MANAGEMENT SCIENCE
Vol. 50, No. 7, July 2004, pp. 869–882
ISSN 0025-1909 | EISSN 1526-5501 | 04 | 5007 | 0869

informs

DOI 10.1287/mnsc.1030.0156

Pricing and Priority Auctions in Queueing Systems with a Generalized Delay Cost Structure

Philipp Afèche
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208-2009,
p-afeche@kellogg.northwestern.edu

Haim Mendelson
Graduate School of Business, Stanford University, Stanford, California 94305, haim@stanford.edu

This paper studies alternative price-service mechanisms for a provider that serves customers whose delay cost depends on their service valuations. We propose a generalized delay cost structure that augments the standard additive model with a multiplicative component, capturing the interdependence between delay cost and values. We derive and compare the revenue-maximizing and socially optimal equilibria under uniform pricing, preemptive, and nonpreemptive priority auctions with an admission price. We find that the delay cost structure has a paramount effect on system behavior. The classical result that the revenue-maximizing admission price is higher and the utilization lower than is socially optimal can be reversed under our generalized structure, and we identify the conditions driving this reversal under each mechanism. We show that the conditional bid equilibria are unique and induce the socially optimal allocations. The auctions yield gains in system net value and provider profit over uniform pricing, which are dramatically larger for the preemptive mechanism. Both auctions perform better under multiplicative compared to additive delay costs. The highest-value customers always gain under the preemptive, but may lose under the nonpreemptive auction. The lowest-value customers always gain in either auction.

Key words: auctions; congestion; delay cost; incentive compatibility; pricing; priority; queueing; quality of service; revenue management; scheduling; service differentiation
History: Accepted by Paul Glasserman, stochastic models and simulation; received November 26, 2001. This paper was with the authors 2 months for 2 revisions.

1. Introduction

Delay is a key dimension of service quality, playing an important role whenever a capacity-constrained provider faces delay-sensitive customers. Service examples include call centers, data communications, transportation, and government services. Delay is also important in manufacturing, particularly in make-to-order operations. Naor's (1969) classical analysis of resource allocation and pricing when congestion determines service quality spawned a sizable literature on the topic. We focus on two aspects that have received limited attention to date, the delay cost structure and the use of auctions to allocate priorities. We propose a generalized delay cost structure where, unlike in the standard model, the delay cost is intertwined with the value of service. We show that, in this case, revenue maximization may yield a lower price and higher utilization than social optimization, contrary to the standard results, and we identify for alternative price-service mechanisms (uniform pricing, preemptive or nonpreemptive priority auctions) the conditions for this reversal. We further show how the delay cost structure and the priority discipline affect the benefits of using auctions versus uniform pricing. We highlight our model and results below.

Generalized Delay Cost Structure. Following Naor (1969), customer utility u has been modeled as additive in value and delay cost: $u(v, t) = v - C(t)$, where v is the service value absent delay, t is the delay and $C(t)$ is the delay cost, a nondecreasing (often linear) function of t (cf., Yechiali 1972, De Vany 1976, Knudsen 1972, Lippman and Stidham 1977, Mendelson 1985, Westland 1992, Mendelson and Whang 1990, Loch 1991, Kalai et al. 1992, Li and Lee 1994, Lederer and Li 1997, Masuda and Whang 1999, Cachon and Harker 2002). Even in studies that allow for nonlinear $C(t)$, value and delay cost are additive (Dewan and Mendelson 1990, Van Mieghem 2000). This model, where the delay cost is independent of a job's value v, captures customers' productivity loss, but fails to address situations where the delay cost is intertwined with the value of service.

For example, in financial or industrial markets, a delay in the execution of a trade deflates the investor's expected profit if part or all of the anticipated price change occurs prior to execution. Thus, the delay cost is proportional to the transaction value in the absence of delay. Examples of a similar cost structure include the quality deterioration of video in data communications and the depreciation of

perishable or short life-cycle goods during delivery. To capture the interaction between delay cost and values, we model customer utility as $u(v, t) = v \cdot D(t) - C(t)$, where $D(t) \leq 1$ is decreasing in t and deflates the value v. This *generalized delay cost structure* encompasses both value decay and productivity losses. We show that the delay cost structure has a paramount effect on system behavior.

Priority Auctions. When customers have different valuations and delay costs, uniform pricing is suboptimal. Virtually all studies of price and service differentiation in queueing systems analyze *centralized* pricing, where the provider chooses a price-service menu (cf., Kleinrock 1967, Dolan 1978, Mendelson and Whang 1990, Rao and Petersen 1998, Van Mieghem 2000, Ha 2001, Gupta et al. 1996). In settings with many different customers whose valuations are not known to the provider, it may be beneficial to use auctions, where the provider allocates service resources based on customers' bids. While the auction literature is extensive (cf., Klemperer 1999), few studies address priority auctions in queueing systems. Balachandran (1972) derives the equilibrium bidding strategy for identical customers,[1] and Lui (1985), Hassin (1995), and Glazer and Hassin (1985) derive them when heterogeneous customers can bid for preemptive priorities under additive delay costs. We analyze both preemptive (*AP*) and nonpreemptive (*AN*) auctions under our generalized delay cost structure and study their benefits compared to uniform pricing.

Summary of Results. Priority auctions yield a lower admission price, a higher utilization, and gains under both social optimization and revenue maximization compared to uniform pricing. The percentage gains are dramatically larger under *AP* compared to *AN*, while both mechanisms perform better under multiplicative compared to additive delay costs. Compared to uniform pricing, the highest-value customers always gain under *AP* but may lose under *AN*, while in both auctions, those with the lowest values always gain and those in between may gain or lose.

We show that the standard results comparing revenue maximization and social optimization can be reversed under our generalized delay cost structure. Naor (1969) showed that revenue maximization yields a higher price and lower utilization than social optimization, similar to the classical result from economics. The same holds for different models (cf., De Vany 1976, Mendelson 1985), but we show that all these results hinge on the assumed additive delay cost structure and may be reversed when value and delay cost are intertwined. For uniform pricing and for both *AP* and *AN* auctions, the comparison between revenue maximization and social optimization depends on both the delay cost structure and the elasticity of demand. For example, when demand has constant elasticity and the delay cost is multiplicative in service values, *revenue maximization is socially optimal* under uniform pricing and priority auctions.

In what follows, §2 presents the basic model and §3 analyzes the uniform pricing mechanism. Section 4 analyzes priority auctions and §5 compares *AP*, *AN*, and uniform pricing. Section 6 offers our concluding remarks. All proofs are in the appendix.

2. Model

We consider a capacity-constrained firm, modeled as a single-server queueing system that serves delay-sensitive customers. Service times are i.i.d. with unit mean. For simplicity of exposition, we assume that the marginal cost of serving a customer is zero. Customers arrive following an exogeneous renewal process (independent of service times) with rate or market size Λ ($\Lambda \geq 1$). In this section, we consider the *uniform price mechanism*, whereby the firm charges all customers the same price P and serves them in first-in-first-out (FIFO) order. Upon arrival, customers decide whether or not to request service at the posted price P, and they cannot renege.

Customer Values. Customers are infinitesimal relative to the market size. They differ in their service values v, i.e., their willingness to pay for service in the absence of delay. Values are i.i.d. draws from a continuous distribution Φ (independent of arrival and service times) with p.d.f. ϕ, assumed strictly positive and continuous on $[\underline{v}, \overline{v}]$, where $0 \leq \underline{v} < \overline{v} \leq \infty$. Let $\overline{\Phi} = 1 - \Phi$. If all jobs with values $\geq v$ join the system, the arrival (or demand) rate will be $\lambda = \Lambda\overline{\Phi}(v)$. Conversely, when the arrival rate is λ, the marginal value v is equal to $\overline{\Phi}^{-1}(\lambda/\Lambda)$, where $\overline{\Phi}^{-1}$ is the inverse of $\overline{\Phi}$. Let $V(\lambda)$ denote the expected aggregate (gross) value generated by the system per unit time. Then, it follows that the downward-sloping *marginal value* (or inverse gross demand) *function* $V'(\lambda) := \overline{\Phi}^{-1}(\lambda/\Lambda)$ defines a one-to-one mapping between the demand rate λ ($\lambda < \Lambda$) and the marginal value $V'(\lambda)$ (cf., Lippman and Stidham 1977, Mendelson 1985). Hence, $V'(\lambda) > 0$ and $V''(\lambda) < 0$ for $\lambda < \Lambda$. Define the gross revenue function $R(\lambda) := \lambda \cdot V'(\lambda)$, assumed continuously differentiable and strictly concave. It measures the revenue that the provider could collect in the absence of delay if each user were charged the same amount equal to the marginal value $V'(\lambda)$.

Generalized Delay Cost Structure. The key innovation in this section is in the *structure* of delay costs. In the literature following Naor (1969), the utility of a customer with value v who pays P and experiences a delay t (between her service request and

[1] In his model, customers can observe the queue length.

its completion) has typically been assumed *additive*, $u(v, t, P) = v - C(t) - P$, where $C(t)$ is an increasing delay cost function satisfying $C(0) = 0$. In particular, a job's delay cost was assumed independent of its value. We propose a generalized structure that allows delay costs and values to be interdependent through a multiplicative term

$$u(v, t, P) = v \cdot D(t) - C(t) - P. \tag{1}$$

The *delay discount function* $D(t)$ is nonincreasing and satisfies $D(0) = 1$ and $\lim_{t\to\infty} D(t) \le 0$.

Information Structure. The arrival process, service time and value distributions, and the functions C and D are common knowledge. Actual service times are not known ex ante. Each customer's value is her private information and is observed neither by the provider nor by other customers, and arriving customers cannot observe the system state.

Expected Utility. Each customer is self-interested and maximizes her own expected utility, which she forecasts using the distribution of the steady-state delay $\widetilde{W}(\lambda)$. It depends on the set of paying customers only through the resulting aggregate demand rate λ ($\lambda < 1$), and is not affected by the actions of an individual customer. Let $\overline{D}(\lambda) := E[D(\widetilde{W}(\lambda))]$ and $\overline{C}(\lambda) := E[C(\widetilde{W}(\lambda))]$ be the expected delay discount and delay cost functions, respectively. Given λ, a customer with value v who pays P has expected net value $v \cdot \overline{D}(\lambda) - \overline{C}(\lambda)$ and expected utility $u(v \mid P, \lambda) := v \cdot \overline{D}(\lambda) - \overline{C}(\lambda) - P$.

Assumption 1. *To allow for nontrivial equilibria, we require* $R'(0) \cdot \overline{D}(0) - \overline{C}(0) > 0$ *and* $\overline{D}(0) > 0$.

Assumption 2. *The function* $\overline{D}$ ($\overline{C}$) *is strictly decreasing* (*increasing*, *respectively*), *twice continuously differentiable, and concave* (*convex*, *respectively*) *in* λ.

Assumption 3. *Expected net values are nonpositive at full utilization*: $\lim_{\lambda\to 1} \overline{D}(\lambda) \le 0$.

Discussion. One way to visualize the model is to consider a retail setting where a firm serves a large pool of small, "atomistic" customers who arrive at random. They are indistinguishable to the provider who, thus, considers their values and service times as random samples from a given continuous distribution Φ and a given service time distribution (cf., Mendelson 1985, Dewan and Mendelson 1990, Westland 1992). The actions of any infinitesimal customer do not affect the distribution of delay in the system. This model does not fit a setting with a small number of large customers, each with her own demand curve and service time distribution, and so large that her individual decisions significantly affect the system's delay distribution. Similarly, our i.i.d. assumptions mean that a given customer is either served infrequently or cannot be individually tracked by the provider.[2]

The *net* value of a customer with value v who experiences delay t is $v \cdot D(t) - C(t)$, and her *delay cost* is $v \cdot [1 - D(t)] + C(t)$. The additive term $C(t)$ captures a cost driven by time but not value, e.g., the productivity loss of a customer waiting for a process to end. The multiplicative term $v \cdot [1 - D(t)]$ is well suited when delay deflates values. A variety of important phenomena lead to delay-driven value losses.

(i) *Delayed information*: A delay in the receipt or use of information adversely affects its value. For example, a delay in the execution of an order to trade a stock deflates the trader's expected profit if part or all of the anticipated price change occurs prior to execution (cf., Dewan and Mendelson 1998). The increased price volatility, the rise of electronic markets, and dynamic pricing extend the relevance of this scenario to many industrial markets.

(ii) *Physical decay*, e.g., during transportation delays for perishable goods such as produce.

(iii) *Audio or video signal distortions* caused by delays in real-time transmissions over the Internet.

(iv) *Technological obsolescence* of short life-cycle products such as computer chips.

Examples. The following examples illustrate the building blocks of our model:

Customer values: (1) If values are uniformly distributed on $[\underline{v}, \overline{v}]$, then $V'(\lambda)$ is linear: $V'(\lambda) = \overline{v} - (\lambda/\Lambda)(\overline{v} - \underline{v})$. (2) If values are Pareto distributed, i.e., $\phi(v) = a\underline{v}^a v^{-(1+a)}$ for $v \ge \underline{v} > 0$ and $a > 1$, then $V'(\lambda)$ is an isoelastic or power function: $V'(\lambda) = \underline{v}(\lambda/\Lambda)^{-1/a}$.

