Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Implications of Hyperbolic Discounting for Optimal Pricing and Scheduling of Unpleasant Services That Generate Future Benefits

Erica L. Plambeck, Qiong Wang,

To cite this article:
Erica L. Plambeck, Qiong Wang, (2013) Implications of Hyperbolic Discounting for Optimal Pricing and Scheduling of Unpleasant Services That Generate Future Benefits. Management Science 59(8):1927-1946. http://dx.doi.org/10.1287/mnsc.1120.1673

MANAGEMENT SCIENCE
Vol. 59, No. 8, August 2013, pp. 1927–1946
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1673

Implications of Hyperbolic Discounting for Optimal Pricing and Scheduling of Unpleasant Services That Generate Future Benefits

Erica L. Plambeck
Operations, Information and Technology, Graduate School of Business, Stanford University, Stanford, California 94305, elp@stanford.edu

Qiong Wang
Industrial and Enterprise Systems Engineering, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801, qwang04@illinois.edu

People tend to lack the self-control to undergo an unpleasant service that would generate future benefits. This paper derives a tractable quasi-hyperbolic discounting model of that behavioral tendency (for a queueing system in which service time is short relative to the time horizon for its benefits). Planning in advance, people may naively overestimate their self-control. This paper shows how customers' lack of self-control and naivete affect optimal pricing and scheduling. The welfare-maximizing usage fee and the revenue-maximizing usage fee decrease with customers' lack of self-control. Charging for subscription, in addition to or instead of per use, increases revenue, especially when subscribers are naive. If the manager can charge for subscription or per use, subscription is optimal for revenue maximization, whereas usage-based pricing is optimal for welfare maximization. If customers are heterogeneous in their self-control and naivete, priority scheduling can dramatically increase welfare and revenue.

Key words: organizational studies; behavior; queues; optimization; probability; stochastic model applications; pricing; service
History: Received November 22, 2009; accepted August 15, 2012, by Christian Terwiesch, operations management. Published online in *Articles in Advance* March 18, 2013.

1. Introduction

In many service systems, a customer's experience is less pleasant than an alternative activity, but necessary to achieve some long-term benefit. Examples include an emergency room, flu shot clinic, exercise facility, barber shop, car wash or oil-change center, and the Department of Motor Vehicles (DMV). The psychology and economics literature on hyperbolic discounting (e.g., Ainslie 1992, Frederick et al. 2002, DellaVigna and Malmendier 2006) and interviews with managers (Ault 2010, Hurn 2011) provide evidence that people lack the self-control to undertake such services as frequently as they should. This paper deduces implications for optimal pricing and scheduling. It helps to explain the fact that car washes, barber shops, and exercise facilities sell memberships, and people pay more for membership than is justified by their actual use of these services (Ault 2010, China Commodity Net 2009, Hurn 2011, DellaVigna and Malmendier 2006).

When people must choose between a smaller reward and a larger reward that will occur a fixed amount of time later, their preferences tend to change over time. People tend to prefer the larger-later reward (e.g., good health) when both are in the future. However, when the smaller-sooner reward (e.g., relaxing rather than going to the gym) is imminent, it is relatively more attractive. These preferences are best represented by a hyperbolic discount function: a reward at time t is discounted by the factor $(1+\alpha t)^{-\gamma/\alpha}$, where the parameters α and γ are strictly positive. The corresponding discount rate $\gamma/(1+\alpha t)$ decreases with respect to time, in contrast to the standard exponential discount function in which the discount rate is constant. Surveys of the empirical evidence and theory regarding hyperbolic discounting are in Ainslie (1992) and Frederick et al. (2002).

Economists have found the hyperbolic discount function $(1+\alpha t)^{-\gamma/\alpha}$ to be analytically intractable, and therefore have adopted a discrete-time "quasi-hyperbolic" discount function in which rewards in the current time period are undiscounted, and rewards occurring t periods in the future are discounted by $\beta\delta^t$, where $\beta, \delta \in (0,1)$. This discrete-time quasi-hyperbolic discount function provides a good fit to experimental data (Angeletos et al. 2001, McClure

et al. 2004) and can well approximate the hyperbolic discount function by appropriate choice of β, δ, and period length (Laibson 1997; see Figure 1). However, a *discrete-time* discount function seems incompatible with the *continuous-time* queueing models used in the service operations management literature. Therefore, this paper derives a similar quasi-hyperbolic discount function for queueing systems in which the service time is short relative to the time horizon over which the future benefits from service occur. That asymptotic regime is representative of all our motivating examples (emergency room, flu shot clinic, dentist's office, exercise facility, barber shop, car wash or oil-change center, and the DMV). In these systems, many people undergo a service that takes less than an hour, but generates benefits for weeks, months, or even years to come.

To the extent that the "hyperbolic discounting" phenomenon is strong, the parameter β is small. Economists have estimated β to be substantially smaller than 1 in field studies of gym membership and usage (DellaVigna and Malmendier 2006), household consumption, savings and financial management (Choi et al. 2011, Laibson et al. 2007, Skiba and Tobacman 2008, Ausubel and Shui 2005, Madrian and Shea 2001), unemployment and job search (Fang and Silverman 2009, Paserman 2008), and investment in fertilizer by poor farmers (Duflo et al. 2009).

Economists commonly interpret small β pejoratively, as a lack of self-control. For example, see Laibson (1997) and O'Donoghue and Rabin (2001). Although time inconsistency in an individual's preferences raises the question of whether to define social welfare in terms of an individual's short-run or long-run preferences, the standard practice is to define social welfare in terms of an individual's long-run preferences. This practice is discussed and defended by O'Donoghue and Rabin (1999) and Bernheim and Rangel (2005). Consistent with the economics literature, we will use both of the terms "lack of self-control" and "hyperbolic discounting" to refer to time inconsistency in preferences as represented by either a hyperbolic or quasi-hyperbolic discount function.

Are individuals aware of their self-control problems? This is an open research question. A person might be completely "naive" and believe that his preferences will not change over time. Alternatively, he could be "sophisticated" and perfectly anticipate how his preferences will change. Frederick et al. (2002) surveyed the limited behavioral evidence and concluded that people are partially naive, that is, they lie somewhere between these two extremes. To identify sophistication, researchers look for costly commitments that constrain future choices. For example, DellaVigna and Malmendier (2006) found that many people commit to membership in a health club (pay for subscription) rather than pay a fee per use, even though membership is far more expensive than if they paid a usage fee on each visit to the club. By purchasing a costly membership, a person can motivate his or her future self to exercise more frequently. They concluded that their data were best explained by a model of quasi-hyperbolic discounting with partial naivete; that is, individuals anticipate that they will exercise too little (that β is smaller than 1) but underestimate the extent of the self-control problem (overestimate β).

This paper deduces implications of hyperbolic discounting for the management of services that are unpleasant but generate long-term future benefits for customers. It addresses both of the alternative objectives of social welfare and revenue maximization. Section 2 describes our model and derives the quasi-hyperbolic discount function. For the case that customers are homogeneous, §3 addresses optimal pricing. For the case that customers are heterogeneous in their self-control and naivete, §4 characterizes the optimal pricing scheme and conditions under which priority scheduling strictly outperforms first-in, first-out (FIFO) scheduling. Section 5 summarizes the main conclusions. All proofs are in the appendix.

In the nascent literature on behavioral operations management (reviewed by Bendoly et al. 2006, Gino and Pisano 2008, Loch and Wu 2007), this paper is one of the first to address hyperbolic discounting. In an airline revenue-management problem, Su (2009) allowed for a fraction of customers to refrain from early purchase because they overweighted the immediate cost relative to the future benefit from purchase. The optimal revenue decreases with the strength of this tendency, but may increase with the fraction of customers that exhibit it. In the context of project management, Wu et al. (2009) showed how heterogeneous levels of hyperbolic discounting (propensities to procrastinate) among workers influence the optimal composition of project teams and structure of incentive payments. Regarding product design, Ulku et al. (2009) found in lab experiments that subjects undervalue (or overvalue) modular upgradeability for products in categories with rapid (or slow) innovation in component functionality. The degree of hyperbolic discounting in these biases is stronger than in decisions involving cash. Hassin and Haviv (2003) reviewed the literature on queueing systems (without hyperbolic discounting) that sets the stage for our analysis. We will discuss specific papers from that literature in relation to our results in §§3 and 4.

2. Model Formulation

This section describes a queueing system with hyperbolic discounting of utility by customers. For the asymptotic parameter regime with short service times

(relative to the time horizon for the benefits from service) and a large number of potential customers, it shows that customers' decisions regarding subscription and actual use of the service are governed by simple equations. These equations correspond to a quasi-hyperbolic model of discounting that we will adopt, for purposes of tractability, in subsequent analysis.

A single station offers service to a market of M homogeneous potential customers. Each customer experiences a need for service according to a Poisson process with rate $\bar{\lambda}$. Service time requirements are i.i.d. exponential with rate μ. Queueing and the service itself are unpleasant; a customer incurs cost c per unit time in the system. However, service completion generates benefits for a customer for the next L units of time at an expected rate of r/L per unit time; r is the undiscounted expected total reward. Neither the customer nor the manager knows the realization of a customer's service time in advance, and the manager cannot interrupt a customer's service to begin serving a different customer. Therefore, the manager optimally uses FIFO scheduling. We restrict attention to symmetric equilibria in which all customers have the same probability $\lambda/\bar{\lambda}$ of going for service when a need arises, so that each customer actually goes for service according to a Poisson process with rate $\lambda \le \bar{\lambda}$. (This is without loss of generality in maximizing welfare or revenue.) Let W denote the resulting steady-state waiting time, including service. A customer cannot see the queue and knows only the steady-state distribution of W. (Our main results hold when the customer sees the queue, as explained at the end of §3.) Therefore, for a given usage fee u, a customer is willing to go for service if and only if

$$\mathrm{E}\left[-\int_0^W c(1+\alpha t)^{-\gamma/\alpha}\,dt + \int_W^{W+L} r/L(1+\alpha t)^{-\gamma/\alpha}\,dt\right] - u \ge 0. \tag{1}$$

An equilibrium must satisfy $\lambda = 0$ or (1), and if $\lambda \in (0, \bar{\lambda})$, then (1) must hold with equality. No customer will renege after entering the system. The reason is that the waiting time has an exponential distribution, so the posterior distribution of remaining waiting time for a customer that has already waited t units is identical for all t. Hence, the utility from continuing to wait is exactly the first term $\mathrm{E}[-\int_0^W c(1+\alpha t)^{-\gamma/\alpha}\,dt + \int_W^{W+L} r/L(1+\alpha t)^{-\gamma/\alpha}\,dt]$ in (1). Indeed, as the term in brackets in (1) decreases with W, to ensure that no customer reneges, we require only that distribution of residual waiting time is stochastically nonincreasing with time in system.

In scenarios with subscription, to access the service, a customer must commit at time 0 to pay for subscription at a constant rate $s \ge 0$ per unit time. At time 0, the expected hyperbolically discounted utility of going for service at time τ (conditional on a need for service arising at time τ) is

$$U_\tau = \mathrm{E}\left[\int_W^{W+L} r/L(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\right] - (1+\alpha\tau)^{-\gamma/\alpha}u - \mathrm{E}\left[\int_0^W c(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\right].$$

Note that a subscriber must still pay a usage fee u to actually go for service (though the manager might set this to zero). A customer will subscribe at time 0 if and only if

$$\mathrm{E}\left[\sum_{i=1}^{\infty} U_{\tau_i}\right] \ge \int_0^\infty (1+\alpha t)^{-\gamma/\alpha}\,s\,dt, \tag{2}$$

where the $\{\tau_i\}_{i=1,\ldots,\infty}$ are the arrival times for that Poisson process with rate $\hat{\lambda}$, the rate at which the customer believes he will go for service. A customer that subscribes at time 0 will never thereafter wish to unsubscribe. If the customer could commit himself in advance, at time $t = -T$, to subscribe during the time period $t \in [0, \infty)$, he would like do so if and only if

$$\mathrm{E}\left[\sum_{i=1}^{\infty} U_{T+\tau_i}\right] \ge \int_T^\infty (1+\alpha t)^{-\gamma/\alpha}\,s\,dt. \tag{3}$$

Unfortunately, he would renege on that commitment and cancel the subscription at time 0 unless (2) was satisfied. Note that U_τ simplifies to the left-hand side of (1) at $\tau = 0$. Furthermore, if $U_\tau \ge 0$, then $U_{\tau'} > 0$ for all $\tau' > \tau \ge 0$, which means that when planning in advance, the customer puts greater weight on future rewards relative to the initial costs of service. Therefore, a customer will pay a positive amount for subscription even if he anticipates being indifferent about paying the usage fee and going for service when a need arises and, from the long-term perspective of $t = -T$, subscription is even more attractive: (3) is weaker than (2). In writing (2) and (3), like Cachon and Feldman (2011), we allow a subscriber to generate a new request for service while still waiting for a previous request to be fulfilled, and we assume that he disregards the possible effect of a request for service on the delay he will experience with future requests for service. When service times are short (μ is large), the likelihood that a customer generates a new request for service while waiting in queue is negligible, as are interaction effects between his arrival and waiting times.

