

Service Competition with General Queueing Facilities

Gad Allon

Kellogg School of Management, Northwestern University, Evanston, Illinois 60208, g-allon@kellogg.northwestern.edu

Awi Federgruen

Graduate School of Business, Columbia University, New York, New York 10027, af7@columbia.edu

In many service industries, companies compete with each other on the basis of the waiting time their customers experience, along with the price they charge for their service. A firm's waiting-time standard may either be defined in terms of the expected value or a given, for example 95%, percentile of the steady state waiting-time distribution. We investigate how a service industry's competitive behavior depends on the characteristics of the service providers' queueing systems. We provide a unifying approach to investigate various standard single-stage systems covering the spectrum from $M/M/1$ to general $G/GI/s$ systems, along with open Jackson networks to represent multistage service systems. Assuming that the capacity cost is proportional with the service rates, we refer to its dependence on (i) the firm's demand rate, and (ii) the waiting-time standard as the capacity cost function. We show that across the above broad spectrum of queueing models, the capacity cost function belongs to a specific four-parameter class of function, either exactly or as a close approximation. We then characterize how this capacity cost function impacts the equilibrium behavior in the industry. We give separate treatments to the case where the firms compete in terms of (i) prices (only), (ii) their service level or waiting-time standard (only), and (iii) simultaneously in terms of both prices and service levels. The firms' demand rates are given by a general system of equations of the prices and waiting-time standards in the industry.

Subject classifications: queues; multichannel games; noncooperative marketing; competitive strategy.

Area of review: Stochastic Models.

History: Received April 2004; revisions received April 2006, August 2006; accepted September 2006.

1. Introduction and Summary

In many service industries, companies compete with each other on the basis of the waiting time their customers experience, along with the price they charge for their service. Often, specific waiting-time standards or guarantees are advertised. For example, Ameritrade has increased its market share in the online discount brokerage market by “guaranteeing” that trades take no more than 10 seconds to be executed; the guarantee is backed up with a complete waiver of commissions in case the time limit is violated. This has led most major online brokerage firms (E-trade, Fidelity) to offer and aggressively advertise even more ambitious waiting-time standards. Lucky and other supermarket chains have started to guarantee that no line in front of a checkout counter has more than three customers waiting. It is this service guarantee which is emphasized in their “3’s a crowd” advertising campaign. Various call centers promise that the customer will be helped within one hour, say, possibly by a callback. See Allon and Federgruen (2007) for a longer list of examples.

A firm's waiting-time standard may either be defined in terms of the expected value or a given, for example, 95%, percentile of the steady-state waiting-time distribution. Firms commit themselves to a given waiting-time guarantee by selecting appropriate capacity levels. In the economics

literature, Luski (1976), Levhari and Luski (1978), and De Vany and Saving (1983) initiated the study of the competitive behavior of oligopolies of service providers, incorporating the dependence of both revenues and capacity-related costs on the firms' waiting-time standards. These seminal papers have been followed by many others, both in the economics and the operations management literature. Virtually without exception, these papers model a firm's service facility as an $M/M/1$ queueing system, i.e., a system with a single service stage, a single server, Poisson arrivals, and independent exponential service times. The facility's capacity level is given by the service rate; it is well known that in an $M/M/1$ system, the required service rate to meet a given waiting-time standard is given by a simple linear function of (i) the firm's demand rate, and (ii) the reciprocal of the waiting-time standard (irrespective of whether the expected waiting time or a given percentile of the waiting-time distribution is used as a standard). Clearly, this linear relationship ceases to apply when the service facility needs to be represented by a more general queueing system. Consider, for example, the case where service times fail to be exponential, resulting in an $M/G/1$ system, while the waiting-time standard is expressed in terms of the expected waiting time. It can easily be shown that the service rate varies convexly with the demand rate when the coefficient of variation of

the service-time distribution is less than one, but concavely otherwise. In other words, the coefficient of variation determines whether the cost structure exhibits economies of scale or diseconomies of scale. The importance of this distinction in the context of service competition has been emphasized by Cachon and Harker (2002).

In this paper, we investigate how the industry's competitive behavior depends on the characteristics of the service providers' queueing systems. We provide a unifying approach to investigate various standard single-stage systems covering the spectrum from M/M/1 to general G/GI/s systems, along with open Jackson networks to represent multistage service systems. (A G/GI/s system has a general stationary arrival process and independent and identically distributed (i.i.d.) generally distributed service times.) Assuming that each firm i 's capacity cost is proportional with the service rate, we refer to its dependence on (i) the firm's demand rate λ_i , and (ii) the waiting-time standard w_i or its service level $\theta_i = w_i^{-1}$ as the *capacity cost function*. We show that across the above broad spectrum of queueing models, the capacity cost functions belong to the following specific four-parameter class of function, either exactly or as a close approximation:

$$(\mathcal{C}) \quad C_i(\lambda_i, \theta_i) = B_1\lambda_i + B_2\theta_i + \sqrt{B_3\lambda_i^2 + B_4\lambda_i\theta_i + B_2^2\theta_i^2}$$

for positive B_1, B_2, B_3 , and B_4 ,
a (possibly negative) constant.

We then characterize how this capacity cost function impacts the equilibrium behavior in the industry. We give separate treatments to the case where the firms compete in terms of (i) prices (only), (ii) their service level or waiting-time standard (only), and (iii) simultaneously in terms of both prices and service levels. To avoid repetitions, refer to the Conclusions section for a summary of the qualitative properties of the equilibrium in the three types of competitive models.

Although the characteristics of the queueing systems of the service providers and the resulting type of capacity cost function are a key determinant for the industry's equilibrium behavior, the same is true for the consumer-choice model, which prescribes how the firms' demand volumes depend on the various price and service levels offered. Most prior work on service competition has assumed that customers, when evaluating a firm, consider a single aggregation of its price and its waiting-time standard, usually referred to as the *full price*. The full price is usually defined as the price plus a multiple of the waiting-time standard under the explicit or implicit assumption that the "cost of waiting" is strictly proportional to the waiting time, and that consumers are able to identify the waiting-time cost rate per unit of time. Furthermore, the full price is often considered the only criterion according to which a service provider is selected, leaving all other attributes aside. Such attributes include the reputation of the service provider, the customers' awareness of his existence, the quality of

the service, the location of the service provider, and the user friendliness of the service process. See Allon and Federgruen (2007) for concrete examples.

These authors have argued that the consumer-choice model should allow for general trade-offs among (i) the price of service, (ii) the waiting-time standard, and (iii) the "other attributes." This would result in the firms' demand rates being specified by a system of general functions of the price and service-level vectors. In Allon and Federgruen (2007), the authors proceed to characterize the equilibrium behavior in the competitive settings listed above for demand equations that are separable functions of the prices and service levels, which in addition are linear in the price vector.

The consumer-choice literature (see Anderson et al. 1992 or Leeflang et al. 2000) has demonstrated that nonlinear relationships often prevail. Among such nonlinear demand functions, the class of *attraction models*, which includes the class of multinomial logit functions as an important special case, is particularly prevalent and supported by econometric work as well as axiomatic foundations (see Bell et al. 1975). Here, a firm's attractiveness is characterized by an attraction value specified as a function of his price and/or his service level, and each firm's market share is entirely based on the relative gap between this attraction value and the average value in the industry. In characterizing the equilibrium behavior in the various abovementioned competitive settings, we therefore provide comparisons between the case where the demand functions are linear in the prices and that where they are given by an attraction model.

The remainder of this paper is organized as follows. In §2, we give a brief review of the relevant literature. In §3, we introduce the models and demonstrate that across the above broad spectrum of single and multistage queueing systems, the capacity cost function belongs to the specific class of four parameter functions (\mathcal{C}). Section 4 characterizes the equilibrium behavior in the various competition models under demand functions that are price-linear. Section 5 does the same under demand functions specified by an attraction model. Section 6 reports on a numerical study, and §7 summarizes our conclusions.

2. Literature Review

As mentioned in the introduction, Luski (1976) and Levhari and Luski (1978) were the first to model competition between service providers. Both papers address a *duopoly*, where each of the firms acts as an M/M/1 system, with given identical service rates. Customers select their service provider strictly on the basis of the *full price*, defined as the direct price plus the expected steady-state waiting time multiplied with the waiting-time cost rate. In Luski (1976), all customers share a common cost rate, while in Levhari and Luski (1978), these are assumed to be independent and identically distributed (i.i.d.) with a given distribution. The question whether a price equilibrium exists in these models

remained an open question, until, for the basic model with a uniform cost rate, it was recently resolved in the affirmative by Chen and Wan (2003). These authors show, however, that the Nash equilibrium may fail to be unique. More recent variants of the Levhari and Luski models include Li and Lee (1994) and Armony and Haviv (2003); see the survey text by Hassin and Haviv (2003) for details.

In the above papers, firms compete in terms of their price (only), with fixed exogenously specified capacity levels (§7 in Chen and Wan (2003) relaxes this assumption; see below). Several other papers assume, alternatively, that prices are fixed while firms compete in terms of their capacity levels. Kalai et al. (1992) consider, again, a duopoly with Poisson arrivals and exponential service times. A fixed customer population joins, upon arrival, a single queue from which they are served on a FIFO basis by the first available server. (When a customer arrives to an empty queue, he is randomly assigned to one of the two providers.) In this model, asymmetric Nash equilibria of service-rate pairs may arise. Christ and Avi-Itzhak (2002) show that a unique symmetric equilibrium exists in a variant of this model in which the servers are equally expensive, but only a queue length dependent fraction of arriving customers actually joins the queue. Gilbert and Weng (1997) show that a unique symmetric equilibrium arises in the variant of the model where, upon arrival, customers are routed to one of the two service providers with a probability that equates expected waiting times at each. Cachon and Zhang (2003) generalizes Gilbert and Weng (1997) to allow for routing probabilities that depend on the providers' service rates according to more general (allocation) schemes.

Chen and Wan (2005) characterize the equilibrium behavior in the Luski model in which the two firms select a capacity level under fixed prices. De Vany and Saving (1983) and Reitman (1998) consider variants of the Luski model with an arbitrary number of identical firms, which simultaneously compete in terms of prices and service rates. Section 7 in Chen and Wan (2003) provides a complete characterization of the equilibrium behavior under simultaneous price and capacity competition for the case of two identical service providers.

To our knowledge, Loch (1991) and Lederer and Li (1997) are the only service competition papers which model the service provider via more general than basic M/M/1 queueing systems. Loch (1991) considers a variant of the Luski model in which the service times of the two providers have a general, although still identical, distribution, i.e., in which each provider is modelled as an M/G/1 system. Assuming that the total demand rate for service is given by a general function of the full price, the author shows that a symmetric equilibrium pair of prices exists, irrespective of whether the two firms target prices directly (Bertrand competition), or indirectly, via demand rates (Cournot competition). Lederer and Li (1997) generalize Loch (1991) to allow for an arbitrary number of service providers and a finite number of customer classes, each with a given

waiting cost rate. Here, the total demand rate for each customer class is given by a general function of the full price that applies to this class. Each firm optimally prioritizes among the customer classes on the basis of the cost rate, in accordance with the well-known $c\mu$ rule. In the case of Cournot competition, the authors show that a Nash equilibrium exists, in which no customer has an incentive to switch to a different provider to reduce his full price or to misrepresent his class identity. Mendelson and Shneerson's (2003) model for the competition between Internet service network operators may be viewed as an adaptation of the Lederer and Li (1997) model with a single-customer class.

Cachon and Harker (2002) and So (2000) analyzed the first models in which customers choose a service provider on the basis of criteria other than the lowest full price. Both confined themselves, again, to M/M/1 service providers. Cachon and Harker (2002) considered the case of two firms, where demand rates are given as either linear or multinomial logit functions of the two full prices. So (2000) considers an arbitrary number of competing firms and a different class of attraction models in which the logarithm of the firms' attraction value is specified as a common linear combination of the logarithm of the price and the logarithm of the waiting time standard, plus a firm-dependent constant. This specification of So (2000) continues to imply that the price and waiting time are aggregated into a single, albeit different full-price measure. Allon and Federgruen (2007) appears to be the first model to treat the price and waiting-time standard as completely independent firm attributes, which different customers may trade off in different ways. Nevertheless, this paper confined itself to systems of demand rates that are linear in the prices and to M/M/1 service providers. We refer to Hassin and Haviv (2003) for a recent survey text on queueing models with competition.

Several papers have modeled industries of consumer goods distributors who compete with a "quality" or "service" instrument along with their price. Banker et al. (1998) and Tsay and Agrawal (2000) characterize the equilibrium behavior in a single-period duopoly. The former (latter) assumes that the price and quality choices are made sequentially (simultaneously). Demand functions are assumed to be linear in all prices and quality variables; they therefore represent a special case of the quasi-separable specification (5) below. Both papers assume that each retailer's cost increases quadratically with the service or quality level provided. This structural assumption does not follow from an underlying operational infrastructure, but is substantiated by the fact that "under the assumptions of standard inventory models, moving from, say a 97% to a 99% fill rate, typically requires a greater incremental investment than moving from 95% to 97%" (see Footnote 3 in Tsay and Agrawal, p. 375). Anderson et al. (1992) consider an oligopoly model with identical firms and demand functions of the multinomial logit type, a special case of the second class of demand models, addressed below (i.e., the so-called attraction models; see (10)). The costs are

assumed to grow proportionally with the sales volume, at a rate which depends on the chosen quality level according to a general convex cost function. Bernstein and Federgruen (2004b) address an infinite-horizon model with an arbitrary number of competing firms, in which the service measure is given by the steady-state fill rate. Each firm faces a sequence of i.i.d. demands; the demand in each given period is the product of a mean, which is given by a function of the vector of prices and fill rates in the industry, and a random component whose distribution is independent of the price and fill rate choices. The authors show that this competition model is equivalent to a single-stage model in which the cost grows linearly with the sales volume and convexly with the chosen fill rate. As explained above, there are therefore important differences between the costs that arise in service industries with firms pursuing given waiting-time standards, and those in consumer goods industries, in which firms commit to a given fill rate. (Most of the Bernstein and Federgruen 2004a paper confines itself to the quasi-separable and attraction model demand functions addressed in this paper.)