Delay cost: (1) Linear delay costs: $D(t) = 1 - d \cdot t$ with $d > 0$ and $C(t) = c \cdot t$ with $c > 0$ yield $\overline{D}(\lambda) = 1 - d \cdot \overline{W}(\lambda)$ and $\overline{C}(\lambda) = c \cdot \overline{W}(\lambda)$, where $\overline{W}(\lambda)$ is the mean delay. These satisfy Assumptions 1–3 if $d < 1$ and $R'(0) \cdot (1 - d) > c$. This commonly used structure assumes that customers are risk neutral and their instantaneous delay sensitivity is independent of their elapsed time in the system. (2) Exponential delay costs: $D(t) = 1 - K_1(e^{dt} - 1)$ and $C(t) = K_2(e^{ct} - 1)$. This structure allows for both risk aversion ($K_1, K_2, c, d > 0$) and increasing impatience ($K_1, K_2, c, d < 0$), and makes the marginal delay cost dependent on the delay experienced so far. For $K_1 = -1$, $d < 0$, $K_2 = 1$, and $c > 0$, $\overline{D}(\lambda) = W^*(-d; \lambda)$ and $\overline{C}(\lambda) = W^*(-c; \lambda) - 1$, where $W^*(x; \lambda)$ is the Laplace transform of the delay given λ, and Assumptions 1–3 hold if $R'(0) \cdot W^*(-d; 0) > W^*(-c; 0) - 1$.

3. Uniform Pricing

We compare the socially optimal and revenue-maximizing equilibrium admission prices and demand rates.

[2] Similar assumptions are implicit in most queueing models.

We show that under the generalized delay cost structure, *the revenue-maximizing* (*or monopoly*) *price may be at or below the socially optimal level*.

3.1. Equilibrium for Fixed P

For given P, there is an unique Nash equilibrium, with demand rate $\lambda(P) < \Lambda$, if the highest-value customer has positive expected utility at zero utilization ($V'(0) \cdot \bar{D}(0) - \bar{C}(0) > P$).[3] The marginal customer has value $V'(\lambda(P))$ and zero expected utility; that is, $\lambda(P)$ is the unique solution to

$$u(V'(\lambda) \mid P, \lambda) = V'(\lambda) \cdot \bar{D}(\lambda) - \bar{C}(\lambda) - P = 0. \quad (2)$$

Customers join if, and only if, their values exceed $V'(\lambda(P))$. The inverse demand function $P(\lambda) := V'(\lambda) \cdot \bar{D}(\lambda) - \bar{C}(\lambda)$ is well defined ($\lambda(P)$ is continuous, strictly decreasing) and maps the equilibrium demand rate λ to the admission price $P(\lambda)$. A customer with value $v \geq V'(\lambda)$ has expected net value $v \cdot \bar{D}(\lambda) - \bar{C}(\lambda)$.

3.2. Social Optimization vs. Revenue Maximization

Social Optimization. Let $AV(\lambda) := V(\lambda)/\lambda$ be the system's expected average value and $NV(\lambda)$ be its expected aggregate net value (net of delay costs) per unit time. For social optimization, the provider solves[4]

$$\max_{\lambda \in [0, 1)} NV(\lambda) = V(\lambda) \cdot \bar{D}(\lambda) - \lambda \cdot \bar{C}(\lambda) = \lambda \cdot [AV(\lambda) \cdot \bar{D}(\lambda) - \bar{C}(\lambda)]. \quad (3)$$

Assumptions 1–3 imply that (3) has a unique solution that satisfies the first-order condition

$$P(\lambda) = V'(\lambda) \cdot \bar{D}(\lambda) - \bar{C}(\lambda) = \lambda \cdot [-AV(\lambda) \cdot \bar{D}'(\lambda) + \bar{C}'(\lambda)]. \quad (4)$$

At the socially optimal (or efficient) rate λ^*, the marginal customer's expected net value equals the *net value externality*,[5] i.e., the expected net value loss she inflicts on the system. A marginal increase in λ raises the average value discount by $-AV(\lambda) \cdot \bar{D}'(\lambda)$ and the average additive delay cost by $\bar{C}'(\lambda)$. The aggregate external effect equals $\lambda \cdot [-AV(\lambda) \cdot \bar{D}'(\lambda) + \bar{C}'(\lambda)]$. It follows that if the price is set equal to the externality, then the Nash equilibrium is socially optimal.

[3] This follows because the expected utility $u(v \mid P, \lambda)$ is continuous and strictly increasing in v, and because $u(V'(\lambda) \mid P, \lambda)$ is continuous and strictly decreasing in λ with $u(V'(0) \mid P, 0) > 0$ and $\lim_{\lambda \to 1} u(V'(\lambda) \mid P, \lambda) < 0$.

[4] For convenience, we perform the analysis in terms of the demand rate λ.

[5] An externality is a cost or benefit of an action that affects parties other than the one taking it.

Revenue Maximization. The provider maximizes expected revenue, denoted by $\Pi(\lambda)$,

$$\max_{\lambda \in [0, 1)} \Pi(\lambda) = \lambda \cdot P(\lambda) = \lambda \cdot [V'(\lambda) \cdot \bar{D}(\lambda) - \bar{C}(\lambda)]. \quad (5)$$

Assumptions 1–3 imply that (5) has a unique solution that satisfies the first-order condition

$$R'(\lambda) \cdot \bar{D}(\lambda) - \bar{C}(\lambda) = \lambda \cdot [-V'(\lambda) \cdot \bar{D}'(\lambda) + \bar{C}'(\lambda)]. \quad (6)$$

At the revenue-maximizing rate λ^M, the expected marginal net revenue under constant delay, the left-hand side of (6), equals the marginal delay-induced expected revenue loss per unit time. This revenue externality equals λ, multiplied by the marginal customer's expected net value loss. By (3) and (5), the expected aggregate net value $NV(\lambda)$ equals λ multiplied by the expected *average* customer net value, while the expected aggregate revenue $\Pi(\lambda)$ equals λ multiplied by the expected net value of the *marginal* customer. The difference between the efficient and revenue-maximizing equilibria depends on how the delay cost of the average customer compares to that of the marginal customer, which, in turn, depends on the delay cost structure. Table 1 summarizes the marginal customer's net benefit and the externality she imposes on the system under the additive and multiplicative delay cost structures.

Additive Delay Cost ($\bar{D}(\lambda) \equiv 1$)**.** Because $V'(\lambda) > R'(\lambda)$, the marginal customer contributes more to the aggregate net value than to revenues. On the other hand, all customers' expected delay cost, $\bar{C}(\lambda)$, is the *same* and the associated net value and revenue externalities both equal $\lambda \cdot \bar{C}'(\lambda)$. Hence, $NV(\lambda)$ is steeper than $\Pi(\lambda)$, and the revenue-maximizing price (demand) is higher (lower) than socially optimal, leading to surplus losses. This result (cf., Naor 1969, De Vany 1976, Mendelson 1985) follows the standard intuition, whereby a monopoly sets a higher price and lower consumption compared to efficient levels.

Multiplicative Delay Cost ($\bar{C}(\lambda) \equiv 0$)**.** The marginal customer contributes $V'(\lambda) \cdot \bar{D}(\lambda)$ to the aggregate net value and $R'(\lambda) \cdot \bar{D}(\lambda)$ to the aggregate

Table 1 Social Optimization vs. Revenue Maximization

Delay cost	Objective	Marginal net benefit	Externality
Additive	Social optimization	$V'(\lambda) - \bar{C}(\lambda)$	$\lambda \cdot \bar{C}'(\lambda)$
	Revenue maximization	$R'(\lambda) - \bar{C}(\lambda)$	$\lambda \cdot \bar{C}'(\lambda)$
Multiplicative	Social optimization	$V'(\lambda) \cdot \bar{D}(\lambda)$	$-\lambda \cdot AV(\lambda) \cdot \bar{D}'(\lambda)$
	Revenue maximization	$R'(\lambda) \cdot \bar{D}(\lambda)$	$-\lambda \cdot V'(\lambda) \cdot \bar{D}'(\lambda)$

Note. Marginal net benefit and externality for additive and multiplicative delay cost.

revenue. Unlike the additive case, here the average delay cost *exceeds* the marginal customer's delay cost, because the average value $AV(\lambda)$ exceeds the marginal value $V'(\lambda)$. Hence, the net value externality is larger than the revenue externality. Therefore, *under revenue maximization*, the marginal customer not only contributes less but also imposes a smaller externality on the system *than under social optimization*. To see which effect dominates, compare the respective *ratios*

$$\frac{R'(\lambda)\cdot\bar{D}(\lambda)}{-R(\lambda)\cdot\bar{D}'(\lambda)}-\frac{V'(\lambda)\cdot\bar{D}(\lambda)}{-V(\lambda)\cdot\bar{D}'(\lambda)}=\frac{r(\lambda)-v(\lambda)}{\lambda}\cdot\frac{\bar{D}(\lambda)}{-\bar{D}'(\lambda)}, \tag{7}$$

where $r(\lambda) := R'(\lambda)/V'(\lambda)$ is the ratio of marginal to average (gross) revenue (average gross revenue equals marginal value) and $v(\lambda) := V'(\lambda)/AV(\lambda)$ is the ratio of marginal to average value. The marginal benefit to delay cost ratio is *proportional* to $r(\lambda)$ under revenue maximization and to $v(\lambda)$ under social optimization. The following equivalence relationship holds:

$$\begin{aligned} r(\lambda^*) > v(\lambda^*) \quad &\Leftrightarrow \quad \Pi'(\lambda^*) > NV'(\lambda^*) = 0 \\ &\Leftrightarrow \quad \lambda^M > \lambda^*. \end{aligned} \tag{8}$$

Under social optimization, the marginal multiplicative benefit to delay cost ratio *equals* one at the efficient rate λ^* because $NV'(\lambda^*) = 0$ and $\bar{C}(\lambda) \equiv 0$. By (7), the respective ratio under revenue maximization *exceeds* one if $r(\lambda^*) > v(\lambda^*)$, i.e., the marginal customer at rate λ^* is "more beneficial" for revenue maximization than for social optimization. Hence, the slope of $\Pi(\lambda)$ is positive at λ^*, and $\lambda^M > \lambda^*$ because $\Pi(\lambda)$ is strictly concave. Similar arguments apply if $r(\lambda^*) \le v(\lambda^*)$. The *slope* of $r(\lambda)$ determines the *sign* of the difference $r(\lambda) - v(\lambda)$:

$$r(\lambda) \text{ strictly increases} \quad \Rightarrow \quad r(\lambda) - v(\lambda) > 0 \quad \text{for all } \lambda. \tag{9}$$

An increasing $r(\lambda)$ characterizes a $V'(\lambda)$ that gets "relatively flatter" as λ increases, i.e., the percentage drop in marginal value $V''(\lambda)/V'(\lambda)$ gets smaller as λ increases, relative to the percentage increase in quantity $1/\lambda$. Because the average $AV(\lambda)$ remains relatively large (it includes high values that correspond to small λ), $r(\lambda)$ exceeds the ratio of marginal to average value $v(\lambda)$. The ratio $r(\lambda)$ is related to the price elasticity of $V'(\lambda)$, $\epsilon(\lambda) := V'(\lambda)/(-V''(\lambda)\cdot\lambda)$, which is the percentage increase in quantity relative to a percentage drop in marginal value. If $V'(\lambda)$ is *isoelastic*, these changes are proportional. We have

$$r(\lambda) = \frac{R'(\lambda)}{V'(\lambda)} = 1 - \frac{1}{\epsilon(\lambda)}. \tag{10}$$

The marginal (gross) revenue at λ is positive if, and only if, $V'(\lambda)$ is elastic at λ, i.e., if $\epsilon(\lambda) > 1$. By (10), $r(\lambda)$ is increasing in $\epsilon(\lambda)$, and their slopes have the *same sign*. Hence, (8)–(10) imply

$$\epsilon(\lambda) \text{ strictly increases} \quad \Rightarrow \quad \lambda^M > \lambda^*. \tag{11}$$

Similar arguments apply to nonincreasing $r(\lambda)$. In summary, the slope of $\epsilon(\lambda)$ determines the difference between the marginal benefit to delay cost ratios under revenue maximization and social optimization and, in turn, the slope of the revenue function at the socially optimal rate λ^*.

Generalized Delay Cost Structure ($\bar{C}'(\lambda) > 0$, $\bar{D}'(\lambda) < 0$)**.** Introducing an additive delay cost $\bar{C}(\lambda)$ "pushes" the system equilibria toward the standard result $\lambda^M < \lambda^*$. Because $\bar{C}(\lambda)$ is the same for all customers, it reduces the relative difference between the average delay cost and that of the marginal customer. How λ^M compares to λ^* depends on the magnitude of the additive, relative to the multiplicative, delay cost. Proposition 1 summarizes the results.