In our motivating examples, service times are indeed short, relative to the time horizon over which benefits occur, and the number of prospective customers M is large. Let us therefore consider a system in which we scale up the service rate, number of prospective customers, and waiting cost by a factor σ,

$$\mu_\sigma = \sigma\mu, \quad M_\sigma = \sigma M, \quad c_\sigma = \sigma c, \tag{4}$$

while holding all other parameter constant. This preserves the system utilization $\lambda M_\sigma/\mu_\sigma = \lambda M/\mu$ and the expected waiting cost $\mathrm{E}[c_\sigma W_\sigma] = c_\sigma/(\mu_\sigma - \lambda M_\sigma) = c/(\mu - \lambda M)$ for any given arrival rate per customer λ. In the limit as $\sigma \to \infty$, (1) for system σ simplifies to

$$\beta r - u - c\,\mathrm{E}[W] \geq 0, \tag{5}$$

with $\mathrm{E}[W] = (\mu - \lambda M)^{-1}$, and (2) simplifies to

$$\hat{\lambda}\big(\eta r - u - c\,\mathrm{E}[W]\big) \geq s, \tag{6}$$

where

$$\beta = \frac{1-(1+\alpha L)^{1-\gamma/\alpha}}{(\gamma-\alpha)L} \in (0,1), \tag{7}$$

$$\eta = \begin{cases} 1 & \text{if } \gamma \leq \alpha, \\ \dfrac{\ln(1+\alpha L)}{\alpha L} \in (\beta, 1) & \text{if } \gamma = 2\alpha, \\ \dfrac{1-(1+\alpha L)^{2-\gamma/\alpha}}{(\gamma-2\alpha)L} \in (\beta, 1) & \\ & \text{if } \gamma > \alpha \text{ and } \gamma \neq 2\alpha. \end{cases} \tag{8}$$

In the limit as $\sigma \to \infty$ and $T \to \infty$, (3) simplifies to

$$\hat{\lambda}\big(r - u - c\,\mathrm{E}[W]\big) \geq s; \tag{9}$$

this means that from an individual's long-term perspective, the instantaneous expected utility from subscription is

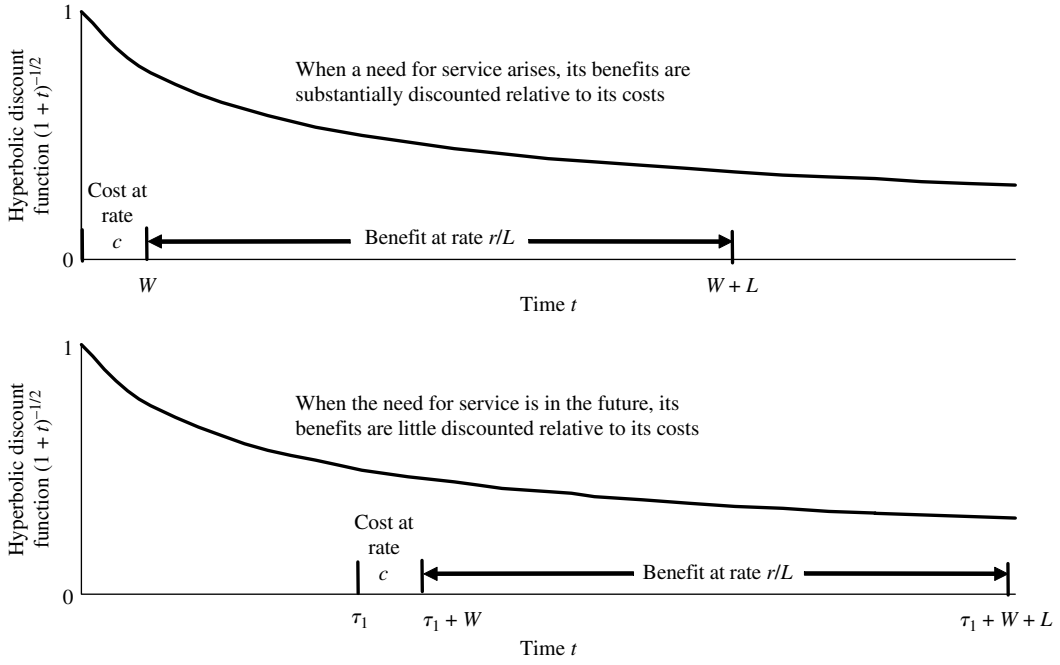
$$\hat{\lambda}\big(r - u - c\,\mathrm{E}[W]\big) - s. \tag{10}$$

Note that (5)–(9) hold regardless of whether a customer pays the usage fee u upon joining the queue, upon completing service, or at any time in between. The proof of (5)–(9) is in the appendix. Figure 1 provides the intuition. When the service rate is large, so that the waiting time W is small, the discount function is approximately constant within the waiting time. When the need for service is in the future (bottom panel) rather than immediate (top panel), benefits are less discounted relative to the costs of service.

The representation of customer choice in (5)–(9) generalizes that used by Naor (1969) and related literature on optimal pricing in queues. As $\alpha \to 0$, the hyperbolic discount function $(1+\alpha t)^{-\gamma/\alpha}$ converges to an exponential discount function $e^{-\gamma t}$ and β converges to η. As $\alpha \to 0$ and $\gamma \to 0$, both $\beta \to 1$ and $\eta \to 1$. Substitution of $\beta = 1$ and $\eta = 1$ in (5) and (6) gives the undiscounted formulation of customer choice used by Naor (1969).

Like in Laibson's (1997) discrete-time model of quasi-hyperbolic discounting, β should be interpreted as the customer's level of self-control. To the extent that β is small, a customer discounts the future reward relative to the cost of service more greatly when that cost is imminent. Observe that $\eta > \beta$, which reflects the fact that a customer takes a longer-term perspective when deciding whether or not to subscribe; his needs for service will arise in the future, so the future rewards are discounted by less, relative to costs of service. However, in the case that $\eta < 1$, which corresponds to $\gamma > \alpha$, the prospective customer exhibits insufficient self-control even in the decision of whether or not to subscribe for the service.

Figure 1 **Moving a Service Into the Future Equalizes the Discount Factor Applied to Its Costs and Future Benefits**



In our exercise facility, car wash, and barber shop examples, subscribers believe that they will use the service more than they actually do (DellaVigna and Malmendier 2006, Ault 2010, Hurn 2011), which motivates us to generalize (5). In deciding whether or not to subscribe, a customer believes that in future he will choose to go for service when a need arises if and only if

$$\hat{\beta}r - u - c\,\mathrm{E}[W] \geq 0 \tag{11}$$

for some $\hat{\beta} \in [\beta, 1]$. If $\hat{\beta} = \beta$, each customer is "sophisticated" in that he perfectly predicts his future decisions. If $\hat{\beta} > \beta$, each customer is "naive" in that he overestimates his self-control. In summary, a customer will subscribe if and only if

$$\hat{\lambda}(\eta r - u - c\mathrm{E}[W]) \geq s \quad \text{and} \quad \hat{\beta}r - u - c\mathrm{E}[W] \geq 0, \tag{12}$$

where, for a sophisticated customer, $\hat{\lambda} = \lambda$, the true equilibrium usage rate. However, for a naive customer, $\hat{\lambda} = \bar{\lambda}$. In an equilibrium with usage rate $\lambda \in (0, \bar{\lambda})$, the expected waiting time must satisfy (5) with equality, which implies that a naive customer anticipates having strictly positive expected utility $\hat{\beta}r - u - c\,\mathrm{E}[W] > 0$ and hence always going for service when a need arises, at rate $\bar{\lambda}$.

Henceforth, we assume that a customer is willing to go for service when a need arises if and only if (5) holds, and that a customer subscribes at time zero if and only if (12) holds. Outside of the asymptotic regime (4), this is equivalent to assuming a quasi-hyperbolic discount function in which, at the time a need for service arises, the future benefits from service are discounted by β relative to the costs of service, whereas when the customer decides whether or not to subscribe, he places relatively greater weight $\eta > \beta$ on the benefits from service relative to the costs of service, because both are in the future, and may overestimate his propensity to go for service, according to the parameter $\hat{\beta} \geq \beta$. We assume $\beta r > c$; otherwise, no customers would go for service, even if it were free. Under these assumptions, we will prove results for general positive μ, M, and c.

We will consider expected social welfare maximization and expected revenue maximization. Our welfare objective function is exactly the same as that of Naor (1969), the expected rate at which customers benefit from service less the rate at which they incur waiting costs,

$$\Omega = \lambda M\big(r - c\,\mathrm{E}[W]\big). \tag{13}$$

Our analysis above shows that this formulation of social welfare is consistent with the economics literature on hyperbolic discounting, which defines consumer utility and hence social welfare in terms of an individual's long-term perspective. The welfare objective in (13) is the sum of two terms, the rate at which the system generates expected utility for subscribers,

$$\big(\lambda(r - u - c\,\mathrm{E}[W]) - s\big)M,$$

according to their long-term perspective (10) and true equilibrium usage rate $\hat{\lambda} = \lambda$, plus the expected rate at which the system generates revenue,

$$(s + \lambda u)M. \tag{14}$$

Proposition 2 and Lemma 1 below confirm that expected social welfare and expected revenue are indeed maximized by having all M prospective customers subscribe for the service. Our revenue objective function, maximizing the expected rate at which the system manager collects money from customers (14), is also the same as that of Naor (1969). In all cases considered in this paper, the objective function is unimodal in the fee(s). Hence, if the manager does not know the parameters of customer demand λ, M, β, $\hat{\beta}$, and η, he can nevertheless converge to an optimal pricing scheme through straightforward experimentation with the fee(s).

Observe in (5)–(12) that for any pricing scheme, equilibrium in subscription and usage requires that customers know the expected waiting time $\mathrm{E}[W]$, but not the full distribution of waiting time W. In reality, prospective subscribers may obtain waiting time information through a free trial membership (as is common for fitness centers), official website (many health maintenance organizations post the current waiting time for urgent care), and word of mouth or Internet postings by other users. People may dynamically subscribe, unsubscribe, and decide how frequently to use the service while updating their beliefs about waiting times after each service experience. Bitran et al. (2008) accounted for such dynamics in a model similar to ours, and proved that the total arrival rate decreases with the waiting time distribution. One may therefore expect to see convergence to the full-information equilibrium over time.

In reality, waiting in the system may be more or less unpleasant than the service itself. For example, in a flu shot clinic, the shot is painful and acutely unpleasant for most people, whereas the wait is only annoying and inconvenient. Waiting in a crowded emergency room while suffering from an undiagnosed, acute injury, or illness may be more unpleasant than receiving treatment. All our propositions hold when a different cost per unit time $c' \neq c$ applies to service time than to waiting time (with more complex expressions for thresholds).

Our exponential single-server assumption is used directly only in the proof of Proposition 8 and to guarantee that customers do not renege. Our derivation

of the quasi-hyperbolic discounting model, (5)–(8), requires only that the coefficient of variation of the waiting time, $\mathrm{VAR}[W]/\mathrm{E}[W]$, is bounded; $c\,\mathrm{E}[W]$ converges to a constant in the asymptotic regime; and customers do not renege. The proofs of all propositions in §3 require only that $\mathrm{E}[W]$ is a convex function of λ and customers do not renege.

3. Optimal Pricing with Homogeneous Customers

In the absence of hyperbolic discounting, the revenue-maximizing usage fee achieves the optimal social welfare and leaves customers with zero surplus; subscription is unnecessary (Hassin and Haviv 2003, pp. 45–51). This section shows how hyperbolic discounting breaks those well-known results.

Initially, suppose that the system manager charges only a usage fee, not for subscription. One may think of the manager as choosing the throughput $\Lambda = \lambda M$, then setting $u = \beta r - c/(\mu - \Lambda)$ to implement that Λ according to (5) with $\mathrm{E}[W] = 1/(\mu - \Lambda)$. Let

$$\bar{\Lambda} \equiv \bar{\lambda} M$$

denote the maximum throughput. Optimal social welfare

$$\max_{\Lambda \le \bar{\Lambda}} \{\Lambda(r - c/(\mu - \Lambda))\} \tag{15}$$

and the welfare-maximizing throughput, the solution to (15), are invariant with customers' self-control β and naivete $\hat{\beta}$, whereas optimal revenue

$$\max_{\Lambda \le \bar{\Lambda}} \{\Lambda(\beta r - c/(\mu - \Lambda))\} \tag{16}$$

and the revenue-maximizing throughput, the solution to (16), increase with customer's self-control β but are invariant with naivete $\hat{\beta}$. Hence, in both scenarios, the optimal usage fee strictly increases with customers' self-control β but is invariant with naivete $\hat{\beta}$.

Proposition 1. *The welfare-maximizing usage fee*[1] *strictly increases with β. It is strictly negative if and only if $\beta < \bar{\beta}$, where $\bar{\beta} \in (0, 1)$. If the usage fee must be nonnegative, then for $\beta < \bar{\beta}$, the (constrained) welfare-maximizing usage fee is zero, and the resulting throughput and welfare strictly increase with β. The revenue-maximizing usage fee and resulting revenue strictly increase with β. If $\bar{\Lambda} > \mu - \sqrt{c\mu/(\beta r)}$, then the revenue-maximizing usage fee is strictly greater than the welfare-maximizing usage fee, and results in throughput and welfare that strictly increase with β. Otherwise, for $\bar{\Lambda} \le \mu - \sqrt{c\mu/(\beta r)}$, the revenue-maximizing usage fee results in the maximum throughput $\bar{\Lambda}$, which is socially optimal.*

[1] When the maximum throughput $\bar{\Lambda}$ is sufficiently small, it is socially optimal and can be achieved with a range of usage fees. Then, we choose the welfare-maximizing usage fee to be the maximum one that implements $\bar{\Lambda}$. Otherwise, it is unique.

Hyperbolic discounting reduces the welfare-maximizing usage fee as does a positive network externality in the queueing models of Westland (1992), Giridharan and Mendelson (1994), and Johari and Kumar (2010). When $\beta < \bar{\beta}$, a welfare-maximizing manager must reward people for using the service (set $u < 0$) to increase throughput. However, budget constraints are likely to prevent the manager from doing so. This motivates charging for subscription. Under a break-even budget constraint, Propositions 1 and 2 establish that subscription strictly increases social welfare if and only if $\beta < \bar{\beta}$. The manager sets the usage fee to induce the welfare-maximizing throughput and sets the subscription fee sufficiently high to have nonnegative expected revenue, and, in equilibrium, all prospective customers subscribe. If the usage fee must be nonnegative for other reasons (e.g., to prevent abuse by customers entering the system simply to collect money), charging for subscription can only reduce throughput and welfare.

Proposition 2. *If the usage fee is constrained to be nonnegative, a welfare-maximizing manager should not charge for subscription. However, subject only to a break-even budget constraint, the manager achieves the welfare-maximizing throughput by charging for subscription in addition to a usage fee.*

If a welfare-maximizing manager could charge only for subscription or per use, not both, charging per use would always be optimal.

A revenue-maximizing manager, in charging only a usage fee, sacrifices potential revenue for two reasons. First, throughput (and hence welfare) is lower than optimal. Second, customers have relatively low willingness to pay at the time of service. When a need for service arises, the usage fee and waiting costs are immediate, whereas the long-term future benefits generated by the service are discounted by β. When a person decides whether or not to subscribe, he takes a long-term perspective with greater weight $\eta > \beta$ on the reward relative to the costs of service. Therefore, let us focus on revenue maximization with subscription for the remainder of this section. Our analysis is simplified by the observation that under a revenue-maximizing pair of subscription and usage fees, everyone subscribes.

Lemma 1. *Revenue-maximizing subscription and usage fees result in full subscription.*

The intuition is that a sophisticated customer has higher willingness to pay at the time of subscription than at the time of service, and a naive customer is willing to pay even more at the time of subscription than a sophisticated customer would. Hence, for any target level of throughput and expected waiting time, the manager obtains more revenue by

restricting usage per subscriber (raising the usage fee), rather than restricting the number of subscribers. The proof of Lemma 1 translates the problem of finding revenue-maximizing fees into a choice of throughput. An equilibrium throughput $\Lambda \le \bar{\Lambda}$ means that each subscriber goes for service at rate $\Lambda/\bar{\Lambda} = \lambda/\bar{\lambda}$; recall that $\lambda/\bar{\lambda}$ is also the probability that a subscriber goes for service when a need arises. When customers are sophisticated ($\hat{\beta} = \beta$), the manager sets usage and subscription fees

$$\begin{aligned} u &= \beta r - c/(\mu - \Lambda) \quad \text{and} \\ s &= \lambda(\eta r - c/(\mu - \Lambda) - u) \end{aligned} \tag{17}$$

to implement an equilibrium in which all subscribe, go for service at rate $\lambda = \Lambda/M$, and experience expected waiting time $1/(\mu - \Lambda)$, so that the s in (17) is the maximum amount that they will pay for subscription. Hence, revenue maximization with sophisticated customers simplifies to choosing throughput according to

$$\max_{\Lambda \le \bar{\Lambda}} \{\Lambda(\eta r - c/(\mu - \Lambda))\}. \tag{18}$$

and substituting it into (17) to obtain the subscription and usage fees. Revenue maximization (18) is equivalent to welfare maximization (15) with substitution of ηr for r. When $\eta < 1$, customers retain strictly positive expected utility from a long-term perspective (10) from subscribing and using the service, which leads a revenue-maximizing manager to set throughput Λ too low. This implies Proposition 3.

