Inequality (33) is equivalent to $\mu < r_1(\widetilde{\mu} - N_2\gamma_2)^2/c$, which is the upper bound of μ in the current case. Therefore, we can conclude that the derivative (30) is positive, and $\widehat{\Pi}_1$ is increasing with $\widehat{n}_1$. As a result, $\widehat{\Pi}_1$ is maximized at $\widehat{n}_1 = (\mu - \widetilde{\mu})/\gamma_1$. □

Lemma 6 *(a)* $\partial\widehat{n}_1/\partial\mu > \partial\overline{n}_1/\partial\mu > 0$ *and (b)* $\overline{n}_1 < \widehat{n}_1$ *implies* $u_2(\overline{n}_1) < 0$.

Proof of Lemma 6 (a) By Lemma 5, $\widehat{n}_1 = (\mu - \widetilde{\mu})/\gamma_1$. Therefore,

$$\frac{\partial\widehat{n}_1}{\partial\mu} = \frac{1}{\gamma_1} > 0. \tag{34}$$

Recall from Lemma 1, that for $\overline{n}_1 < N_1$, $\overline{n}_1$ is defined as the solution of the following implicit equation:

$$r_1 - \frac{c\mu}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^2} - \frac{N_2 c(\gamma_1 - \gamma_2)}{(\mu - \overline{n}_1\gamma_1)^2} = 0.$$

We use this equation to calculate implicitly the derivative of $\overline{n}_1$ with respect to μ:

$$-\frac{c(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^2 - 2c\mu(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)(1 - \gamma_1\frac{\partial\overline{n}_1}{\partial\mu})}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^4} + \frac{2N_2 c(\gamma_1 - \gamma_2)(1 - \gamma_1\frac{\partial\overline{n}_1}{\partial\mu})}{(\mu - \overline{n}_1\gamma_1)^3} = 0,$$

then

$$\frac{\partial \overline{n}_1}{\partial \mu} = \frac{1}{\gamma_1}\left(1 - \frac{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)(\mu - \overline{n}_1\gamma_1)^3}{2\mu(\mu - \overline{n}_1\gamma_1)^3 + 2N_2(\gamma_1 - \gamma_2)(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^3}\right),$$

and since

$$\begin{aligned} 0 <& \frac{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)(\mu - \overline{n}_1\gamma_1)^3}{2\mu(\mu - \overline{n}_1\gamma_1)^3 + 2N_2(\gamma_1 - \gamma_2)(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^3} \\ &= \frac{(\mu - \overline{n}_1\gamma_1)^3}{2\frac{\mu}{(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)}(\mu - \overline{n}_1\gamma_1)^3 + 2N_2(\gamma_1 - \gamma_2)(\mu - \overline{n}_1\gamma_1 - N_2\gamma_2)^2} \\ &< \frac{(\mu - \overline{n}_1\gamma_1)^3}{2(\mu - \overline{n}_1\gamma_1)^3} = \frac{1}{2} \end{aligned}$$

where the second inequality follows since $\mu > \mu - \overline{n}_1\gamma_1 - N_2\gamma_2$ and $\gamma_1 > \gamma_2$. Therefore

$$\frac{1}{\gamma_1} > \frac{\partial \overline{n}_1}{\partial \mu} > 0$$

and from (34), $\partial \widehat{n}_1/\partial\mu > \partial \overline{n}_1/\partial\mu > 0$.