Our analysis depends heavily on so-called exponential approximations for the tail probability of steady-state waiting times, also referred to as Cramer-Lundberg approximations. Thus, if W is the steady-state waiting time (i.e., the delay or sojourn time) in a queueing system, the exponential approximation states

$$\mathbb{P}(W > x) \sim \alpha e^{-\eta x} \quad (1)$$

for sufficiently large x , where α and η are constants that depend on the characteristics of the queueing system, i.e., $\lim_{x \rightarrow \infty} e^{\eta x} \mathbb{P}(W > x) = \alpha$. A voluminous literature supports this approximation. The identity is known to be exact for the GI/M/s system with arrivals arising from an arbitrary renewal process with general interarrival times. Smith (1953) already established for the GI/GI/1 system with service-time distributions whose Laplace transform is rational that the ratio of the left-hand and right-hand side expressions in (1) goes to one as x tends to infinity. (A GI/GI/1 system has a renewal arrival process with an arbitrary interarrival time distribution and i.i.d. service times, again, with an arbitrary common distribution.) Later treatments and refinements of this limit result can be found, for example, in Feller (1971), Borovkov (1976), and Asmussen (1987). Abundant numerical support for the remarkable accuracy of the exponential approximation (1) has been provided by Tijms (1986) and Seelen et al. (1985). Starting with Kingman's (1962) seminal paper, many heavy-traffic limit theorems substantiated the accuracy of the exponential approximation in (1) for any finite x when the utilization rate approaches one. See Abate et al. (1995, 1996) and the references cited there. These authors have developed simple approximations for the constants α and η in (1), which we will show to be especially useful when representing the capacity cost function in our various competition models.¹

Finally, our capacity cost functions bear resemblance to those employed by a variety of authors when the service rate of each individual server is given, but the number of servers can be varied to guarantee a given “no-wait” probability. Assuming that the servers' cost is proportional to the number of servers, the capacity cost is of the form

$$C_i = B\lambda_i + \Gamma\sqrt{\lambda_i} \quad (2)$$

(see, for example, Newell 1973, Halfin and Whitt 1981, Whitt 1992, and the references cited therein). The functions in (2) viewed as functions of λ_i are a subset of the class \mathcal{C} , with $B_2 = B_3 = 0$. At the same time, while the functions in (2) are always concave, exhibiting economies of scale, those in the general class \mathcal{C} permit concave as well as convex instances, as discussed above (see also Lemma 1 below). As mentioned, the concavity/convexity properties of the cost functions have important implications for the ability to guarantee the existence of equilibrium; see §7. The coefficient Γ depends on the “no wait” probability in a much more complex, nonlinear fashion compared to the dependence of the functions in \mathcal{C} on the service level θ . This creates further complications for the equilibrium analysis, in particular of the service competition models and the combined price and service competition model.

3. The Model and the Capacity Cost Function

We consider a service industry with N competing service providers.² Each firm i positions itself in the market by selecting a price p_i , as well as a service level θ_i . The price p_i has to be chosen from an interval $[p_i^{\min}, p_i^{\max}]$. Depending upon the application, customers may be particularly concerned about the time they spend waiting for service or their complete sojourn time in the system, which includes the time spent in service. For example, in the restaurant industry, customers are sensitive to the time they need to wait until their food is served. The service time (i.e., the time spent eating at their table) is usually experienced as a pleasure, rather than a nuisance. It is for this reason that restaurant chains such as “Black Angus” specify their service level in terms of the delay, offering a free lunch if it is not served within 10 minutes of arrival. In contrast, in the overnight delivery industry, customers are clearly concerned with the complete sojourn time of their packages (i.e., the time until ultimate delivery to the recipient). As mentioned in the introduction, here companies specify their service levels in terms of “guaranteed” delivery times at the final destination.

A second important distinction is whether the service levels are expressed in terms of the expected steady-state delay or sojourn time versus a ϕ th, (e.g., 0.95) fractile of the delay or sojourn-time distribution. Let

$$D_i = \text{steady-state delay in firm } i\text{'s service facility, } i = 1, \dots, N,$$

$$T_i = \text{steady-state sojourn time in firm } i\text{'s service facility, } i = 1, \dots, N.$$

Firm i 's waiting-time standard w_i may thus be expressed as (a) $w_i = \mathbb{E}(D_i)$, (b) $w_i = \mathbb{E}(T_i)$, (c) w_i is the value for which $\mathbb{P}(D_i > w_i) = 1 - \phi$, and (d) w_i is the value for which $\mathbb{P}(T_i > w_i) = 1 - \phi$. The service level θ_i is defined as the reciprocal of the actual waiting-time guarantee, i.e., $\theta_i = 1/w_i$, $i = 1, \dots, N$.

We assume that each firm i experiences arrivals of customers, generated by a stationary counting process, with a well-defined long-run average rate λ_i . Let $\{A_{\lambda_i}^i(t) : t > 0\}$ denote the arrival process experienced by firm i under arrival rate λ_i . These counting processes are time-scaled versions of a standardized counting process $\{A_1^i(t) : t \geq 0\}$, i.e., $\{A_{\lambda_i}^i(t) : t \geq 0\} = \{A_1^i(\lambda_i t) : t \geq 0\}$, where the processes are stochastically equivalent.

3.1. The Demand Model

As argued in the introduction, in general, the demand rate λ_i may depend on all of the industry's prices and service levels, i.e.,

$$\lambda_i = \lambda_i(\mathbf{p}, \theta), \quad i = 1, \dots, N. \quad (3)$$

As demonstrated below, the system of demand equations can be derived from one or several underlying consumer-utility models. Nevertheless, we treat the system (3) as a primitive of the model. Earlier work on service competition models has assumed that the price and service level can be aggregated into a single full-price measure $F_i = F(p_i, \theta_i)$, such that the demand rates depend on the vector of full prices only:

$$\lambda_i = \lambda_i(F_1, \dots, F_N), \quad i = 1, \dots, N. \quad (4)$$

Below, we will show that, in general, the specification in (3) is more general than (4). We focus on the following frequently used classes of demand functions:

(I) Separable demand functions

$$\lambda_i = a_i(\theta_i) - \sum_{i \neq j} \alpha_{ij}(\theta_j) - b_i p_i + \sum_{i \neq j} \beta_{ij} p_j, \quad i = 1, \dots, N. \quad (5)$$

Here, a_i is an increasing concave function, reflecting the fact that service-level improvements result in an increase in demand volume, however, with nonincreasing marginal returns to scale. The functions α_{ij} are general increasing functions, again reflecting the fact that firm i 's demand volume can only decrease in response to service-level improvements by any of its competitors.

Without loss of practical generality, we assume that a uniform price increase by all N firms cannot result in an increase in any firm's demand volume, i.e.,

$$(D) \quad b_i > \sum_{j \neq i} \beta_{ij}, \quad i = 1, \dots, N, \quad (6)$$

a condition that is usually referred to as the *dominant diagonal condition*. Similarly, a price increase by a given firm

cannot result in an increase of the industry's aggregate demand volume, i.e.,

$$(D') \quad b_i > \sum_{j \neq i} \beta_{ji}, \quad i = 1, \dots, N. \quad (7)$$

(5) may, e.g., be derived from a representative consumer model with utility function $U(\lambda, \theta) \equiv C + \frac{1}{2} \lambda^T B^{-1} \lambda - \lambda^T B^{-1} \bar{a}(\theta)$, where the $N \times N$ matrix B has $B_{ii} = -b_i$, and $B_{ij} = \beta_{ij}$, $i \neq j$, $\bar{a}(\theta) \equiv a_i(\theta_i) - \sum_{j \neq i} \alpha_{ij}(\theta_j)$, and $C > 0$. ((D) ensures that B^{-1} exists and is negative semidefinite, giving rise to a jointly concave utility function.) The demand functions (5) arise by optimizing the utility function subject to a budget constraint.

As mentioned, it is, in general, not possible to aggregate the price p_i and the service level θ_i into a single full-price measure $F_i = F(p_i, \theta_i)$, such that the system of Equations (5) is of the form (4).

To demonstrate this, consider the following two-firm example with demand functions of the quasi-separable type (5) and linear $a(\cdot)$ and $\alpha(\cdot)$ functions:

EXAMPLE 3.1. Let $N = 2$ and

$$\lambda_1 = 1,000 + 2\theta_1 - \theta_2 - 8p_1 + 6p_2, \quad (8)$$

$$\lambda_2 = 1,000 + 3\theta_2 - 2\theta_1 - 7p_2 + 5p_1. \quad (9)$$

To allow for aggregation of (p_1, θ_1) in (8) into a single full-price explanatory variable $F(p_1, \theta_1)$, the aggregation scheme needs to be proportional to $p_1 - 0.25\theta_1$, but such schemes fail to aggregate (p_2, θ_2) into a single full price. Similarly, λ_2 depends on both p_1 and θ_1 , not just on $F(p_1, \theta_1)$. Alternatively, the existence of a full-price aggregation scheme implies that any increase of a firm's price can be offset by an increase of the service level such that all market effects remain unchanged. However, if firm i increases its price p_i by one unit, an increase of θ_1 by four units is required to leave the demand rate of firm 1 unchanged, but this changes λ_2 by three units.

(II) Demand functions given by an attraction model

In an attraction model, each firm's market share (of a given potential number of M customers) is determined by the so-called attractiveness value v_i , itself, a general function of the firm's price p_i and service level θ_i , i.e., $v_i = v_i(p_i, \theta_i)$. For given positive constants M and v_0 , the demand rates of the firms are thus given by the system of equations

$$\lambda_i = M \frac{v_i(p_i, \theta_i)}{\sum_{j=1}^N v_j(p_j, \theta_j) + v_0}, \quad i = 1, \dots, N. \quad (10)$$

Without loss of generality, we assume that

$$\frac{\partial v_i}{\partial p_i} \leq 0, \quad \frac{\partial v_i}{\partial \theta_i} \geq 0. \quad (11)$$

Assuming once again, that a uniform price increase cannot result in an increase of any firm's demand, we obtain as the analogue of (D):

$$\frac{v_i}{\sum_{j=1}^N v_j + v_0} < \frac{\partial v_i / \partial p_i}{\sum_{j=1}^N \partial v_j / \partial p_j} \quad \forall i = 1, \dots, N. \quad (12)$$

The attraction model may be derived from the following consumer-utility model. Assume, for example, that there is a given potential market size Λ and an arbitrary customer derives a random utility $V_i = \log v_i(p_i, \theta_i) + \varepsilon_i$, $i = 1, \dots, N$, from receiving service from firm i , with ε_i a random variable with $\mathbb{E}(\varepsilon_i) = 0$ and with a distribution that is independent of the firm's price and service level. Similarly, the utility from receiving no service is given by $V_0 = \log v_0 + \varepsilon_0$, $\mathbb{E}(\varepsilon_0) = 0$. A customer patronizes firm i if $V_i = \max_{0 \leq j \leq N} V_j$ and foregoes service if $V_0 = \max_{0 \leq j \leq N} V_j$. It is well known that if the $\{\varepsilon_j, j = 0, \dots, N\}$ random variables are independent with a double exponential distribution, we obtain the demand functions (10). Note, however, that even when the demand functions can be conceived as resulting from this random utility model, aggregation of price and waiting time into a single full-price measure F_i is feasible only if all v_i functions coincide, i.e., $v_i(p, \theta) \equiv u(p, \theta)$. (In the latter case, any aggregation scheme $F(p, \theta) = A - Bv(p, \theta)$ may be used with $A, B > 0$ given constants.) Thus, in general, the demand model (10) cannot be specified as a full-price model (4).

The choice $v_i = e^{a_i(\theta_i) - b_i p_i}$ gives rise to the popular multinomial logit specification. As mentioned in the introduction, So (2000) focuses on the Cobb-Douglas specification $v_i = c_i w_i^{-a_i} p_i^{-b_i} = c_i (\bar{w}/\theta_i)^{-a_i} p_i^{-b_i}$. (So 2000, in fact, confines himself to the case where $a_i = a$, $b_i = b$ for all firms $i = 1, \dots, N$, which, as mentioned, implies that all customers aggregate the price and service-level attributes into a single full-price measure.) See Leeflang et al. (2000) for an axiomatic foundation of the class of attraction models, alternative specifications of the attraction functions, and specific applications.

Besides allowing for considerably more general dependencies of the firms' demand volumes with respect to the industry choices, the demand model (3) has several advantages over (4). First, demand volumes, prices, and waiting-time standards are directly observable so that the demand equations (3) can be estimated directly (assuming a specific structural form such as (5) or (10)). In contrast, full prices $\{F_i\}$ are not observable and depend on unknown aggregation parameters. Under certain structural forms, it is possible to ensure that the estimated demand functions (3) reduce to the special case (4), but only by imposing a large set of constraints on the parameters of the demand equations.

Second, beyond stipulating that the demand equations (3) are of the special type (4), full-price competition models are based on the assumption that firms select and advertise a single full-price measure and that consumers find the aggregate measure sufficient to compare alternative possibilities,

behavioral findings to the contrary, notwithstanding (see Allon and Federgruen 2007 for a review of economics, marketing, and psychology papers documenting these behavioral findings). Full-price competition models also do not enable us to characterize the equilibrium behavior in the industry, when firms compete only in terms of their prices or only in terms of their service levels, under exogenously specified service levels or prices, respectively. While in general the analysis of noncooperative games with multidimensional strategy spaces is more complex than those with one-dimensional strategy sets, as in the full-price models, in our case it has turned out to be easier. As a consequence, we have been able to establish the existence of Nash equilibria under minor parameter conditions, while hitherto, the existence of an equilibrium has remained an open question even in very special full-price models, such as the two-firm model with linear demand functions and M/M/1 queueing facilities; see Cachon and Harker (2002).

Finally, it is well known that the classical Bertrand price-competition model with homogenous goods has a plausible but rather unappealing equilibrium behavior: for example, when all providers have identical cost structures, their profits are reduced to zero and a slight price change can cause a firm to lose its entire market share. Similar discontinuities have been observed in the equilibrium behavior models in various models in which customers are assumed to select a provider on the basis of the (full) price only, patronizing only those whose full price is the lowest in the industry. Chen and Wan (2003), for example, give an example where no (pure) equilibrium exists. The example shows that the equilibrium behavior is very unstable: as the total market varies from 1.2 to 1.3, the industry moves from a unique equilibrium to no-equilibrium, to an infinite number of equilibria. In addition to treating price and service level as truly independent strategic instruments, which, in general, cannot be aggregated into a full-price measure, we address settings with heterogenous service providers, i.e., service is differentiated on the basis of attributes other than prices and waiting times. Thus, even if a firm offers the lowest price and service level in the industry, it does not capture 100% of the market. We have found that, as with classical price-competition models, the equilibrium behavior with heterogenous service providers is robust with respect to small parameter changes, while that for models with homogenous service providers often is not.

3.2. The Cost Structure

The cost structure consists of two components. First, each firm i incurs a given cost c_i per customer served. Second, it incurs capacity-related costs. When the service process consists of a single service stage, these costs are proportional with the service rate, at a cost rate γ_i . If customers (potentially) go through multiple stages of service, the service facility needs to be modelled as a queueing network, in which case the capacity costs are assumed to be proportional with the service rates of the various nodes of

the network. Thus, the service rates characterize the firm's capacity. These need to be selected so as to satisfy the firm's waiting-time standard w , under the given demand rate λ_i .