Proposition 3. *Suppose that customers are sophisticated. Revenue is strictly higher under subscription than with a usage fee only. As η increases, a revenue-maximizing manager reduces the usage fee (which increases throughput) and increases the subscription fee, such that at $\eta = 1$, welfare is maximized, and customers are left with zero expected utility. From a long-term perspective, customers have strictly positive utility from subscribing to the service if and only if $\eta < 1$. If the usage fee must be nonnegative, then for $\beta < \underline{\beta}$ the (constrained) revenue-maximizing usage fee is zero, and the resulting revenue, throughput, and welfare strictly increase with β. For $\beta \ge \underline{\beta}$, revenue is constant, the revenue-maximizing usage fee increases with β, and the revenue-maximizing subscription fee decreases with β. The threshold $\underline{\beta} \in (0, \bar{\beta}]$*

With naive customers ($\hat{\beta} > \beta$), the manager could implement full subscription and throughput of Λ with the fees (17). However, that would leave money on the table. With fees (17) and corresponding expected waiting time $1/(\mu - \Lambda)$, a naive subscriber erroneously believes that he will always go for service when a need arises ((5) is satisfied as a strict inequality) rather than with the equilibrium rate λ. A naive subscriber would therefore be willing to pay an incremental fee of $(\eta r - u - c/(\mu - \Lambda))(\bar{\lambda} - \lambda)$ per unit time for subscription, which amounts to additional revenue of $(\eta - \beta)(\bar{\Lambda} - \Lambda)r$. Hence, revenue is maximized with naive customers ($\hat{\beta} > \beta$) by choosing throughput Λ to

$$\max_{\Lambda \le \bar{\Lambda}} \{\bar{\Lambda}(\eta - \beta)r + \Lambda(\beta r - c/(\mu - \Lambda))\} \tag{19}$$

and setting the corresponding fees

$$\begin{aligned} u &= \beta r - c/(\mu - \Lambda) \quad \text{and} \\ s &= \bar{\lambda}(\eta r - u - c/(\mu - \Lambda)) = \bar{\lambda}(\eta - \beta)r. \end{aligned} \tag{20}$$

Remarkably, with naive customers, the revenue-maximizing throughput and usage fee are the same with and without subscription; compare (19) and (16). Therefore, the results in Proposition 1 regarding the revenue-maximizing usage fee and resulting throughput and social welfare hold when the manager charges both for subscription and per use.

Naive customers end up paying more and using the service less than if they were sophisticated. As explained above, for any given level of throughput, naive subscribers pay a higher subscription fee than if they were sophisticated; compare the expressions for s in (17) and (20). Moreover, the additional $(\eta - \beta)(\bar{\Lambda} - \Lambda)r$ in the objective of (19) compared to (18) reduces the optimal Λ and correspondingly increases the usage fee. Lower throughput implies a shorter waiting time, which (amplified by the fact that naive subscribers expect to always go for service) enables the manager to charge an even higher subscription fee to naive subscribers, in addition to the higher usage fee.

Proposition 4. *Suppose that customers are naive. Revenue is strictly higher under subscription than with a usage fee only. The subscription fee, usage fee, and revenue are higher and throughput is lower than if customers were sophisticated, strictly so if $\bar{\Lambda} > \mu - \sqrt{c\mu/(\beta r)}$. Revenue decreases with β, whereas throughput increases with β.*

When customers are sophisticated, revenue increases with their self-control β because, for any given usage fee, they anticipate using the service more frequently. In contrast, when customers are naive, revenue decreases with their self-control β because they continue to erroneously expect to always use the service, whereas actual throughput and waiting increase. Longer waiting times imply lower willingness to pay for subscription according to (20).

A nonnegativity constraint on u (that the manager will not pay customers to use the service) implies that we must add the constraint $\beta r \ge c/(\mu - \Lambda)$ in (18) and in (19). That constraint is binding, meaning that the optimal usage fee is zero in (17) and (20), respectively, when β is small. It follows that when the hyperbolic discounting effect is strong (β is small) and

the manager cannot pay customers to use the service, she should charge only for subscription, not per use. This may help to explain the observation by Cachon and Feldman (2011) that many firms charge only for subscription. An alternative reason to charge only for subscription (not represented in our model) could be that customers incur a mental "pain of paying" each time they pay a usage fee versus only once upon subscription (Prelec and Loewenstein 1998).

Assuming that a manager must choose either subscription or per-use pricing, Cachon and Feldman (2011) showed that, in the absence of hyperbolic discounting, subscription can be optimal because a subscriber's benefit from service r is stochastic. Randhawa and Kumar (2008) found that subscription outperforms per-use pricing for rentals of reusable products. Proposition 5 shows that hyperbolic discounting causes subscription to always strictly outperform per-use pricing.

Proposition 5. *Suppose that a revenue-maximizing manager must either charge for subscription or charge a fee per use, not both. Subscription is strictly optimal.*

In summary, if the manager charges only a usage fee, both the welfare- and revenue-maximizing usage fees decrease with hyperbolic discounting (increase with β). The welfare-maximizing usage fee is lower than the revenue-maximizing usage fee, and is strictly negative for small β. Hence, for small β, under a budget-balance constraint, charging for subscription increases social welfare. Charging for subscription always strictly increases revenue, and is particularly lucrative when customers are naive, in which case throughput is lower than the welfare-maximizing level. More revenue is generated by charging only for subscription than charging only per use. This contrasts with the well-known result that, absent hyperbolic discounting, both welfare and revenue are maximized by charging a usage fee only (Hassin and Haviv 2003, pp. 45–51).

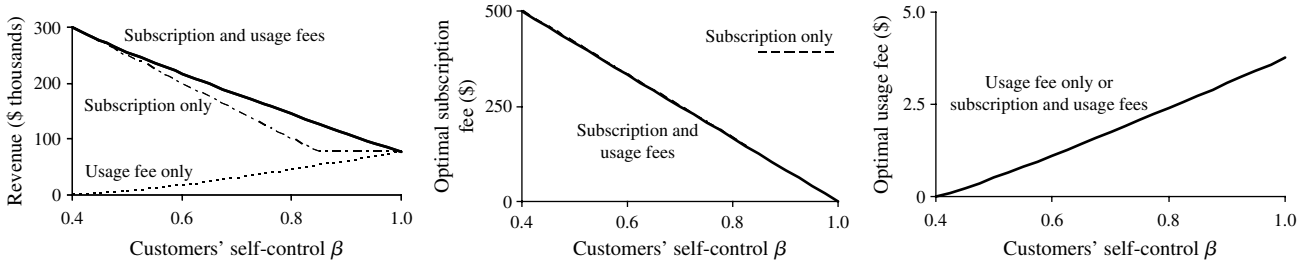
Hyperbolic discounting helps to explain the fact that Fort Car Wash in Fort Atkinson, Wisconsin, sells subscriptions. According to the proprietor, this is surprisingly more profitable than just charging per wash, because subscribers use the car wash less than half as often as they expect to (Ault 2010). In our parlance, those subscribers lack the self-control to get a wash as often as would be in their long-term best interest, and are naive about that lack of self-control.

Consider a numerical example based on the data for Fort Car Wash (assuming that customers are homogeneous with $\eta = 1$ and do not balk or renege after arriving at the car wash). We obtained the data by interviewing the proprietor of the car wash (Ault 2010). A car moves through the washing tunnel in a deterministic service time of four minutes. A maximum of four cars may be lined up in the tunnel, so the queueing time before entering the tunnel ($\mathrm{E}[W] - 4$) follows the distribution for an $M/D/1$ system with service time of one minute. The car wash removes salt and grime from the road, which prevents rust and thus ultimately increases the resale value of the vehicle. Driving a clean car is more pleasant. These benefits amount to $r = \$8$. (This is a reasonable assumption, though in reality, customers differ somewhat in their valuation of a wash.) The other system parameters are $c = \$0.8$/minute, $M = 600$, $\bar{\Lambda} = 2.2$ cars/minute, $\beta = 0.6$, and $\hat{\beta} > \beta$. (These are implied by the subscription fee, number of subscribers, their naively expected and actual utilization, and the mean waiting time reported by the proprietor.[2]) The revenue-maximizing policy and throughput are identical for all $\hat{\beta} > \beta$, so we need not estimate $\hat{\beta}$ exactly.

For insight, let us vary customers' self-control β while holding the other car wash parameters constant. The left graph of Figure 2 shows optimal revenue in the scenarios with a usage fee only, subscription fee only, and combination of subscription and usage fees. The right graph shows the optimal usage fee, which is the same whether or not the proprietor charges for subscription in addition to the usage fee. The middle graph shows the optimal subscription fee, which for $\beta < 0.84$ is the same whether or not the manager also charges per use. Customers with little self-control ($\beta \leq 0.4$) fail to use the service even at zero usage fee, so optimal revenue with a usage fee only is zero, whereas optimal revenue with subscription is maximal, \$300,000. As β increases, throughput and waiting times increase, the optimal usage fee increases, and the optimal subscription fee decreases. Increasing self-control makes naive subscribers better off, because

[2] According to the proprietor, the car wash does almost all of its business during nine "peak" hours each week, during which the arrival rate is 0.8 cars/minute and the expected queueing time is 2 minutes, so $\mathrm{E}[W] = 6$ minutes. Subscribers pay \$299 per year per car (billed monthly to their credit cards) for unlimited access to the car wash ($u = 0$). Subscribers expect to use the service twice per week, but actually use it only 3.2 times per month. For simplicity, we assume that the car wash operates only during peak hours and has a stationary arrival process and the corresponding steady-state waiting time distribution. Scaling the actual arrival rate up by subscribers' ratio of anticipated to actual usage gives us the maximum arrival rate of $\bar{\lambda}M = (0.8 \text{ cars/min}) \times (2 \text{ washes/week}) \times (52 \text{ weeks/year})/(3.2 \text{ washes/month} \times 12 \text{ months/year}) = 2.2$ cars/min. The fact that customers use the service less than anticipated tells us that $\hat{\beta} > \beta$. Assuming $\eta = 1$, we can then solve for $c = \$0.8$/min in (12), which must hold with equality at an optimal subscription fee. We can then solve for $\beta = 0.6$ in (5). Assuming that only subscribers use the car wash implies $M = 600$. In reality, the car wash has only 125 subscribers, and the remainder of demand is from drop-in customers. We will account for these two types (subscribers and drop-in customers) in §5.

Figure 2 Optimal Revenue and Fees Vary with Customers' Self-Control



they pay less but actually use the service more frequently. For $\beta \in (0, 0.84]$, the optimal subscription fee (which is the same whether or not the manager also charges per use) results in full subscription. In contrast, for $\beta \in (0.84, 1]$, if the manager can charge only for subscription, he sets a higher subscription fee to limit the number of subscribers and hence usage. Only if customers have perfect self-control ($\beta = 1$) can the manager achieve the same revenue with a usage fee as by charging for subscription.

At the true $\beta = 0.6$ for the car wash, revenue is nearly as high with only a fee for subscription as with both subscription and usage fees. In reality, the proprietor charges no usage fee for subscribers. In constructing the numerical example, we did not account for a transaction cost or "pain of paying" associated with the usage fee. According to the proprietor, these favor charging only for subscription, as opposed to charging per use, or charging for subscription and per use.

What If Customers Observe the Queue? We have assumed that customers cannot observe the system occupancy. For the case that customers can observe the system occupancy, all of our results on optimal pricing (Propositions 1–4 and Lemma 1) hold, with only minor modifications in Propositions 1 and 4.[3] Their proofs are qualitatively the same, with maximum system occupancy substituted for throughput as the decision variable. However, Proposition 5 changes: When a revenue-maximizing manager is constrained to use either a usage fee or subscription, not both, there exists a threshold $\tilde{\beta} \in [2c/r, \eta]$ such that subscription is optimal if and only if $\beta \le \tilde{\beta}$. Revealing the queue favors a usage fee relative to a subscription fee only, because the usage fee is useful in moderating the maximum system occupancy and in extracting customers' surplus when the queue is short. Maximum system occupancy is lower under the revenue-maximizing usage fee than the revenue-maximizing subscription fee.

4. Optimal Pricing and Scheduling with Heterogeneous Customers

Mendelson and Whang (1990) proved that when customers are heterogeneous in their cost of waiting c, priority scheduling outperforms FIFO scheduling. Furthermore, absent hyperbolic discounting but allowing for heterogeneity in both r and c, subscription is never needed; optimal social welfare is always achieved with usage fees contingent on the choice of service priority.

This section incorporates heterogeneity in customers' self-control. It shows that even when all customers have exactly the same cost of waiting c, priority scheduling can strictly increase welfare and revenue. Priority scheduling for customers with low self-control is valuable as a substitute for the manager paying them to use the service. Moreover, in contrast to Mendelson and Whang's (1990) theorem, subscription is necessary to maximize welfare or revenue with priority scheduling.

Suppose that there are two types of customers. Type 1s lack self-control ($\beta_1 < \eta_1$) and may be either sophisticated ($\hat{\beta}_1 = \beta_1$) or somewhat naive ($\hat{\beta}_1 \in (\beta_1, \eta_1)$) about that lack of self-control. Type 2s have time-consistent preferences ($\beta_2 = \hat{\beta}_2 = \eta_2 = 1$). The two types are homogeneous in all other characteristics. In particular, we assume $\eta_1 = \eta_2$, meaning that when deciding whether or not to subscribe, type 1s discount the future benefits from service in the same manner as do type 2s. Hence, for a given pricing and scheduling policy, if type 1s naively expect to use the service as much as a type 2 would, then there exist multiple equilibria in terms of the number of subscribers of each type and their use of the service. Therefore, in the case that type 1s are naive, we focus on equilibria in which if any type 1s subscribe and/or use the service, then all type 2s use the

[3] The analog of Proposition 1 with a revealed queue is as follows: The welfare-maximizing usage fee increases with β and, if constrained to be nonnegative, is zero for $\beta \le \tilde{\beta}$, in which case the maximum system occupancy and hence social welfare increase with β. The revenue-maximizing usage fee is greater than the socially optimal usage fee, and it strictly increases with β (except at jump points for the maximum system occupancy). The maximum revenue and the corresponding utilization also increase with β. The analog of Proposition 4 is as follows: Revenue increases with subscribers naivete $\hat{\beta}$. Revenue decreases with β, whereas utilization increases with β.

service. This is the most plausible choice of equilibrium because type 2s have greater propensity to use the service and hence greater actual utility from subscription. Let $m_k > 0$ denote the number of type k customers so that $m_1 + m_2 = M$. Then $\bar{\Lambda}_k = \bar{\lambda} m_k$ is the maximum throughput of the service by prospective subscribers of type k for $k = 1, 2$. Let $\bar{\Lambda} = \bar{\Lambda}_1 + \bar{\Lambda}_2$. Finally, for brevity of exposition, we focus on the parameter region $\bar{\Lambda}_2 < \mu - \sqrt{c\mu/r}$, wherein, for purposes of welfare maximization or revenue maximization with sophisticated customers, the manager serves at least some type 1s.