(b) Recall from (18) the definition of u_2:

$$u_2 = \frac{c(\gamma_1 - \gamma_2)}{\mu - \overline{n}_1\gamma_1} - (r_1\gamma_1 - r_2\gamma_2) > 0 \quad \text{if} \quad \overline{n}_1 > 0.$$

When substituting $\widehat{n}_1 = (\mu - \widetilde{\mu})/\gamma_1$ into u_2, we get:

$$\begin{aligned} u_2(\widehat{n}_1) &= \frac{c(\gamma_1 - \gamma_2)}{\mu - \widehat{n}_1\gamma_1} - (r_1\gamma_1 - r_2\gamma_2) = \frac{c(\gamma_1 - \gamma_2)}{\mu - \frac{\mu - \widetilde{\mu}}{\gamma_1}\gamma_1} - (r_1\gamma_1 - r_2\gamma_2) \\ &= \frac{c(\gamma_1 - \gamma_2)}{\widetilde{\mu}} - (r_1\gamma_1 - r_2\gamma_2) = (r_1\gamma_1 - r_2\gamma_2) - (r_1\gamma_1 - r_2\gamma_2) = 0. \end{aligned}$$

When $\overline{n}_1 > \widehat{n}_1$, then

$$u_2(\overline{n}_1) = \frac{c(\gamma_1 - \gamma_2)}{\mu - \overline{n}_1\gamma_1} - (r_1\gamma_1 - r_2\gamma_2) < \frac{c(\gamma_1 - \gamma_2)}{\mu - \widehat{n}_1\gamma_1} - (r_1\gamma_1 - r_2\gamma_2) = 0.$$

Therefore when $\overline{n}_1 > \widehat{n}_1$, $u_2(\overline{n}_1) < 0$. $u_2 < 0$ is a contradiction to the IR constraint (9f) and therefore is not a feasible solution. □

Lemma 7 *If $\overline{n}_1(\widetilde{\mu}) = 0$, then it is optimal to serve $\widehat{n}_1$ class-1 customers. Otherwise, if $\overline{n}_1(\widetilde{\mu}) > 0$, there exists $\widehat{\mu}$ such that for $\mu < \widehat{\mu}$ it is optimal to serve $\overline{n}_1$ class-1 customers, and for $\mu > \widehat{\mu}$ it is optimal to serve $\widehat{n}_1$ class-1 customers.*

Proof of Lemma 7 By the definition of $\widehat{n}_1$ given in (26), $\widehat{n}_1(\widetilde{\mu}) = 0$. So

$$\widehat{\Pi}_1(\widetilde{\mu}) = N_2\gamma_2(r_2 - cW(\widetilde{\mu})) = \Pi_0(\widetilde{\mu}).$$

We first consider the case where $\overline{n}_1(\widetilde{\mu}) = 0$. The definition of $\overline{\mu}$ implies $\overline{\Pi}_1(\widetilde{\mu}) = \Pi_0(\widetilde{\mu})$. Therefore, $\Pi_1(\widetilde{\mu}) = \widehat{\Pi}_1(\widetilde{\mu})$. By Lemma 6(a), $\partial \widehat{n}_1/\partial\mu > \partial \overline{n}_1/\partial\mu > 0$. By Lemma 6(b) it is not feasible to serve $\overline{n}_1$. Thus, for $\widetilde{\mu} < \mu < K_1$ the provider serves $\widehat{n}_1$.

Next, we consider the case where $\overline{n}_1(\widetilde{\mu}) > 0$. The definition of $\overline{\mu}$ implies $\overline{\Pi}_1(\widetilde{\mu}) > \Pi_0(\widetilde{\mu})$. Therefore, $\overline{\Pi}_1(\widetilde{\mu}) > \widehat{\Pi}_1(\widetilde{\mu})$, so it is optimal to serve $\overline{n}_1$ at $\widetilde{\mu}$. By Lemma 6a, $\partial\widehat{n}_1/\partial\mu > \partial\overline{n}_1/\partial\mu > 0$. Therefore, there exists $\widehat{\mu} > \widetilde{\mu}$ such that $\overline{n}_1(\widehat{\mu}) = \widehat{n}_1(\widehat{\mu})$, and for $\widetilde{\mu} < \mu < \widehat{\mu}$, $\overline{n}_1 > \widehat{n}_1$ and for $\mu > \widehat{\mu}$, $\overline{n}_1 < \widehat{n}_1$. We now show that $\widehat{\mu}$ also satisfies $\overline{\Pi}_1(\widehat{\mu}) = \widehat{\Pi}_1(\widehat{\mu})$. Recall from (19):