Proposition 1. *Under Assumptions* 1–3, *revenue maximization* (M) *and social optimization* ($*$) *compare as follows. Let* $\bar{C}(\lambda; c) := c\cdot\bar{C}(\lambda)$, *where* $c \ge 0$ *is a scale parameter and* $c = 0$ *in the multiplicative case.*

(1) *If* $\epsilon(\lambda)$ *is strictly increasing in* λ, *then for any function* $\bar{D}$, *there is a unique* $c^* > 0$ *for which revenue maximization is socially optimal. For* $c \in [0, c^*)$, *the revenue-maximizing price* (*demand rate*) *is lower* (*higher*) *than socially optimal, and vice versa for* $c > c^*$.

(2) *If* $\epsilon(\lambda)$ *is constant in* λ, *then revenue maximization is socially optimal for* $c = 0$. *For* $c > 0$, *the revenue-maximizing price* (*demand rate*) *is higher* (*lower*) *than socially optimal.*

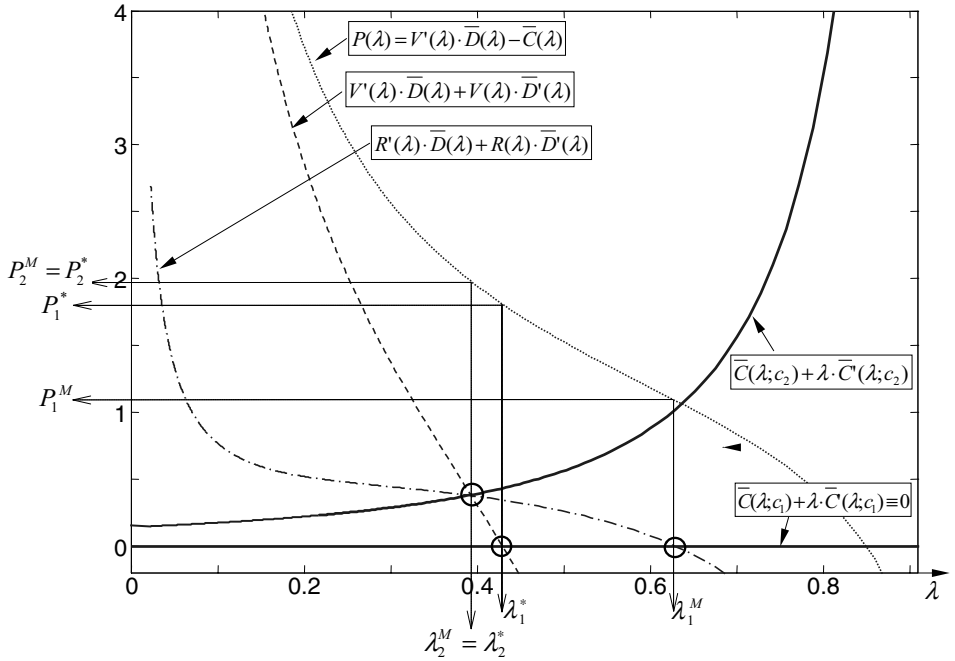
(3) *If* $\epsilon(\lambda)$ *is strictly decreasing in* λ, *the revenue-maximizing price* (*demand*) *is higher* (*lower*) *than socially optimal.*

It is important to emphasize that, unlike other cases in the literature, here the revenue-maximizing provider *never captures all surplus, even when it sets the socially optimal price.*

Example for Case 1 of Proposition 1. Consider an $M/M/1$ system. The market has two unit size segments ($\Lambda = 2$), each with a different isoelastic marginal value function $V_i'(\lambda_i) = \lambda_i^{-\alpha_i}$ with elasticity $\epsilon_i(\lambda) = 1/\alpha_i$, where $\alpha_1 = 0.3$ and $\alpha_2 = 0.9$.[6] Hence, V_1' is more elastic than V_2'. The total demand rate as a function of the marginal value v is $\lambda(v) = v^{-(1/\alpha_1)} +$

[6] That is, values in segment i exceed 1 and are Pareto distributed with mean equal to $1/(1-\alpha_i)$.

Figure 1 Revenue-Maximizing and Socially Optimal Equilibria for the $M/M/1$ System



Note. $V'(\lambda)$ is the inverse of $\lambda(v)=v^{-(10/3)}+v^{-(10/9)}$, $R(\lambda)=\lambda V'(\lambda)$, $\bar{D}(\lambda)=(1-\lambda)/(1-\lambda+0.1)$, and $\bar{C}(\lambda;c)=c/(1-\lambda)$ for $c_1=0$ and $c_2=c^*=0.14$.

$v^{-(1/\alpha_2)}$. It is straightforward to verify that its inverse, $V'(\lambda)$, has strictly increasing elasticity $\epsilon(\lambda)$. The function $D(t)$ is exponential: $D(t)=e^{d\cdot t}$, where $d=-0.1$ is the delay discount rate and $\bar{D}(\lambda)=(1-\lambda)/(1-\lambda-d)$. The additive delay cost function is linear: $C(t;c)=c\cdot t$ and $\bar{C}(\lambda;c)=c/(1-\lambda)$. Figure 1 shows the equilibria in the purely multiplicative case ($c_1=0$) and with additive delay cost ($c_2=0.14$). The downward-sloping functions $V'(\lambda)\cdot\bar{D}(\lambda)+V(\lambda)\cdot\bar{D}'(\lambda)$ and $R'(\lambda)\cdot\bar{D}(\lambda)+R(\lambda)\cdot\bar{D}'(\lambda)$ capture the marginal multiplicative effects under social optimization and revenue maximization, respectively. They intersect at $\lambda=0.38$. Their difference equals

$$NV'(\lambda)-\Pi'(\lambda)=[V'(\lambda)-R'(\lambda)]\cdot\bar{D}(\lambda)+[V(\lambda)-R(\lambda)]\cdot\bar{D}'(\lambda). \quad (12)$$

For $\lambda<0.38$, $NV'(\lambda)>\Pi'(\lambda)$, because the negative difference $[V(\lambda)-R(\lambda)]\cdot\bar{D}'(\lambda)$ is small at low λ and is offset by the fact that $V'(\lambda)>R'(\lambda)$. The converse holds for $\lambda>0.38$. By (4) and (6), the socially optimal and revenue-maximizing rates are determined where the respective downward-sloping function intersects the marginal additive delay cost function $\bar{C}(\lambda;c)+\lambda\cdot\bar{C}'(\lambda;c)$. It is identically zero in the multiplicative case for $c=c_1=0$, where $\lambda_1^M>\lambda_1^*>0.38$ ($P_1^M<P_1^*$), and scales up with c, reducing the equilibrium rates. For $c<c^*=0.14$, we have $\lambda^M>\lambda^*$. The converse holds for $c>c^*$ and $\lambda_2^*=\lambda_2^M=0.38$ ($P_2^*=P_2^M=1.95$) for $c=c_2=c^*$.

4. Priority Auctions

Uniformly charging and serving customers who differ in their delay sensitivity is suboptimal. With many or a continuum of unobserved customer types, and when customer segments are not clearly defined, the provider may find it beneficial to *let customers bid* for service instead of posting a (menu of) price(s). We study two such auctions: one for preemptive priorities (AP) and one for nonpreemptive priorities (AN). Indeed, preemption can reduce the delay cost but may be too costly, time consuming, or impair quality, e.g., in call centers, transportation services, or data transmissions. We thus study how the priority discipline affects the results.

The provider first sets an admission or reserve price $\underline{P}$ and announces how he will schedule customers based on their payments. Upon arrival, customers choose and pay their bids. They are served based on a static head-of-line priority rule. Under AP, a customer who pays $p>\underline{P}$ gets priority over *all* those with strictly lower bids, and equal bidders are served FIFO. Under AN, service interruptions are precluded. We use the subscripts AN and AP for the respective variables and functions; $\mathcal{A}$ refers to either auction.

Additional Assumptions. We supplement the assumptions of §2 as follows.

Assumption 4. *Service and arrival times give rise to an $M/M/1$ system.*

Assumption 5. *The functions $D(t)$ and $C(t)$ are linear: $D(t) = 1 - d \cdot t$ and $C(t) = c \cdot t$ with delay sensitivity parameters $d > 0$ and $c \geq 0$. Thus, customers are risk neutral as discussed in §2.*

The steady-state delay distribution depends on customers' strategies. We focus on symmetric pure strategies, represented by a payment or bid function $b\colon [\underline{v}, \overline{v}] \to \mathbb{R}_+$. Let $I(\cdot)$ denote the indicator function, and $q^+(p \mid b) := \Lambda \cdot \int_{\underline{v}}^{\overline{v}} \phi(v) \cdot I(b(v) > p)\, dv$ and $q^-(p \mid b) := \Lambda \cdot \int_{\underline{v}}^{\overline{v}} \phi(v) \cdot I(b(v) \geq p)\, dv$, respectively, be the rate of customers whose bid strictly (weakly, respectively) exceeds p under bid strategies b. Note that $q^+(p \mid b)$ and $q^-(p \mid b)$ are decreasing in p. The total demand rate corresponding to a reserve price $\underline{P}$ is $\lambda(\underline{P} \mid b) := q^-(\underline{P} \mid b)$. Given bid strategies b for auction $\mathcal{A}$ with reserve price $\underline{P}$, an individual customer's expected delay depends on her own bid p. Let $\overline{W}_{\mathcal{A}}(p \mid \underline{P}, b)\colon [\underline{P}, \infty) \to \mathbb{R}_+$ denote the function that maps a customer's bid into her expected steady-state delay. If $\lambda(\underline{P} \mid b) < 1$, then the system is stable and $\overline{W}_{\mathcal{A}}(p \mid \underline{P}, b)$ is finite for all p, consistent with bidders' expectations, and given by (cf., Kleinrock 1967, Theorems 1 and 2)

$$\overline{W}_{\mathcal{A}}(p \mid \underline{P}, b) = \begin{cases} 1 + \dfrac{\lambda(\underline{P} \mid b)}{[1 - q^+(p \mid b)][1 - q^-(p \mid b)]}, \\ \qquad \mathcal{A} = AN, p \geq \underline{P}, \\ \dfrac{1}{[1 - q^+(p \mid b)][1 - q^-(p \mid b)]}, \\ \qquad \mathcal{A} = AP, p \geq \underline{P}. \end{cases} \tag{13}$$

A customer's expected priority increases and her expected delay decreases in her bid p. Under auction $\mathcal{A}$, reserve price $\underline{P}$, and bid function b, a customer with value v has expected utility

$$u_{\mathcal{A}}(v \mid \underline{P}, b) = \left\{v - (v \cdot d + c) \cdot \overline{W}_{\mathcal{A}}(b(v) \mid \underline{P}, b)\right\} \cdot I(b(v) \geq \underline{P}) - b(v). \tag{14}$$

Note that the bid of an infinitesimal customer does not affect other customers' expected utility.

4.1. Fixed $\underline{P}$: Conditional Priority Auction Equilibria

A Nash equilibrium for auction $\mathcal{A}$ at reserve price $\underline{P}$ is any bid function $b_{\mathcal{A}}(v \mid \underline{P})$ that satisfies the participation constraints, i.e., all customers get nonnegative expected utility, and the incentive-compatibility constraints, i.e., none has an incentive to unilaterally change her bid:

$$u_{\mathcal{A}}(v \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) \geq \{v - (v \cdot d + c) \cdot \overline{W}_{\mathcal{A}}(p \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P}))\} \cdot I(p \geq \underline{P}) - p \quad \forall\, v, \forall\, p \geq 0. \tag{15}$$

By (13), the system is stable in equilibrium, else $u_{\mathcal{A}}(v \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) = -\infty$ for some v.