Recall from §2 that neither the customer nor the manager knows the realization of a customer's service time in advance, and the manager cannot interrupt a customer's service to begin serving a different customer. In other words, we restrict attention to scheduling policies that are nonpreemptive and nonanticipating with respect to a customer's service time.

If the manager can pay customers to use the service (set $u < 0$), then an optimal policy for pricing and scheduling has qualitatively the same structure as in the setting of §3 with homogeneous customers (with the exception that if customers are sophisticated, Lemma 1 does not hold). The manager should not price discriminate between the different types of customers, but instead offer one pricing scheme for all customers and serve arriving customers FIFO.

Proposition 6. *Suppose that the usage fee may be strictly negative. FIFO scheduling is optimal for welfare maximization and for revenue maximization. Welfare is maximized with only a usage fee, which strictly increases with β_1 and is strictly negative for*

$$\beta_1 < \min\bigl[c/(r(\mu - \bar{\Lambda})), \sqrt{c/(r\mu)}\bigr]. \tag{21}$$

When type 1*s are sophisticated, revenue is maximized by combining that welfare-maximizing usage fee with a subscription fee that is strictly positive and strictly decreases with β_1, such that a fraction of type* 1*s subscribe, all type* 2*s subscribe, all subscribers always go for service when a need arises, throughput is at the socially optimal level, and revenue equals social welfare. When type* 1*s are instead naive, revenue is maximized by charging the same subscription fee but a strictly higher usage fees, which strictly increases with β_1 and is strictly negative if and only if*

$$\beta_1 < \frac{c}{r(\mu - \bar{\Lambda}_2)}; \tag{22}$$

those fees induce all types 1*s and* 2*s to subscribe and plan to always go for service when a need arises, but type* 1*s fail to do so.*

In revenue maximization with type 1s that are naive and have self-control as low as in (22), the negative usage fee entices the type 1s to pay a higher subscription fee, anticipating "money back" when they go for service, though they will never actually do so.

Now consider a priority scheduling policy: when a customer completes service, next serve the high priority customer who has been waiting longest or, if no high priority customer is waiting, serve the low priority customer who has been waiting longest. The manager must use subscription to discriminate between the two types to implement priority scheduling. If she simply charged a usage fee contingent on the service priority as is optimal in Mendelson and Whang (1990), all customers would prefer the same service priority, and the scheduling policy would degenerate to FIFO. Under subscription and priority scheduling, each prospective subscriber decides whether or not to subscribe and, if he does so, chooses a service priority level. He must pay at rate s^i for subscription to priority level i, and then pay an additional fee of u^i when he uses the service, where $i = h$ and $i = l$ represent high and low service priority, respectively. This formulation allows for charging only a usage fee for service with priority i, by setting $s^i = 0$. Let m_k^i denote the number of type k ($k = 1, 2$) customers who subscribe for service priority level i, let Λ_k^i denote their collective throughput, and let $\Lambda^i = \Lambda_1^i + \Lambda_2^i$ denote total throughput of priority i service ($i = 1, k$). The expected waiting for high and low priority services are

$$\begin{aligned} \mathrm{E}[W^h(\Lambda^h, \Lambda^l)] &\equiv \frac{\mu + \Lambda^l}{\mu(\mu - \Lambda^h)} \quad \text{and} \\ \mathrm{E}[W^l(\Lambda^h, \Lambda^l)] &\equiv \frac{\mu^2 - \Lambda^h(\mu - \Lambda^h - \Lambda^l)}{\mu(\mu - \Lambda^h)(\mu - \Lambda^h - \Lambda^l)}, \end{aligned} \tag{23}$$

respectively. The manager's objective is to maximize revenue $\Pi^p = \Sigma_{i=h,l}\Sigma_{k=1,2}(m_k^i s^i + \Lambda_k^i u^i)$ or social welfare $\Omega^p = \Sigma_{i=h,l}\Sigma_{k=1,2}(\Lambda_k^i(r - c\,\mathrm{E}[W^i]))$, subject to the constraints that for $i = h, l$ and $k = 1, 2$,

$$\begin{gathered} \Lambda_k^i = 0 \quad \text{or, with complementary slackness,} \\ \Lambda_k^i \le m_k^i/\mu \quad \text{and} \quad \beta_k r \ge u^i + c\,\mathrm{E}[W^i]; \end{gathered} \tag{24}$$

$$\begin{gathered} m_k^i = 0 \quad \text{or both} \\ \bar{\lambda}\bigl(r - c\,\mathrm{E}[W^i] - u^i\bigr)\mathbf{1}\bigl(\hat{\beta}_k r \ge c\,\mathrm{E}[W^i] + u^i\bigr) \ge s^i; \\ \text{and, for } j \ne i, \end{gathered} \tag{25}$$

$$\begin{aligned} &\bar{\lambda}\bigl(r - c\,\mathrm{E}[W^i] - u^i\bigr)\mathbf{1}\bigl(\hat{\beta}_k r \ge c\,\mathrm{E}[W^i] + u^i\bigr) - s^i \\ &\quad \ge \bar{\lambda}\bigl(r - c\,\mathrm{E}[W^j] - u^j\bigr)\mathbf{1}\bigl(\hat{\beta}_k r \ge c\,\mathrm{E}[W^j] + u^j\bigr) - s^j. \end{aligned} \tag{26}$$

With or without an additional constraint that the usage fee must be nonnegative, $u^i \ge 0$ for $i = h, l$, Proposition 7 shows that the solution has a simple, intuitive structure, and that priority scheduling is optimal.

Proposition 7. *Priority scheduling is optimal for welfare maximization and for revenue maximization. The corresponding optimal pricing scheme has a strictly positive subscription fee for high priority service and zero subscription fee for low priority service, and motivates type* 1*s to subscribe for high priority service* (*or not use the service at all*) *and all type* 2*s to use the low priority service whenever a need arises.*

For purposes of implementation, the zero subscription fee means that type 2s need not subscribe, but instead could "drop in" for low priority service when a need arises. This stands in contrast to the observation in Proposition 6 that under FIFO scheduling, revenue maximization requires that all customers pay a strictly positive subscription fee.

Together, Propositions 6 and 7 show that the manager can optimally use either FIFO or priority scheduling if the usage fee is not constrained to be nonnegative, or if that constraint is not binding under FIFO. However, when the nonnegativity constraint on the usage fee is binding under FIFO (which occurs when type 1's self-control β_1 is low, as specified in (21) and (22) of Proposition 6), Proposition 8 shows that priority scheduling can strictly outperform FIFO. Priority scheduling is optimal because it minimizes the expected waiting time for type 1s, which best mitigates their self-control problem. Reducing the expected waiting time is a substitute for paying customers to use the service. However, for priority scheduling to yield strictly greater welfare than FIFO scheduling, type 1s must have sufficiently high self-control (27) to use the high priority service at zero usage fee; otherwise priority scheduling would degenerate to FIFO. In revenue maximization with naive customers, condition (27) weakens to (28), meaning that at the time of subscription, type 1s believe that they will have sufficiently high self-control to use the high priority service at zero usage fee.

Proposition 8. *Suppose that usage fees must be nonnegative. Priority scheduling yields strictly greater welfare than does FIFO scheduling if* (21) *holds and*

$$\beta_1 > (1+\bar{\Lambda}_2/\mu)c/(r\mu), \tag{27}$$

in which case it also yields strictly greater revenue with sophisticated customers. When type 1*s are naive, priority scheduling yields strictly greater revenue than does FIFO scheduling if* (22) *holds and*

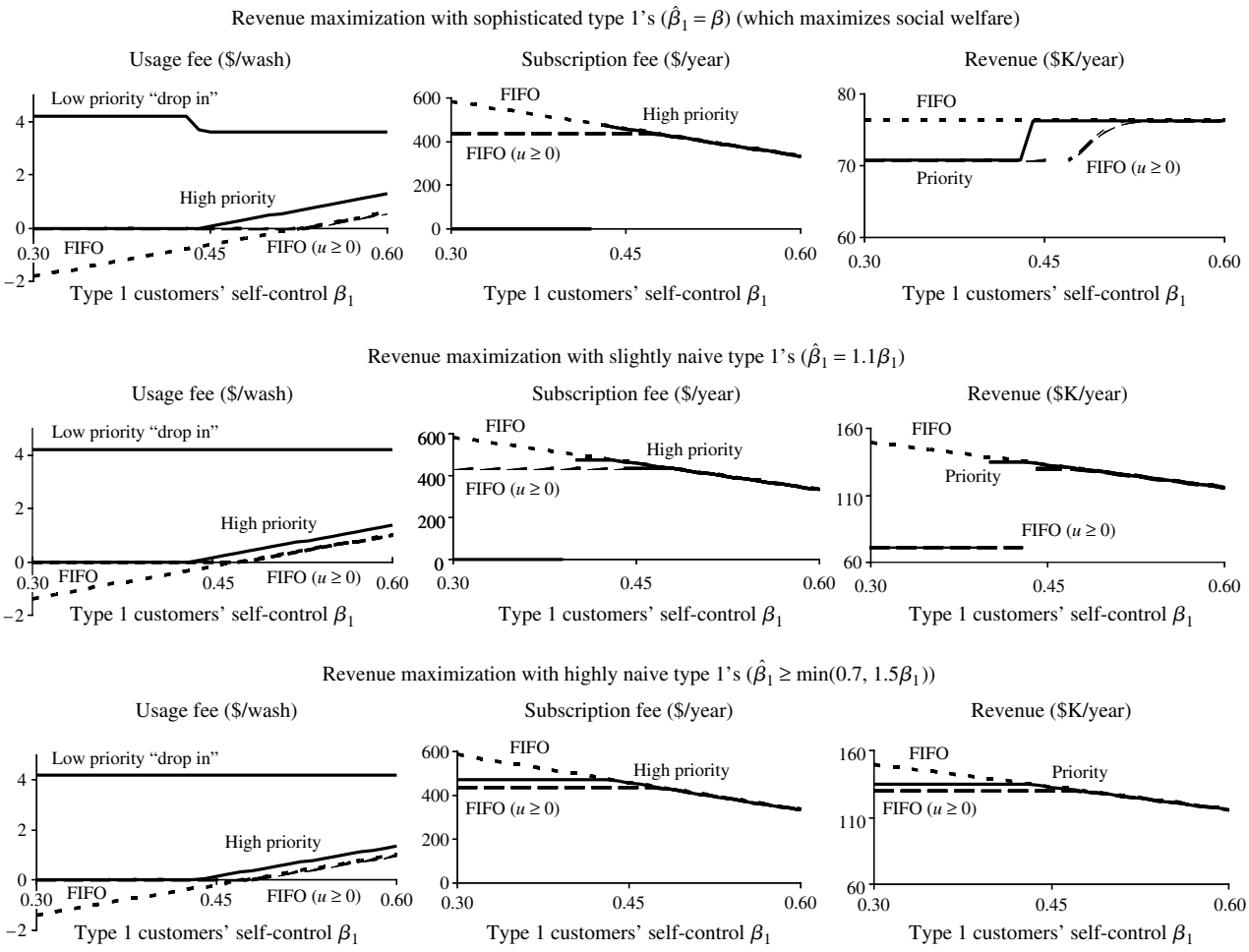
$$\hat{\beta}_1 \geq (1+\bar{\Lambda}_2/\mu)c/(r\mu). \tag{28}$$

Now let us revisit our car wash example, but assume that 125 of the prospective customers are of type 1 (with the parameters described in §3, notably $\beta_1 = 0.6$ and $\hat{\beta}_1 > \beta_1$) and 165 are of type 2 (with the same parameters as type 1, except that $\hat{\beta}_2 = \beta_2 = 1$), so that $\bar{\Lambda}_1 = 0.5$ and $\bar{\Lambda}_2 = 0.6$. Revenue is maximized with FIFO scheduling, a subscription fee of \$330 per annum, and a \$1 usage fee. Under that policy, all 290 prospective customers subscribe, but only the type 2s use the service. The maximum revenue (\$116,000 per year) can also be achieved by allowing customers to pay \$4.20 to "drop in" for a car wash; then, all type 2s drop in when they need a car wash, whereas all the type 1s subscribe but do not use the service. If the manager offered high priority service for subscribers and low priority service for drop-ins, it would be optimal to raise the usage fee for subscribers to \$1.40 to prevent them from actually using the high priority service. The waiting time and hence optimal fee for drop-ins remains the same as under FIFO scheduling. The subscription fee also remains the same, because the increase in the usage fee for subscribers cancels out the expected benefit from reduced waiting time with priority service. Hence, priority scheduling results in exactly the same maximum revenue of \$116,000 per year as FIFO. (In reality, for the reasons discussed in §3, Fort Car Wash does not charge subscribers a usage fee. In our numerical example, under the constraint that $u = 0$ for subscribers, an optimal policy is FIFO scheduling, a subscription fee of \$330 per annum, and a usage fee of \$3.2 for drop-ins. Type 1s subscribe and use the service, so waiting time is higher than under the unconstrained optimal policy, which explains why the subscription fee remains the same and the usage fee for drop-ins decreases. Annual revenue decreases by 17%.)

Now let us vary the self-control β_1 and naivete $\hat{\beta}_1$ of the type 1 customers to see how this affects optimal pricing and scheduling. For all $\beta_1 \geq 0.6$, FIFO scheduling is optimal, and the optimal policy is identical for all $\hat{\beta}_1 > \beta_1$. Figure 3 shows that for $\beta_1 < 0.6$, priority scheduling can increase revenue, and the optimal policy and revenue vary with $\hat{\beta}_1$. With priority scheduling, the manager optimally charges only a usage fee for the low priority service, which is depicted by the solid line on the top in the left graphs. The optimal usage fee and subscription fee for high priority service are depicted by the lower solid line in the left graphs and the solid line in the middle graphs, respectively. For comparison, the dotted lines show the optimal usage fee and subscription fee if the manager may pay customers to use the service (i.e., set $u < 0$, in which case FIFO is always optimal) and the resulting revenue. The dashed lines show the optimal usage fee, subscription fee, and resulting revenue under FIFO scheduling and the constraint that $u \geq 0$.