$$\overline{\Pi}_1 = (n_1 + N_2)r_1\gamma_1 - \frac{c(\overline{n}_1\gamma_1 + N_2\gamma_2)}{\mu - \overline{n}_1\gamma_1 - N_2\gamma_2} - \frac{cN_2(\gamma_1 - \gamma_2)}{\mu - \overline{n}_1\gamma_1}.$$

and from (29):

$$\widehat{\Pi}_1 = N_2\gamma_2 r_2 + \widehat{n}_1\gamma_1 r_1 - \frac{c(\widehat{n}_1\gamma_1 + N_2\gamma_2)}{\mu - \widehat{n}_1\gamma_1 - N_2\gamma_2}.$$

Then for $\mu = \widehat{\mu}$,

$$\overline{\Pi}_1(\widehat{\mu}) - \widehat{\Pi}_1(\widehat{\mu}) = N_2(r_1\gamma_1 - r_2\gamma_2) - \frac{cN_2(\gamma_1 - \gamma_2)}{\mu - \widehat{n}_1\gamma_1} = N_2\left(r_1\gamma_1 - r_2\gamma_2 - \frac{c(\gamma_1 - \gamma_2)}{\widetilde{\mu}}\right) = 0.$$

where from (11), $\widetilde{\mu} = c(\gamma_1 - \gamma_2)/(r_1\gamma_1 - r_2\gamma_2)$.

We now prove by contradiction that $\widehat{\mu}$ is the only intersection point between $\overline{\Pi}_1$ and $\widehat{\Pi}_1$. Assume that there exists $\widetilde{\mu} < \dot{\mu} < \widehat{\mu}$ such that $\dot{\mu}$ is also an intersection point between $\overline{\Pi}_1$ and $\widehat{\Pi}_1$, i.e. $\overline{\Pi}_1(\dot{\mu}) = \widehat{\Pi}_1(\dot{\mu})$. Recall $\overline{n}_1(\dot{\mu}) > \widehat{n}_1(\dot{\mu})$. Let $\overline{W}_1$ be the waiting time of class-1 when serving $\overline{n}_1$, let $\widehat{W}_1$ be the waiting time of class-1 when serving $\widehat{n}_1$, and let $\overline{W}_2$ and $\widehat{W}_2$ be the complementary waiting time of class-2 customers, respectively. Recall the revenues:

$$\begin{aligned}
\overline{\Pi}_1 &= \overline{n}_1(r_1 - c\overline{W}_1)\gamma_1 + N_2[(r_2 - c\overline{W}_2)\gamma_2 - u_2], \\
\widehat{\Pi}_1 &= \widehat{n}_1(r_1 - c\widehat{W}_1)\gamma_1 + N_2(r_2 - c\widehat{W}_2)\gamma_2.
\end{aligned}$$

Note that when the number of prioritized customers increases, the waiting time for both prioritized and non-prioritized customers increases. Further, as the waiting time increases, the price paid by each customer decreases. Therefore if $\overline{n}_1 > \widehat{n}_1$, $(r_1 - c\overline{W}_1)\gamma_1 < (r_1 - c\widehat{W}_1)\gamma_1$. Because $u_2 > 0$, $(r_2 - c\overline{W}_2)\gamma_2 - u_2 < (r_2 - c\overline{W}_2)\gamma_2 < (r_2 - c\widehat{W}_2)\gamma_2$. Therefore for $\widetilde{\mu} < \mu < \dot{\mu}$, $\overline{\Pi}_1(\mu) > \widehat{\Pi}_1(\mu)$ is a result of $\overline{n}_1 > \widehat{n}_1$. For $\dot{\mu} < \mu \leq \widehat{\mu}$, by Lemma 6a, both $\widehat{n}_1$ and $\overline{n}_1$ increase, but the increase in $\widehat{n}_1$ is larger then the increase in $\overline{n}_1$. Therefore for $\dot{\mu} < \mu \leq \widehat{\mu}$, $\widehat{\Pi}_1 > \overline{\Pi}_1$, which is a contradiction to the equality of the revenues at $\widehat{\mu}$. Therefore such $\dot{\mu} < \widehat{\mu}$ does not exist and for $\widetilde{\mu} < \mu < \widehat{\mu}$ it is optimal to serve $\overline{n}_1$.