Proposition 2. *Under Assumptions 4 and 5 and if $\overline{v} \cdot (1 - d) - c > \underline{P}$, there is a unique symmetric bid equilibrium for $\mathcal{A}$.*

(1) *The equilibrium function satisfies $b_{\mathcal{A}}(v \mid \underline{P}) = 0$ for $v < v_{\mathcal{A}}(\underline{P})$ and*

$$b_{\mathcal{A}}(v \mid \underline{P}) = \int_{v_{\mathcal{A}}(\underline{P})}^{v} -(x \cdot d + c) \cdot \overline{W}'_{\mathcal{A}}(x \mid \underline{P})\, dx + \underline{P}, \quad v \geq v_{\mathcal{A}}(\underline{P}), \tag{16}$$

where the marginal value $v_{\mathcal{A}}(\underline{P})$ is the unique solution of

$$\begin{aligned} \underline{P} &= v - (v \cdot d + c) \cdot \left(1 + \frac{\Lambda \cdot \overline{\Phi}(v)}{(1 - \Lambda \cdot \overline{\Phi}(v))^2}\right), \quad \mathcal{A} = AN, \\ \underline{P} &= v - (v \cdot d + c) \cdot \frac{1}{(1 - \Lambda \cdot \overline{\Phi}(v))^2}, \quad \mathcal{A} = AP, \end{aligned} \tag{17}$$

the demand rate is $\lambda_{\mathcal{A}}(\underline{P}) := \Lambda \cdot \overline{\Phi}(v_{\mathcal{A}}(\underline{P}))$, and the mean delay for $v \geq v_{\mathcal{A}}(\underline{P})$ is

$$\overline{W}_{\mathcal{A}}(v \mid \underline{P}) := \overline{W}_{\mathcal{A}}(b_{\mathcal{A}}(v \mid \underline{P}) \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) = \begin{cases} 1 + \dfrac{\lambda_{\mathcal{A}}(\underline{P})}{(1 - \Lambda \cdot \overline{\Phi}(v))^2}, & \mathcal{A} = AN, \\ \dfrac{1}{(1 - \Lambda \cdot \overline{\Phi}(v))^2}, & \mathcal{A} = AP. \end{cases} \tag{18}$$

(2) *Given $\underline{P}$, $b_{AP}(\cdot \mid \underline{P})$ ($b_{AN}(\cdot \mid \underline{P})$) is socially optimal over all (nonpreemptive) allocations.*

Write $u_{\mathcal{A}}(v \mid \underline{P}) := u_{\mathcal{A}}(v \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P}))$ for the expected equilibrium utility of a bidder with value v. Her expected net value is $v - (v \cdot d + c) \cdot \overline{W}_{\mathcal{A}}(v \mid \underline{P})$. By (15), the following holds in equilibrium:

$$u_{\mathcal{A}}(v_H \mid \underline{P}) - u_{\mathcal{A}}(v_L \mid \underline{P}) \geq (v_H - v_L)(1 - d \cdot \overline{W}_{\mathcal{A}}(v_L \mid \underline{P}))I(b_{\mathcal{A}}(v_L \mid \underline{P}) \geq \underline{P}) \quad \forall\, v_H > v_L. \tag{19}$$

The difference between the high- and low-value bidder's expected utility exceeds their expected net value difference if *both* bid $b_{\mathcal{A}}(v_L \mid \underline{P})$. It is positive because a bidder's net value parameter, $v(1 - d) - c$, strictly increases in v. Hence, all customers with value $v \geq v_{\mathcal{A}}(\underline{P})$ bid, their utility strictly increases in value, and the provider *cannot* extract all surplus. Using (19) twice yields

$$(v_H - v_L) \cdot d \cdot \overline{W}_{\mathcal{A}}(v_H \mid \underline{P}) \leq (v_H - v_L) \cdot d \cdot \overline{W}_{\mathcal{A}}(v_L \mid \underline{P}) \quad \text{for } v_H > v_L \geq v_{\mathcal{A}}(\underline{P}), \tag{20}$$

i.e., the difference between the high- and low-value customer's delay cost must be larger if both bid like the lower-value bidder. Hence, the equilibrium mean delays are decreasing, and (by (13)) the bids increasing in values. By (16) and (18), the bid and mean delay functions are strictly monotone: if $b_{\mathcal{A}}(v_L \mid \underline{P}) =$

$b_{\mathcal{A}}(v_H \mid \underline{P})$, a customer with $v_0 \in (v_L, v_H]$ could do better by upping her bid by a negligible amount to get priority over customers with $v \in (v_L, v_H]$. By (16), a marginal increase in the bid of a customer with value v, $b'_{\mathcal{A}}(v \mid \underline{P})$ equals the resulting change in her expected net value, $-(v \cdot d + c)\overline{W}'_{\mathcal{A}}(v \mid \underline{P})$. The bid premium $b_{\mathcal{A}}(v \mid \underline{P}) - \underline{P}$ equals her *priority externality*, i.e., the marginal net value losses she inflicts on all lower-priority customers. By (17), the lowest bidder has value $v_{\mathcal{A}}(\underline{P})$, zero expected utility, and pays $\underline{P}$.

Conditional on $\underline{P}$, the equilibrium bids yield the socially optimal allocation by revealing the order of customers' values. Because the expected net values for any delay strictly increase in v, it is socially optimal to admit the highest-value customers for any λ and service rule. Because delay cost parameters strictly increase in v, it is socially optimal for any λ to schedule admitted customers using a continuum of priorities in the order of their values: by the $c\mu$-rule, this allocation in the AP (AN) equilibrium minimizes the expected aggregate delay cost (cf., Kleinrock 1967, pp. 304–318) and, hence, maximizes the expected aggregate net value, over all (nonpreemptive) rules.

For convenience, we perform the subsequent analysis in terms of λ. By (17), $\lambda_{\mathcal{A}}(\underline{P})$ is strictly decreasing and continuous in $\underline{P}$, yielding the inverse demand function

$$\underline{P}_{\mathcal{A}}(\lambda) = V'(\lambda) - (V'(\lambda) \cdot d + c) \cdot \overline{W}_{\mathcal{A}}(V'(\lambda) \mid \underline{P}_{\mathcal{A}}(\lambda)), \quad \mathcal{A} \in \{AN, AP\}. \tag{21}$$

If $\underline{P}_{\mathcal{A}}(\lambda) \geq 0$, then λ is the equilibrium demand rate, $v_{\mathcal{A}}(\underline{P}_{\mathcal{A}}(\lambda)) = V'(\lambda)$ is the marginal value, and the expected delay, the expected net value and the bid of a bidder with value $V'(q)$ are

$$\overline{W}_{\mathcal{A}}(q, \lambda) := \overline{W}_{\mathcal{A}}(V'(q) \mid \underline{P}_{\mathcal{A}}(\lambda)) = \begin{cases} 1 + \dfrac{\lambda}{(1-q)^2}, & \mathcal{A} = AN, \\ \dfrac{1}{(1-q)^2}, & \mathcal{A} = AP, \end{cases} \tag{22}$$

$$nv_{\mathcal{A}}(q, \lambda_{\mathcal{A}}) := V'(q) - (V'(\lambda) \cdot d + c) \cdot \overline{W}_{\mathcal{A}}(q, \lambda_{\mathcal{A}}), \tag{23}$$

$$b_{\mathcal{A}}(q, \lambda_{\mathcal{A}}) := b_{\mathcal{A}}(V'(q) \mid \underline{P}_{\mathcal{A}}(\lambda)) = \underline{P}_{\mathcal{A}}(\lambda_{\mathcal{A}}) + \int_q^{\lambda_{\mathcal{A}}} (V'(\lambda) \cdot d + c) \frac{\partial \overline{W}_{\mathcal{A}}(q, \lambda_{\mathcal{A}})}{\partial q}\, dq. \tag{24}$$

The bidder's priority level is q ($q = 0$ is the highest priority and $q = \lambda$ the lowest priority) and her expected utility $u_{\mathcal{A}}(q, \lambda) := u_{\mathcal{A}}(V'(q) \mid \underline{P}_{\mathcal{A}}(\lambda))$ equals $nv_{\mathcal{A}}(q, \lambda) - b_{\mathcal{A}}(q, \lambda)$.

4.2. Social Optimization vs. Revenue Maximization

Social Optimization. The provider maximizes the expected aggregate net value per unit time as

$$\max_{\lambda \in [0,1)} NV_{\mathcal{A}}(\lambda) := \int_0^{\lambda} nv_{\mathcal{A}}(q, \lambda)\, dq = \int_0^{\lambda} V'(q) \cdot (1 - d \cdot \overline{W}_{\mathcal{A}}(z, \lambda)) - c \cdot \overline{W}_{\mathcal{A}}(q, \lambda)\, dq. \tag{25}$$

Our assumptions imply that (25) has a unique solution $\lambda^*_{\mathcal{A}}$ that satisfies the first-order condition

$$\underline{P}_{\mathcal{A}}(\lambda) = V'(\lambda)[1 - d\overline{W}_{\mathcal{A}}(\lambda, \lambda)] - c\overline{W}_{\mathcal{A}}(\lambda, \lambda) = \int_0^{\lambda} [V'(q) \cdot d + c] \frac{\partial \overline{W}_{\mathcal{A}}(q, \lambda)}{\partial \lambda}\, dq. \tag{26}$$

The optimal reserve price $\underline{P}^*_{\mathcal{A}} := \underline{P}_{\mathcal{A}}(\lambda^*_{\mathcal{A}}) = b_{\mathcal{A}}(\lambda^*_{\mathcal{A}}, \lambda^*_{\mathcal{A}})$ equals the marginal (lowest priority) customer's expected net value and her *admission net value externality*, i.e., the expected net value loss her admission inflicts on the system. In the AP auction, $\underline{P}^*_{AP} = 0$ because this loss is zero: no customer is delayed by those with lower priority. In the AN case, $\underline{P}^*_{AN} > 0$: all customers' expected delay increases in λ because they cannot interrupt a lower-priority customer. A customer's bid reflects the social cost she causes (i) by being admitted—it equals the reserve price $\underline{P}^*_{\mathcal{A}}$, and (ii) by being prioritized—it equals her bid premium $b_{\mathcal{A}}(q, \lambda^*_{\mathcal{A}}) - \underline{P}^*_{\mathcal{A}}$, as in (24).

Revenue Maximization. By (21)–(24), a customer with value $V'(q)$ has expected utility

$$u_{\mathcal{A}}(q, \lambda) = \int_q^{\lambda} \frac{\partial u_{\mathcal{A}}(z, \lambda)}{\partial z}\, dz = \int_q^{\lambda} -V''(z) \cdot (1 - d \cdot \overline{W}_{\mathcal{A}}(z, \lambda))\, dz = nv_{\mathcal{A}}(q, \lambda) - b_{\mathcal{A}}(q, \lambda). \tag{27}$$

In equilibrium the integrand $\partial u_{\mathcal{A}}(z, \lambda)/\partial z$, i.e., the difference in the expected utility of customers with infinitely close values $V'(z - \varepsilon) > V'(z)$, must equal their net value difference if both bid $b_{\mathcal{A}}(z, \lambda)$ (see (19) with $v_H \to v_L$). By (27), the revenue-maximization problem is

$$\max_{\lambda \in [0,1)} \Pi_{\mathcal{A}}(\lambda) = \int_0^{\lambda} b_{\mathcal{A}}(q, \lambda)\, dq = \int_0^{\lambda} [V'(q) + qV''(q)] \cdot (1 - d\overline{W}_{\mathcal{A}}(q, \lambda)) - c\overline{W}_{\mathcal{A}}(q, \lambda)\, dq. \tag{28}$$

Our assumptions imply that (28) has a unique solution $\lambda_{\mathcal{A}}^M$ that satisfies the first-order condition

$$R'(\lambda)[1-d\overline{W}_{\mathcal{A}}(\lambda,\lambda)]-c\overline{W}_{\mathcal{A}}(\lambda,\lambda) = \int_0^{\lambda}[R'(q)\cdot d+c]\frac{\partial \overline{W}_{\mathcal{A}}(q,\lambda)}{\partial \lambda}\,dq, \quad (29)$$

where $V'(q)+qV''(q)=R'(q)$. At the revenue-maximizing rate $\lambda_{\mathcal{A}}^M$, the expected marginal net revenue under constant delays (the left-hand side of (29)) equals the *admission revenue externality* (the right-hand side), i.e., the delay-induced expected revenue loss caused by admitting the marginal customer.

Proposition 3. *Under Assumptions* 4 *and* 5, *revenue maximization* (M) *and social optimization* ($*$) *compare as follows*:

(1) *AN auction*: $\underline{P}_{AN}^* > 0$ *and* $\underline{P}_{AN}^M > 0$. (a) *If* $\epsilon(\lambda)$ *is strictly increasing in* λ, *then for* $d>0$, *there is a unique* $c^*(d)>0$ *at which* $\underline{P}_{AN}^M = \underline{P}_{AN}^*$. *If* $c<c^*(d)$, *then* $\underline{P}_{AN}^M < \underline{P}_{AN}^*$, *and conversely if* $c>c^*(d)$. (b) *If* $\epsilon(\lambda)$ *is constant in* λ, *then* $\underline{P}_{AN}^M = \underline{P}_{AN}^* \Leftrightarrow c=0$ *and* $\underline{P}_{AN}^M > \underline{P}_{AN}^* \Leftrightarrow c>0$. (c) *If* $\epsilon(\lambda)$ *is strictly decreasing in* λ, *then* $\underline{P}_{AN}^M > \underline{P}_{AN}^*$ *for all* c.

(2) *AP auction*: $\underline{P}_{AP}^* = 0$. *If* $\epsilon(\lambda_{AP}^*) \geq 1$ *and* $c=0$, *then* $\underline{P}_{AP}^M = \underline{P}_{AP}^*$. *Else*, $\underline{P}_{AP}^M > \underline{P}_{AP}^*$.

As under uniform pricing, the additive components are the same for $NV_{\mathcal{A}}(\lambda)$ and $\Pi_{\mathcal{A}}(\lambda)$, while for each integrand, the multiplicative component is larger under social optimization (see (25) and (28)). The results for AN are structurally the same as for uniform pricing: contribution and admission externality of the marginal customer under social optimization are larger than their counterparts under revenue maximization (compare (26) and (29)). The slope of the elasticity function and the relative magnitudes of the delay cost components determine which effect dominates. However, under AP, the admission externality is zero for both objective functions. With additive delay cost, the revenue-maximizing reserve price is thus higher than socially optimal. Otherwise, it is socially optimal only if the marginal (gross) revenue function $R'(\lambda)$ is nonnegative (equivalently, if $V'(\lambda)$ is elastic) at λ_{AP}^*. In this case, $\underline{P}_{AP}^M = 0$, the AP auction is *self-regulating*. In summary, for revenue maximization to yield social optimality under AP requires stronger conditions on the delay cost structure and weaker conditions on the elasticity function, compared to AN.