First consider revenue maximization with sophisticated type 1s (top row of Figure 3). In all cases, all type 2s subscribe, a fraction of type 1s subscribe, and all subscribers always go for service when a need

Figure 3 Optimal Revenue, Fees, and Scheduling Vary with Customers' Naivete and Self-Control



arises. With FIFO scheduling, the usage fee is strictly positive if and only if $\beta_1 > 0.53$. In that parameter region or if the manager may set $u < 0$, throughput is constant at the socially optimal level (with 25% of type 1s subscribing), the usage fee strictly increases with β_1 and the subscription fee strictly decreases with β_1, such that revenue stays constant at the level of optimal social welfare. The optimal prices are not unique; the manager could reduce the usage fee and increase the subscription fee accordingly. The optimal scheduling rule is also not unique; the manager could use priority scheduling, charge only a usage fee for low priority service (for type 2s), and charge the same subscription fee but a higher usage fee for the high priority service (for type 1s). Now consider the case that the usage fee must be nonnegative. Under FIFO scheduling, the constraint $u \geq 0$ is binding for $\beta_1 < 0.53$, so throughput is below the socially optimal level. For $\beta_1 \in (0.48, 0.53)$, the number of type 1s that subscribe and hence revenue strictly increase with β_1, though waiting time and hence the subscription fee strictly decrease with β_1. For $\beta_1 \leq 0.48$, only type 2s subscribe, so revenue and prices are invariant with respect to β_1. For $\beta_1 \in (0.44, 0.53)$, priority scheduling yields strictly greater revenue than FIFO because of the constraint $u \geq 0$. The drop-in usage fee for low priority service is \$3.6, which is the maximum that type 2s are willing to pay when throughput is at the socially optimal level and extracts all their surplus. The socially optimal throughput is achieved, with 25% of type 1s subscribing for the high priority service. The optimal usage fee for high priority service is positive and decreases with β_1, and the subscription fee correspondingly increases, to extract all of the type 1 subscribers' surplus. As β_1 decreases below 0.44, the constraint $u \geq 0$ is binding even under priority scheduling, so the number of type 1 subscribers decreases, as do revenue and waiting times; the subscription fee for high priority service and usage fee for low priority service increase, to extract all the surplus associated with shorter waiting times. For $\beta_1 \leq 0.42$, only type 2s use the service under priority scheduling, so the drop-in usage fee remains constant and achieves exactly the same revenue as under FIFO scheduling.

Now consider revenue maximization with naive type 1s (middle row and bottom row of Figure 3). In all cases, all type 2s and all type 1s subscribe and plan to always go for service when a need arises, but only the type 2s actually do so. Under FIFO scheduling, for $\beta_1 \leq 0.66$, the usage fee is just high enough that type 1 subscribers never actually use the service. This reduces the waiting time and enables the manager to charge a correspondingly higher subscription fee. The usage fee is the same for all levels of naivete $\hat{\beta}_1 > \beta_1$. It increases with type 1 subscribers' self-control β_1, which (as the waiting time remains constant with only type 2s actually using the service) implies that the subscription fee and hence revenue decrease with β_1. The constraint $u \geq 0$ is binding for $\beta_1 < 0.48$, and then priority scheduling increases revenue. The usage fee for subscribers to the high priority service (type 1s) is set sufficiently high that they do not actually use the service. As a result, the usage fee for drop-in low priority service is invariant with respect to β_1 and $\hat{\beta}_1$, and extracts all of the type 2 subscribers' surplus. The subscription fee for high priority service decreases with β_1 because type 1 subscribers anticipated paying the higher usage fee, and therefore revenue decreases with β_1. When type 1s are only slightly naive, the increase in revenue from priority scheduling exceeds 90% for $\beta_1 \in [0.4, 0.44]$, wherein the type 1s erroneously anticipate always using the high priority service and pay an accordingly high subscription fee, but correctly anticipate that they will not use service so do not subscribe under FIFO scheduling. For $\beta_1 \in [0.44, 0.48)$, priority scheduling improves revenue, but by only 1%–3%; type 1s subscribe but do not use the service under either FIFO or priority scheduling, but are willing to pay a higher subscription fee for the high priority. For $\beta_1 < 0.4$, the slightly naive type 1s realize that they will not use the priority service, so they do not subscribe; hence revenue with priority scheduling drops to the same level as with FIFO ($u \geq 0$), the revenue derived from serving only type 2s. In contrast, highly naive type 1s erroneously anticipate using the high priority service or FIFO service, so revenue remains constant for $\beta_1 < 0.4$ and is 4% higher under priority scheduling than FIFO simply because the type 1s pay a higher subscription fee for the high priority service. A final important observation from Figure 3 is that revenue is dramatically higher with naive type 1s than with sophisticated type 1s under either FIFO or priority scheduling (except when the usage fee must be nonnegative, and the type 1 subscribers' self-control β_1 and naivete $\hat{\beta}_1$ are so low that they correctly anticipate never using the service).

Consistent with our numerical example in which, for $\beta_1 \geq 0.6$, revenue is maximized with FIFO scheduling, Fort Car Wash uses FIFO scheduling. As a real-world example of priority scheduling in subscription services, Kennedy's Club Barbershops sell subscriptions for four levels of membership, all of which provide for unlimited free haircuts. The most expensive level of membership also confers nonpreemptive priority in scheduling. A typical priority member is a busy executive who plans to go for a haircut twice per month, but frequently postpones that service. According to the Kennedy's proprietor, managers of most other barbershops fail to maximize revenue through subscription out of a "failure to understand customers' psychology. They overestimate how frequently a man might go for a free haircut." (Hurn 2011).

5. Conclusions

This paper shows how hyperbolic discounting affects optimal pricing and scheduling for a service that is less pleasant for a customer than alternative activities, but generates long-term future benefits.

Some of our findings are valid for systems with or without congestion and queueing. Hyperbolic discounting decreases the revenue-maximizing usage fee and the welfare-maximizing usage fee. When the hyperbolic discounting effect is strong (β is small), the welfare-maximizing usage fee is strictly negative, meaning that customers must be rewarded for using the service. Then, to achieve the optimal social welfare with a balanced budget, the system manager must charge for subscription. However, if the usage fee is constrained to be nonnegative, a welfare-maximizing manager should not charge for subscription. In revenue maximization, hyperbolic discounting favors charging for subscription (in addition to or instead of a usage fee), consistent with the empirical observations of DellaVigna and Malmendier (2006).

Our other prescriptions pertain only to service systems with congestion and queueing. The revenue-maximizing (but not welfare-maximizing) pricing policy depends on whether subscribers are sophisticated about their hyperbolic discounting or naively overestimate their propensity to go for service. A revenue-maximizing manager charges a higher subscription fee and higher usage fee to naive customers than to sophisticated customers, which reduces throughput and welfare. If the revenue-maximizing manager must charge either for subscription or per use, subscription is always strictly optimal. In contrast, welfare is better maximized with a usage fee than subscription. When customers are heterogeneous in their degree of hyperbolic discounting, priority scheduling can dramatically increase the maximum achievable welfare or revenue; it is a substitute for paying the customer with low self-control to use the service.

A parallel paper (Plambeck and Wang 2012) addresses capacity investment and information management. It shows that, for purposes of welfare maximization, hyperbolic discounting favors *hiding* the queue, to prevent people from balking when they see a long line. In contrast, hyperbolic discounting motivates a profit-maximizing system manager to *reveal* the queue (to buy less capacity and still charge a higher usage fee), which reduces throughput and hence social welfare. The welfare loss from revealing the queue and underinvesting in capacity is worsened when the system (e.g., an emergency room) must serve some arriving customers for little or no compensation. One implication is that policy makers should consider prohibiting hospitals from publicizing the emergency room queue and waiting time via the Internet.

Acknowledgments

The authors thank Hunt Allcott and Mor Armony for conversations that helped to identify this topic and develop some of the authors' main results. The authors thank their review team for extraordinarily insightful, constructive and detailed suggestions that greatly improved the paper.

Appendix

Proof of (5)–(9). For the system with scaled-up parameters (4), let $f_\sigma(W)$ denote the probability density function, and let $\bar{F}_\sigma(W)$ denote the complementary cumulative distribution function for the steady-state waiting time W_σ. Expanding the expected discounted reward term in U_τ and then integrating by parts,

$$\begin{aligned}
&\mathrm{E}\left[\int_{W_\sigma}^{W_\sigma+L} \frac{r}{L}(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\right]\\
&\quad=\frac{r}{L}\int_0^\infty f_\sigma(W)\int_W^{W+L}(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\,dW\\
&\quad=\frac{r}{L}\left[-\bar{F}_\sigma(W_\sigma)\int_W^{W+L}(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\right]\Bigg|_0^\infty\\
&\qquad+\frac{r}{L}\int_0^\infty \bar{F}_\sigma(W)\big[(1+\alpha(\tau+W+L))^{-\gamma/\alpha}\\
&\qquad\qquad-(1+\alpha(\tau+W))^{-\gamma/\alpha}\big]\,dW\\
&\quad=\frac{r}{L}\int_0^L(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt+O(\mathrm{E}[W_\sigma]).
\end{aligned}\tag{29}$$

Expanding the expected waiting cost term in U_τ, integrating by parts twice, and using $\mathrm{E}[W_\sigma]=\int_0^\infty \bar{F}_\sigma(t)\,dt$,

$$\begin{aligned}
&\mathrm{E}\left[\int_0^{W_\sigma} c_\sigma(1+\alpha(\tau+t))^{-\gamma/\alpha}dt\right]\\
&\quad=c_\sigma\int_0^\infty f_\sigma(W)\int_0^W(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\,dW\\
&\quad=c_\sigma\left[-\bar{F}_\sigma(W)\int_0^W(1+\alpha(\tau+t))^{-\gamma/\alpha}dt\right]\Bigg|_0^\infty\\
&\qquad+c_\sigma\int_0^\infty \bar{F}_\sigma(W)(1+\alpha(\tau+W))^{-\gamma/\alpha}dW\\
&\quad=-c_\sigma(1+\alpha(\tau+W))^{-\gamma/\alpha}\int_W^\infty \bar{F}_\sigma(t)\,dt\Big|_0^\infty\\
&\qquad-c_\sigma\int_0^\infty\gamma(1+\alpha(\tau+W))^{-(1+\gamma/\alpha)}\int_W^\infty\bar{F}_\sigma(t)\,dt\,dW\\
&\quad=c_\sigma(1+\alpha\tau)^{-\gamma/\alpha}\mathrm{E}[W_\sigma]\\
&\qquad-c_\sigma\int_0^\infty\gamma(1+\alpha(\tau+W))^{-(1+\gamma/\alpha)}\int_W^\infty\bar{F}_\sigma(t)\,dt\,dW.
\end{aligned}\tag{30}$$

In an $M/M/1$ queue, the waiting time W_σ is exponentially distributed, and therefore

$$\begin{aligned}
2(\mathrm{E}[W_\sigma])^2&=\mathrm{VAR}[W_\sigma]+(\mathrm{E}[W_\sigma])^2=\mathrm{E}[W_\sigma^2]\\
&=\int_0^\infty W^2 f_\sigma(W)\,dW=2\int_0^\infty\int_W^\infty\bar{F}_\sigma(t)\,dt\,dW,
\end{aligned}$$

which leads to the following bound on the second term in (30):

$$\begin{aligned}
&c_\sigma\int_0^\infty\gamma(1+\alpha(\tau+W))^{-(1+\gamma/\alpha)}\int_W^\infty\bar{F}_\sigma(t)\,dt\,dW\\
&\quad\le c_\sigma\gamma\int_0^\infty\int_W^\infty\bar{F}_\sigma(t)\,dt\,dW=c_\sigma\gamma(\mathrm{E}[W_\sigma])^2.
\end{aligned}$$

In the asymptotic regime with $\sigma\to\infty$, in all stable systems where $\lambda M_\sigma<\mu_\sigma$,

$$\lim_{\sigma\to+\infty}\mathrm{E}[W_\sigma]=\lim_{\sigma\to+\infty}(\mu_\sigma-\lambda M_\sigma)^{-1}=\lim_{\sigma\to+\infty}\sigma^{-1}\mathrm{E}[W]=0$$

$$\text{and}\quad c_\sigma\,\mathrm{E}[W_\sigma]=c\,\mathrm{E}[W],\quad\text{where }\mathrm{E}[W]=(\mu-\lambda M)^{-1}.$$

Thus, we conclude that when $\sigma\to+\infty$, the second terms in (29) and (30) vanish, and

$$\begin{aligned}
U_\tau&\to\frac{r}{L}\int_0^L(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\\
&\qquad-(1+\alpha\tau)^{-\gamma/\alpha}(u+c\,\mathrm{E}[W]).
\end{aligned}\tag{31}$$

$$U_0\to\beta r-u-c\,\mathrm{E}[W],\quad\text{where}$$

$$\beta=L^{-1}\int_0^L(1+\alpha t)^{-\gamma/\alpha}\,dt=\frac{1-(1+\alpha L)^{1-\gamma/\alpha}}{(\gamma-\alpha)L}.$$

Note that U_0 is the left-hand side of (1), a customer's immediate utility from going for service when an opportunity arises. This completes the proof that in the asymptotic regime, (1) simplifies to (5) with β given by (7). One can verify through Taylor expansion of $(1+\alpha L)^{1-\gamma/\alpha}$ that $\beta\in(0,1)$.