Next, we prove that for $\widehat{\mu} < \mu < K_1$, it is optimal to serve $\widehat{n}_1$. By Lemma 6(a), when $\widehat{\mu} < \mu < K_1$, then $\widehat{n}_1 > \overline{n}_1$. By Lemma 6(b) it is not feasible to serve $\overline{n}_1$ in this region. Thus, the optimal solution is to serve $\widehat{n}_1$. □

Lemma 8 $\widehat{\Pi} < \Pi^*$.

Proof of Lemma 8 From (32) and (33), $\widehat{n}_1 < n_1^*$ when $\widetilde{\mu} \le \mu < K_1$. Therefore, $\widehat{n}_1$ is a feasible solution for the FI problem. Noting n_1^* is the optimal solution of the FI problem

$$\begin{aligned}\Pi^* &= n_1^*\gamma_1 + N_2\gamma_2 - \frac{c(n_1^*\gamma_1 + N_2\gamma_2)}{\mu - n_1^*\gamma_1 + N_2\gamma_2} \\ &< \widehat{n}_1\gamma_1 + N_2\gamma_2 - \frac{c(\widehat{n}_1\gamma_1 + N_2\gamma_2)}{\mu - \widehat{n}_1\gamma_1 + N_2\gamma_2} \\ &= \widehat{\Pi}\end{aligned}$$

□

It is left to show that when $K_1 \le \mu < K_2$ solution (v) holds. Since $\mu \ge K_1 = r_1(\widetilde{\mu} - N_2\gamma_2)^2/c$, then $\widetilde{\mu} \le N_2\gamma_2 + \sqrt{c\mu/r_1}$. In the FI setting, for $\mu > \mu_2$, $r_1 = c\mu/(\mu - n_1^*\gamma_1 - N_2\gamma_2)^2$. Therefore we can write:

$$n_1^*\gamma_1 = \mu - N_2\gamma_2 - \sqrt{\frac{c\mu}{r_1}}, \tag{35}$$

and

$$W_1 = \frac{1}{\mu - n_1^*\gamma_1} = \frac{1}{N_2\gamma_2 + \sqrt{\frac{c\mu}{r_1}}} \le \frac{1}{\widetilde{\mu}} = \widetilde{W}. \tag{36}$$

Therefore, for $\mu \ge K_1$, it is possible to satisfy constraint (9a) when serving n_1^*. Note that $W_1 = \widetilde{W}$ only when $\mu = K_1$. Following (36), it is possible to serve n_1^* class-1 customers by giving them priority W_1^* within the interval: $[W_1, \widetilde{W}]$. Letting W_2^* be the minimum complementary required waiting time, we find

$$W_2 = \frac{\mu}{(\mu - n_1^*\gamma_1)(\mu - n_1^*\gamma_1 - N_2\gamma_2)}.$$

Next, we show that for $\mu < K_2$, $W_1 < W$. Recall that W is the FIFO waiting time, where

$$W = \frac{1}{\mu - n_1\gamma_1 - N_2\gamma_2}.$$

Substituting in n_1^* from (35),

$$W = \frac{1}{\mu - n_1^*\gamma_1 - N_2\gamma_2} = \frac{1}{\sqrt{\frac{c\mu}{r_1}}}.$$

But when $\mu < K_2 = r_1\widetilde{\mu}^2/c$, then $\widetilde{\mu} > \sqrt{c\mu/r_1}$. Therefore

$$W = \frac{1}{\sqrt{\frac{c\mu}{r_1}}} > \frac{1}{\widetilde{\mu}} = \widetilde{W} \ge W_1.$$

Therefore $W_1 < W$.