The specific shape of the $C_i(\lambda_i, \theta_i)$ function depends, of course, on the characteristics of the queueing system which arises from the firm's service process. Here, we will consider fully general G/GI/s systems to represent service processes with a single stage and general open Jackson networks to represent processes with multiple stages. For the single-stage process, $\{A_i^\lambda(t): t \geq 0\}$ represents the arrival process and we shall assume that the number of servers, s , is exogenously given, but their service times can be scaled upwards or downwards by adjusting the service rate μ_i . Below we will demonstrate that in spite of the large generality of queueing systems considered here, the capacity cost function, C_i , can be represented, either exactly or as a close approximation, as a member of the four-parameter class of functions \mathcal{C} . Under \mathcal{C} , the cost structure exhibits economies of scales or diseconomies of scale. More specifically:

LEMMA 1. Assume that the cost function is of type \mathcal{C} .

$$(a) \quad \frac{\partial^2 C_i}{\partial \lambda_i^2} = -\frac{1}{4} \frac{(B_4 \theta_i + 2B_3 \lambda_i)^2}{(B_3 \lambda_i^2 + B_4 \lambda_i \theta_i + B_2^2 \theta_i^2)^{3/2}} + \frac{1}{2} \frac{2B_3}{(B_3 \lambda_i^2 + B_4 \lambda_i \theta_i + B_2^2 \theta_i^2)^{1/2}}, \quad (13)$$

$$\frac{\partial^2 C_i}{\partial \theta_i^2} = -\frac{1}{4} \frac{(B_4 \lambda_i + 2B_2^2 \theta_i)^2}{(B_3 \lambda_i^2 + B_4 \lambda_i \theta_i + B_2^2 \theta_i^2)^{3/2}} + \frac{1}{2} \frac{2B_2^2}{(B_3 \lambda_i^2 + B_4 \lambda_i \theta_i + B_2^2 \theta_i^2)^{1/2}}. \quad (14)$$

(b) $C_i(\lambda_i, \theta_i)$ is jointly convex (concave) if and only if $|B_4| \leq (\geq) 2B_2 \sqrt{B_3}$.

$$(c) \quad \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} = -\frac{1}{4} \frac{(B_4 \theta_i + 2B_3 \lambda_i)(B_4 \lambda_i + 2B_2^2 \theta_i)}{(B_3 \lambda_i^2 + B_4 \lambda_i \theta_i + B_2^2 \theta_i^2)^{3/2}} + \frac{1}{2} \frac{B_4}{(B_3 \lambda_i^2 + B_4 \lambda_i \theta_i + B_2^2 \theta_i^2)^{1/2}} \leq (\geq) 0 \quad (15)$$

if and only if $|B_4| \leq (\geq) 2B_2 \sqrt{B_3}$.

PROOF. (a) and (c) are by straightforward calculus.

(b) Assume that $\Delta = B_2^2 - B_4^2/4B_3 > 0$. To show that C_i is jointly convex in (λ_i, θ_i) , it suffices to show that the function $f(x, y) = \sqrt{x^2 + \Delta y^2}$ with $x = \sqrt{B_3}(\lambda_i + (B_4/2B_3)\theta_i)$ and $y = \theta_i$ is jointly convex in (x, y) . But $\partial f/\partial x = x/\sqrt{x^2 + \Delta y^2}$, $\partial f/\partial y = \Delta y/\sqrt{x^2 + \Delta y^2}$, and $\partial^2 f/\partial x^2 = (x^2 + \Delta y^2)^{-3/2} \Delta y^2 \geq 0$, $\partial^2 f/\partial y^2 = (x^2 + \Delta y^2)^{-3/2} \Delta x^2 \geq 0$. Because $\partial^2 f/\partial x \partial y = -(x^2 + \Delta y^2)^{-3/2} \Delta y x \leq 0$, it follows that the Hessian of $f(x, y)$ has zero as its determinant, completing the proof that f is jointly convex. The proof for the case where $\Delta \leq 0$ is analogous. \square

Thus, when $|B_4| = 2B_2 \sqrt{B_3}$, the capacity cost function is in fact linear, i.e., of the type

$$(\mathcal{C}^{\mathcal{L}, \mathcal{F}, \mathcal{N}}) \quad C_i(\lambda_i, \theta_i) = B'_1 \lambda_i + B'_2 \theta_i \quad \text{for positive } B'_1, B'_2. \quad (16)$$

3.2.1. Capacity Cost Functions Associated with Expectation-Based Guarantee. In this subsection, we consider various queueing systems with waiting-time guarantees, stated in terms of the expected sojourn time or delay. Starting with the former, the most elementary setting in which the simple affine approximation $\mathcal{C}^{\mathcal{L}, \mathcal{F}, \mathcal{N}}$ arises is where the service system can be modelled as an M/M/1 system, either under individual service or under processor sharing. Thus, let $w_i = \mathbb{E}(T_i)$. It is well known that

$$w_i = \frac{1}{\mu_i - \lambda_i}. \quad (17)$$

It follows that the service rate μ_i required to satisfy the sojourn time guarantee is a positive linear combination of the demand rate λ_i and the reciprocal of the waiting-time guarantee, and the same applies therefore to the capacity cost function C_i .

A simple affine capacity cost function of the type $\mathcal{C}^{\mathcal{L}, \mathcal{F}, \mathcal{N}}$ also arises when the service process consists of multiple stages, and the service facility is described as an open Jackson network. Let V be the set of vertices of the network, π_j the fraction of customers who start their service process at vertex j , and P_{jk} the probability that a customer moves to node k after completing his service at node j . Let $\pi = (\pi_j: j \in V)$. The matrix $P = (P_{jk}: j, k \in V)$ is assumed to be substochastic, i.e., $\lim_{n \rightarrow \infty} P^n = 0$. Finally, let μ_j denote the service rate at node $j \in V$ and γ_j the capacity cost per unit of service rate at node j . The vector of aggregate arrival rates Λ satisfies the system of equations $\Lambda = \lambda \pi + P^T \Lambda$, i.e., $\Lambda = \lambda \xi$, where $\xi = (I - P^T)^{-1} \pi$. It is well known that the expected number of customers at node j is given by $\Lambda_j/(\mu_j - \Lambda_j)$, so that the expected total number of customers in the system is given by $\lambda \sum_{j \in V} \xi_j/(\mu_j - \Lambda_j)$; by an application of the $L = \lambda W$ identity, we conclude that $w = \mathbb{E}(T) = \sum_{j \in V} \xi_j/(\mu_j - \lambda \xi_j)$. The minimal capacity costs are obtained by adopting a vector $\mu = (\mu_j: j \in V)$ of service rates which optimizes the convex program

$$\min \sum_{j \in V} \gamma_j \mu_j \quad (18)$$

$$\text{s.t.} \quad \sum_{j \in V} \frac{\xi_j}{\mu_j - \lambda \xi_j} \leq w, \quad (19)$$

$$\mu_j > \lambda \xi_j. \quad (20)$$

With $\nu > 0$, the Lagrange multiplier associated with (20), μ^* is optimal if it satisfies the Kuhn-Tucker conditions

$$\gamma_j = \frac{\nu \xi_j}{(\mu_j^* - \lambda \xi_j)^2}, \quad j \in V, \quad (21)$$

as well as (20), which is satisfied as an equality at the optimal service-rate vector μ^* . It follows from (21) that

$$\mu_j^* - \lambda \xi_j = \sqrt{\nu} \sqrt{\frac{\xi_j}{\gamma_j}}, \quad j \in V. \tag{22}$$

Substituting (22) into (20), we obtain $\sqrt{\nu} = (\sum_{j \in V} \sqrt{\xi_j \gamma_j})/w$, so that the minimum capacity cost value $C(\lambda, w) = \sum_{j \in V} \gamma_j \mu_j^*$ is indeed of the form $\mathcal{C}^{\mathcal{L}, \mathcal{F}, \mathcal{N}}$, with

$$B'_1 = \sum_{j \in V} \xi_j \gamma_j, \tag{23}$$

$$B'_2 = \left(\sum_{j \in V} \sqrt{\xi_j \gamma_j} \right)^2 > B'_1. \tag{24}$$

In the special case where all $\gamma_j = \gamma$, the optimization problem (18), (19) may be viewed as the “dual” of Kleinrock’s (1976, §5.7) well-known capacity allocation problem: in the latter, the expected number of customers in the system, i.e., λ times the left-hand side of (19), is minimized subject to a bound on the capacity cost in (18). Indeed, substituting the cost value C_i in (16) and solving for w_i , we obtain the same dependence structure as in Kleinrock’s allocation problem.

In an M/G/1 service system (where service times follow an arbitrary distribution), let S_i^1 denote the service time when the service rate is normalized to be one, with coefficient of variation c_s . Recall that for a general service rate $\mu \neq 1$, the service time $S_i^\mu = (1/\mu)S_i^1$. $w_i = \mathbb{E}(T_i)$ is, of course, given by the well-known Pollaczek-Khintchine formula

$$w_i = \frac{1}{\mu_i} + \frac{1 + c_s^2}{2} \frac{\lambda_i}{\mu_i(\mu_i - \lambda_i)}. \tag{25}$$

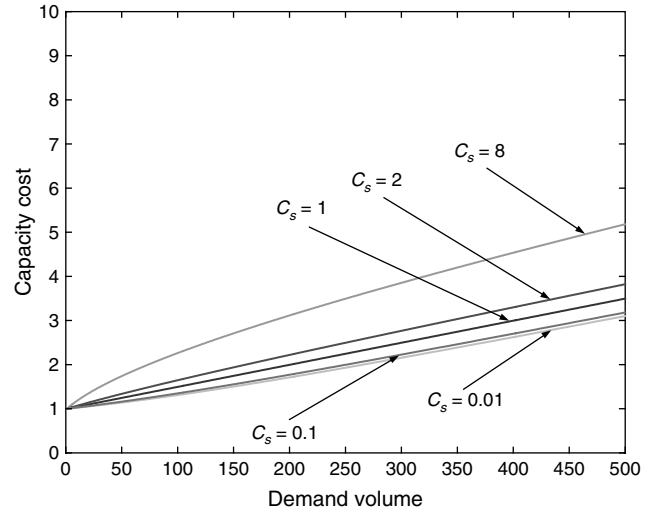
Rewriting (25) as a quadratic equation in μ_i , we obtain that the unique service rate μ_i , which results in a given value w_i , is given by the expression

$$\mu_i = \frac{\lambda_i}{2} + \frac{1}{2w_i} + \frac{1}{2} \sqrt{\left(\lambda_i + \frac{1}{w_i} \right)^2 - 4 \frac{\lambda_i}{w_i} \left(1 - \frac{1 + c_s^2}{2} \right)}. \tag{26}$$

Because the capacity cost C_i is proportional with the service rate μ_i , it follows that $C_i(\lambda_i, w_i)$ is of the form \mathcal{C} with $B_1 = \frac{1}{2}\gamma_i$, $B_2 = \frac{1}{2}\gamma_i$, $B_3 = \frac{1}{4}\gamma_i^2$, $B_4 = \frac{1}{2}c_s^2\gamma_i^2$. The M/G/1 system is thus the simplest setting in which the capacity cost function C_i fails to be affine in the demand rate λ_i . Note from Lemma 1 that the cost function is convex (concave) if and only if the coefficient of variation, c_s , is less (greater) than one; see also Figure 1. Under non-Poisson arrivals, no exact formula exists for the expected sojourn time. However, for the GI/GI/1 queue, with arrivals generated by a general renewal process with an interarrival time distribution with coefficient of variation c_a , the well-known Kingman bound applies:

$$w_i \leq \frac{1}{\mu_i} + \frac{c_a^2 + c_s^2}{2} \frac{\lambda_i}{\mu_i(\mu_i - \lambda_i)}. \tag{27}$$

Figure 1. Capacity cost functions for the M/G/1 queuing system.



The bound holds with equality when customers arrive according to a Poisson process (in which case it reduces to the Pollaczek-Khintchine formula) and it is asymptotically tight in heavy traffic, i.e., when $\rho_i = \lambda_i/\mu_i \rightarrow 1$. Following the above derivation, we obtain that the service rate μ_i , which results in a guaranteed value for $w_i = \mathbb{E}(T_i)$, is now given by the expression

$$\mu_i = \frac{\lambda_i}{2} + \frac{1}{2w_i} + \frac{1}{2} \sqrt{\left(\lambda_i + \frac{1}{w_i} \right)^2 - 4 \frac{\lambda_i}{w_i} \left(1 - \frac{c_a^2 + c_s^2}{2} \right)}. \tag{28}$$

Once again, the resulting cost function is of the form \mathcal{C} with $B_1 = \frac{1}{2}\gamma_i$, $B_2 = \frac{1}{2}\gamma_i$, $B_3 = \frac{1}{4}\gamma_i^2$, $B_4 = ((c_a^2 + c_s^2 - 1)/2)\gamma_i^2$. It follows, again from Lemma 1, that the cost function is convex (concave) if and only if $(c_a^2 + c_s^2)/2 \leq (\geq) 1$.

When the waiting-time guarantee is stated in terms of the *expected delay*, it equals the second term in (25) for M/G/1 systems and is bounded by the second term in (27) for GI/GI/1 systems. Both cases result in quadratic equations in μ_i , and a capacity cost function of the type \mathcal{C} with $B_2 = 0$, which, again by Lemma 1, implies that the cost function is always concave, regardless of the characteristics of the service and interarrival time distributions. (Even in the M/M/1 case, the capacity cost function is nonlinear and concave.)

3.2.2. Capacity Cost Functions Associated with Fractile Guarantee. In this subsection, we consider queueing systems with waiting-time guarantees stated in terms of a given fractile of their sojourn time or delay distributions. We start with the latter. As mentioned in §2, the tail of the steady-state delay distribution is exactly of the exponential form given in (1), when the service system is of the GI/M/s type; see Chapter 6 in Kleinrock (1975). For more general queueing systems, the exponential form in (1) holds as a close approximation, supported by two types of limit results.

First, for general GI/GI/s systems, there exist under minimal regularity conditions with respect to the interarrival and service time distributions, constants α and η such that

$$\lim_{x \rightarrow \infty} e^{\eta x} \mathbb{P}(D > x) = \alpha. \quad (29)$$

See Abate et al. (1995) and the references cited therein. As mentioned in §2, in the case of a single server with a service-time distribution whose Laplace transform is rational, the limit result goes back to Smith (1953). Second, the exponential tail approximation is completely accurate in heavy traffic for any value of x , i.e.,

$$\lim_{\rho \uparrow 1} \mathbb{P}(D^\rho > x) e^{(1-\rho)\eta x} = \alpha \quad \forall x > 0. \quad (30)$$

This limit result holds, again, for general GI/GI/s systems, as long as the service-time and interarrival-time distributions have finite second moments; see Kingman (1962). More generally, as described in Abate et al. (1995, 1996), the arrival process does not need to be a renewal process. All that is required is that the normalized process $\{A_i^1(t); t > 0\}$ satisfies a functional central limit theorem (FCLT), i.e.,

$$(nc_a^2)^{-1/2} [A_i^1(nt) - nt] \Rightarrow B(t), \quad (31)$$

where $B = \{B(t); t > 0\}$, s is standard Brownian motion, $c_a^2 = \lim_{t \rightarrow \infty} \text{Var}A(t)/\mathbb{E}A(t)$ (assumed to exist), and where \Rightarrow denotes convergence in distribution. See Chapters 5 and 7 in Whitt (2002) for a discussion of various stochastic processes that satisfy an FCLT, including Markov modulated renewal processes and superpositions of renewal processes.

The approximation is also supported by the following rigorous bound: $\mathbb{P}(D > x) \leq e^{-\eta x}$ for all x , at least in GI/GI/1 systems under minor distributional conditions similar to those guaranteeing the limit result (29). In other words, for GI/GI/1 systems, the exponential tail approximation holds as a rigorous bound when choosing $\alpha = 1$. Last, but not least, the exponential tail approximation in (1) is supported by extensive numerical comparisons; see in particular Seelen et al. (1985) and Tijms (1986).