Tonelli's theorem and our assumption that a customer's anticipated arrival process (the Poisson process $\{\tau_i\}_{i=1,..,\infty}$ with rate $\hat{\lambda}$) and waiting times are independent imply that

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^\infty U_{\tau_i}\right]&=\mathrm{E}\left[\int_0^\infty U_\tau 1\big[\tau_i=\tau\text{ for some }i\in\{1,\ldots,\infty\}\big]\,d\tau\right]\\
&=\int_0^\infty\hat{\lambda}U_\tau\,d\tau.
\end{aligned}\tag{32}$$

Combining (31) and (32), in the asymptotic regime,

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^\infty U_{\tau_i}\right]&\to\int_0^\infty\hat{\lambda}\left[rL^{-1}\int_0^L(1+\alpha(\tau+t))^{-\gamma/\alpha}\,dt\right.\\
&\qquad\left.-(1+\alpha\tau)^{-\gamma/\alpha}(u+c_\sigma\,\mathrm{E}[W_\sigma])\right]d\tau\\
&=\hat{\lambda}(\eta r-u-c\,\mathrm{E}[W])\int_0^\infty(1+\alpha\tau)^{-\gamma/\alpha}\,d\tau.
\end{aligned}$$

Thus, (2) simplifies to (6), with

$$\Delta = \int_0^\infty (1+\alpha\tau)^{-\gamma/\alpha}\, d\tau$$
$$\eta = (L\Delta)^{-1} \int_0^\infty \int_0^L (1+\alpha(\tau+t))^{-\gamma/\alpha}\, dt d\tau. \tag{33}$$

To see that η simplifies to (8), note that when $\gamma = 2\alpha$, $\Delta = \alpha^{-1}$, and

$$\eta = -\frac{[\ln(1+\alpha(\tau+L)) - \ln(1+\alpha\tau)]_0^\infty}{\alpha L} = \frac{\ln(1+\alpha L)}{\alpha L}.$$

When $\gamma \le \alpha$, by applying L'Hôpital's rule,

$$\eta = \frac{\int_0^\infty \int_0^L (1+\alpha(\tau+t))^{-\gamma/\alpha} dt d\tau}{L\int_0^\infty (1+\alpha\tau)^{-\gamma/\alpha} d\tau}$$
$$= \lim_{\tau\to\infty} \frac{\int_0^L (1+\alpha(\tau+t))^{-\gamma/\alpha} dt}{L(1+\alpha\tau)^{-\gamma/\alpha}}$$
$$= \lim_{\tau\to\infty} L^{-1} \int_0^L \left(1+\frac{\alpha t}{1+\alpha\tau}\right)^{-\gamma/\alpha} dt = 1.$$

When $\gamma > \alpha$ and $\gamma \ne 2\alpha$, $\Delta = (\gamma-\alpha)^{-1}$, and

$$\eta = \frac{[(1+\alpha(\tau+L))^{2-\gamma/\alpha} - (1+\alpha\tau)^{2-\gamma/\alpha}]_0^\infty}{(\gamma-2\alpha)L}$$
$$= \frac{1-(1+\alpha L)^{2-\gamma/\alpha}}{(\gamma-2\alpha)L} + \lim_{\tau\to+\infty} \frac{(1+\alpha(\tau+L))^{2-\gamma/\alpha} - (1+\alpha\tau)^{2-\gamma/\alpha}}{(\gamma-2\alpha)L}.$$

Applying Taylor expansion to $(1+\alpha(\tau+L))^{2-\gamma/\alpha}$ with respect to L,

$$\eta = \frac{1-(1+\alpha L)^{2-\gamma/\alpha}}{(\gamma-2\alpha)L} + \lim_{\tau\to+\infty} O((1+\alpha\tau)^{1-\gamma/\alpha})$$
$$= \frac{1-(1+\alpha L)^{2-\gamma/\alpha}}{(\gamma-2\alpha)L}.$$

We now prove (9). Define

$$\Delta(T) = \int_T^\infty (1+\alpha\tau)^{-\gamma/\alpha}\, d\tau \quad \text{and}$$
$$\eta(T) = \frac{1}{L\Delta(T)} \int_T^\infty \int_0^L (1+\alpha(\tau+t))^{-\gamma/\alpha} dt d\tau.$$

Then, following similar steps from (32) to (33), (3) holds if and only if

$$\hat\lambda(\eta(T)r - u - c\,\mathrm{E}[W]) \ge s.$$

Applying l'Hôpital's rule and then the Taylor expansion,

$$\lim_{T\to\infty} \eta(T) = \lim_{T\to\infty} \frac{\int_0^L (1+\alpha(T+t))^{-\gamma/\alpha} dt}{L(1+\alpha T)^{-\gamma/\alpha}} = 1. \quad \square$$

Proof of Proposition 1. The welfare-maximizing throughput is $\Lambda_\Omega^* = \min(\bar\Lambda, \Lambda_\Omega)$, and the corresponding usage fee is $u_\Omega^* = \beta r - c\,\mathrm{E}[W(\Lambda_\Omega^*)]$, where Λ_Ω is the unique solution to the first-order condition

$$\frac{d\Omega}{d\Lambda} = r - c\,\mathrm{E}[W(\Lambda)] - c\Lambda \frac{d\,\mathrm{E}[W(\Lambda)]}{d\Lambda} = 0. \tag{34}$$

Because Λ_Ω^* does not depend on β, u_Ω^* strictly increases with β and is strictly negative if and only if

$$\beta < \bar\beta \equiv c\,\mathrm{E}\big[W(\min(\Lambda_\Omega, \bar\Lambda))\big]/r.$$

That $\bar\beta \in (0,1)$ is implied by (34) and the fact that $d\,\mathrm{E}[W(\Lambda)]/d\Lambda > 0$. If u $(=\beta r - c\,\mathrm{E}[W(\Lambda)])$ must be nonnegative, then concavity of Ω in Λ implies that $\Lambda_\Omega^* = \min(\bar\Lambda, \Lambda_\Omega, \Lambda_0)$, where Λ_0 is the unique solution to $\beta r - c\,\mathrm{E}[W(\Lambda)] = 0$, which strictly increases with β. When $\beta < \bar\beta$, $\Lambda_\Omega^* = \Lambda_0$, so $u_\Omega^* = \beta r - c\,\mathrm{E}[W(\Lambda_0)] = 0$. Welfare, which strictly increases in Λ when $\Lambda < \Lambda_\Omega$, strictly increases in β. In revenue maximization

$$\max_u \big\{\Pi = \Lambda(u)u = \Lambda(u)\big(\beta r - c\,\mathrm{E}[W(\Lambda(u))]\big)\big\}, \tag{35}$$

the optimal throughput is $\Lambda_\Pi^* = \min(\bar\Lambda, \Lambda_\Pi)$, where Λ_Π is the unique solution to the first-order condition

$$\frac{d\Pi}{d\Lambda} = \beta r - c\left(\mathrm{E}[W(\Lambda)] + \Lambda \frac{d\,\mathrm{E}[W(\Lambda)]}{d\Lambda}\right) = 0. \tag{36}$$

By comparison of (36) with (34), $\Lambda_\Pi < \Lambda_\Omega$, so $\Lambda_\Pi^* \le \Lambda_\Omega^*$ with strict inequality if and only if $\Lambda_\Pi < \bar\Lambda$, which occurs for $\bar\Lambda > \mu - \sqrt{c\mu/(\beta r)}$. The revenue-maximizing usage fee is

$$u_\Pi^* = \beta r - c\,\mathrm{E}[W(\Lambda_\Pi^*)] = \max\left(\beta r - c\,\mathrm{E}[W(\bar\Lambda)], c\Lambda_\Pi \frac{d\,\mathrm{E}[W(\Lambda_\Pi)]}{d\Lambda}\right),$$

where $d\,\mathrm{E}[W(\Lambda_\Pi)]/d\Lambda$ is $d\,\mathrm{E}[W(\Lambda)]/d\Lambda$ evaluated at $\Lambda = \Lambda_\Pi$, which is strictly positive and, because $\mathrm{E}[W(\Lambda)]$ is convex, increases in Λ, which strictly increases with β. Hence, u_Π^* strictly increases with β. That revenue strictly increases in β is immediate from (35). $\square$

Proof of Proposition 2. Building on the proof of Proposition 1, throughput under the (constrained) welfare-maximizing nonnegative usage fee is lower than welfare-maximizing level $\Lambda_\Omega^* = \min(\bar\Lambda, \Lambda_\Omega, \Lambda_0)$ if and only if $\beta < \bar\beta$. In that parameter region, the optimal usage fee is zero, and charging for subscription would reduce throughput and hence welfare. With only a break-even budget constraint $s + \lambda u \ge 0$, the welfare-maximizing throughput $\Lambda_\Omega^* = \min(\bar\Lambda, \Lambda_\Omega)$ is achieved in the region $\beta < \bar\beta$ with the strictly negative usage fee u_Ω^* (characterized in the proof of Proposition 1) and $s = (\Lambda_\Omega^*/M)(\eta r - u_\Omega^* - c\,\mathrm{E}[W(\Lambda_\Omega^*)])$, which is the maximum subscription fee that satisfies (12), so all customers subscribe. The equilibrium arrival rate per customer is $\lambda = \Lambda_\Omega^*/M$. Our assumption $\beta r > c$ and the fact that $\eta > \beta$ imply that $\eta r - c\,\mathrm{E}[W(\Lambda_\Omega^*)] > 0$, so the budget constraint is satisfied. $\square$

Proof of Lemma 1. The manager sets the subscription fee s and usage fee u to determine number of subscribers m and throughput Λ to

$$\max \quad \{\Pi^s = ms + \Lambda u\} \tag{37}$$
$$\text{s.t.} \quad 0 \le \Lambda \le \bar\lambda m, \tag{38}$$
$$u = \beta r - c\,\mathrm{E}[W(\Lambda)], \tag{39}$$
$$s \le \hat\lambda\big(\eta r - c\,\mathrm{E}[W(\Lambda)] - u\big). \tag{40}$$

The upper limit on Λ in (38) corresponds to having every subscriber go for service at the maximum rate $\bar\lambda$. In equilibrium, subscribers adopt a mixed strategy of entering the system with probability $\lambda/\bar\lambda$, where $\lambda = \Lambda/m$, such that the discounted reward is balanced with the expected cost of waiting and the usage fee (39). As in (12), $\hat\lambda$ is the rate that

a customer expects to use the service: $\hat{\lambda} = \lambda$ if customers are sophisticated, and otherwise $\hat{\lambda} = \bar{\lambda}$. Note that the second inequality in (12) is implied by (39) and $\hat{\beta} \geq \beta$. Clearly, (40) is binding at an optimal solution. Hence, with sophisticated customers, the objective in (37) simplifies to

$$\Pi^s = \Lambda\big(\eta r - c\,\mathrm{E}[W(\Lambda)]\big), \tag{41}$$

and the manager optimally sets $s = (\Lambda/M)(\eta - \beta)r$ and achieves full subscription ($m = M$). Having $m = M$ does not affect the objective function and makes the constraints (38) least binding. With naive customers, the manager optimally sets $s = \bar{\lambda}(\eta r - c\,\mathrm{E}[W(\Lambda)] - u) = \bar{\lambda}(\eta - \beta)r$ which, when substituted into (37), yields

$$\begin{aligned}\Pi^s &= m\bar{\lambda}(\eta - \beta)r + \Lambda u \\ &= m\bar{\lambda}(\eta - \beta)r + \Lambda\big(\beta r - c\,\mathrm{E}[W(\Lambda)]\big).\end{aligned} \tag{42}$$

Increasing m relaxes the constraints (38) and increases the revenue in (42), so full subscription ($m = M$) is optimal. □

Proof of Proposition 3. That revenue is strictly higher under subscription is immediate by comparison of the revenue maximization problem with a usage fee only (Equation (16)) and the revenue maximization problem with subscription ((41) for sophisticated customers and (42) for naive customers), and the fact that $\eta > \beta$.

When the manager can set $u < 0$, the revenue-maximizing throughput, usage fee, and subscription fee are

$$\begin{aligned}&\Lambda^*_\Pi = \min\left(\bar{\Lambda}, \mu - \sqrt{\frac{c\mu}{\eta r}}\right), \quad u^*_\Pi = \beta r - \frac{c}{\mu - \Lambda^*_\Pi}, \quad \text{and} \\ &s^*_\Pi = \frac{\Lambda^*_\Pi}{M}\left(\eta r - u^*_\Pi - \frac{c}{\mu - \Lambda^*_\Pi}\right) = \frac{\Lambda^*_\Pi}{M}(\eta - \beta)r,\end{aligned} \tag{43}$$

and the resulting revenue, social welfare, and consumer surplus are

$$\begin{aligned}&Ms^*_\Pi + u^*_\Pi\Lambda^*_\Pi = \Lambda^*_\Pi\left(\eta r - \frac{c}{\mu - \Lambda^*_\Pi}\right), \\ &\Lambda^*_\Pi\left(r - \frac{c}{\mu - \Lambda^*_\Pi}\right), \quad \text{and} \quad \Lambda^*_\Pi(1 - \eta)r,\end{aligned} \tag{44}$$

respectively.

When the manager is constrained to set a nonnegative usage fee $u \geq 0$,

$$\Lambda^*_\Pi = \min\left(\bar{\Lambda}, \mu - \sqrt{\frac{c\mu}{\eta r}}, \mu - \frac{c}{\beta r}\right),$$

where $\mu - c/(\beta r)$ is the value that holds $\beta r = c\,\mathrm{E}[W] = c/(\mu - \Lambda)$ at equality. Let $\underline{\beta}$ be the unique value that satisfies

$$\min\left(\bar{\Lambda}, \mu - \sqrt{\frac{c\mu}{\eta r}}\right) = \mu - \frac{c}{\beta r}.$$

When $\beta \geq \underline{\beta}$, throughput, fees, revenue, social welfare, and surplus are given by (43) and (44). When $\beta < \underline{\beta}$,

$$\begin{aligned}&\Lambda^*_\Pi = \mu - \frac{c}{\beta r}, \quad u^*_\Pi = 0, \quad \text{and} \\ &s^*_\Pi = \frac{\Lambda^*_\Pi}{M}\left(\eta r - \frac{c}{\mu - \Lambda^*_\Pi}\right) = \frac{\Lambda^*_\Pi}{M}(\eta - \beta)r.\end{aligned}$$

Revenue, social welfare, and consumer surplus are

$$\begin{aligned}&Ms^*_\Pi = \Lambda^*_\Pi(\eta - \beta)r, \\ &\Lambda^*_\Pi\left(r - \frac{c}{1 - \Lambda^*_\Pi}\right) = \Lambda^*_\Pi(1 - \beta)r, \quad \text{and} \quad \Lambda^*_\Pi(1 - \eta)r,\end{aligned}$$

respectively. Observe that $\bar{\beta}$ is the value of β at which

$$\min\left(\bar{\Lambda}, \mu - \sqrt{\frac{c\mu}{r}}\right) = \mu - \frac{c}{\beta r},$$

wherein the left-hand side is the socially optimal throughput level. Hence, $\underline{\beta} < \bar{\beta}$ when $\eta < 1$. All other statements in the proposition follow immediately from the above expressions. □

Proof of Proposition 4. Immediate from (16), (41), and (19) and as described in the text surrounding Proposition 4. Observe from (19) that the optimal revenue decreases in β because $\Lambda \leq \bar{\Lambda}$. □

Proof of Proposition 5. Let Λ^*_Π denote the revenue-maximizing throughput when the manager can charge only a usage fee; the resulting optimal revenue is

$$\Pi(u^*_\Pi) = \Lambda^*_\Pi\big(\beta r - c\,\mathrm{E}[W(\Lambda^*_\Pi)]\big). \tag{45}$$

The manager can induce the same throughput Λ^*_Π by charging only the subscription fee

$$s = \bar{\lambda}\big(\eta r - c\,\mathrm{E}[W(\Lambda^*_\Pi)]\big)$$

to induce a mixed strategy equilibrium with $m = \Lambda^*_\Pi/\bar{\lambda}$ subscribers that use the service whenever a need arises, so the expected cost of waiting is $c\,\mathrm{E}[W(\Lambda^*_\Pi)]$. The fact that subscribers will use the service whenever a need arises follows from

$$\beta r = u^*_\Pi + c\,\mathrm{E}[W(\Lambda^*_\Pi)] > c\,\mathrm{E}[W(\Lambda^*_\Pi)].$$