It is left to show that the value of the solution of the PI problem is equal to the value of the solution of the FI problem:

$$\begin{aligned}\Pi &= n_1^*\gamma_1(r_1 - cW_1^*) + N_2\gamma_2(r_2 - cW_2^*) = n_1^*\gamma_1 r_1 + N_2\gamma_2 r_2 - c(n_1^*\gamma_1 W_1^* + N_2\gamma_2 W_2^*) \\ &= n_1^*\gamma_1 r_1 + N_2\gamma_2 r_2 - cW = \Pi^*.\end{aligned}$$

Step 3. Recall n_1^* as the number of class-1 customers that are optimally served in the FI problem.

If $\widetilde{\mu} < \mu_2$, then for $\mu > \mu_2$, $\mu - n_1^*\gamma_1 \geq \mu_2 > \widetilde{\mu}$. Therefore, $W_1 = \frac{1}{\mu - n_1^*\gamma_1} < \frac{1}{\widetilde{\mu}} = \widetilde{W}$ and constraint (9a) is satisfied.

Step 4. As in case **3.**, when the provider serves n_1^*,

$$W = \frac{1}{\mu - n_1^*\gamma_1 + N_2\gamma_2} = \frac{1}{\sqrt{\frac{c\mu}{r_1}}}.$$

Because $\mu \geq K_2 = \frac{r_1\widetilde{\mu}^2}{c}$, $\sqrt{\frac{c\mu}{r_1}} \geq \widetilde{\mu}$. Therefore

$$W = \frac{1}{\sqrt{\frac{c\mu}{r_1}}} \leq \frac{1}{\widetilde{\mu}} = \widetilde{W}.$$

So, when $\mu \geq K_2$, constraint (9a) is satisfied with $W_1 = W$. As a result, $\Pi = \Pi^*$. ■

Appendix B: Details of cases in Proposition 5

The solution classes are:

(i) $n_1 = n_1^*,\ n_2 = n_2^*,\ u_2 = 0,\ W_1 = W$. Prices are fixed and equal to those of the FI solution, equation 2:

$$P_i^* = (r_i - cW_i)\gamma_i, \quad i = 1, 2,$$

and the revenue is $\Pi^*(\mu)$, which is equal to the revenue of the FI solution.

(ii) $n_1 = 0$ and therefore W_1 is not defined, $n_2 = N_2$, $u_2 = 0$. Since $r_2 > r_1$ then $r_2 - CW > r_1 - CW$, and the provider can achieve this solution by publishing a single per-use price:

$$P(\gamma) = (r_2 - cW)\gamma.$$

This price per-use does not satisfy the IR constraint of class-1, and therefore they will not join, and the revenue under this solution is designated as $\Pi_0(\mu)$. We show that $\Pi_0(\mu) < \Pi^*(\mu)$.

(iii) $n_2 = N_2,\ u_2 > 0$ and $W_1 < W$. This solution is the same as the solution presented in Prop. 4 case 2. In the same way, we define $\overline{n}_1$ as was defined in Lemma 1, and therefore $\overline{n}_1 < n_1^*$. The price functions are the same as in Lemma 3:

$$\begin{aligned}
\overline{P}_1(\gamma) &= (r_1 - cW_1)\gamma_1 \geq P_1^*(\gamma_1), \\
\overline{P}_2(\gamma) &= (r_2 - cW_2)\gamma - \frac{c(\gamma_1 - \gamma_2)}{\mu - n_1\gamma_1} + (r_1\gamma_1 - r_2\gamma_2) \leq P_2^*(\gamma_2),
\end{aligned}$$

which means subscription price for class-1 and two-part tariff for class-2. Let the revenue under this solution be $\overline{\Pi}_1(\mu)$. We show that $\overline{\Pi}_1(\mu) < \Pi^*(\mu)$.