The remaining difficulty is to characterize how the constants α and η in (1) depend on the characteristics of the service and arrival processes, and in particular how they depend on the service and arrival rates. For GI/GI/s systems, Abate et al. (1995) have developed a Taylor series expansion for the constant η in terms of powers of $(1 - \rho_i)$, with $\rho_i = \lambda_i / \mu_i s_i < 1$. (This expansion holds, again, under the minor distributional assumptions underlying the limit result (29).) Moreover, these authors also demonstrated, numerically, that very accurate approximations are obtained by using only the first two terms in the expansion:

$$\eta_i = \frac{2\mu_i(1 - \rho_i)}{c_a^2 + c_s^2} - \frac{2\mu_i}{c_a^2 + c_s^2} (1 - \rho_i)^2 \eta^* + O((1 - \rho_i)^3), \quad \rho_i \rightarrow 1, \quad (32)$$

where $\eta^* = ((2v_3 - 3c_s^2(c_s^2 + 2)) - (2u_3 - 3c_a^2(c_a^2 + 2))) / 3(c_a^2 + c_s^2)^2$, with c_s and c_a the coefficient of variation of the (normalized) service and interarrival time distributions, and v_3 and u_3 their respective third moments. (Accurate approximations are obtained even when using only the linear term in $(1 - \rho_i)$, i.e., $\eta_i = \mu_i(1 - \rho_i)/(c_a^2 + c_s^2)$.) As to the constant $\alpha = \mathbb{P}(D_i > 0)$, Abate et al. (1995) shows, at least for GI/GI/1 systems,

$$\alpha_i = \eta_i \mathbb{E}(D_i) + O((1 - \rho)^2) \quad \text{as } \rho \uparrow 1. \quad (33)$$

To obtain an approximation for α_i , the authors suggest replacing η_i in (33) by the first-order approximation in its Taylor series expansion (32) and $\mathbb{E}(D_i)$ by the second term in the Kingman bound (27), thus resulting in the simple expression $\tilde{\alpha}_i = \rho_i$. Note that $\alpha_i = \mathbb{P}(D_i > 0) = \rho_i$ for M/G/1 and GI/M/1 systems (see Wolff 1989) while for general GI/GI/1 systems, ρ_i equals the time average likelihood of observing a busy server. As an even simpler approximation, the authors suggest the simple choice $\hat{\alpha}_i = 1$, which is accurate in heavy traffic ($\lim_{\rho \uparrow 1} \hat{\alpha}_i / \alpha_i = 1$ by (32) and (33)). Moreover, as Abate et al. (1995) point out, in an environment where service rates are given, “the relative error in an approximation for a percentile is typically substantially less than the relative error for the corresponding tail probability itself.” By a similar argument, it is clear that the same applies to the relative error in an approximation for the required service rate when the percentile is given.

The simplest capacity cost function C_i is thus obtained when combining the approximation $\hat{\alpha}_i = 1$ with the first term in (32) as an approximation for η_i . This gives rise to a cost function of type $\mathcal{E}^{\mathcal{L}\mathcal{F}\mathcal{N}}$, with $C_i(\lambda_i, w_i) = \lambda_i + \log(1/(1 - \phi))(c_a^2 + c_s^2)/2(1/w_i)$. Using $\hat{\alpha}$, and approximating η_i by the first two terms in (32), gives rise to a quadratic equation in μ_i , and in addition, to a capacity cost function C_i of type \mathcal{E} :

$$\begin{aligned} C_i(\lambda_i, w_i) &= \lambda_i \frac{\gamma(1 - 2\eta_i^*)}{2(1 - \eta_i^*)} + \frac{\gamma}{w_i} \frac{c^2 \log(1/(1 - \phi))}{2(1 - \eta_i^*)} + \frac{\gamma}{2(1 - \eta_i^*)} \\ &\quad \cdot \sqrt{\left[\lambda_i(1 - 2\eta_i^*) + \frac{c^2 \log(1/(1 - \phi))}{w_i} \right]^2 + 4(1 - \eta_i^*) \eta_i^* \lambda_i^2}, \end{aligned} \quad (34)$$

where $c^2 \equiv (c_a^2 + c_s^2)/2$. It follows from Lemma 1 that the capacity cost function is convex (concave) if and only if $\eta^* \geq (\leq) 0$. Indeed, if $\eta^* = 0$, which arises, for example, when the service and interarrival times have the same distribution, the capacity cost function reduces to the affine structure $\mathcal{E}^{\mathcal{L}\mathcal{F}\mathcal{N}}$. Note also that if the service- and interarrival-time distributions have identical third moments, $\eta^* \geq 0 \Leftrightarrow c_s \leq c_a$. As the service time becomes more variable (c_s increases), the total capacity cost increases, as well as the marginal cost. As shown, the capacity cost function

for an M/G/1 system with a waiting-time guarantee based on the expected sojourn time exhibits the same monotonicity properties. For $c_s < c_a$, the marginal cost increases with the demand volume, while it decreases for $c_s > c_a$. Conversely, combining the superior approximation $\tilde{\alpha}_i = \rho_i$ with the first-order approximation for η_i in (32), we obtain the following equation in μ_i :

$$\log(1 - \phi) = \log\left(\frac{\lambda_i}{\mu_i s_i}\right) - \frac{2(\mu_i - \lambda_i/s_i)}{c_a^2 + c_s^2} \frac{1}{w_i}. \tag{35}$$

To obtain a closed-form expression for μ_i , we use the approximation

$$\log(x) \approx 2\frac{x-1}{x+1}, \quad 0 < x < 2, \tag{36}$$

which is highly accurate for $0.6 < x = \rho_i < 2$, say, with relative error of less than 3%. This gives rise to the equation

$$\log(1 - \phi) = \frac{2(\alpha_i/\mu_i s_i - 1)}{\alpha_i/\mu_i s_i + 1} - \frac{2(\mu_i - \lambda_i/s_i)}{c_a^2 + c_s^2} \frac{1}{w_i}, \tag{37}$$

which can be written as a quadratic equation in μ_i , and in addition, again, gives rise to a capacity cost function C_i of type \mathcal{C} :

$$\begin{aligned} C_i(\lambda_i, w_i) &= \frac{c\gamma}{2w} \log\left(\frac{1}{1-\phi} - 2\right) \\ &+ \gamma \sqrt{\frac{c^2}{w^2} \left(2 - \log\left(\frac{1}{1-\phi}\right)\right)^2 + 4\frac{c}{w} \left(\lambda_i \log\left(\frac{1}{1-\phi}\right) + \lambda_i^2 \frac{w_i}{c}\right)}. \end{aligned} \tag{38}$$

Note from Lemma 1 that under this approximation, the capacity cost function is concave irrespective of the shapes of the service- and interarrival-time distributions. We omit the last possibility where the superior approximation $\tilde{\alpha}_i = \rho_i$ is combined with the second-order approximation for η_i in (32) because it gives rise to a cubic equation in μ_i , even when employing the approximation of the logarithmic function in (36).

While the Seelen et al. (1985) and Tijms (1986) studies focus on assessing the accuracy of the approximated waiting-time distribution for a given capacity level, it is equally true that, even under moderate utilization rates, the capacity required to meet a given waiting standard with a given likelihood is accurately approximated in terms of the exponential approximation. To illustrate this, see Table 1. Here, we compare the approximate capacity level μ^* with the actual capacity level μ^{ex} required to meet a waiting-time standard $w = 0.1$ with 95% probability when the service facility acts like a G/G/1 system, with either Erlang 3 service and interarrival times, or with hyperexponential service and interarrival times. (The hyperexponential distribution mixes with equal probability an exponential with a mean of one, and one with a mean of two; the mean of the

Table 1. Accuracy level of the exponential approximation.

Distribution	λ	μ^*	μ^{ex}	Error (%)	ρ
Erlang 3	10	19.99	18.2	8.94	0.500
	100	109.99	109.7	0.26	0.901
	1,000	1,009.99	1,009.9	0.01	0.990
	10,000	10,009.99	10,009.9	0.00	0.999
Hyper-exponential	10	46.61	41.2	11.62	0.215
	100	136.61	140	-2.48	0.732
	1,000	1,036.61	1,044	-0.71	0.965
	10,000	10,036.61	10,031	0.06	0.996

Erlang distribution is one.) The Erlang 3 distribution has a coefficient of variation $cv = 0.58$, while that of the hyperexponential is 1.22. μ^* is computed, employing the most basic approximation, with $\hat{\alpha}_i = 1$ and η_i given by the first term in (32).

For each type of distribution, we have evaluated the capacity requirement for an average arrival rate $\lambda = 10, 100, 1,000, 10,000$ using a high-precision simulation. Note that for the instances where $\rho \geq 0.9$, the approximate capacity level is within 0.25% of the actual capacity requirement. Even for low-utilization rates ($\rho = 0.5$ or $\rho = 0.2$), the accuracy is remarkably good.

We complete this subsection with a brief discussion of (approximate) capacity cost functions that arise when the waiting-time guarantees are specified in terms of fractiles of the sojourn time rather than the delay distribution. Once again, the exponential tail approximation (1) continues to apply for general GI/GI/s systems, again, supported by small tail asymptotics such as (29), i.e., there exist constants $\eta^{(T)}$ and $\alpha^{(T)}$ such that

$$\lim_{x \rightarrow \infty} e^{\eta^{(T)} x} \mathbb{P}(T > x) = \alpha^{(T)}. \tag{39}$$

Indeed, the asymptotic decay rate $\eta^{(T)}$ is identical to that pertaining to the delay distribution, so that the Taylor series expansion (32) continues to apply to it. Analogous to their recommendation in the case of the delay distribution, Abate et al. (1996) suggest using $\eta^{(T)} \mathbb{E}(T)$ for $\alpha^{(T)}$, once again, employing the first-order approximation in (32) for $\eta^{(T)}$ and (at least in single-server systems) the Kingman bound for $\mathbb{E}(T)$. Combining this approximation for $\alpha^{(T)}$ with the first-order approximation for η in (32) results, again, in a capacity cost function of type \mathcal{C} . (As mentioned in Allon and Federgruen 2007, in the special case of an M/M/1 system, an exact capacity cost function can be derived and it is of the simple affine structure $\mathcal{C}^{S,F,N}$, i.e., $C_i(\lambda_i, w_i) = \lambda_i + \log(1/(1-\phi))(1/w_i)$.)

4. Separable Demand Model

In this section, we characterize the equilibrium behavior in the linear demand model under a capacity cost function of the general type \mathcal{C} , which, of course, includes

the cost structure of the affine type $\mathcal{C}^{E,F,N}$ as a special case. Thus, let $C_i(\lambda_i, w_i) = B_{1i}\lambda_i + B_{2i}(1/w_i) + \sqrt{B_{3i}\lambda_i^2 + B_{4i}(\lambda_i/w_i) + B_{2i}^2(1/w_i^2)}$. We confine ourselves, at first, to the case where each capacity cost function is convex in the demand volume, which by Lemma 1 is equivalent to assuming that

$$|B_{4i}| \leq 2B_{2i}\sqrt{B_{3i}}. \quad (40)$$

In §3, we showed that in many queueing models, the convexity condition is equivalent to a specific bound for the coefficient of variation of the facility's service and/or arrival process.

Assume that the firms compete in terms of their price choices, under given service levels θ . The profit earned by firm i can be expressed as

$$\begin{aligned} \pi_i &= (p_i - c_i)\lambda_i - C_i(\lambda_i, \theta_i) \\ &= (p_i - c_i)\lambda_i - B_{1i}\lambda_i - B_{2i}\theta_i \\ &\quad - \sqrt{B_{3i}\lambda_i^2 + B_{4i}\lambda_i\theta_i + B_{2i}^2\theta_i^2}. \end{aligned} \quad (41)$$

The capacity cost function per unit of demand ($\bar{C}_i = C_i(\lambda_i, \theta_i)/\lambda_i$) is clearly bounded from below by $B_{1i} + \sqrt{B_{3i}}$. We therefore assume, without loss of generality, that $p_i^{\min} = c_i + B_{1i} + \sqrt{B_{3i}}$, while p_i^{\max} is sufficiently large as to have no impact on the equilibrium prices.

THEOREM 1 (PRICE COMPETITION). *Assume that the capacity cost function C_i is convex in the demand volume λ_i , i.e., $|B_{4i}| \leq 2B_{2i}\sqrt{B_{3i}}$, $i = 1, \dots, N$. There exists a unique price equilibrium $p^*(\theta)$, which is the unique solution to the system of equations*

$$\begin{aligned} 0 &= \frac{\partial \pi_i}{\partial p_i} = \lambda_i - b_i(p_i - c_i) + b_i \frac{\partial C_i}{\partial \lambda_i} \\ &= \lambda_i - b_i(p_i - c_i) \\ &\quad + b_i \left[B_{1i} + \frac{\frac{1}{2}(B_{4i}/w_i + 2B_{3i}\lambda_i)}{\sqrt{B_{3i}\lambda_i^2 + B_{4i}(\lambda_i/w_i) + B_{2i}^2(1/w_i^2)}} \right]. \end{aligned} \quad (42)$$

PROOF. A straightforward adaptation of Lemma 1 in Bernstein and Federgruen (2004a) shows that the price competition game is supermodular, i.e.,

$$\frac{\partial^2 \pi_i}{\partial p_i \partial p_j} \geq 0 \quad (43)$$

because C_i is a convex function. Moreover, by (D),

$$-\frac{\partial^2 \pi_i}{\partial p_i^2} > \sum_{i \neq j} \frac{\partial^2 \pi_i}{\partial p_i \partial p_j}, \quad (44)$$

a condition guaranteeing that the equilibrium is unique; see Vives (2000). The choice $p_i = p^{\min}$ clearly results in negative profits for firm i . The equilibrium point $p^*(\theta)$ is therefore an interior point of the price space, satisfying the

first-order conditions $0 = \partial \pi_i / \partial p_i$. Moreover, (43) and (44) imply that the function π_i is strictly concave in p_i ; thus, (42) has at most one solution. We conclude that $p^*(\theta)$ is the unique solution of (42). \square

As it is easily verified from the proof, the results in Theorem 1 continue to apply under concave cost functions as long as $\partial^2 C_i / \partial \lambda_i^2 \geq -1/b_i$. The optimality conditions (42) can be used, in conjunction with the implicit function theorem, to compute the marginal impact a firm's service-level improvement has on its prices and those of its competitors.

Consider now a setting where the firms compete in terms of their service levels under exogenously given prices. We assume that firm i chooses his service level from an interval $[0, \theta_i^{\max}]$ with θ_i^{\max} sufficiently large, so as to have no impact on the industry equilibrium behavior. The following theorem establishes that an equilibrium exists in this service-level competition model as well. In addition, the theorem shows that under mild conditions, the service-competition game is of the special *supermodular type*. Because each firm's feasible action set is a closed interval, the game is supermodular if each firm i 's marginal profit function $(\partial \pi_i / \partial \theta_i)(\theta)$ can be shown to be increasing in any of the competitors' service levels. In case more than one Nash equilibrium (may) exist, we know that when the game is supermodular, a componentwise largest equilibrium $\bar{\theta}$ and a componentwise smallest equilibrium $\underline{\theta}$ exist. The full set of Nash equilibria is a sublattice of \mathbb{R}^n . Moreover, an equilibrium is easily computed with the following tâtonnement scheme: starting with an arbitrary service-level vector θ^0 , we determine in the k th iteration of the scheme the best-response service levels for each firm, assuming all its competitors maintain their service levels according to the vector determined in the $(k-1)$ st iteration. The sequence of service-level vectors $\{\theta^k\}$ converges to a Nash equilibrium: if $\theta^0 = 0$, it converges to $\underline{\theta}$; if $\theta^0 = \theta^{\max}$, it converges to $\bar{\theta}$.

Let $dC'_i/d\theta_i \equiv \partial(\partial C_i / \partial \lambda_i) / \partial \theta_i = a'_i(\theta_i)(\partial^2 C_i / \partial \lambda_i^2) + \partial^2 C_i / \partial \lambda_i \partial \theta_i$ represent the total marginal impact of a service improvement on the firm's marginal per-unit cost.