Sophisticated prospective subscribers know β, and naive ones overestimate β, and therefore all correctly anticipate that they will use the service whenever a need arises, so will pay a maximum of $\bar{\lambda}(\eta r - c\,\mathrm{E}[W(\Lambda^*_\Pi)])$ for subscription. The fact that only $m = \Lambda^*_\Pi/\bar{\lambda}$ people subscribe in equilibrium follows from the observation that exceeding that number would drive the expected waiting cost above $c\,\mathrm{E}[W(\Lambda^*_\Pi)]$ and the value of the subscription below the subscription fee, which would prevent people from subscribing. Conversely, having fewer subscribers would decrease the waiting cost, which would motivate all to subscribe. The resulting revenue rate is

$$ms = \Lambda^*_\Pi\big(\eta r - c\,\mathrm{E}[W(\Lambda^*_\Pi)]\big), \tag{46}$$

which strictly dominates $\Pi(u^*_\Pi)$ because $\beta < \eta$. □

Proof of Proposition 6. Any candidate optimal scheduling policy ψ satisfies the conditions for Kleinrock's (1965) conservation law: preemption is not allowed, scheduling has no impact on the realizations of arrival or service completion times, and it is never optimal for the server to be idle when a customer is waiting. Kleinrock's (1965) conservation law says that

$$\frac{\Lambda_1}{\mu}\,\mathrm{E}[W^\psi_1] + \frac{\Lambda_2}{\mu}\,\mathrm{E}[W^\psi_2] = \frac{(\Lambda_1 + \Lambda_2)/\mu}{\mu - \Lambda_1 - \Lambda_2}, \tag{47}$$

where Λ_k is throughput by type ks, and $\mathrm{E}[W_k^\psi]$ is the mean waiting time for type ks ($k = 1, 2$) that use the service. Therefore, welfare optimization simplifies to choosing throughput Λ_1 and Λ_2 in

$$\Omega^* = \max\left\{(\Lambda_1 + \Lambda_2)\left(r - \frac{c}{\mu - \Lambda_1 - \Lambda_2}\right)\right\}.$$

Without loss of social optimality, we assume that $\Lambda_1 > 0$ (some type 1s are served) only if $\Lambda_2 = \bar{\Lambda}_2$ (all type 2s are served), so

$$\Omega^* = \max\{\Omega_1^*, \Omega_2^*\}, \quad \text{where}$$
$$\Omega_1^* = \max_{0\le\Lambda_1\le\bar{\Lambda}_1}\left\{(\Lambda_1 + \bar{\Lambda}_2)r - \frac{c(\Lambda_1 + \bar{\Lambda}_2)}{\mu - \Lambda_1 - \bar{\Lambda}_2}\right\} \quad \text{and} \tag{48}$$
$$\Omega_2^* = \max_{0\le\Lambda_2\le\bar{\Lambda}_2}\left\{\Lambda_2\left(r - \frac{c}{\mu - \Lambda_2}\right)\right\}$$

are, respectively, the optimal welfare when serving both types and the optimal welfare when serving only type 2s. One may easily verify that within our assumed parameter region where $\bar{\Lambda}_2 < \mu - \sqrt{c\mu/r}$, $\Omega_1^* > \Omega_2^*$. In this case, the socially optimal throughput by type 1s is

$$\Lambda_{1\Omega}^* = \min\left(\bar{\Lambda}_1, \mu - \bar{\Lambda}_2 - \sqrt{c\mu/r}\right). \tag{49}$$

Socially optimal throughput ($\Lambda_2 = \bar{\Lambda}_2$ and $\Lambda_1 = \Lambda_{1\Omega}^*$) is induced under FIFO scheduling by charging the usage fee

$$\begin{aligned} u_\Omega^* &= \beta_1 r - c\,\mathrm{E}\left[W(\Lambda_{1\Omega}^* + \bar{\Lambda}_2)\right] \\ &= \beta_1 r - \frac{c}{\mu - \min(\bar{\Lambda}, \mu - \sqrt{c\mu/r})}, \end{aligned} \tag{50}$$

which strictly increases in β_1 and is strictly negative if and only if (21) holds.

When type 1s are sophisticated, revenue cannot exceed social welfare. The manager earns revenue equal to the optimal social welfare by charging a subscription fee of

$$s_\Omega^* = \bar{\lambda}\left(r - c\,\mathrm{E}[W(\Lambda_{1\Omega}^* + \bar{\Lambda}_2)] - u_\Omega^*\right) = \bar{\lambda}(1 - \beta_1)r > 0 \tag{51}$$

plus the usage fee (50) and using FIFO scheduling. The sophisticated type 1s will pay (51) to subscribe if and only if they will use the service whenever an opportunity arises (at rate $\bar{\lambda}$), and only $\Lambda_{1\Omega}^*/\bar{\lambda}$ of them subscribe in equilibrium. Note that the subscription fee is set to give them exactly zero utility from subscription. All type 2s subscribe and always go for service whenever a need arises. The resulting throughput is socially optimal and the resulting revenue,

$$\begin{aligned} &s_\Omega^*\left(\Lambda_{1\Omega}^*/\bar{\lambda} + m_2\right) + u_\Omega^*\left(\Lambda_{1\Omega}^* + \bar{\Lambda}_2\right) \\ &\quad = \left(\Lambda_{1\Omega}^* + \bar{\Lambda}_2\right)\left(r - c\,\mathrm{E}[W(\Lambda_{1\Omega}^* + \bar{\Lambda}_2)]\right), \end{aligned}$$

is exactly the optimal social welfare.

Consider revenue maximization with naive type 1s. By assumption, if any type 1s subscribe and/or use the service, then all type 2s use the service. Hence, the maximum revenue that can be collected from type 2s is

$$\bar{\Lambda}_2(r - c\,\mathrm{E}[W_2^\psi]), \tag{52}$$

because they use the service at the maximum rate $\bar{\Lambda}_2$ and have expected waiting time $\mathrm{E}[W_2^\psi]$. With Λ_1 denoting type 1 subscribers' equilibrium throughput and u_1 their usage fee, the maximum revenue that can be collected from type 1s is

$$\Lambda_1 u_1 + m_1\bar{\lambda}(r - c\,\mathrm{E}[W_1^\psi] - u_1), \tag{53}$$

which is achieved when all m_1 type 1s subscribe and plan to always go for service, and pay the subscription fee $\bar{\lambda}(r - c\,\mathrm{E}[W_1^\psi] - u_1) > 0$ that exactly extracts their expected utility.

Observe that $u_1 \le \beta_1 r - c\,\mathrm{E}[W_1^\psi(\Lambda_1, \bar{\Lambda}_2)]$ if $\Lambda_1 < \bar{\Lambda}_1$, so revenue is maximized by setting $u_1 = \beta_1 r - c\,\mathrm{E}[W_1^\psi(\Lambda_1, \bar{\Lambda}_2)]$ in (53). Then, adding (52) and (53) and applying Kleinrock's (1965) conservation law (47), total revenue is

$$\begin{aligned} \max_{0\le\Lambda_1\le\bar{\Lambda}_1}\Big\{&\Pi(\Lambda_1) = (\bar{\Lambda}_1 + \bar{\Lambda}_2)(1 - \beta_1)r \\ &+ (\Lambda_1 + \bar{\Lambda}_2)\left(\beta_1 r - \frac{c}{\mu - \Lambda_1 - \bar{\Lambda}_2}\right)\Big\}. \end{aligned} \tag{54}$$

Because (54) is greater than Ω_1^* in (48) and $\Omega_1^* > \Omega_2^*$, (54) exceeds the maximum revenue that can be achieved by serving only type 2s, Ω_2^*.

The throughput of type 1s that maximizes (54) is

$$\Lambda_{1\Pi}^* = \min\left(\bar{\Lambda}_1, (\mu - \bar{\Lambda}_2 - \sqrt{c\mu/(\beta_1 r)})^+\right),$$

which is induced under FIFO scheduling with the usage fee

$$u_\Pi^* = \beta_1 r - c\,\mathrm{E}[W(\Lambda_{1\Pi}^* + \bar{\Lambda}_2)] = \beta_1 r - \frac{c}{\mu - \Lambda_{1\Pi}^* - \bar{\Lambda}_2}. \tag{55}$$

Because $\Lambda_{1\Pi}^* \le \bar{\Lambda}_1$, although type 1s believe they will always go for service when a need arises, they fail to do so unless β_1 is sufficiently large. Revenue maximization requires subscription by type 1s; their utility cannot be fully captured by a usage fee because $\beta_1 < 1$, and they anticipate using the system more than they actually do. The revenue-maximizing subscription fee is

$$s_\Pi^* = \bar{\lambda}\left(r - \mathrm{E}[W(\Lambda_{1\Pi}^* + \bar{\Lambda}_2)] - u_\Pi^*\right) = \bar{\lambda}(1 - \beta_1)r.$$

All type 1s subscribe, and the resulting revenue is the same as that in (54), i.e.,

$$\begin{aligned} &(m_1 + m_2)s_\Pi^* + \left(\Lambda_{1\Pi}^* + \bar{\Lambda}_2\right)u_\Pi^* \\ &\quad = \left(\bar{\Lambda}_1 + \bar{\Lambda}_2\right)(1 - \beta_1)r \\ &\qquad + \left(\Lambda_{1\Pi}^* + \bar{\Lambda}_2\right)\left(\beta_1 r - \frac{c}{\mu - \Lambda_{1\Pi}^* - \bar{\Lambda}_2}\right). \end{aligned} \tag{56}$$

Comparing $\Lambda_{1\Pi}^*$ with (49), $\Lambda_{1\Pi}^* \le \Lambda_{1\Omega}^*$ and strictly so if $\Lambda_{1\Pi}^* < \bar{\Lambda}_1$. Thus, by (50) and (55), the revenue-maximizing usage fee is strictly higher than the socially optimal usage fee. The latter equality also shows that u_Π^* strictly increases with β_1, and is strictly negative if and only if (22) holds. □

Proof of Propositions 7 and 8. Proposition 8 assumes that the usage fee is constrained to be nonnegative, whereas Proposition 7 holds regardless of whether or not the usage fee is constrained to be nonnegative. In the proof below, we assume the usage fee is constrained to be nonnegative, $u^i \ge 0$ for $i = l, h$ as the default case, and insert additional explanation as needed to prove Proposition 7 for the case that the usage fee may be negative.

In welfare maximization, the objective is the same as in (48). In the parameter region where $\bar{\Lambda}_2 < \mu - \sqrt{c\mu/r}$, the manager serves both types, the optimal welfare is

$$\max_{0\le\Lambda_1\le\bar{\Lambda}_1}\left\{(\Lambda_1+\bar{\Lambda}_2)\left(r-\frac{c}{\mu-\Lambda_1-\bar{\Lambda}_2}\right)\right\}, \tag{57}$$

subject to the constraints that, because the usage fee must be nonnegative,

$$\begin{aligned}&\beta_1 r \ge c\,\mathrm{E}[W_1^{\psi}] \quad \text{or} \quad \Lambda_1 = 0, \quad \text{and}\\ &(\beta_1 r - c\,\mathrm{E}[W_1^{\psi}])(\bar{\Lambda}_1 - \Lambda_1) = 0,\end{aligned} \tag{58}$$

where $\mathrm{E}[W_1^{\psi}]$ is the expected waiting time for type 1s under a general scheduling policy ψ, at throughput levels $(\Lambda_1, \bar{\Lambda}_2)$. Out of all nonpreemptive policies ψ, priority scheduling for type 1s (indexed by p) minimizes $\mathrm{E}[W_1^{\psi}]$ (Gelenbe and Mitrani 1980, Lemma 6.5, p. 199) and therefore optimizes (57)–(58) by making the constraint (58) least binding.

Following (23), $\mathrm{E}[W_1^p] = \mathrm{E}[W^h(\Lambda_1, \bar{\Lambda}_2)] = (1 + \bar{\Lambda}_2/\mu)/(\mu - \Lambda_1)$. Using this expression to solve (57)–(58), the welfare-maximizing throughput by type 1s is

$$\begin{aligned}\Lambda_{1\Omega}^p = \min\big(&\bar{\Lambda}_1, \mu - \bar{\Lambda}_2 - \sqrt{c\mu/r},\\ &\big(\mu - (1+\bar{\Lambda}_2/\mu)c/(\beta_1 r)\big)^+\big).\end{aligned} \tag{59}$$

Consider the following fee schedule for inducing the welfare-maximizing throughput:

$$\begin{aligned}&u^h = \big(\beta_1 r - c\,\mathrm{E}[W^h]\big)^+, \quad s^h = \bar{\lambda}(r - c\,\mathrm{E}[W^h] - u^h),\\ &u^l = r - c\,\mathrm{E}[W^l], \quad \text{and} \quad s^l = 0,\end{aligned} \tag{60}$$

where $\mathrm{E}[W^h]$ and $\mathrm{E}[W^l]$ are given by (23) with $\Lambda^h = \Lambda_{1\Omega}^p$ and $\Lambda^l = \bar{\Lambda}_2$. These fees are feasible if they are nonnegative. That $s^h > 0$ follows from the observations that if $\beta_1 r \ge c\,\mathrm{E}[W^h(0, \bar{\Lambda}_2)]$ or $\Lambda_{1\Omega}^p > 0$, then $u^h = \beta_1 r - c\,\mathrm{E}[W^h]$ and therefore $s^h = \bar{\lambda}(1-\beta_1)r > 0$. Otherwise, $\Lambda_{1\Omega}^p = 0$ and

$$s^h = \bar{\lambda}(r - c\,\mathrm{E}[W^h(0, \bar{\Lambda}_2)]) = \bar{\lambda}\left(r - c\frac{1+\bar{\Lambda}_2/\mu}{\mu}\right),$$

which is strictly positive because $\bar{\Lambda}_2 < \mu - \sqrt{c\mu/r}$ (which also implies that $r\mu > c$). That $u^l > 0$ follows from the observation that (59) implies $\Lambda_{1\Omega}^p < \mu - \bar{\Lambda}_2 - \sqrt{c\mu/r}$, and hence $\sqrt{r\mu/c} > 1$. Therefore, by (23),

$$\begin{aligned}\mathrm{E}[W^l(\Lambda_{1\Omega}^p, \bar{\Lambda}_2)] &= \frac{1}{\mu} + \frac{\Lambda_{1\Omega}^p + \bar{\Lambda}_2}{(\mu - \Lambda_{1\Omega}^p)(\mu - \Lambda_{1\Omega}^p - \bar{\Lambda}_2)}\\ &< \frac{1}{\mu} + \frac{\mu - \sqrt{c\mu/r}}{c\mu/r} \le \frac{r}{c}.\end{aligned}$$

Under (60),

$$\begin{aligned}&m_1^h = \Lambda_{1\Omega}^p/\bar{\lambda}, \quad m_2^l = m_2, \quad m_1^l = m_2^h = 0, \quad \Lambda_1^h = \Lambda_{1\Omega}^p,\\ &\Lambda_2^l = \bar{\Lambda}_2, \quad \text{and} \quad \Lambda_1^l = \Lambda_2^h = 0\end{aligned}$$

satisfy (24)–(26), and therefore constitute an equilibrium.