5. Priority Auctions vs. Uniform Pricing

Auctions create more value than uniform pricing (henceforth identified by U) by serving impatient customers faster, and extract more value by charging them more. We study these aggregate gains and their distribution among customers for an $M/M/1$ system with linear $C(t)$ and $D(t)$.

5.1. Aggregate Auction Gains

The auctions yield lower admission prices, higher demand rates, and gains in expected system net value (or surplus) and expected provider revenue, compared to uniform pricing.

Proposition 4. *Under Assumptions* 4 *and* 5, *the priority auction and uniform pricing equilibria compare as follows*:

(1) *Social optimization*: *If*

$$0<d<1-\frac{c}{V'(0)}\colon \underline{P}_{AP}^* < \underline{P}_{AN}^* < P_U^*,\ \lambda_{AP}^* > \lambda_{AN}^* > \lambda_U^*, \quad (30)$$

$$NV_{AP}(\lambda_{AP}^*) > NV_{AN}(\lambda_{AN}^*) > NV_U(\lambda_U^*). \quad (31)$$

(2) *Revenue maximization*: *If*

$$0<d<1-\frac{c}{R'(0)}\colon \underline{P}_{AP}^M < \underline{P}_{AN}^M < P^M,\ \lambda_{AP}^M > \lambda_{AN}^M > \lambda_U^M, \quad (32)$$

$$\Pi_{AP}(\lambda_{AP}^M) > \Pi_{AN}(\lambda_{AN}^M) > \Pi_U(\lambda_U^M). \quad (33)$$

The auction gains are due to heterogeneity in the delay cost parameters. Under U, they equal $d\cdot V'(z)+c$ for $z\in[0,\lambda_U^*]$, and the delay *cost* heterogeneity is maximized because all customers' expected delay is the same. The surplus gain[7] from auction $\mathcal{A}\in\{AN, AP\}$ consists of a service differentiation (SD) gain on customers served in the equilibria of both $\mathcal{A}$ and U, and a market expansion (ME) gain from customers added under the auction. The SD gain equals

$$SD_{\mathcal{A}}(\lambda_U^*) := NV_{\mathcal{A}}(\lambda_U^*) - NV_U(\lambda_U^*) = d\int_0^{\lambda_U^*} V'(x)\left(\frac{1}{1-\lambda_U^*}-\overline{W}_{\mathcal{A}}(x,\lambda_U^*)\right)dx, \quad (34)$$

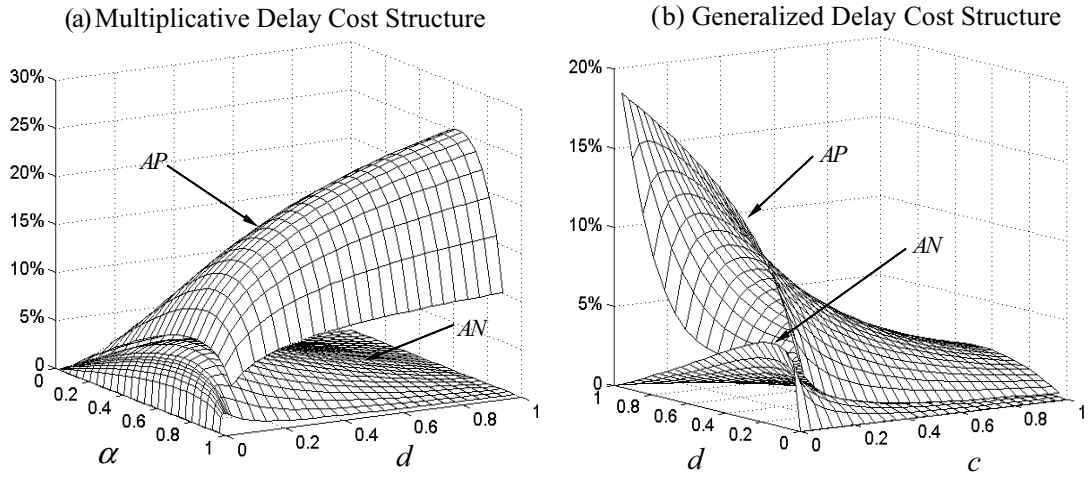
the reduction in the expected aggregate delay cost given λ_U^*. Prioritizing customers reduces (increases) the expected delay cost of those with high (low) delay cost parameter. It also reduces the marginal customer's net value externality, which leads to market expansion. The ME gain approximately equals the market expansion $\lambda_{\mathcal{A}}^* - \lambda_U^*$, times the marginal SD gain, i.e., the difference in marginal delay costs at λ_U^* between uniform pricing and auction

$$ME_{\mathcal{A}}(\lambda_{\mathcal{A}}^*,\lambda_U^*) := NV_{\mathcal{A}}(\lambda_{\mathcal{A}}^*) - NV_{\mathcal{A}}(\lambda_U^*) \simeq (\lambda_{\mathcal{A}}^* - \lambda_U^*)\cdot SD'_{\mathcal{A}}(\lambda_U^*), \quad \mathcal{A}\in\{AN, AP\}. \quad (35)$$

Under AP, the marginal customer creates no externality. Its ME gain thus exceeds AN's. Because the SD and ME gains are realized by reducing the delay cost heterogeneity, they are larger for AP than for AN, and their magnitude is sensitive to the delay cost struc-

[7] The auction revenue gains are driven by similar effects with the function V replaced by R.

Figure 2 Priority Auctions vs. Uniform Pricing: Net Value Gains Under Social Optimization



Note. Isoelastic functions $V'(\lambda) = (1-\alpha)\lambda^{-\alpha}$. (a) $(\alpha, d) \in [0, 0.99] \times [0, 1]$, $c = 0$. (b) $(d, c) \in [0, 1] \times [0, 1]$, $\alpha = 0.4$.

ture. Let $\Delta NV^*_{\mathcal{A}}(d, c) := NV_{\mathcal{A}}(\lambda^*_{\mathcal{A}}; d, c) - NV_U(\lambda^*_U; d, c)$ be the surplus gain and $\Delta\Pi^M_{\mathcal{A}}(d, c) := \Pi_{\mathcal{A}}(\lambda^M_{\mathcal{A}}; d, c) - \Pi^M_U(\lambda^M_U; d, c)$ the revenue gain of auction $\mathcal{A}$ versus U as a function of c and d.

Proposition 5. *Under Assumptions* 4 *and* 5, *the priority auction gains are decreasing in c but nonmonotone in d*:

(1) *Additive delay cost*: *For* $d < 1$, $\Delta NV^*_{\mathcal{A}}(d, c)$ *is strictly decreasing in* $c \in [0, V'(0)(1-d)]$, *and* $\Delta\Pi^M_{\mathcal{A}}(d, c)$ *is strictly decreasing in* $c \in [0, R'(0)(1-d)]$.

(2) *Multiplicative delay cost*: *For* $c < V'(0)$, $\Delta NV^*_{\mathcal{A}}(0, c) = 0$ *and* $\Delta NV^*_{\mathcal{A}}(d, c)$ *is maximized at some* $d^*_{\mathcal{A}}(c) \in (0, 1 - c/V'(0))$. *For* $c < R'(0)$, $\Delta\Pi^M_{\mathcal{A}}(0, c) = 0$ *and* $\Delta\Pi^M_{\mathcal{A}}(d, c)$ *is maximized at some* $d^M_{\mathcal{A}}(c) \in (0, 1 - c/R'(0))$.

To see why the auction gains are nonmonotone in d, note that while the *SD* gain (34) is zero when $d = 0$ (all customers are equally delay sensitive), an increase in d has two opposite effects. First, it increases the *dispersion* of delay cost parameters for fixed λ, $d \cdot V'(z) + c$, $z \in [0, \lambda]$, increasing *SD* gain: the delay cost gain of high-value, high-priority customers is larger than the loss of low-value, low-priority customers. Second, it increases the *average* delay cost parameter, yielding a lower demand λ^*_U under U, which reduces the *SD* gain. Increasing d has a positive (negative) net effect for small (large) d, where the utilization and the mean delay differences are high (low). By contrast, increasing c always reduces the *SD* gain: increasing c only reduces λ^*_U while leaving the delay cost heterogeneity constant for any fixed λ. The magnitude of the *ME* gain (35) and, hence, of the total auction gain, are governed by the same effects.

Numerical Study. Consider the isoelastic functions $V'(\lambda; \alpha) = (1-\alpha)\lambda^{-\alpha}$, $\alpha \in [0, 1)$, where $\epsilon(\lambda; \alpha) = \alpha^{-1}$ is the elasticity and $V(1; \alpha) = 1$ at full utilization. We study the percentage gains in system net value under the auctions, relative to U, for two parameter sets: (i) $(\alpha, d) \in [0, 0.99] \times [0, 1]$ and $c = 0$, which yields all combinations of isoelastic demand and multiplicative delay cost (Figure 2a; here, Propositions 1 and 3 imply that the percentage revenue gains and percentage net value gains are equal), and (ii) $(d, c) \in [0, 1] \times [0, 1]$ and $\alpha = 0.4$ (Figure 2b), showing the effect of the delay cost structure. The *AP* gains are significant, with a maximum of 26% for $(\alpha, d, c) = (0.68, 0.99, 0)$. They dramatically exceed the *AN* gains, which peak at 6.4% for $(\alpha, d, c) = (0.64, 0.11, 0)$. The delay cost structure and the marginal value curve impact the auction gains as follows.

Multiplicative delay cost $(d > 0, c = 0)$. *AN* gains the most (6.4%) for low d, while *AP* attains its maximum gain (26%) in the *most* delay-sensitive setting as $d \to 1$.[8] While both attain their largest absolute gains for low d (Proposition 5), the *AN* gains decrease much more sharply in d compared to *AP*, where adding customers creates no externality.

Additive delay cost $(d > 0, c > 0)$. For given demand curve and fixed d, the percentage gains in system net value under *AN* and *AP* are decreasing in c. The absolute gain $\Delta NV^*_{\mathcal{A}}(d, c)$ (the numerator) decreases in c (Proposition 5) at a faster rate than $NV_U(\lambda^*_U; d, c)$ (the denominator).

Marginal value curve $(c = 0)$. Integrating (34) by parts, the *SD* gain can be written as

$$SD_{\mathcal{A}}(\lambda^*_U) = d \cdot \int_0^{\lambda^*_U} x \int_x^{\lambda^*_U} |AV'(z)|\, dz \frac{\partial \overline{W}_{\mathcal{A}}(x, \lambda^*_U)}{\partial x}\, dx, \quad \mathcal{A} \in \{AN, AP\}, \quad (36)$$

[8] The gains are zero for $d = 0$. For the sake of expository clarity, Figure 2 does not show these points for *AP*.

where $AV'(z)$ is the (negative) slope of the average value curve. By inspection of (36), the *SD* gain is increasing in $|AV'(z)|$. The steeper AV, the larger the difference between average and marginal delay cost parameter. Here, $AV'(\lambda) = -\alpha\lambda^{-(1+\alpha)}$. For fixed λ, the auction gains are zero in homogeneous markets ($\alpha = 0$) and increase in α, but λ_U^* decreases in α because $V'(\lambda;\alpha)/V(\lambda;\alpha) = (1-\alpha)/\lambda$ decreases in α. Hence, the auctions gain the most at intermediate values of α.

5.2. Auction Impact on Customers: Winners and Losers

The difference in the expected utilities of a customer with value $V'(q)$ under the socially optimal auction and uniform price equilibria equals her delay cost difference minus her price difference:

$$\begin{aligned} u_{\mathcal{A}}^*(q) &:= u_{\mathcal{A}}(q, \lambda_{\mathcal{A}}^*) - u_U(q, \lambda_U^*) \\ &= (V'(q)d + c)[\overline{W}_U(\lambda_U^*) - \overline{W}_{\mathcal{A}}(q, \lambda_{\mathcal{A}}^*)] \\ &\quad - [b_{\mathcal{A}}(q, \lambda_{\mathcal{A}}^*) - P_U^*]. \end{aligned} \tag{37}$$

Let $u_{\mathcal{A}}^M(q)$ be the respective expected utility difference under the revenue-maximizing equilibria.