THEOREM 2. *Assume that firm i 's capacity cost function is convex in the demand volume, λ_i , i.e., $|B_{4i}| \leq 2B_{2i}\sqrt{B_{3i}}$, $i = 1, \dots, N$.*

(a) *There exists a service equilibrium θ^* which satisfies the equations*

$$\left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} \right) = \frac{\partial C_i / \partial \theta_i}{a'_i(\theta_i)}, \quad i = 1, \dots, N. \quad (45)$$

(b) *The service-competition game is supermodular if and only if $dC'_i/d\theta_i \geq 0$, $i = 1, \dots, N$.*

PROOF. (a) To establish the existence of a Nash equilibrium, it suffices to show that π_i is concave in θ_i . Note that

$$\frac{\partial \pi_i}{\partial \theta_i} = a'_i(\theta_i)(p_i - c_i) - \left(\frac{\partial C_i}{\partial \lambda_i} a'_i(\theta_i) + \frac{\partial C_i}{\partial \theta_i} \right), \quad (46)$$

and therefore,

$$\begin{aligned} \frac{\partial^2 \pi_i}{\partial \theta_i^2} &= a_i''(\theta)(p_i - c_i) - \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} (a_i'(\theta_i))^2 + 2 \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} a_i'(\theta_i) \right. \\ &\quad \left. + a_i''(\theta_i) \frac{\partial C_i}{\partial \lambda_i} + \frac{\partial^2 C_i}{\partial \theta_i^2} \right) \\ &= a_i''(\theta_i) \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} \right) \\ &\quad - \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} (a_i'(\theta_i))^2 + 2 \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} a_i'(\theta_i) + \frac{\partial^2 C_i}{\partial \theta_i^2} \right). \end{aligned} \quad (47)$$

Lemma 1 shows that C_i is either jointly concave or jointly convex in λ_i and θ_i . Thus, because C_i is convex in λ_i , it is jointly convex in λ_i and θ_i . The first term is negative because $a_i''(\theta_i) \leq 0$ and $p_i - c_i - \partial C_i / \lambda_i \geq p_i^{\min} - c_i - \partial C_i / \lambda_i \geq p_i^{\min} - c_i - (B_{1i} + \sqrt{B_{3i}}) = 0$, where the second inequality follows from the fact that $\partial C_i / \partial \lambda_i$ increases to the limit value $\lim_{\lambda_i \rightarrow \infty} \partial C_i / \partial \lambda_i = B_{1i} + \sqrt{B_{3i}}$. To show that the second term in (47) is negative as well, note that the quadratic function $(\partial^2 C_i / \partial \lambda_i^2)x^2 + 2(\partial^2 C_i / \partial \lambda_i \partial \theta_i)x + \partial^2 C_i / \partial \theta_i^2$ is uniformly positive because the convexity of C_i in λ_i implies that the coefficient of the quadratic term is positive while the discriminant is negative. (Joint convexity of C_i implies that the determinant of the Hessian is nonnegative.) Thus, $\partial^2 \pi_i / \partial \theta_i^2 \leq 0$, i.e., π_i is concave in θ_i . Because the choice $\theta_i = 0$ is clearly inferior for firm i , and given the choice of θ_i^{\max} , the equilibrium vector θ^* is an interior point of the feasible region and therefore satisfies the first-order conditions (45).

(b) $[0, \theta_i^{\min}]$ is the action space of firm i . To show that the service-competition model is a supermodular game, it thus suffices to verify that $\forall i = 1, \dots, N$, $\partial \pi_i / \partial \theta_i \partial \theta_j \geq 0$. Thus, differentiating both sides of (46) with respect to θ_j , we obtain because $\alpha'_{ij}(\theta_j) \geq 0$, that

$$\begin{aligned} \frac{\partial \pi_i}{\partial \theta_i \partial \theta_j} &= a_i'(\theta_i) \alpha'_{ij}(\theta_j) \frac{\partial^2 C_i}{\partial \lambda_i^2} + \frac{\partial^2 C_i}{\partial \theta_i \partial \lambda_i} \alpha'_{ij}(\theta_j) \\ &= \alpha'_{ij}(\theta_j) \frac{dC'_i}{d\theta_i} \geq 0 \end{aligned} \quad (48)$$

if and only if $dC'_i / d\theta_i > 0$. \square

Thus, an equilibrium always exists in the service-competition model. For the game to be supermodular with the special properties listed above, we need the condition $\partial C'_i / \partial \theta_i \geq 0$. While intuitive, this condition needs to hold for all feasible service-level combinations. Indeed, using (13) and (15), we obtain after some algebra

$$\frac{\partial C'_i}{\partial \theta_i} \geq 0 \Leftrightarrow \lambda_i \leq \theta_i a_i'(\theta_i). \quad (49)$$

For example, when the intercept function $a_i(\theta_i)$ is of the form $a_i(\theta_i) = a_i^0 + a_i^1 \ln(\theta_i)$, the condition in (49) is equivalent to $a_i(\theta_i) \leq \sum_{j \neq i} \alpha_{ij}(\theta_j) + b_i p_i - \sum_{j \neq i} \beta_{ij} p_j + a_i^1$.

Because the functions $\alpha_{ij}(\cdot)$ are increasing, this inequality is satisfied if $a_i^1 \ln(\theta_i) \leq \sum_{j \neq i} \alpha_{ij}(0) + b_i p_i - \sum_{j \neq i} \beta_{ij} p_j + a_i^1 - a_i^0$, i.e., if $\theta_i \leq \theta_i^{\max}(p) = \exp[(\sum_{j \neq i} \alpha_{ij}(0) + b_i p_i - \sum_{j \neq i} \beta_{ij} p_j + a_i^1 - a_i^0) / a_i^1]$.

Finally, assume that the firms engage in simultaneous price and service competition, i.e., each firm i selects a price $p_i \in [p_i^{\min}, p_i^{\max}]$, and a service level $\theta_i \in [0, \theta_i^{\max}]$. All firms make their choices simultaneously. Theorem 3 below shows that an equilibrium exists in this simultaneous-competition model, provided each firm i 's feasible price range starts at a minimum price $\underline{p}_i \geq p_i^{\min}$, guaranteeing a minimum gross profit margin, $m_i = p_i - c_i - \partial C_i / \partial \lambda_i$.

THEOREM 3 (SIMULTANEOUS PRICE AND SERVICE COMPETITION). *Assume that firm i 's capacity cost function is convex in the demand volume, λ_i , i.e., $|B_{4i}| \leq 2B_{2i}\sqrt{B_{3i}}$, $i = 1, \dots, N$. There exists a minimum price vector $\underline{p} \geq p^{\min}$ such that under the price range $[\underline{p}, p^{\max}]$, an equilibrium price vector p^* , and an equilibrium service-level vector θ^* exist in the simultaneous-competition model. The pair (p^*, θ^*) satisfies the system of equations*

$$\begin{aligned} \lambda_i - b_i(p_i - c_i) \\ + b_i \left[B_{1i} + \frac{\frac{1}{2}(B_{4i}/w_i + 2B_{3i}\lambda_i)}{\sqrt{B_{3i}\lambda_i^2 + B_{4i}(\lambda_i/w_i) + B_{2i}^2(1/w_i^2)}} \right] &= 0, \\ i = 1, \dots, N, \end{aligned} \quad (50)$$

$$\left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} \right) = \frac{\partial C_i / \partial \theta_i}{a_i'(\theta_i)}, \quad i = 1, \dots, N. \quad (51)$$

PROOF. Because each firm's feasible action set is a closed rectangle in \mathbb{R}^2 , it suffices to show that the profit function π_i is jointly concave in (p_i, θ_i) . Concavity of π_i in p_i was shown at the end of the proof of Theorem 1, as a direct consequence of (43) and (44). Concavity of π_i in θ_i was shown in the proof of Theorem 2(a). It thus suffices to show that the determinant of the Hessian of π_i with respect to (p_i, θ_i) is positive. Note that

$$\begin{aligned} \frac{\partial^2 \pi_i}{\partial p_i^2} &= -2b_i - b_i^2 \frac{\partial^2 C_i}{\partial \lambda_i^2}, \\ \frac{\partial^2 \pi_i}{\partial \theta_i^2} &= a_i''(\theta_i) \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} \right) \\ &\quad - \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} (a_i'(\theta_i))^2 + 2 \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} a_i'(\theta_i) + \frac{\partial^2 C_i}{\partial \theta_i^2} \right), \\ \frac{\partial^2 \pi_i}{\partial \lambda_i \partial \theta_i} &= a_i'(\theta_i) + b_i \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} a_i'(\theta_i) + \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} \right). \end{aligned}$$

The determinant of the Hessian is nonnegative if and only if

$$\begin{aligned} \left[2b_i + b_i^2 \frac{\partial^2 C_i}{\partial \lambda_i^2} \right] \left[-a_i''(\theta_i) \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} \right) \right. \\ \left. + \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} (a_i'(\theta_i))^2 + 2 \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} a_i'(\theta_i) + \frac{\partial^2 C_i}{\partial \theta_i^2} \right) \right] \end{aligned}$$

$$\begin{aligned}
 &\geq \left[a'_i(\theta_i) + b_i \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} a'_i(\theta_i) + \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} \right) \right]^2 \\
 &\Leftrightarrow -a''_i(\theta_i) \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} \right) \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} b_i^2 + 2b_i \right) \\
 &\quad + b_i^2 \left(\frac{\partial^2 C_i}{\partial \theta_i^2} \frac{\partial^2 C_i}{\partial \lambda_i^2} - \left(\frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} \right)^2 \right) \\
 &\quad + 2b_i \left(a'_i(\theta_i) \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} + \frac{\partial^2 C_i}{\partial \theta_i^2} \right) \geq (a'_i(\theta_i))^2 \\
 &\Leftrightarrow -a''_i(\theta_i) \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} \right) \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} b_i^2 + 2b_i \right) \\
 &\quad + 2b_i \left(\frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} a'_i(\theta_i) + \frac{\partial^2 C_i}{\partial \theta_i^2} \right) \geq (a'_i(\theta_i))^2. \quad (52)
 \end{aligned}$$

The first equivalence follows after some algebra and the second one from the fact that the determinant of the Hessian of C_i is zero, as shown in the proof of Lemma 1. A sufficient condition for (52) is obtained by replacing $\partial C_i / \partial \lambda_i$ by its upper bound $B_{1i} + \sqrt{B_{3i}}$. Thus, because $p_i^{\min} = c_i + B_{1i} + \sqrt{B_{3i}}$, the following is a sufficient condition:

$$\begin{aligned}
 &\left(\frac{\partial^2 C_i}{\partial \lambda_i^2} b_i^2 + 2b_i \right) (p_i - p_i^{\min}) + \frac{2b_i}{-a''_i(\theta_i)} \\
 &\cdot \left(\frac{\partial^2 C_i}{\partial \theta_i^2} + \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} a'_i(\theta_i) \right) \geq \frac{(a'_i(\theta_i))^2}{-a''_i(\theta_i)}. \quad (53)
 \end{aligned}$$

Clearly, a minimum price level $\underline{p} \geq p_i^{\min}$ can be found such that (53) is satisfied for all θ .

We conclude that an equilibrium pair (p^*, θ^*) exists. Note that in particular p^* is a price equilibrium in the price-competition game, which arises when the service levels are prespecified according to the vector θ^* . Conversely, θ^* is an equilibrium service-level vector in the service-competition game, which arises when prices are fixed according to the vector p^* . It thus follows from Theorems 1 and 2 that an equilibrium (p^*, θ^*) exists such that both p^* and θ^* are interior points of their respective feasible regions. As a consequence, this pair (p^*, θ^*) satisfies the first-order conditions (42) and (45), which are equivalent to (50) and (51). \square

5. Demand Specified by an Attraction Model

In this section, we investigate whether and how the equilibrium behavior in the various competitive models changes when the demand functions are specified by the general attraction model (10). As mentioned in §3, this class of nonlinear demand functions is one of the most frequently used classes and is supported by a general axiomatic foundation. (For example, it includes the popular multinomial logit functions as a special case.)

As in the previous section, we start with a characterization of the price-competition model, which arises when the

firms' service levels are exogenously given. Let $\tilde{v}_i = \log v_i$, $\tilde{\lambda}_i = \log \lambda_i$, and $e_i = (\partial v_i / \partial p_i)(p_i / v_i) = (\partial \tilde{v}_i / \partial p_i) p_i$, the price elasticity of firm i 's attraction value, a dimensionless index.

THEOREM 4 (PRICE COMPETITION). *Assume that firm i 's capacity cost function is convex in the demand volume, λ_i , i.e., $|B_{4i}| \leq 2B_{2i}\sqrt{B_{3i}}$, $i = 1, \dots, N$. Assume that the demand functions are given by the general attraction model (10).*

(a) *Assume that the price elasticities e_i are decreasing in p_i for all $i = 1, \dots, N$. There exists a price equilibrium vector p^* which satisfies the first-order conditions*

$$\begin{cases} \frac{1}{p_i^* - c_i - \partial C_i / \partial \lambda_i} + \frac{\partial \tilde{\lambda}_i}{\partial p_i}(p^*) = 0 & \text{if } p_i^* > p_i^{\min}, \\ p_i^* = p_i^{\min} & \text{otherwise.} \end{cases} \quad (54)$$

(b) *Assume that the price elasticity e_i decreases in p_i while $\lim_{p_i \uparrow \infty} e_i < -1$, $i = 1, \dots, N$. The price equilibrium p^* is unique.*

(c) *Assume that the capacity cost functions are of the affine type $\mathcal{C}^{\mathcal{L}, \mathcal{F}, \mathcal{N}}$. There exists under fully general attraction functions a price equilibrium p^* , which satisfies*

$$\frac{1}{p_i^* - c_i - B_{1i}} + \frac{\partial \tilde{\lambda}_i}{\partial p_i} = 0. \quad (55)$$

This equilibrium is unique if the attraction functions are log-concave in p_i .

PROOF. (a) Let $\pi_i(p, \theta) = (p_i - c_i)\lambda_i(p, \theta) - C_i(\lambda_i, \theta_i)$ denote firm i 's profit function. Lemma 1 in Gallego et al. (2006) shows that this function is quasi-concave in p_i , with

$$\frac{\partial \pi_i}{\partial p_i} = \frac{(1 - \lambda_i / M)v'_i}{\sum_{j=1}^N v_j + v_0} \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i} + \frac{1}{\partial \tilde{\lambda}_i / \partial p_i} \right).$$

The existence of an equilibrium thus follows from the Nash-Debreu Theorem (see, for examples, Theorem 2.1 in Vives 2000).

(b) Uniqueness, under the slightly stronger conditions for the attraction functions, is shown in Proposition 1 of Gallego et al. (2006).