(iv) $n_2 = N_2,\ u_2 = 0$, and $W_1 < W$. Let $\widehat{n}_1$ be the optimal number of class-1 customers that are served: $0 < \widehat{n}_1 < n_1^*$. Then the price functions are:

$$\begin{aligned}
\widehat{P}_1(\gamma) &= (r_1 - cW_1)\gamma_1 \\
\widehat{P}_2(\gamma) &= (r_2 - cW_2)\gamma,
\end{aligned}$$

which means subscription price for class-1 and per-use price for class-2. Let the revenue under this solution be $\widehat{\Pi}_1(\mu)$. We show that $\widehat{\Pi}_1(\mu) < \Pi^*(\mu)$.

(v) $n_1 = n_1^*,\ n_2 = N_2,\ u_2 = 0,\ W_1 < W$. Then prices are defined as in solution (iv):

$$\begin{aligned}
\widehat{P}_1(\gamma) &= (r_1 - cW_1)\gamma_1 \\
\widehat{P}_2(\gamma) &= (r_2 - cW_2)\gamma,
\end{aligned}$$

but since the number of customers that are served from both classes are the same as in the FI solution, we show that the revenue equals to $\Pi^*(\mu)$.

References

Afèche, P. 2004. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delay tactics. Tech. rep., University of Toronto, Toronto.

Afèche, P. 2013. Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management* **15**(3) 423–443.

Afèche, P., O. Baron, Y. Kerner. 2013. Pricing time-sensitive services based on realized performance. *Manufacturing & Service Operations Management* **15** 492–506.

Allon, G., A. Federgruen. 2009. Competition in service industries with segmented markets. *Management Science* **55**(4) 619634.

Boyaci, T., S. Ray. 2003. Product differentiation and capacity cost interaction in time and price sensitive markets. *Manufacturing & Service Operations Management* **5**(1) 1836.

Cachon, G., P. Feldman. 2011. Pricing services subject to congestion: charge per-use fees or sell subscriptions? *Manufacturing & Service Operations Management* 244–260.

Fudenberg, D., J. Tirole. 1991. *Game Theory*. MIT Press.

Hassin, R. 1995. Decentralized regulation of a queue. *Management Science* **41**(1) 163–173.

Hassin, R., M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*, vol. 59. Springer.

Hsu, V., S. Xu, B. Jukic. 2009. Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing & Service Operations Management* **11**(3) 375396.

Katta, A., J. Sethuraman. 2005. Pricing strategies and service differentiation in queuesa profit maximization perspective. Tech. rep., Columbia University, New York.

Lederer, P., L. Li. 1997. Pricing, production, scheduling and delivery-time competition. *Operations Research* **45**(3) 407–420.

Maglaras, C., J. Yao, A. Zeevi. 2014. Revenue maximization in queues via service differentiation. Tech. rep., Columbia University, New York.

Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research* **53**(2) 242262.

Masuda, Y., S. Whang. 2006. On the optimality of fixed-up-to tariff for telecommunications service. *Information System Research* **17** 247–253.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Communincations of the ACM* **28** 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research* **38** 870–883.

Myerson, R.B. 1997. *Game Theory: Analysis of Conflict*. Harvard University Press.

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

Plambeck, E.L., Q. Wang. 2013. Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Science* **59** 1927–1946.

Randhawa, R.S., S. Kumar. 2008. Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing & Service Operations Management* **10** 429–447.

Rao, S., E.R. Petersen. 1998. Optimal pricing of priority services. *Operations Research* **46** 46–56.

Van Mieghem, J.A. 2000. Price and service discrimination in queuing systems: Incentive compatibility of gcμ scheduling. *Management Science* **46** 1249–1267.

Yahalom, T., J.M. Harrison, S. Kumar. 2006. Designing and pricing incentive compatible grades of service in queueing systems. Tech. rep., Stanford University, Stanford, CA.

Zhao, X., K.E. Stecke, A. Prasad. 2012. Lead time and price quotation mode selection: Uniform or differentiated? *Production and Operations Management* **21**(1) 177–193.