Proposition 6. *Under Assumptions* 4 *and* 5, *customers benefit as follows from an auction* ($\mathcal{R} = *$ *or* $\mathcal{R} = M$):

(1) *For* $\mathcal{A} \in \{AN, AP\}$: *The set of losers,* $L_{\mathcal{A}}^{\mathcal{R}} := \{q \in [0, \lambda_U^{\mathcal{R}}]: u_{\mathcal{A}}^{\mathcal{R}}(q) < 0\}$, *is an interval. Let* $\underline{l}_{\mathcal{A}}^{\mathcal{R}} := \inf L_{\mathcal{A}}^{\mathcal{R}}$ *and* $\bar{l}_{\mathcal{A}}^{\mathcal{R}} := \sup L_{\mathcal{A}}^{\mathcal{R}}$. *Customers with value* $V'(q_{\mathcal{A}}^{\mathcal{R}})$ *benefit the least*:

$$q_{\mathcal{A}}^{\mathcal{R}} = \arg\min_{q \in [0, \lambda_U^{\mathcal{R}}]} u_{\mathcal{A}}^{\mathcal{R}}(q), \quad \text{where } q_{AP}^{\mathcal{R}} = 1 - \sqrt{1 - \lambda_U^{\mathcal{R}}},$$

$$q_{AN}^{\mathcal{R}} = \max\left(1 - \sqrt{\lambda_{AN}^{\mathcal{R}} \tfrac{(1-\lambda_U^{\mathcal{R}})}{\lambda_U^{\mathcal{R}}}}, 0\right),$$

and

$$\overline{W}_{AP}(q_{AP}^{\mathcal{R}}, \lambda_{AP}^{\mathcal{R}}) = \overline{W}_U(\lambda_U^{\mathcal{R}}), \tag{38}$$

$$\overline{W}_{AN}(q_{AN}^{\mathcal{R}}, \lambda_{AN}^{\mathcal{R}}) = (>)\overline{W}_U(\lambda_U^{\mathcal{R}}) \quad \text{and} \quad q_{AN}^{\mathcal{R}} \geq (=) 0 \quad \text{if } \lambda_{AN}^{\mathcal{R}} \leq (>) \frac{\lambda_U^{\mathcal{R}}}{1 - \lambda_U^{\mathcal{R}}}. \tag{39}$$

The utility difference $u_{\mathcal{A}}^{\mathcal{R}}(q)$ *is decreasing* (*increasing*) *in* q *on* $[0, q_{\mathcal{A}}^{\mathcal{R}})$ (*on* $(q_{\mathcal{A}}^{\mathcal{R}}, \lambda_U^{\mathcal{R}}]$).

(2) *AP*: *For convex* V': *If* $b_{AP}(q_{AP}^{\mathcal{R}}, \lambda_{AP}^{\mathcal{R}}) > P_U(\lambda_U^{\mathcal{R}})$, *all but intermediate-value customers gain, i.e.,* $0 < \underline{l}_{AP}^{\mathcal{R}} < \bar{l}_{AP}^{\mathcal{R}} < \lambda_U^{\mathcal{R}}$, $u_{AP}^{\mathcal{R}}(\underline{l}_{AP}^{\mathcal{R}}) = u_{AP}^{\mathcal{R}}(\bar{l}_{AP}^{\mathcal{R}}) = 0$. *Else, all are* (*weakly*) *better off*: $L_{AP}^{\mathcal{R}} = \varnothing$.

(3) *AN*: (a) *Linear marginal value function* $V'(\lambda) = 1 - B\lambda$, $B \in (0, \Lambda^{-1}]$: *If* $\lambda_{AN}^{\mathcal{R}} < \lambda_U^{\mathcal{R}}/(1 - \lambda_U^{\mathcal{R}})$ *and* $b_{AN}(q_{AN}^{\mathcal{R}}, \lambda_{AN}^{\mathcal{R}}) > P_U(\lambda_U^{\mathcal{R}})$, *all but intermediate-value customers benefit*: $0 < \underline{l}_{AN}^{\mathcal{R}} < \bar{l}_{AN}^{\mathcal{R}} < \lambda_U^{\mathcal{R}}$ *and* $u_{AN}^{\mathcal{R}}(\underline{l}_{AN}^{\mathcal{R}}) = u_{AN}^{\mathcal{R}}(\bar{l}_{AN}^{\mathcal{R}}) = 0$. *Else, all are* (*weakly*) *better off*: $L_{AN}^{\mathcal{R}} = \varnothing$.

(b) *Isoelastic marginal value function* $V'(\lambda) = K\lambda^{-\alpha}$, $\alpha \in (0, 1)$, $K > 0$: *If* $\lambda_{AN}^{\mathcal{R}} < \lambda_U^{\mathcal{R}}/(1 - \lambda_U^{\mathcal{R}})$, *all but intermediate-value customers benefit if* $b_{AN}(q_{AN}^{\mathcal{R}}, \lambda_{AN}^{\mathcal{R}}) > P_U(\lambda_U^{\mathcal{R}})$ *and all are* (*weakly*) *better off if* $b_{AN}(q_{AN}^{\mathcal{R}}, \lambda_{AN}^{\mathcal{R}}) \leq P_U(\lambda_U^{\mathcal{R}})$. *If* $\lambda_{AN}^{\mathcal{R}} > \lambda_U^{\mathcal{R}}/(1 - \lambda_U^{\mathcal{R}})$, *high-* (*low-*) *value customers are worse* (*better*) *off*: $0 = \underline{l}_{AN}^{\mathcal{R}} < \bar{l}_{AN}^{\mathcal{R}} < \lambda_U^{\mathcal{R}}$, $u_{AN}^{\mathcal{R}}(\underline{l}_{AN}^{\mathcal{R}}) < 0$, $u_{AN}^{\mathcal{R}}(\bar{l}_{AN}^{\mathcal{R}}) = 0$.

The intuition for Proposition 6 is as follows.

For *any* $\lambda_{\mathcal{A}}^* > \lambda_U^*$: (i) The marginal customer under U has zero expected utility. She is better off in a priority auction: $u_{\mathcal{A}}^*(\lambda_U^*) > 0$ (so are those served only under $\mathcal{A}$). While she waits more and pays less than under uniform pricing, the lower price offsets the higher delay cost because she is relatively delay insensitive. (ii) The "equal-delay customer" with value $V'(q_{\mathcal{A}}^*)$ whose auction delay equals (or if $\lambda_{AN}^* > \lambda_U^*/(1 - \lambda_U^*)$, least exceeds) her FIFO delay benefits the least from an auction: $u_{\mathcal{A}}^*(q_{\mathcal{A}}^*) = \min_{q \in [0, \lambda_U^*]} u_{\mathcal{A}}^*(q)$. The utility difference $u_{\mathcal{A}}^*(q)$ decreases in the priority level from the marginal to the "equal-delay customer" (as $q \downarrow q_{\mathcal{A}}^*$) and goes up from there to the top priority customer (as $q \downarrow 0$): for lower- (higher-) value customers with $q \in (q_{\mathcal{A}}^*, \lambda_U^*]$ ($q \in [0, q_{\mathcal{A}}^*)$), the bid $b_{\mathcal{A}}(q, \lambda_{\mathcal{A}}^*)$ increases more (less) in the priority level, as $q \downarrow q_{\mathcal{A}}^*$ ($q \downarrow 0$), than the negative (positive) delay cost difference $(V'(q)d + c)[\overline{W}_U(\lambda_U^*) - \overline{W}_{\mathcal{A}}(q, \lambda_{\mathcal{A}}^*)]$ drops (goes up). Intermediate- and high-value customers may be better or worse off under the auction as follows.

Small Market Expansion. The marginal auction customer's equilibrium utility is zero and exceeds her expected payoff from choosing the equal-delay customer's bid $b_{\mathcal{A}}(q_{\mathcal{A}}^*, \lambda_{\mathcal{A}}^*)$ (see (15)). If $\lambda_{\mathcal{A}}^* \approx \lambda_U^*$,

$$\begin{aligned} u_{\mathcal{A}}(\lambda_{\mathcal{A}}^*, \lambda_{\mathcal{A}}^*) = 0 &> V'(\lambda_{\mathcal{A}}^*) - (V'(\lambda_{\mathcal{A}}^*)d + c) \\ &\quad \cdot \overline{W}_{\mathcal{A}}(q_{\mathcal{A}}^*, \lambda_{\mathcal{A}}^*) - b_{\mathcal{A}}(q_{\mathcal{A}}^*, \lambda_{\mathcal{A}}^*) \\ &\approx P_U(\lambda_U^*) - b_{\mathcal{A}}(q_{\mathcal{A}}^*, \lambda_{\mathcal{A}}^*). \end{aligned} \tag{40}$$

Her expected net value from bidding $b_{\mathcal{A}}(q_{\mathcal{A}}^*, \lambda_{\mathcal{A}}^*)$ is the same as under uniform pricing because $\overline{W}_{\mathcal{A}}(q_{\mathcal{A}}^*, \lambda_{\mathcal{A}}^*) = \overline{W}_U(\lambda_U^*)$. It equals $P_U(\lambda_U^*)$ because $V'(\lambda_{\mathcal{A}}^*) \approx V'(\lambda_U^*)$ and the marginal customer under uniform pricing has zero expected utility. Hence, the equal-delay customer bids more than $P_U(\lambda_U^*)$ and belongs to a set of intermediate-value customers who are worse off under either auction than under uniform pricing. This holds *for relatively small* $\lambda_{\mathcal{A}}^* - \lambda_U^*$. The highest-value, highest-priority customers are better off under an auction: their delay cost reduction is high enough to offset the higher price they pay.[9]

[9] This holds if $V'(\lambda)$ is convex, which applies for the linear, isoelastic, and other commonly assumed functions.

Large Market Expansion: AP Auction. Increasing the demand rate λ_{AP} shifts the equilibrium bid function down, but leaves customers' expected net values unchanged, increasing their expected utilities. Under a relatively large demand rate difference $\lambda^*_{AP} - \lambda^*_U$, the uniform price exceeds the equal-delay customer's auction bid, and all customers are better off.

Large Market Expansion: AN Auction. Equilibrium bids and net values decrease in λ_{AN} because delays increase. If $V'(\lambda)$ is linear, the bid reductions offset the net value losses and all customers gain from AN for large enough $\lambda^*_{AN} - \lambda^*_U$. If $V'(\lambda)$ is *isoelastic*, the highest-value customers weigh the delay cost difference much more than the price difference. They only gain under AN if the FIFO delay exceeds the top priority AN delay (if $\lambda^*_{AN} < \lambda^*_U/(1-\lambda^*_U)$). Conversely, if $\lambda^*_{AN} > \lambda^*_U/(1-\lambda^*_U)$, their AN delays exceed their FIFO delays, making them worse off than under U.

6. Concluding Remarks

This paper evaluates alternative price-service mechanisms for a capacity-constrained provider serving delay-sensitive customers. Three elements characterize the system: (i) A generalized delay cost structure that augments the standard additive model with a multiplicative component. (ii) The price-service mechanism: uniform pricing, preemptive (AP), or nonpreemptive priority auction (AN). (iii) The objective function: social optimization or revenue maximization. We compare the equilibria under alternative mechanisms and objective functions and find that the delay cost structure has a paramount effect on system behavior.

Known results for the additive structure can be reversed under our generalized delay cost structure: the revenue-maximizing uniform or admission price (utilization) is equal to or lower (higher, respectively) than is socially optimal in several cases, resulting in no or only a small surplus loss. We characterize the delay cost structure and demand elasticity that drive this reversal. The results of Propositions 1 and 3 can be extended to multiserver, multiresource networks under U and AP; the latter using Afèche's (2003) analysis of networks with preemptive priorities. The results also hold for multiple markets with equal and constant elasticity, each with its own multiplicative delay cost ($c = 0$); and if a positive direct marginal cost is included in the additive delay cost.

In both priority auctions, for any given admission price, bids are strictly increasing in values and induce the efficient allocations. Compared to uniform pricing, the auctions yield a lower admission price, a higher utilization, and objective function gains. The percentage gains are much larger under AP than under AN, but both mechanisms perform better under multiplicative compared to additive delay costs. Compared to uniform pricing, the highest-value customers always gain under AP but may be worse off under AN, while in both auctions, those with the lowest values always gain and those in between may gain or lose.

There are various avenues for future work. One is to consider a small set of "large" customers; or multiple customer segments, each with its own delay cost, value, and service time distribution. A second is to study dynamic versions of this problem. Another is to study capacity setting under the generalized delay cost structure (cf., Lui 1985 and Hassin 1995 for the additive model).

Acknowledgments

The authors thank the associate editor and anonymous referees for their insightful comments.

Appendix. Proofs

Proof of Proposition 1. The proof builds on three observations regarding the functions $NV(\lambda)$, $\Pi(\lambda)$, $\epsilon(\lambda)$, $r(\lambda)$, and $v(\lambda)$.

Observation 1. The functions $NV(\lambda)$ and $\Pi(\lambda)$ are continuously differentiable and strictly concave because $V''(\lambda) < 0$ and by Assumptions 1–3. Hence, the Problem (3) ((5), respectively) has a unique interior solution λ^* (λ^M), which solves the first-order condition (4) ((6), respectively). As a result, $\Pi'(\lambda^*) > NV'(\lambda^*) = 0 \Leftrightarrow \lambda^M > \lambda^*$.

Observation 2. If $\epsilon(\lambda)$ is strictly increasing, then $r(\lambda) > v(\lambda)$ for all λ: Equation (10) implies that $r(\lambda)$ is strictly increasing, and $R'(q)V'(\lambda) < V'(q)R'(\lambda)$ for all $q < \lambda$ implies that $R(\lambda)V'(\lambda) < V(\lambda)R'(\lambda)$. Similarly, if $\epsilon(\lambda)$ is constant (strictly decreasing) in λ, then $r(\lambda) \equiv v(\lambda)$ ($r(\lambda) < v(\lambda)$ for all λ).