(c) Under affine capacity cost functions of type $\mathcal{C}^{\mathcal{L}, \mathcal{F}, \mathcal{N}}$, $\pi_i(p, \theta) = (p_i - c_i - B_{1i})\lambda_i(p, \theta) - B_{2i}\theta_i$. Define $\tilde{\pi}_i = \log[\pi_i + B_{2i}\theta_i] = \log(p_i - c_i - B_{1i}) + \tilde{\lambda}_i$. It suffices to demonstrate that for all $i = 1, \dots, N$, $\tilde{\pi}_i(p, \theta)$ is supermodular in the pair (p_i, p_j) for all $j \neq i$. It is easily verified that

$$\frac{\partial \tilde{\lambda}_i}{\partial p_i} = \frac{\partial \tilde{v}_i}{\partial p_i} \left(1 - \frac{\lambda_i}{M} \right), \quad \frac{\partial \lambda_i}{\partial p_j} = -\frac{\partial \tilde{v}_j}{\partial p_j} \frac{\lambda_i \lambda_j}{M}. \quad (56)$$

Note that

$$\frac{\partial \tilde{\pi}_i}{\partial p_i} = \frac{1}{p_i - c_i - B_{1i}} + \frac{\partial \tilde{\lambda}_i}{\partial p_i}. \quad (57)$$

with the price charged for the service. This is immediate from

$$\frac{\partial^2 \tilde{\lambda}_i}{\partial p_i \partial \theta_i} = \frac{\partial^2 \tilde{v}_i}{\partial p_i \partial \theta_i} \left(1 - \frac{\lambda_i}{M}\right) + \frac{-\partial \tilde{v}_i}{\partial \theta_i} \frac{\partial \tilde{v}_i}{\partial p_i} \frac{\lambda_i}{M} \left(1 - \frac{\lambda_i}{M}\right). \quad (63)$$

Typically, the diagonal elements in A_θ are not only positive, but dominate the off-diagonal elements so that an increase in a firm's service level is followed by equilibrium price increases for all firms. However, price decreases may occur, and are in fact guaranteed to occur in case the attraction functions v_i are strictly log-submodular, i.e., $\partial^2 \tilde{v}_i / \partial p_i \partial \theta_i < 0$, with sufficiently small values for these cross-partial derivatives. Admittedly, this situation is unlikely to arise in practice.

While the sign of the equilibrium price service-level sensitivities in (60) is somewhat ambiguous, a similar application of the implicit function theorem to (54) reveals that an increase in one of the variable cost parameters c_i or B_{1i} for some firm i results in an across the board increase of all equilibrium prices in the industry.

We now turn our attention to the *service-competition model*, which arises when all firms select their service levels simultaneously under a given price vector p .

THEOREM 5 (SERVICE COMPETITION). *Assume that the capacity cost function C_i is convex in the demand volume λ_i , i.e., $|B_{4i}| \leq 2B_{2i}\sqrt{B_{3i}}$, $i = 1, \dots, N$. Assume that each of the attraction functions v_i is concave in θ_i . There exists an equilibrium vector of service levels $\theta^*(p)$ for any given price vector p , which satisfies the first-order conditions $\partial \lambda_i / \partial \theta_i = (\partial C_i / \partial \theta_i) / (p_i - c_i - \partial C_i / \partial \lambda_i)$.*

PROOF. Let $\pi_i = \lambda_i(p_i - c_i) - C_i(\lambda_i, \theta_i)$. It suffices to show that each function π_i is concave in θ_i . We first show that λ_i is concave in θ_i . Analogous to (56), one verifies that

$$\frac{\partial \lambda_i}{\partial \theta_i} = \frac{\partial \tilde{v}_i}{\partial \theta_i} \lambda_i \left(1 - \frac{\lambda_i}{M}\right)$$

and

$$\frac{\partial[\lambda_i(1 - \lambda_i/M)]}{\partial \theta_i} = \frac{\partial \tilde{v}_i}{\partial \theta_i} \lambda_i \left(1 - \frac{\lambda_i}{M}\right) \left(1 - \frac{2\lambda_i}{M}\right).$$

Thus,

$$\begin{aligned} \frac{\partial^2 \lambda_i}{\partial \theta_i^2} &= \frac{\partial^2 \tilde{v}_i}{\partial \theta_i^2} \lambda_i \left(1 - \frac{\lambda_i}{M}\right) + \left(\frac{\partial \tilde{v}_i}{\partial \theta_i}\right)^2 \lambda_i \left(1 - \frac{\lambda_i}{M}\right) \left(1 - \frac{2\lambda_i}{M}\right) \\ &= \lambda_i \left(1 - \frac{\lambda_i}{M}\right) \left[\frac{\partial^2 \tilde{v}_i}{\partial \theta_i^2} + \left(\frac{\partial \tilde{v}_i}{\partial \theta_i}\right)^2 \left(1 - \frac{2\lambda_i}{M}\right) \right]. \end{aligned}$$

Because v_i is concave in θ_i , $\partial v_i / \partial \theta_i = v_i(\partial \tilde{v}_i / \partial \theta_i)$ is decreasing in θ_i , so that

$$\frac{\partial v_i}{\partial \theta_i} \frac{\partial \tilde{v}_i}{\partial \theta_i} + v_i \frac{\partial^2 \tilde{v}_i}{\partial \theta_i^2} = v_i \left[\left(\frac{\partial \tilde{v}_i}{\partial \theta_i}\right)^2 + \frac{\partial^2 \tilde{v}_i}{\partial \theta_i^2} \right] < 0,$$

i.e., $\partial^2 \tilde{v}_i / \partial \theta_i^2 < -(\partial \tilde{v}_i / \partial \theta_i)^2$. Replacing $\partial^2 \tilde{v}_i / \partial \theta_i^2$ by this upper bound, we obtain that

$$\frac{\partial^2 \lambda_i}{\partial \theta_i^2} < -\lambda_i \left(1 - \frac{\lambda_i}{M}\right) \frac{2\lambda_i}{M}. \quad (64)$$

Now,

$$\frac{\partial \pi_i}{\partial \theta_i} = \frac{\partial \lambda_i}{\partial \theta_i} \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i}\right) - \frac{\partial C_i}{\partial \theta_i} \quad (65)$$

and

$$\begin{aligned} \frac{\partial^2 \pi_i}{\partial \theta_i^2} &= \frac{\partial^2 \lambda_i}{\partial \theta_i^2} \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i}\right) - \frac{\partial \lambda_i}{\partial \theta_i} \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} \frac{\partial \lambda_i}{\partial \theta_i} + \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i}\right) \\ &\quad - \frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} \frac{\partial \lambda_i}{\partial \theta_i} - \frac{\partial^2 C_i}{\partial \theta_i^2} \\ &= \frac{\partial^2 \lambda_i}{\partial \theta_i^2} \left(p_i - c_i - \frac{\partial C_i}{\partial \lambda_i}\right) \\ &\quad - \left(\frac{\partial^2 C_i}{\partial \lambda_i^2} \left(\frac{\partial \lambda_i}{\partial \theta_i}\right)^2 + 2\frac{\partial^2 C_i}{\partial \lambda_i \partial \theta_i} \frac{\partial \lambda_i}{\partial \theta_i} + \frac{\partial^2 C_i}{\partial \theta_i^2}\right). \quad (66) \end{aligned}$$

The first term to the right is negative by the definition of p_i^{\min} . By Lemma 1, we have that the function $((\partial^2 C_i / \partial \lambda_i^2)x^2 + 2(\partial^2 C_i / \partial \lambda_i \partial \theta_i)x + \partial^2 C_i / \partial \theta_i^2)$ is uniformly nonnegative because the coefficient of the quadratic term is positive and the discriminant is zero. Choosing $x = \partial \lambda_i / \partial \theta_i$ thus implies that the second term to the right of (66) is negative. It completes the proof that π_i is concave in θ_i . Moreover, because each profit function is concave in θ_i and each equilibrium θ^* is an interior point of the interval $[0, \theta_i^{\max}]$, it must satisfy the first-order conditions. \square

It is worth noting from the proof of Theorem 5 that an equilibrium in the service-competition model is guaranteed to exist when each demand function λ_i is concave in the firm's service level θ_i . (Similarly, a price equilibrium is guaranteed to exist when λ_i is concave in the price p_i .) Finally, the case of simultaneous price and service competition can be analyzed in analogy to the proof of Theorem 3. However, because in the case of attraction models the sufficient conditions guaranteeing the existence of an equilibrium become significantly less elegant, we omit the details.

6. Numerical Study

In this section, we report on a numerical study conducted to illustrate various properties of the equilibrium behavior in the price-competition, service-competition, and simultaneous-competition models. The study is built around two base instances: one with separable demand functions (as in §4), and one with demand functions specified by an attraction model (as in §5). All instances represent industries with $N = 3$ firms.

Each of the three providers is modeled as an M/G/1 queueing facility and the expected sojourn time represents

a firm’s service level. In the base scenario, the coefficient of variation (c.v.) of the service times equals one, but this c.v. value is varied in many of the instances. Finally, $c_1 = c_2 = 20$ and $c_3 = 5$, while $\gamma_1 = \gamma_2 = 35$ and $\gamma_3 = 50$. Recall from Lemma 1 that for all three firms,

$$\begin{cases} \frac{\partial C_i}{\partial \lambda_i} \geq c_i + \gamma_i = 55 & \text{if c.v.}_i \leq 1, \\ \frac{\partial C_i}{\partial \lambda_i} = c_i + \gamma_i = 55 & \text{if c.v.}_i = 1, \\ \frac{\partial C_i}{\partial \lambda_i} \leq c_i + \gamma_i = 55 & \text{if c.v.}_i \geq 1, \end{cases} \quad (67)$$

i.e., all three firms experience the same total marginal cost per customer served (at least when the c.v. value equals one), but firms 1 and 2 have relatively low-capacity costs and higher noncapacity (for example, communication) costs. Thus, firms 1 and 2 may represent (outsourced) foreign service providers, and firm 3 a domestically-based provider.

The base instance with separable demands has the following demand functions:

$$\begin{aligned} \lambda_1 &= 145 - 10p_1 + 4.5p_2 + 4.5p_3 + 100\log(\theta_1) \\ &\quad - 40\log(\theta_2) - 50\log(\theta_3), \\ \lambda_2 &= 145 - 10p_2 + 4.5p_1 + 4.5p_3 + 100\log(\theta_2) \\ &\quad - 40\log(\theta_1) - 50\log(\theta_3), \\ \lambda_3 &= 235 - 10p_3 + 4.5p_1 + 4.5p_2 + 100\log(\theta_3) \\ &\quad - 40\log(\theta_1) - 40\log(\theta_2). \end{aligned}$$

In Table 2, we characterize the equilibria which arise under price competition for eight distinct but common c.v. values, and two distinct but common service levels. Because firms 1 and 2 are interchangeable, we only report

the equilibrium price, demand volume, capacity, and profit level of firms 1 and 3. Focusing first on the case of a moderate service-level guarantee $\theta_i = 5$, observe that in the base case with c.v. = 1, firm 3 is able to charge a somewhat higher price, even though his marginal cost per customer, which in this case equals $c_3 + \gamma_3 = 55$ throughout, is identical to that of his competitors. If all firms were to adopt identical prices along with identical service guarantees, firm 3 would, presumably because of attributes other than price and service guarantees, enjoy an incremental demand value of 106 beyond those experienced by his competitors ($\lambda_3 - \lambda_1 = 90 + 10\log(\theta) \simeq 106$). Instead, firm 3 positions himself with a 6.5% higher price ending up with a demand volume which is only 43 instead of 106 units larger than that of his competitors. Nevertheless, the 6.5% higher price and the 38% larger demand volume contribute to ensuring this firm of almost double the profit earned by his competitors.

Observe next that the equilibrium is rather insensitive to the c.v. value, as long as it is below one, i.e., when the capacity cost function is convex in the demand volume (the case treated in Theorem 1). At the same time, the equilibrium behavior is considerably more sensitive to the service variability when the c.v. values are larger than one. This corresponds with the case where the capacity cost functions are concave in the demand volume (see Lemma 1), where it is considerably harder to provide sufficient conditions guaranteeing the existence of a price equilibrium. Nevertheless, the reported equilibria were found by applying the tatônnement scheme described in §4. Note that when the scheme converges, its limit point is necessarily a Nash equilibrium. Moreover, we have been able to check that each of the reported equilibria is unique by verifying that it arises as a limit point, irrespective of the scheme’s starting vector of prices. However, when the service time becomes extremely volatile (i.e., the c.v. value becomes very large), an equilibrium may fail to exist or multiple equilibria may

Table 2. Price competition with separable demand: Equilibria under different c.v. values.

θ_i	c.v. _i	p_1	λ_1	μ_1	π_1	p_3	λ_3	μ_3	π_3
5.00	0.20	66.41	114.17	174.06	1,209.76	70.74	157.48	239.00	2,346.74
	0.40	66.41	114.17	174.66	1,199.32	70.74	157.48	239.60	2,331.79
	0.60	66.41	114.17	175.65	1,182.03	70.74	157.48	240.59	2,307.01
	0.80	66.41	114.17	177.02	1,158.07	70.74	157.48	241.96	2,272.57
	1.00	66.42	114.16	178.74	1,128.70	70.75	157.47	243.70	2,230.13
	2.00	66.58	113.95	192.02	906.72	70.90	157.41	257.55	1,904.30
	3.00	67.08	113.27	210.96	606.22	71.36	157.31	278.46	1,448.69
	4.00	67.97	112.25	233.51	272.25	72.22	156.72	303.71	920.49
10.00	0.20	67.11	121.71	188.37	1,284.72	71.73	167.74	257.31	2,538.90
	0.40	67.12	121.70	189.54	1,265.05	71.74	167.73	258.49	2,510.64
	0.60	67.12	121.74	191.56	1,231.43	71.74	167.73	260.45	2,461.48
	0.80	67.14	121.63	194.08	1,185.57	71.75	167.81	263.29	2,396.61
	1.00	67.17	121.60	197.40	1,129.88	71.78	167.78	266.67	2,315.38
	2.00	67.63	121.01	221.78	736.04	72.21	167.62	292.79	1,725.88
	3.00	68.82	119.64	254.42	254.04	73.36	166.83	329.17	964.97
	4.00	70.67	117.79	291.19	-245.78	75.21	164.98	370.53	132.76

Table 3. Centralized solution under different c.v. values.

θ_i	c.v. _{<i>i</i>}	p_1	p_2	p_3	λ_1	λ_2	λ_3	π
5.00	0.20	124.50	124.50	128.00	52.34	52.34	107.69	14,813.96
	0.40	124.50	124.50	128.00	52.34	52.34	107.69	14,778.34
	0.60	124.50	124.50	128.00	52.34	52.34	107.69	14,719.64
	0.80	124.50	124.50	128.00	52.34	52.34	107.69	14,638.84
	1.00	124.50	124.50	128.00	52.34	52.34	107.69	14,537.15
	2.00	125.00	125.00	128.50	51.84	51.84	107.19	13,766.72
	3.00	124.50	129.50	127.50	72.59	0.0	135.19	12,836.96
	4.00	134.00	134.00	128.00	0.0	0.0	193.19	11,897.54
10.00	0.20	129.00	129.00	133.00	57.03	57.03	112.05	16,514.58
	0.40	129.00	129.00	133.00	57.03	57.03	112.05	16,444.28
	0.60	129.00	129.00	133.00	57.03	57.03	112.05	16,329.49
	0.80	129.00	129.00	133.00	57.03	57.03	112.05	16,173.40
	1.00	129.00	129.00	133.00	57.03	57.03	112.05	15,979.86
	2.00	131.00	136.50	135.00	79.78	0.0	134.80	14,594.08
	3.00	131.50	136.50	134.50	72.53	0.0	142.05	13,157.57
	4.00	141.00	141.00	135.00	0.0	0.0	200.05	11,979.01

arise. For example, when $\theta = 5$ and c.v. = 6, no equilibrium exists because the tatōnement scheme oscillates between the following three price vectors: (82.37, 72.9, 77.6), (71.54, 83.21, 77.5), and (73.4, 71.61, 75.67), irrespective of its starting point. When $\theta = 5$ and c.v. = 8, the following two price vectors are the equilibria $p^{*1} = (76.55, 87.08, 81.2)$ and $p^{*2} = (87.08, 76.55, 81.2)$. Note that, even though firms 1 and 2 share identical characteristics, in equilibrium one of them charges the highest price of \$87.08 and the other the lowest price of \$76.55, while firm 3 in both equilibria charges an intermediate price of \$81.2.