Now let us compare priority scheduling with FIFO under (21) and (27). Maximizing (57) subject to (58) with $\mathrm{E}[W_1^{\psi}] = 1/(\mu - \Lambda_1 - \bar{\Lambda}_2)$, the optimal throughput of type 1s under FIFO is

$$\Lambda_{1\Omega}^F = \min\big(\bar{\Lambda}_1, \mu - \bar{\Lambda}_2 - \sqrt{c\mu/r}, \big(\mu - \bar{\Lambda}_2 - c/(\beta_1 r)\big)^+\big). \tag{61}$$

Comparing (61) with (59), under (21) and (27),

$$\begin{aligned}\Lambda_{1\Omega}^F &= \big(\mu - \bar{\Lambda}_2 - c/(\beta_1 r)\big)^+\\ &< \min\big(\bar{\Lambda}_1, \mu - \bar{\Lambda}_2 - \sqrt{c\mu/r}, \big(\mu - (1+\bar{\Lambda}_2/\mu)c/(\beta_1 r)\big)^+\big)\\ &= \Lambda_{1\Omega}^p.\end{aligned} \tag{62}$$

Because social welfare

$$(\Lambda_1 + \bar{\Lambda}_2)r - \frac{c(\Lambda_1 + \bar{\Lambda}_2)}{\mu - \Lambda_1 - \bar{\Lambda}_2}$$

strictly increases with Λ_1 when $\Lambda_1 + \bar{\Lambda}_2 < \mu - \sqrt{c\mu/r}$, we conclude that priority scheduling strictly improves social welfare if (21) and (27) hold. Both conditions are necessary for welfare to improve under priority scheduling: socially optimal throughput is not achievable under FIFO with a nonnegative usage fee only if (21) holds, and type 1s have sufficient self-control to go for service only if (27) holds.

The maximum revenue when serving sophisticated types 1s and 2s is the maximum welfare in (57), which is achieved with priority scheduling and the fees in (60). Therefore, the optimal revenue under priority scheduling strictly exceeds that under FIFO under the same necessary and sufficient conditions as in the case of welfare maximization, i.e., (21) and (27).

If the usage fee can be negative, (58) becomes irrelevent, so instead of (59), optimal throughput is

$$\Lambda_{1\Omega}^p = \min\big(\bar{\Lambda}_1, \mu - \bar{\Lambda}_2 - \sqrt{c\mu/r}\big).$$

Similar to (60), the fee schedule

$$\begin{aligned}&u^h = \beta_1 r - c\,\mathrm{E}[W^h]^+, \quad s^h = \bar{\lambda}(r - c\,\mathrm{E}[W^h] - u^h) = \bar{\lambda}(1-\beta_1)r,\\ &u^l = r - c\,\mathrm{E}[W^l], \quad \text{and} \quad s^l = 0\end{aligned}$$

induces all type 1s subscribe for high priority service and all type 2s use low priority service. One may verify, using details from the proof of Proposition 6, that the resulting optimal welfare and revenue with sophisticated type 1s are the same as under FIFO.

Now consider revenue maximization with naive type 1s. By assumption, if any type 1s subscribe and/or use the service, then all type 2s use the service. Given that type 2 subscribers' throughput is $\bar{\Lambda}_2$, two situations may arise. In the first situation, $\beta_1 r < c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)] \le \hat{\beta}_1 r$, then $\Lambda_1 = 0$, and it is optimal to charge type 1s zero usage fee and a subscription fee of

$$s_1 = \bar{\lambda}\big(r - c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)]\big) \ge \bar{\lambda}(1-\hat{\beta}_1)r > 0$$

to exactly extract their expected utility. All type 1s subscribe, but none uses service. The maximum revenue is the following sum of type 1 subscribers' subscription payments and all utility of type 2s:

$$\bar{\Lambda}_1\big(r - c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)]\big) + \bar{\Lambda}_2\big(r - c\,\mathrm{E}[W_2^{\psi}(0, \bar{\Lambda}_2)]\big). \tag{63}$$

In the second situation, $\beta_1 r \ge c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)]$, so $\beta_1 r = c\,\mathrm{E}[W_1^{\psi}(\Lambda_1, \bar{\Lambda}_2)] + u_1$, where Λ_1 is type 1 subscribers' throughput in equilibrium. The manager optimally sets the subscription fee,

$$s_1 = \bar{\lambda}\big[r - c\,\mathrm{E}[W_1^{\psi}(\Lambda_1, \bar{\Lambda}_2)]\big] - u_1 = \bar{\lambda}(1-\beta_1)r > 0,$$

to exactly extract type 1 subscribers' expected utility. The resulting revenue is

$$\begin{aligned} & s_1 m_1 + u_1\Lambda_1 + \bar{\Lambda}_2\big(r - c\,\mathrm{E}[W_2^{\psi}(\Lambda_1, \bar{\Lambda}_2)]\big) \\ &\quad = \bar{\Lambda}_1(1-\beta_1)r + \Lambda_1\big(\beta_1 r - c\,\mathrm{E}[W_1^{\psi}(\Lambda_1, \bar{\Lambda}_2)]\big) \\ &\qquad + \bar{\Lambda}_2\big(r - c\,\mathrm{E}[W_2^{\psi}(\Lambda_1, \bar{\Lambda}_2)]\big). \end{aligned} \tag{64}$$

In both cases, all type 2 subscribers' utility can be extracted by charging a zero subscription fee and usage fee

$$u_2 = \big(r - c\,\mathrm{E}[W_2^{\psi}(\Lambda_1, \bar{\Lambda}_2)]\big) > 0.$$

Using Kleinrock's (1965) conservation law (47), the revenue optimization problems (63) and (64) simplify to

$$\begin{aligned} & \bar{\Lambda}_1\big(r - \max\big(\beta_1 r, c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)]\big)\big) + \bar{\Lambda}_2 r \\ &\quad + \max_{0\le\Lambda_1\le\bar{\Lambda}_1} \big\{\Lambda_1\beta_1 r - (\Lambda_1 + \bar{\Lambda}_2)c\,\mathrm{E}[W(\Lambda_1 + \bar{\Lambda}_2)]\big\}, \quad \text{where} \\ & \qquad \hat{\beta}_1 r \ge c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)] \quad \text{and} \quad \Lambda_1 = 0 \quad \text{if} \\ & \qquad\qquad \beta_1 r < c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)]. \end{aligned} \tag{65}$$

Following Lemma 6.5 of Gelenbe and Mitrani (1980), priority scheduling is optimal for (65) because it minimizes $\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)]$. Applying (23) to (65) with $\Lambda^h = 0$ and $\Lambda^l = \bar{\Lambda}_2$, we have

$$\begin{aligned} & \bar{\Lambda}_1\left(r - \max\left[\beta_1 r, c\,\frac{1 + \bar{\Lambda}_2/\mu}{\mu}\right]\right) + \bar{\Lambda}_2 r \\ &\quad + \max_{0\le\Lambda_1\le\bar{\Lambda}_1} \{\Lambda_1\beta_1 r - (\Lambda_1 + \bar{\Lambda}_2)c\,\mathrm{E}[W(\Lambda_1 + \bar{\Lambda}_2)]\}, \end{aligned} \tag{66}$$

subject to $\hat{\beta}_1 r \ge c(1 + \bar{\Lambda}_2/\mu)/\mu$. Applying $\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)] = 1/(\mu - \bar{\Lambda}_2)$ to (65) for FIFO, we have

$$\begin{aligned} & \bar{\Lambda}_1\left(r - \max\left[\beta_1 r, \frac{c}{\mu - \bar{\Lambda}_2}\right]\right) + \bar{\Lambda}_2 r \\ &\quad + \max_{0\le\Lambda_1\le\bar{\Lambda}_1} \big\{\Lambda_1\beta_1 r - (\Lambda_1 + \bar{\Lambda}_2)c\,\mathrm{E}[W(\Lambda_1 + \bar{\Lambda}_2)]\big\}, \end{aligned} \tag{67}$$

subject to $\hat{\beta}_1 r \ge c/(\mu - \bar{\Lambda}_2)$. It is easy to verify that (66) is strictly greater than (67) if and only if (22) holds. To establish that (66) is the maximum achievable revenue under priority scheduling, observe that type 1s are willing to subscribe for high priority service, because they anticipate using it, if and only if (28) holds. Applying the same argument in the proof of Proposition 6 for the naive type 1s case, we conclude that (66) is strictly greater than the optimal revenue when serving only type 2s.

If the usage fee can be negative, then in case $\beta_1 r < c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)]$, the manager achieves higher revenue by setting

$$\begin{aligned} & u_1 = \beta_1 r - c\,\mathrm{E}[W_1^{\psi}(0, \bar{\Lambda}_2)], \quad s_1 = \bar{\lambda}(1-\beta_1)r, \\ & u_2 = r - c\,\mathrm{E}[W_2^{\psi}(0, \bar{\Lambda}_2)], \quad \text{and} \quad s_2 = 0. \end{aligned}$$

All type 1s subscribe for high priority service, and all type 2s use low priority service. Instead of (65), the optimal revenue becomes

$$\bar{\Lambda}_1(1-\beta_1)r + \bar{\Lambda}_2 r + \max_{0\le\Lambda_1\le\bar{\Lambda}_1} \big\{\Lambda_1\beta_1 r - (\Lambda_1 + \bar{\Lambda}_2)c\,\mathrm{E}[W(\Lambda_1 + \bar{\Lambda}_2)]\big\},$$

which is the same as the optimal revenue under FIFO scheduling given by (56). □

References

Ainslie G (1992) *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person* (Cambridge University Press, Cambridge, UK).

Angeletos GM, Laibson D, Repetto A, Tobacman J, Weinberg S (2001) The hyperbolic consumption model: Calibration, simulation, and empirical evaluation. *J. Econom. Perspect.* 15(3):47–68.

Ault T (2010) Interview by Erica L. Plambeck with proprietor of Fort Car Wash, Fort Atkinson, Wisconsin, November 11 and 29.

Ausubel LM, Shui H (2005) Time inconsistency in the credit card market. Working paper, University of Maryland, College Park.

Bendoly E, Donohue K, Schultz K (2006) Behavior in operations management: Assessing recent findings and revisiting old assumptions. *J. Oper. Management* 24(6):737–752.

Bernheim BD, Rangel A (2005) Behavioral public economics: Welfare and public policy analysis with non-standard decision-makers. NBER Working Paper 11518, National Bureau of Economic Research, Cambridge, MA.

Bitran GR, Rocha e Oliveira P, Schilkrut A (2008) Managing customer relationships through price and service quality. Working paper, IESE Business School, University of Navarra, Madrid.

Cachon G, Feldman P (2011) Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing Service Oper. Management* 13(2):244–260.

China Commodity Net (2009) Analysis of the critical path for developing the hair salon industry. Accessed October 2009, http://ccn.mofcom.gov.cn/spbg/show.php?id=9820&ids=.

Choi JJ, Laibson D, Madrian BC (2011) $100 bills on the sidewalk: Suboptimal saving in 401(k) plans. *Rev. Econom. Statist.* 93(3):748–763.

DellaVigna S, Malmendier U (2006) Paying not to go to the gym. *Amer. Econom. Rev.* 96(3):694–718.

Duflo E, Kremer M, Robinson J (2009) Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya. NBER Working Paper 15131, National Bureau of Economic Research, Cambridge, MA.

Fang H, Silverman D (2009) Time-inconsistency and welfare program participation: Evidence from the NILSY. *Internat. Econom. Rev.* 50(4):1043–1077.

Frederick S, Loewenstein G, O'Donoghue T (2002) Time discounting and time preference: A critical review. *J. Econom. Literature* 40(2):351–401.

Gelenbe E, Mitrani I (1980) *Analysis and Synthesis of Computer Systems* (Academic Press, New York).

Gino F, Pisano G (2008) Toward a theory of behavioral operations. *Manufacturing Service Oper. Management* 10(4):676–691.

Giridharan PS, Mendelson H (1994) Free-access policy for internal networks. *Inform. Systems Res.* 5(1):1–21.

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

Hurn CG (2011) Interview by Erica L. Plambeck with proprietor of Kennedy's All-American Barber Club, Orlando, Florida, January 7.

Johari R, Kumar S (2010) Congestible services and network effects. *Proc. 11th ACM Conf. on Electronic Commerce* (ACM, New York).

Kleinrock L (1965) A conservation law for a wide class of queueing disciplines. *Naval Res. Logist.* 12(2):181–192.

Laibson D (1997) Golden eggs and hyperbolic discounting. *Quart. J. Econom.* 112(2):443–477.

Laibson D, Repetto A, Tobacman J (2007) Estimating discount functions with consumption choices over the lifecycle. NBER Working Paper 13314, National Bureau of Economic Research, Cambridge, MA.

Loch C, Wu Y (2007) Behavioral operations management. *Foundations and Trends Tech., Inform. Oper. Management* 1(3): 122–232.

Madrian BC, Shea DF (2001) The power of suggestion: Intertia in 401(k) participation and savings behavior. *Quart. J. Econom.* 116(4):297–310.

McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. *Science* 306(5695):503–507.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* 38(5):870–883.

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

O'Donoghue T, Rabin M (1999) Doing it now or later. *Amer. Econom. Rev.* 89:103–124.

O'Donoghue T, Rabin M (2001) Choice and procrastination. *Quart. J. Econom.* 116(1):121–160.

Paserman MD (2008) Job search and hyperbolic discounting: Structural estimation and policy evaluation. *Econom. J.* 118(531):1418–1452.

Plambeck EL, Wang Q (2012) Hyperbolic discounting and queue-length information management for unpleasant services that generate future benefits. Working paper, Stanford Graduate School of Business, Stanford, CA.

Prelec D, Loewenstein G (1998) The red and the black: Mental accounting of savings and debt. *Marketing Sci.* 17(1):4–28.

Randhawa R, Kumar S (2008) Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing Service Oper. Management* 10(3):429–447.

Skiba P, Tobacman J (2008) Payday loans, uncertainty and discounting: Explaining patterns of borrowing, repayment, and default. Vanderbilt Law and Economics Research Paper 08-33, Vanderbilt Law School, Nashville, TN.

Su X (2009) A model of consumer inertia with applications to dynamic pricing. *Production Oper. Management* 18(4):365–380.

Ulku S, Dimofte CV, Schmidt GM (2009) Customer valuation of product modularity. Working paper, Georgetown University, Washington, DC.

Westland S (1992) Congestion and network externalities in the short run pricing of information system services. *Management Sci.* 38(7):992–1009.

Wu Y, Ramachandran K, Krishnan V (2009) Managing projects with present-biased agents. Working paper, National University of Singapore, Singapore.