Observation 3. Define the functions

$$L(\lambda) = \frac{NV'(\lambda) - \Pi'(\lambda)}{-[V(\lambda) - R(\lambda)]\bar{D}'(\lambda)} = \frac{V'(\lambda)}{-\epsilon(\lambda)[V(\lambda) - R(\lambda)]} \frac{\bar{D}(\lambda)}{\bar{D}'(\lambda)} - 1, \tag{A1}$$

$$Q(\lambda) = \frac{v(\lambda)}{\lambda} \frac{\bar{D}(\lambda)}{-\bar{D}'(\lambda)} - 1. \tag{A2}$$

Note that $r(\lambda) > v(\lambda) \Leftrightarrow L(\lambda) < Q(\lambda)$ for all λ and $r(\lambda) \equiv v(\lambda) \Leftrightarrow L(\lambda) \equiv Q(\lambda)$.

Case 1. If $\epsilon(\lambda)$ is strictly increasing, $L(\lambda)$ is strictly decreasing, $L(0) = \infty$, and $\lim_{\lambda \to 1} L(\lambda) < 0$. So there is a unique $\lambda_1 > 0$ such that $L(\lambda_1) = 0 \Leftrightarrow NV'(\lambda_1) = \Pi'(\lambda_1)$. By Observations 2 and 3, we have $Q(\lambda_1) > L(\lambda_1) = 0$. Let $f(\lambda, c) = c \cdot [\bar{C}(\lambda) + \lambda \bar{C}'(\lambda)]$. For $c = 0$, the first-order condition (4) and $Q(\lambda_1) > 0$ imply that $f(\lambda_1, c) = 0 < NV'(\lambda_1; c)$. Because $f(\lambda_1, c)$ is continuous in c with $f_c(\lambda_1, c) > 0$ and $\lim_{c \to \infty} f(\lambda_1, c) = \infty$, there is a unique $c^* > 0$ such that $NV'(\lambda_1; c^*) = \Pi'(\lambda_1; c^*) = 0$ and $\lambda_1 = \lambda^M(c^*) = \lambda^*(c^*)$. For fixed $c < c^*$, we have $NV'(\lambda_1; c) = \Pi'(\lambda_1; c) > 0$. Because $NV''(\lambda; c) < 0$, this implies that $\lambda^*(c) > \lambda_1$. From Observation 3, it follows that $NV'(\lambda^*(c); c) < \Pi'(\lambda^*(c); c)$ and Observation 1 yields $\lambda^M(c) > \lambda^*(c)$. The converse holds for $c > c^*$.

Case 2. If $\epsilon(\lambda)$ is constant, then $L(\lambda^*) = Q(\lambda^*)$ by Observations 2 and 3. If $c = 0$, then $Q(\lambda^*) = 0$ by the first-order condition (4), establishing $\Pi'(\lambda^*) = NV'(\lambda^*)$ and $\lambda^* = \lambda^M$. If $c > 0$, then $Q(\lambda^*) > 0$ by (4), which implies that $NV'(\lambda^*) > \Pi'(\lambda^*)$, so $\lambda^* > \lambda^M$.

Case 3. If $\epsilon(\lambda)$ is strictly decreasing, then $L(\lambda^*) > Q(\lambda^*)$ by Observations 2 and 3. Because $Q(\lambda^*) \geq 0$ by (4), we have $L(\lambda^*) > 0$, $NV'(\lambda^*) > \Pi'(\lambda^*)$, and $\lambda^* > \lambda^M$. □

Proof of Proposition 2. Part (1). We first show that every equilibrium $b_{\mathcal{A}}(\cdot \mid \underline{P})$ satisfies (16)–(18). Define $u_{\mathcal{A}}(v, p \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) := v - (vd + c)\overline{W}_{\mathcal{A}}(p \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) - p$ and $u_{\mathcal{A}}(v \mid \underline{P}) := u_{\mathcal{A}}(v \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P}))$.

(1) *Monotonicity*. $b_{\mathcal{A}}(v \mid \underline{P})$ is increasing by (13) and (20). Let $v^-_{\mathcal{A}}(p) := \inf\{v\colon b_{\mathcal{A}}(v \mid \underline{P}) \geq p\}$ for $p \geq \underline{P}$. Note that $b_{\mathcal{A}}(v \mid \underline{P})$ strictly increases on $[v^-_{\mathcal{A}}(\underline{P}), \overline{v}]$: else, there is a bid $\hat{p}$ and values $v'' > v' > v^-_{\mathcal{A}}(\underline{P})$ with $b_{\mathcal{A}}(\hat{v} \mid \underline{P}) = \hat{p}$ for all $\hat{v} \in (v', v'')$. But for all $\varepsilon > 0$, $q^+(\hat{p} + \varepsilon \mid b_{\mathcal{A}}(\cdot \mid \underline{P})) \leq q^-(\hat{p} + \varepsilon \mid b_{\mathcal{A}}(\cdot \mid \underline{P})) \leq q^+(\hat{p} \mid b_{\mathcal{A}}(\cdot \mid \underline{P})) \leq q^-(\hat{p} \mid b_{\mathcal{A}}(\cdot \mid \underline{P})) - \Lambda \int_{v'}^{v''} \phi(v)\, dv$. Hence, by (13), $\lim_{\varepsilon \to 0^+} \overline{W}_{\mathcal{A}}(\hat{p} + \varepsilon \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) \leq \overline{W}_{\mathcal{A}}(\hat{p} \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) - k$ for some $k > 0$. A customer with value $\hat{v} \in (v', v'')$ can lower her delay cost by at least $k(\hat{v}d + c)$ by bidding only ε more than $\hat{p}$. Now define $v_{\mathcal{A}}(p) := v^-_{\mathcal{A}}(p) = \inf\{v\colon b_{\mathcal{A}}(v \mid \underline{P}) > p\}$ and $q^-(p \mid b_{\mathcal{A}}(\cdot \mid \underline{P})) = q^+(p \mid b_{\mathcal{A}}(\cdot \mid \underline{P})) = \Lambda\overline{\Phi}(v_{\mathcal{A}}(p))$ for $p \geq \underline{P}$, and $v_{\mathcal{A}}(b_{\mathcal{A}}(v \mid \underline{P})) = v$ for $v > v_{\mathcal{A}}(\underline{P})$. Substitute in (13) to get (18).

(2) *Continuity* of $\overline{W}_{\mathcal{A}}(b_{\mathcal{A}}(v \mid \underline{P}) \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P}))$ follows from (18) and continuity of ϕ. By (15), we have $\lim_{v \to v_{\mathcal{A}}(\underline{P})} \overline{W}_{\mathcal{A}}(b_{\mathcal{A}}(v \mid \underline{P}) \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) = \overline{W}_{\mathcal{A}}(\underline{P} \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) \leq d^{-1} < 1$, so $\Lambda\overline{\Phi}(v_{\mathcal{A}}(\underline{P})) < 1$.

(3) *Differentiability*. Let $I_{\mathcal{A}}(z \mid \underline{P}) = I(b_{\mathcal{A}}(z \mid \underline{P}) \geq \underline{P})$. By (15), for all $v \neq z$,

$$\begin{aligned}(z - v)(1 - d\overline{W}_{\mathcal{A}}(z \mid \underline{P}))I_{\mathcal{A}}(z \mid \underline{P}) &\geq u_{\mathcal{A}}(z \mid \underline{P}) - u_{\mathcal{A}}(v \mid \underline{P}) \\ &\geq (z - v)(1 - d\overline{W}_{\mathcal{A}}(v \mid \underline{P}))I_{\mathcal{A}}(v \mid \underline{P}). \qquad \text{(A3)}\end{aligned}$$

With $z > v > v_{\mathcal{A}}(\underline{P})$, divide by $(z - v)$ and take limits as $z \to v$ to get $u'_{\mathcal{A}}(v \mid \underline{P}) = 1 - d\overline{W}_{\mathcal{A}}(v \mid \underline{P})$ (the limits exist by continuity) and, by (14), $b'_{\mathcal{A}}(v \mid \underline{P}) = -(v \cdot d + c) \cdot \overline{W}'_{\mathcal{A}}(v \mid \underline{P})$ for $v > v_{\mathcal{A}}(\underline{P})$. By continuity and (18), $W'_{\mathcal{A}}(v \mid \underline{P})$ and $b'_{\mathcal{A}}(v \mid \underline{P})$ are continuous at $v > v_{\mathcal{A}}(\underline{P})$ with finite limits $\lim_{v \to v_{\mathcal{A}}(\underline{P})} \overline{W}'_{\mathcal{A}}(v \mid \underline{P})$ and $\lim_{v \to v_{\mathcal{A}}(\underline{P})} b'_{\mathcal{A}}(v \mid \underline{P})$. Next, let $\underline{P}^0 := \inf\{b_{\mathcal{A}}(v \mid \underline{P})\colon v \geq v_{\mathcal{A}}(\underline{P})\}$. By continuity, $\lim_{v \to v_{\mathcal{A}}(\underline{P})}[u_{\mathcal{A}}(v, \underline{P} \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) - u_{\mathcal{A}}(v \mid \underline{P})] = \underline{P}^0 - \underline{P}$, hence, $\underline{P} = \underline{P}^0 = \lim_{v \to v_{\mathcal{A}}(\underline{P})} b_{\mathcal{A}}(v \mid \underline{P})$. Hence,

$$b_{\mathcal{A}}(v \mid \underline{P}) = b_{\mathcal{A}}(v_0 \mid \underline{P}) + \int_{v_0}^{v} -(x \cdot d + c) \cdot \overline{W}'_{\mathcal{A}}(x \mid \underline{P})\, dx \quad \text{for } v > v_0 > v_{\mathcal{A}}(\underline{P}) \qquad \text{(A4)}$$

is well defined because the right-hand side has finite limit as $v_0 \to v_{\mathcal{A}}(\underline{P})$. This yields (16).

(4) *Uniqueness*. In (A3), set $z = v_{\mathcal{A}}(\underline{P})$ and take limits as $v \uparrow v_{\mathcal{A}}(\underline{P})$ to get $u_{\mathcal{A}}(v_{\mathcal{A}}(\underline{P}) \mid \underline{P}) = 0$. Hence, $\lim_{v \downarrow v_{\mathcal{A}}(\underline{P})} b_{\mathcal{A}}(v \mid \underline{P}) = \underline{P} = v_{\mathcal{A}}(\underline{P}) - (v_{\mathcal{A}}(\underline{P})d + c)\overline{W}_{\mathcal{A}}(\underline{P} \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P}))$, and $v_{\mathcal{A}}(\underline{P})$ is the unique solution of (17) because its right-hand side, call it $nv_{\mathcal{A}}(v)$, is continuous with $nv'_{\mathcal{A}}(v) > 0$, $nv_{\mathcal{A}}(\overline{v}) > \underline{P}$, and $\lim_{v \downarrow \overline{\Phi}^{-1}(\Lambda^{-1})} nv_{\mathcal{A}}(v) < \underline{P}$. Note that $u_{\mathcal{A}}(v_{\mathcal{A}}(\underline{P}), p \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P})) = 0$ for $p \in \{0, \underline{P}\}$.

The function (16)–(18) is an equilibrium: for fixed v, (15) holds for $p \leq b_{\mathcal{A}}(\overline{v} \mid \underline{P})$ because (A3) is met for all $z \neq v$, and for $p > b_{\mathcal{A}}(\overline{v} \mid \underline{P})$ because $\overline{W}_{\mathcal{A}}(p \mid \underline{P}, b_{\mathcal{A}}(\cdot \mid \underline{P}) = \overline{W}_{\mathcal{A}}(\overline{v} \mid \underline{P})$.

(For Part (2), see the argument following Proposition 2.) □

Proof of Proposition 3. Because V and R are continuously differentiable and strictly concave, it follows from (22) that $NV_{\mathcal{A}}(\lambda)$ and $\Pi_{\mathcal{A}}(\lambda)$ are continuously differentiable and strictly concave for $\mathcal{A} \in \{AP, AN\}$.

Part (1). *AN* auction. Note that $\Pi'_{AN}(\lambda^*_{AN}) > NV'_{AN}(\lambda^*_{AN}) = 0 \Leftrightarrow \lambda^M_{AN} > \lambda^*_{AN}$. Define

$$\begin{aligned}L_{AN}(\lambda) &:= \frac{NV'_{AN}(\lambda) - \Pi'_{AN}(\lambda)}{d \cdot \int_0^\lambda \big(V'(q) - R'(q)\big)/(1-q)^2\, dq} \\ &= \frac{V'(\lambda)(1 - d + d\lambda)/(1-\lambda)^2}{d \cdot \int_0^\lambda V'(q)(\epsilon(\lambda)/\epsilon(q))/(1-q)^2\, dq} - 1, \qquad \text{(A5)}\end{aligned}$$

$$Q_{AN}(\lambda) := \frac{V'(\lambda)\big(1 - d + d\lambda/(1-\lambda)^2}{d \cdot \int_0^\lambda V'(q)/(1-q)^2\, dq} - 1. \qquad \text{(A6)}$$

Note that $L_{AN}(\lambda) < Q_{AN}(\lambda) \Leftrightarrow \epsilon(\lambda)$ is strictly increasing and that $L_{AN}(\lambda) \equiv Q_{AN}(\lambda) \Leftrightarrow \epsilon(\lambda)$ is a constant. By the first-order condition (26), $Q_{AN}(\lambda^*_{AN}) \geq 0$ with equality iff $c = 0$. For the result, replace L by L_{AN} and Q by Q_{AN} in the proof of Proposition 1.