As the c.v. value increases, all firms respond to the upward shift of their cost function by increasing their prices. The price difference between firm 3 and its competitors is more or less maintained at the \$4.3 level, exhibited in the base case, with c.v. = 1. For c.v. values below one, the equilibrium is rather insensitive to this measure, especially because a relatively low service level is pursued ($\theta = 5$). The changes in the equilibria become more pronounced when the c.v. value is above one, with the marginal impacts of an increase of the c.v. value by one unit becoming increasingly larger. The equilibrium capacity levels are rather insensitive to the c.v. value as long as it is below one; thereafter the required capacity levels grow rapidly and superlinearly.

The same sensitivities with respect to the service-time variability arise when the firms double their service level to $\theta = 10$. In the base case, with c.v. = 1, the firms charge approximately 75 cents more to enable the higher service level, and they enjoy a modest increase in their demand volumes; i.e., the positive impact of the service improvements outweighs the negative impact of the price increases. Nevertheless, even though both the price and the demand volume increase, these benefits are dominated by the increase in the capacity cost, resulting in lower equilibrium profits for all three firms. However, while the firms prefer to operate under the higher service level $\theta = 10$ when c.v. ≤ 1 ,

the opposite is true when the c.v. value is above one. When c.v. ≤ 1 , the increase in capacity costs resulting from higher service levels is more than offset by the ability to raise prices and still attract more customers; when c.v. ≥ 1 , the incremental capacity cost dominates the ability to raise prices.

In Table 3, we report the optimal solution for the same 16 instances, when all three service facilities are owned by the same firm, thus operating as a monopoly. (The value of θ_i continues to denote the service level maintained at facility i .) As is usually the case, in the monopoly solutions, the prices are significantly larger, the demand values are significantly lower, and the aggregate profits are significantly higher than their counterparts in a competitive market. As long as the service-time volatility is not very large, service facilities 1 and 2 (with identical characteristics) charge identical prices. However, when c.v. ≥ 3 for $\theta = 5$ and c.v. ≥ 2 for $\theta = 10$, one or both of facilities 1 and 2 are phased out, even though aggregate demand declines only modestly. All prices increase with the c.v. value and the desired service levels.

Next, it is of interest to compare the results in Table 2 with those in settings where the firms face identical demand functions but maintain the above differentiated cost structure. To this end, we consider the following uniform demand function: $\lambda_i = 145 - 10p_i + 4.5 \sum_{j \neq i} p_j + 100 \log(\theta_i) - 40 \sum_{j \neq i} \log(\theta_j)$, $i = 1, \dots, 3$. The results are exhibited in Table 4. The reduced intercepts of the demand functions induce a price reduction for all three firms, in particular firm 3, whose intercept is reduced most significantly. While firm 3 enjoys the largest profits in the industry in the initial set of instances of Table 2, its profits now become the industry's lowest. All of the qualitative properties reported for our initial set of instances continue to apply (see Table 2). Note that when a relatively large service level ($\theta = 10$) is pursued and the c.v. value is very high (c.v. = 4), firm 3, with its much higher capacity cost rate,

Table 4. Price competition with separable demand: Equilibria under different c.v. values with identical demand functions.

θ_i	c.v. _{<i>i</i>}	p_1	λ_1	μ_1	π_1	p_3	λ_3	μ_3	π_3
5.00	0.20	66.09	111.14	169.53	1,139.62	66.09	111.10	169.46	1,099.28
	0.40	66.09	111.14	170.12	1,129.18	66.09	111.10	170.06	1,084.37
	0.60	66.09	111.14	171.11	1,111.90	66.09	111.10	171.04	1,059.68
	0.80	66.10	111.09	172.40	1,088.46	66.10	111.09	172.40	1,026.47
	1.00	66.11	111.08	174.12	1,059.09	66.11	111.08	174.12	984.09
	2.00	66.30	111.02	187.60	842.25	66.33	110.59	186.94	664.07
	3.00	66.87	110.95	207.37	556.59	67.01	108.92	204.24	224.12
	4.00	67.87	111.03	231.56	247.45	68.25	105.52	222.72	-275.36
10.00	0.20	67.32	123.69	191.33	1,334.79	67.31	123.83	191.55	1,254.34
	0.40	67.32	123.69	192.52	1,314.03	67.31	123.83	192.74	1,224.67
	0.60	67.33	123.68	194.46	1,280.96	67.32	123.82	194.67	1,176.96
	0.80	67.34	123.67	197.13	1,235.08	67.33	123.81	197.34	1,110.94
	1.00	67.37	123.73	200.59	1,180.50	67.37	123.68	200.52	1,029.94
	2.00	67.87	123.68	225.87	798.06	67.97	122.18	223.58	452.00
	3.00	69.13	123.59	260.75	339.80	69.50	118.22	252.14	-281.06
	4.00	74.42	169.06	377.24	1,031.61	86.07	0.13	16.99	-540.69

is driven out of the market, a phenomenon not observed when it enjoys the larger intercept in its demand function.

In Table 5, we characterize the equilibrium behavior in the service-competition model for 24 problem instances obtained from the base instance by considering the above c.v. values of the service times, in combination with three exogenously given and common price levels. The first price level $p_i = 68.75$ is, approximately, the average equilibrium of the prices in the price-competition model discussed above, in the base case with c.v. = 1. Moreover, it

represents a 25% markup above the common variable cost rate $c_i + \gamma_i = 55$. The third price level, $p_i = 71.5$, reflects a 35% markup, and the second price level, $p_i = 70.35$, corresponds with the average equilibrium price charged by the firms in the simultaneous competition model with c.v. = 1, discussed below. While Theorem 2 guarantees the existence of an equilibrium only when the c.v. value is at or below one, the above tatônnement scheme converges to a Nash equilibrium for all c.v. values; moreover, the reported equilibrium is unique in all instances.

Table 5. Service competition with separable demand: Equilibria under different c.v. values.

p_i	c.v. _{<i>i</i>}	θ_1	λ_1	μ_1	π_1	θ_3	λ_3	μ_3	π_3
68.75	0.20	58.89	5,612.52	8,450.89	73,849.93	44.79	4,319.12	6,503.09	56,489.99
	0.40	55.14	5,244.17	7,902.92	68,885.37	41.25	3,970.38	5,983.00	51,803.23
	0.60	49.85	4,725.17	7,130.84	61,890.61	36.47	3,501.19	5,283.31	45,496.64
	0.80	43.94	4,147.44	6,271.39	54,104.68	31.42	3,006.09	4,545.05	38,840.63
	1.00	38.14	3,581.05	5,428.79	46,472.01	26.70	2,545.34	3,858.06	32,645.33
	2.00	18.20	1,656.37	2,565.02	20,545.36	11.97	1,131.09	1,749.57	13,627.17
	3.00	9.75	868.39	1,391.20	9,957.07	6.28	612.09	975.39	6,661.42
	4.00	5.91	530.38	885.96	5,446.97	3.78	402.52	662.20	3,865.77
70.35	0.20	66.29	6,339.24	9,544.99	96,094.80	50.70	4,899.45	7,376.81	73,881.69
	0.40	62.30	5,947.39	8,962.52	90,022.14	46.83	4,517.01	6,806.66	67,972.19
	0.60	56.63	5,389.85	8,133.74	81,381.94	41.58	3,999.29	6,034.88	59,970.91
	0.80	50.23	4,761.65	7,199.90	71,647.14	35.97	3,448.75	5,214.25	51,460.66
	1.00	43.86	4,137.92	6,272.66	61,982.00	30.70	2,932.18	4,444.31	43,473.89
	2.00	21.36	1,956.62	3,029.39	28,194.29	13.96	1,315.86	2,035.54	18,479.88
	3.00	11.54	1,031.34	1,651.97	13,893.20	7.37	706.04	1,126.27	9,065.48
	4.00	7.03	625.23	1,045.29	7,655.40	4.45	454.04	749.71	5,197.31
71.5	0.20	80.16	7,704.90	11,601.06	146,853.27	61.83	5,992.62	9,022.63	113,742.44
	0.40	75.83	7,277.76	10,967.07	138,550.83	57.40	5,554.02	8,369.21	105,243.29
	0.60	69.56	6,659.88	10,049.95	126,541.69	51.32	4,953.71	7,474.93	93,608.13
	0.80	62.34	5,949.19	8,995.04	112,729.42	44.76	4,306.48	6,510.88	81,060.95
	1.00	55.00	5,227.93	7,924.39	98,712.61	38.50	3,690.16	5,592.99	69,110.64
	2.00	27.81	2,574.15	3,984.20	47,156.87	18.00	1,694.23	2,620.95	30,398.94
	3.00	15.28	1,376.72	2,204.08	23,935.66	9.62	904.57	1,444.50	15,101.62
	4.00	9.38	831.29	1,390.26	13,411.87	5.85	566.54	939.68	8,586.04

Table 6. Simultaneous competition with separable demand: Equilibria under different c.v. values.

c.v. _{<i>i</i>}	<i>p</i> ₁	θ_1	λ_1	μ_1	π_1	<i>p</i> ₃	θ_3	λ_3	μ_3	π_3
0.20	67.89	57.39	6,559.18	9,870.01	83,495.54	73.46	58.93	6,822.30	10,265.53	124,406.40
0.40	68.05	54.23	6,244.39	9,402.60	80,408.28	73.50	54.91	6,419.56	9,665.81	117,188.69
0.60	68.28	49.70	5,792.77	8,732.08	75,739.70	73.55	49.31	5,858.29	8,830.03	106,994.74
0.80	68.52	44.52	5,276.37	7,965.42	70,066.95	73.58	43.15	5,239.73	7,908.89	95,582.98
1.00	68.75	39.28	4,752.99	7,188.40	63,969.63	73.58	37.16	4,637.57	7,012.09	84,307.42
2.00	69.33	19.71	2,795.06	4,280.29	38,348.78	73.17	16.94	2,597.08	3,971.02	45,086.60
4.00	69.04	6.29	1,437.43	2,256.59	18,361.68	71.80	4.98	1,367.52	2,131.22	20,909.18

Note that, in contrast to the price-competition model, the equilibrium service levels are highly sensitive to the variability of the service times. For example, to the extent the c.v. value can be reduced from 1 to 0.2, all firms will improve their service levels by approximately 50% under each of the three price levels, and still improve profits by (more than) 50% as well. Equilibrium capacity levels grow by 50%–60%. Due to its higher capacity cost rate and the fact that it charges the same price as its competitors, firm 3 consistently offers a significantly lower service level and is forced to accept the smallest market and profit shares. For example, with $p_i = 68.75$ and c.v. = 1, firm 3’s equilibrium demand volume is 1,036 units below his competitors, while it would exceed the latter by a least 90 units if the firm matched the competitors’ service levels. Finally, equilibrium service levels are quite sensitive to the price levels and increase monotonically with the latter.

Table 6 characterizes the equilibrium positions adopted by the three firms under simultaneous competition, again, for each of the above eight c.v. values. As in the price-only competition model, equilibrium prices are rather insensitive to, and fail to be monotone in the c.v. value. In contrast, and consistent with the service-only competition model, the equilibrium service levels vary largely with the variability of the service process. Under simultaneous competition, Firm 3 adopts a service level close to that of his competitors, compared to the case where the firm is conditioned to match his competitors’ price (see Table 5). The partial matching of the competitors’ service level is enabled by charging a higher price. The equilibrium market shares are much closer to being equal than in the service-competition model. When conditioned to match the competitors’ price, firm 3’s profit is between a third and a quarter lower than that of his competitors. Under simultaneous competition, profit values are much more uniform; in fact, firm 3 is the most profitable firm. Finally, we compare the results with those in Table 5, under the common prespecified price 70.35, which equals the average equilibrium price charged under simultaneous competition with c.v. = 1. Almost invariably, all three firms earn higher profits under simultaneous competition; as mentioned, this is in particular true for firm 3. Interestingly, these higher profits occur with larger demand volumes, i.e., unconditional (simultaneous) competition favors the consumer and all of the competing firms alike. It is, however, unclear whether this comparison applies in general.

The M/G/1 model with varying c.v. values captures one dimension of service variability: all customers are served by the same server, but the spread of the service times of individual customers increases with the c.v. value. Another dimension of service variability arises when customers need to be partitioned into different classes, each with a dedicated pool of servers. (For example, a call center may need to be multilingual while the service agents are proficient in only one language.) This dimension of variability may be captured by a simple Jackson network with *J* parallel nodes, each catering to 1/*J* of the customer base, and each incurring an identical capacity cost rate γ . Recall from §3 that the capacity cost function associated with the Jackson network is linear, i.e., of type \mathcal{E}^{LJN} . While in the M/G/1 model, the c.v. value impacts the B_4 coefficient (only); in the Jackson network, the number of distinct customer classes *J* impacts only the coefficient B_2 . Indeed, it follows from (23) that B_1 is independent of *J*, while B_2 increases linearly. Figure 2 exhibits the equilibrium service levels under simultaneous competition as a function of *J*. These decline more than proportionally with *J*, exemplifying the potential benefit of highly cross-trained agents, within the context of oligopoly competition models.

We now turn to instances with demand functions specified by an attraction model with $M = 15,000$, and with the

Figure 2. Equilibrium service levels as a function of the number of nodes.