Part (2). *AP* auction. If $c = 0$, then $d\overline{W}_{AP}(\lambda^*_{AP}) = 1$, because $V'(\lambda) > 0$ for $\lambda < \Lambda$ and $\Lambda \geq 1$. If $(\lambda^*_{AP}) \geq 1$, then $R'(\lambda^*_{AP}) \geq 0$ and $\lambda^*_{AP} = \lambda^M_{AP}$. Otherwise, $R'(\lambda^*_{AP}) < 0 \Rightarrow \lambda^*_{AP} > \lambda^M_{AP}$. If $c > 0$, then $1 > d\overline{W}_{AP}(\lambda^*_{AP})$ and $R'(\lambda^*_{AP}) < V'(\lambda^*_{AP})$, implying that $\Pi'_{AP}(\lambda^*_{AP}) < 0$ and $\lambda^*_{AP} > \lambda^M_{AP}$. □

Proof of Proposition 4. For NV (for Π, replace V by R), λ^*_U, λ^*_{AN}, and λ^*_{AP} are the solutions of (4) and (26):

$$F_U(\lambda) := V'(\lambda) - d\left[\frac{V'(\lambda)}{1-\lambda} + \frac{V(\lambda)}{(1-\lambda)^2}\right] = \frac{c}{(1-\lambda)^2}, \qquad \text{(A7)}$$

$$\begin{aligned}F_{AN}(\lambda) &:= V'(\lambda) - d\left[V'(\lambda)\left(1 + \frac{\lambda}{(1-\lambda)^2}\right) + \int_0^\lambda \frac{V'(q)}{(1-q)^2}\, dq\right] \\ &= \frac{c}{(1-\lambda)^2}, \qquad \text{(A8)}\end{aligned}$$

$$F_{AP}(\lambda) := V'(\lambda) - d\frac{V'(\lambda)}{(1-\lambda)^2} = \frac{c}{(1-\lambda)^2}. \qquad \text{(A9)}$$

$F_U(\lambda)$, $F_{AN}(\lambda)$, and $F_{AP}(\lambda)$ strictly decrease and $c/(1-\lambda)^2$ strictly increases in λ. Because $F_{AP}(\lambda) > F_{AN}(\lambda) > F_U(\lambda)$ for $\lambda > 0$, $\lambda^*_{AP} > \lambda^*_{AN} > \lambda^*_U$ and $NV_{AP}(\lambda^*_{AP}) > NV_{AN}(\lambda^*_{AN}) > NV_U(\lambda^*_U)$. □

Proof of Proposition 5. Part (1). Fix $c_L < c_H$. For $k = L, H$, let $\lambda^k_{\mathcal{M}} = \lambda^*_{\mathcal{M}}(c_k)$, $\mathcal{M} \in \{AN, AP, U\}$. Then,

$$\Delta NV^*_{\mathcal{A}}(k) = \int_0^{\lambda^k_U} f_{\mathcal{A}}(x)\, dx + \int_{\lambda^k_U}^{\lambda^k_{\mathcal{A}}} F_{\mathcal{A}}(x) - \frac{c_k}{(1-x)^2}\, dx > 0, \quad k = L, H,\ \mathcal{A} \in \{AN, AP\}, \qquad \text{(A10)}$$

where $\Delta NV^*_{\mathcal{A}}(k) = NV_{\mathcal{A}}(\lambda^k_{\mathcal{A}}; c_k) - NV_U(\lambda^k_U; c_k)$ and $f_{\mathcal{A}}(\lambda) := F_{\mathcal{A}}(\lambda) - F_U(\lambda)$. From (A10),

$$\begin{aligned}&\int_{\min(\lambda^H_{\mathcal{A}}, \lambda^L_U)}^{\max(\lambda^H_{\mathcal{A}}, \lambda^L_U)} f_{\mathcal{A}}(x)I(\lambda^H_{\mathcal{A}} < \lambda^L_U) + \frac{c_H - c_L}{(1-x)^2}I(\lambda^H_{\mathcal{A}} > \lambda^L_U)\, dx \\ &\quad + \int_{\lambda^H_U}^{\min(\lambda^H_{\mathcal{A}}, \lambda^L_U)} \frac{c_H}{(1-x)^2} - F_U(x)\, dx \\ &\quad + \int_{\max(\lambda^H_{\mathcal{A}}, \lambda^L_U)}^{\lambda^L_{\mathcal{A}}} F_{\mathcal{A}}(x) - \frac{c_L}{(1-x)^2}\, dx \\ &= \Delta NV^*_{\mathcal{A}}(L) - \Delta NV^*_{\mathcal{A}}(H) > 0,\end{aligned}$$

because $\lambda_U^H < \lambda_U^L$, $\lambda_{\mathscr{A}}^H < \lambda_{\mathscr{A}}^L$ and all integrands are positive.

Part (2). If $d = 0$ or $d = 1 - c/V'(0)$, then $\lambda_{\mathscr{A}}^* = \lambda_U^*$ and $NV_{\mathscr{A}}(\lambda_{\mathscr{A}}^*; d) = NV_U(\lambda_U^*; d)$. For $\mathscr{M} \in \{AN, AP, U\}$, because $NV_{\mathscr{M}}(\lambda; d)$ is strictly concave in λ and differentiable in d, there is a unique differentiable implicit function $\lambda_{\mathscr{M}}^*(d)$ defined by (4) or (26) on the interval $(0, 1-c/V'(0))$. Hence, the difference $NV_{\mathscr{A}}(\lambda_{\mathscr{A}}^*(d); d) - NV_U(\lambda_U^*(d); d)$ is continuous and bounded on $[0, 1 - c/V'(0)]$, and by Proposition 4, it has a maximum at some $d_{\mathscr{A}}^* \in (0, 1 - c/V'(0))$. □

Proof of Proposition 6. Fix $\mathscr{R} = *$ (everything holds for $\mathscr{R} = M$).

Part (1). Equation (22) ⇒ (38) and (39). For $q \in [0, \lambda_U^*]$,

$$u_{\mathscr{A}}^*(q) = -\int_q^{\lambda_{\mathscr{A}}^*} V''(z)(1 - d\overline{W}_{\mathscr{A}}(z, \lambda_{\mathscr{A}}^*))\,dz + \int_q^{\lambda_U^*} V''(z)(1 - d\overline{W}_U(\lambda_U^*))\,dz, \qquad \text{(A11)}$$

hence, $u_{\mathscr{A}}^*(\lambda_U^*) > 0$ and $u_{\mathscr{A}}^{*\prime}(q) = V''(q)d[\overline{W}_U(\lambda_U^*) - \overline{W}_{\mathscr{A}}(q, \lambda_{\mathscr{A}}^*)]$. It follows that $u_{\mathscr{A}}^{*\prime}(q_{\mathscr{A}}^*) = 0$ with $u_{\mathscr{A}}^{*\prime}(q) > 0$ for $q > q_{\mathscr{A}}^*$ and, conversely, for $q < q_{\mathscr{A}}^*$, establish that $u_{\mathscr{A}}^*(q_{\mathscr{A}}^*) = \min_{q \in [0, \lambda_U^*]} u_{\mathscr{A}}^*(q)$.

Part (2). *AP*. By the conservation law and convexity of V', the top-end gain is

$$\begin{aligned} u_{AP}(0, \lambda_U^*) - u_U(0, \lambda_U^*) &= d\int_0^{\lambda_U^*} V''(x)\left[\frac{1}{(1-x)^2} - W_U(\lambda_U^*)\right]dx \\ &\geq d\int_0^{q_{AP}^*} (V''(x) - V''(q_{AP}^*))\left[\frac{1}{(1-x)^2} - \overline{W}_U(\lambda_U^*)\right]dx \geq 0. \end{aligned}$$

Because $\partial u_{AP}(q, \lambda)/\partial\lambda = -V''(\lambda)[1 - d/(1-\lambda)^2] > 0$ and $\lambda_{AP}^* > \lambda_U^*$, it follows that $u_{AP}^*(0) > 0$. If $b_{AP}(q_{AP}^*, \lambda_{AP}^*) \leq P_U(\lambda_U^*)$, then $\min_{q \in [0, \lambda_U^*]} u_{AP}^*(q) \geq 0$. Else, by continuity, there are unique lower and upper bounds, $l_{AP}^* < q_{AP}^* < \bar{l}_{AP}^*$, that specify the interval of losers as stated.

Part (3). *AN*. (a) The result has the same structure as for *AP*: from (A11) we have

$$u_{AN}^*(0) = B\int_{\lambda_U^*}^{\lambda_{AN}^*} \left(1 - \frac{d}{(1-q)^2}\right)dq > 0, \qquad \text{(A12)}$$

and $\min_{q \in [0, \lambda_U^*]} u_{AN}^*(q) < 0$ iff $b_{AN}(q_{AN}^*, \lambda_{AN}^*) > P_U(\lambda_U^*)$ and $q_{AN}^* > 0$.

(b) Integrate (A11) by parts:

$$\begin{aligned} u_{\mathscr{A}}^*(0) = &-\int_{\lambda_U^*}^{\lambda_{AN}^*} V''(x)(1 - d\overline{W}_{AN}(x, \lambda_{AN}^*))\,dx \\ &+ dV'(\lambda_U^*)\left[\overline{W}_{AN}(\lambda_U^*, \lambda_{AN}^*) - \overline{W}_U(\lambda_U^*)\right] \\ &+ d\cdot\left[\lambda_{AN}^*\cdot\int_0^{\lambda_U^*} \frac{6V(x)}{(1-x)^4}\,dx - \lambda_{AN}^*\frac{2V(\lambda_U^*)}{(1-\lambda_U^*)^3}\right. \\ &\left. - V'(0)\left[\overline{W}_{AN}(0, \lambda_{AN}^*) - \overline{W}_U(\lambda_U^*)\right]\right]. \end{aligned}$$

Here, $\lambda_{AN}^* < \lambda_U^*/(1 - \lambda_U^*)$ gives $u_{\mathscr{A}}^*(0) = \infty$, yielding the same result as for *AP*, while $\lambda_{AN}^* < \lambda_U^*/(1 - \lambda_U^*)$ results in $u_{\mathscr{A}}^*(0) = -\infty$. □

References

Afèche, P. 2003. Delay performance in stochastic processing networks with priority service. *Oper. Res. Lett.* **31** 390–400.

Balachandran, K. R. 1972. Purchasing priorities in queues. *Management Sci.* **18** 319–326.

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

De Vany, A. S. 1976. Uncertainty, waiting time, and capacity utilization: A stochastic theory of product quality. *J. Political Econom.* **84** 523–541.

Dewan, S., H. Mendelson. 1990. User delay costs and internal pricing for a service facility. *Management Sci.* **36** 1502–1517.

Dewan, S., H. Mendelson. 1998. Information technology and time-based competition in financial markets. *Management Sci.* **44** 595–609.

Dolan, R. J. 1978. Incentive mechanisms for priority queuing problems. *Bell J. Econom.* **9** 421–436.

Glazer, A., R. Hassin. 1985. Stable priority purchasing in queues. *Oper. Res. Lett.* **4** 285–288.

Gupta, A., D. O. Stahl, A. B. Whinston. 1996. An economic approach to networked computing with priority classes. *J. Organ. Comput.* **6** 71–95.

Ha, A. 2001. Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Sci.* **47** 915–930.

Hassin, R. 1995. Decentralized regulation of a queue. *Management Sci.* **41** 163–173.

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

Kleinrock, L. 1967. Optimum bribing for queue position. *Oper. Res.* **15** 304–318.

Knudsen, N. C. 1972. Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* **40** 515–528.

Lederer, P. J., L. Li. 1997. Pricing, production, scheduling and delivery-time competition. *Oper. Res.* **45** 407–420.

Li, L., Y. S. Lee. 1994. Pricing and delivery-time performance in a competitive environment. *Management Sci.* **40** 633–646.

Lippman, S. A., S. Stidham. 1977. Individual versus social optimization in exponential congestion systems. *Oper. Res.* **25** 233–247.

Loch, C. 1991. Pricing in markets sensitive to delay. Ph.D. dissertation, Stanford University, Stanford, CA.

Lui, F. T. 1985. An equilibrium queuing model of bribery. *J. Political Econom.* **93** 760–781.

Masuda, Y., S. Whang. 1999. Dynamic pricing for network service: Equilibrium and stability. *Management Sci.* **45** 857–869.

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

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

Klemperer, P. 1999. Auction theory: A guide. *J. Econom. Surveys* **13** 227–286.

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

Rao, S., E. Petersen. 1998. Optimal pricing of priority services. *Oper. Res.* **46** 46–56.

Van Mieghem, J. A. 2000. Price and service discrimination in queuing systems: Incentive compatibility of $Gc\mu$ scheduling. *Management Sci.* **46** 1249–1267.

Westland, J. C. 1992. Congestion and network externalities in the short run pricing of information system services. *Management Sci.* **38** 992–1009.

Yechiali, U. 1972. Customers' optimal joining rules for the $GI/M/s$ queue. *Management Sci.* **18** 434–443.