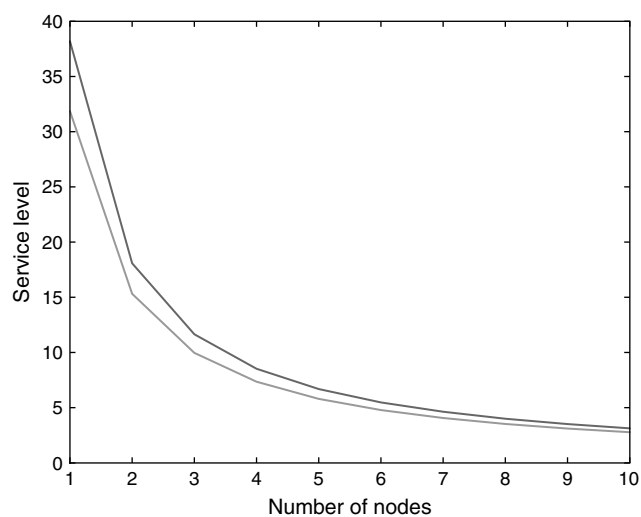


Table 7. Simultaneous competition with attraction demand model: Equilibria under different c.v. values.

c.v. _{<i>i</i>}	<i>p</i> ₁	<i>θ</i> ₁	<i>λ</i> ₁	<i>μ</i> ₁	<i>π</i> ₁	<i>p</i> ₃	<i>θ</i> ₃	<i>λ</i> ₃	<i>μ</i> ₃	<i>π</i> ₃
0.2	91.11	58.17	1,877.18	1,907.88	66,549.38	126.89	72.62	3,324.64	3,362.80	236,822.5
0.4	91.06	50.24	1,875.5	1,904.97	66,385.3	126.84	62.61	3,323.87	3,360.47	236,572.38
0.6	91.01	44.21	1,874.04	1,904.33	66,242.72	126.79	55.03	3,323.2	3,360.82	236,355.25
0.8	90.97	39.47	1,872.74	1,905.23	66,116.68	126.74	49.08	3,322.61	3,362.96	236,163.42
1	90.93	35.65	1,871.58	1,907.23	66,003.73	126.7	44.29	3,322.09	3,366.38	235,991.6
2	90.79	24.01	1,867.08	1,926.00	65,567.71	126.55	29.77	3,320.05	3,393.51	235,328.84
10	90.32	6.62	1,852.43	2,142.39	64,164.62	126.05	8.21	3,313.47	3,686.91	233,196.26
20	90.08	3.46	1,845.05	2,382.96	63,466.62	125.8	4.31	3,310.17	4,022.12	232,133.17
60	89.68	1.19	1,832.78	3,100.01	62,319.06	125.39	1.48	3,304.69	5,049.25	230,380.18

following attraction functions:

$$v_i(p_i, \theta_i) = \begin{cases} 1,800 - 15p_i + 10\log(\theta_i) & \text{if } i = 1, 2, \\ 2,700 - 15p_i + 10\log(\theta_i) & \text{if } i = 3, \end{cases}$$

*v*₀ = 2,000. Maintaining the same cost structures as before, firms 1 and 2 continue to have identical characteristics. Firm 3 continues to enjoy a larger market share than his competitors when offering identical prices and service levels. At the same time, firm 3 continues to have significantly higher capacity costs. Table 7 describes the industry equilibrium under simultaneous competition for the above instances with 8 c.v. values. Once again, the reported equilibria are obtained by the tatônnement scheme, and they are unique. Starting with the base case, c.v. = 1, we note that in this case, firm 3 positions himself as the high-price, high-service provider, capturing almost the same market share as those of his competitors combined, and close to four times the profits each of them makes. His 40% higher

price and intrinsic advantage (due to other “attributes”), as reflected by the larger intercept of the attraction function, allows firm 3 to offer an approximately 25% higher service level in spite of the significantly higher capacity cost rate *γ*. As in the case of separable demand functions, the equilibrium prices are rather insensitive to the c.v. value, while the equilibrium service levels decrease by close to 40% when the c.v. values increases from 0.2 to 1 and by a factor of 2 when the c.v. value increases from 1 to 20.

Table 8 describes the equilibria in the price-competition model under 2 common service levels (the equilibrium service levels in Table 7 for firms 1 and 2 under c.v. = 1 and c.v. = 0.2, respectively). In this case, firm 3 “suffers” only very moderately from the upfront restriction to a common service level, while his two competitors benefit, albeit, again, very moderately. The equilibrium prices are increasing in the c.v. value, as well as in the given service levels. Finally, Table 9 describes the equilibria in the service-level competition model under two common price levels. The first price level *p*_{*i*} = 66 represents a 20% markup above the

Table 8. Price competition with attraction demand model: Equilibria under different c.v. values.

<i>θ</i> _{<i>i</i>}	c.v. _{<i>i</i>}	<i>p</i> ₁	<i>λ</i> ₁	<i>μ</i> ₁	<i>π</i> ₁	<i>p</i> ₃	<i>λ</i> ₃	<i>μ</i> ₁	<i>π</i> ₃
35	0.2	90.93	1,871.64	1,890.00	66,507.38	126.61	3,319.4	3,337.69	236,648.07
	0.4	90.93	1,871.64	1,892.10	66,385.58	126.61	3,319.4	3,339.79	236,473.63
	0.6	90.93	1,871.64	1,895.58	66,264.23	126.61	3,319.4	3,343.28	236,299.55
	0.8	90.93	1,871.64	1,900.44	66,143.32	126.61	3,319.4	3,348.15	236,125.84
	1	90.93	1,871.64	1,906.64	66,022.86	126.61	3,319.4	3,354.40	235,952.49
	2	90.93	1,871.64	1,956.85	65,426.92	126.61	3,319.4	3,405.57	235,091.04
	10	91.02	1,867.98	2,986.58	61,028.65	126.66	3,319.46	4,603.67	228,661.35
	20	91.23	1,859.46	4,672.94	56,209.04	126.78	3,319.43	6,775.35	221,425.2
	40	91.77	1,837.6	8,167.99	48,023.45	127.12	3,317.88	11,456.68	208,719.78
	50	92.07	1,825.43	9,913.31	44,434.89	127.32	3,316.49	13,834.96	203,016.13
60	92.37	1,813.26	1,813.26	41,091.69	127.53	3,314.62	3,314.62	197,627.01	
58	0.2	91.11	1,877.54	1,908.15	66,564.88	126.8	3,322.07	3,352.48	236,772.6
	0.4	91.11	1,877.54	1,911.61	66,363.83	126.8	3,322.07	3,355.96	236,484.15
	0.6	91.11	1,877.54	1,917.37	66,163.99	126.8	3,322.07	3,361.73	236,196.69
	0.8	91.11	1,877.54	1,925.36	65,965.33	126.8	3,322.07	3,369.78	235,910.21
	1	91.11	1,877.54	1,935.54	65,767.84	126.8	3,322.07	3,380.07	235,624.71
	2	91.12	1,877.17	2,016.17	64,802.54	126.81	3,321.87	3,463.32	234,230.31
	10	91.34	1,868.24	3,471.37	57,933.98	126.93	3,322.11	5,210.58	224,082.18
	20	91.76	1,851.31	5,680.42	50,795.3	127.19	3,321.13	8,114.78	213,111.34
	40	92.74	1,811.61	10,147.84	39,221.75	127.85	3,316.3	14,202.21	194,624.8
	50	93.23	1,791.66	12,357.18	34,285.59	128.21	3,312.43	17,270.31	186,505.24
	60	93.73	1,771.03	14,540.98	29,747.54	128.57	3,308.78	20,342.82	178,964.37

Table 9. Service competition with attraction demand: Equilibria under different c.v. values.

p_i	c.v. _{i}	θ_1	λ_1	μ_1	π_1	θ_3	λ_3	μ_3	π_3
91	0.2	51.74	1,648.44	1,675.75	5,8243.85	28.34	4,754.44	4,769.22	170,307.42
	0.4	44.79	1,645.11	1,671.38	58,117.66	24.36	4,754.02	4,768.18	170,290.88
	0.6	39.49	1,642.18	1,669.24	58,007.76	21.36	4,753.68	4,768.23	170,277.15
	0.8	35.31	1,639.59	1,668.65	57,910.41	19.02	4,753.38	4,768.99	170,265.44
	1	31.94	1,637.25	1,669.19	57,823.01	17.15	4,753.13	4,770.28	170,255.25
	2	21.64	1,628.14	1,681.22	57,484.15	11.49	4,752.19	4,780.81	170,218.2
	10	6.09	1,598.05	1,862.76	56,376.83	3.17	4,749.54	4,904.66	170,112.97
	20	3.22	1,582.71	2,075.74	55,815.91	1.67	4,748.3	5,062.46	170,064.06
	40	1.66	1,566.66	2,425.54	55,229.61	0.86	4,747.03	5,357.15	170,014.2
	50	1.34	1,561.38	2,577.13	55,037.06	0.69	4,746.62	5,492.40	169,997.98
	60	1.12	1,557.04	2,714.30	54,878.63	0.58	4,746.29	5,627.19	169,984.69
66	0.2	12.25	2,319.83	2,326.22	25,260.39	6.92	4,804.21	4,807.81	52,638.69
	0.4	10.53	2,317.58	2,323.70	25,235.13	5.94	4,804.02	4,807.47	52,636.4
	0.6	9.23	2,315.62	2,321.90	25,213.21	5.2	4,803.86	4,807.40	52,634.43
	0.8	8.22	2,313.89	2,320.63	25,193.84	4.63	4,803.72	4,807.52	52,632.72
	1	7.41	2,312.33	2,319.74	25,176.48	4.17	4,803.59	4,807.76	52,631.19
	2	4.96	2,306.33	2,318.69	25,109.43	2.79	4,803.11	4,810.08	52,625.41
	10	1.37	2,286.85	2,354.10	24,892.45	0.77	4,801.62	4,840.20	52,607.31
	20	0.72	2,277.03	2,413.28	24,783.21	0.4	4,800.88	4,879.79	52,598.37
	40	0.37	2,266.79	2,531.99	24,669.34	0.21	4,800.11	4,962.71	52,589.09
	50	0.3	2,263.44	2,591.18	24,632	0.17	4,799.86	5,003.79	52,586.05
	60	0.25	2,260.68	2,645.38	24,601.3	0.14	4,799.65	5,039.72	52,583.56

variable cost rate $c_i + \gamma_i = 55$, and the second price level $p_i = 91$ corresponds approximately to the equilibrium price of firms 1 and 2 in the simultaneous-competition model. When forced to match his competitors' price, firm 3 positions himself as the *low service provider*, in contrast to the case of simultaneous and unrestricted competition where this firm arises as the *high-price and high service-level provider*. For example, when tied to a price level of $p_i = 91$, firm 3's equilibrium service level is $\theta_3 = 17.1$, instead of $\theta_3 = 44.2$ in the case of simultaneous competition. All firms are worse off under prespecified price levels, as compared to simultaneous unrestricted competition, with firm 3 experiencing the largest relative profit loss. All equilibrium service levels decrease with the c.v. value and increase with the given price level.

Finally, we have also evaluated a set of instances with attraction functions that are linear in both price and service level. The price-only competition model and the service-only competition model exhibit similar patterns of equilibrium behavior. At the same time, under simultaneous (unrestricted) competition, the firms are driven to adopt the maximum permitted price p^{\max} and associated high service levels. This observation applies consistently to a large set of parameters, and may explain why, for attraction models, it is hard to guarantee, a priori, that an equilibrium exists in the simultaneous-competition model.

7. Conclusions

We have investigated how competing service providers differentiate themselves by selecting price and/or service-level guarantees. Recognizing that customers select a specific service provider by considering all of the firms' prices and

service levels, as well as other attributes, we have analyzed these questions by adopting consumer-choice models which specify the demand rate of each firm as a general, possibly nonlinear, demand function of the complete vector of the industry prices p , and the industry service levels θ . Our analyses have focused on two broad classes of demand functions: (1) demand functions that are separable in (p, θ) and linear in p ; and (2) demand functions specified by an attraction model.

The second major building block of any competition model for service industries consists of an adequate representation of the capacity cost faced by each provider as a function of his demand volume and service level. Close to 100 years of queueing theory have taught us that the performance of a service facility depends critically on a plethora of characteristics: for example, the types of interarrival time and service-time processes, the number of servers, whether service consists of single or multiple tasks, etc. Moreover, rather different performance analysis techniques are used to address different queueing models, and very few allow for exact, let alone, closed-form evaluations. This hybrid and balkanized state of queueing theory notwithstanding, we have shown across a broad spectrum of standard queueing models that the capacity cost functions, either exactly or as a close approximation, belong to a specific four-parameter class of functions \mathcal{C} . This characterization has allowed us to analyze the equilibrium behavior in service-competition models in a unified manner without having to perform a separate analysis for each possible combination of queueing model assumptions. (As mentioned in the literature review, earlier work, perhaps to avoid this complication, has almost invariably confined itself to assuming that

the service providers operate as simple M/M/1 systems.) The class of capacity cost functions \mathcal{C} contains instances where the capacity cost exhibits economies of scale in the demand volume or the service level, as well as cases where it exhibits diseconomies of scale. The class also contains cases where the marginal cost per customer increases with the service level (i.e., the capacity cost is supermodular), as well as cases where it decreases with the service level (i.e., the capacity cost is submodular). A simple inequality involving the four parameters in the class \mathcal{C} determines whether the capacity cost function is jointly convex, linear, or jointly concave in the demand volume and service levels, and whether it is supermodular, separable, or submodular in this pair or variables. (See Lemma 1.) This inequality often reduces to a simple condition with respect to the parameters of the underlying queueing models: for example, in the M/G/1 model, with service levels based on expected waiting times, the capacity cost function is jointly convex (and submodular) in the demand rate and the service level if the c.v. value of the service-time distribution is less than one. It is linear if the c.v. value equals one, and jointly concave (and supermodular) if it is larger than one.

In analyzing the industry's competitive behavior, we systematically consider the case of *price competition* (firms compete in terms of their prices under exogenously given service levels), *service competition* (firms compete in terms of their service levels under exogenously given prices), and *simultaneous price and service competition* (where firms compete in terms of both). We have proven that, as long as all capacity cost functions are convex, an equilibrium exists, both in the price-competition and in the service-competition models, and both under separable demand functions, and demand functions that are specified by an attraction model. (In the case of attraction models, the existence results are shown under some mild conditions with respect to the so-called attraction functions. To guarantee an equilibrium in the price-competition model, it is, for example, sufficient that the attraction functions be log-concave in the price variable—a condition that is satisfied in virtually all commonly used specifications. In the case of service competition, we require that the attraction functions be concave in the service level; while intuitive, the condition fails to hold in traditional multinomial logit specifications.) We characterize the equilibrium through a system of equations, directly exhibiting how it depends on the shape of the capacity cost function, and hence on the characteristics of the industry's queueing models. Through our theoretical and numerical investigations, we have also explored how the equilibrium in the price-competition model depends on the given service levels, and vice versa, how the equilibrium in the service-competition model depends on the exogenously given prices. Here, the shape of the capacity cost function may impact qualitatively on various comparative statics.

When some or all of the capacity cost functions are concave, it is considerably harder to provide simple sufficient conditions that an equilibrium exists; indeed, we

have identified some examples where either no equilibrium exists or multiple equilibria prevail. However, this phenomenon appears to arise only under very large c.v. values for the service-time distribution.

Returning to the case of convex capacity cost functions, we show that under separable demand functions, an equilibrium is also guaranteed to exist in the simultaneous-competition model without additional restrictions on the structure of the demand equations. In contrast, for attraction models, significant conditions on the attraction functions are required in our analysis; moreover, our numerical investigations have shown that even when the attraction functions are linear in the price and service level, the equilibrium in the simultaneous-competition model is always on the boundary of the feasible region, no matter how large p^{\max} and θ^{\max} are chosen. At the same time, an interior point of the feasible region arises as the unique equilibrium when each firm's attraction value grows logarithmically with the service level.

Our numerical study has identified many qualitative insights; here we single out the following: the shape of the capacity cost function may, in some cases, significantly impact the industry's equilibria. For example, if all service facilities operate as an M/G/1 system with a common service-time distribution, the c.v. value tends to have a significant impact on the equilibrium service levels, both under service competition and under simultaneous price and service competition. (Equilibrium prices tend to be much less sensitive to the c.v. value.) The impact on equilibrium demand volumes and equilibrium profits can, if anything, be even larger. We have identified examples where even the relative market share and profit share of individual firms rapidly change as a function of the service-time variability, even though in these instances all firms experience the same c.v. value throughout. Such effects remain, of course, entirely eclipsed when representing the service facilities by simple M/M/1 systems. We have also demonstrated how our model can be used to quantify the benefits of pooling distinct server groups on equilibrium service levels, demand volumes, and profit values.

Endnotes

1. The asymptotic accuracy of the exponential approximation in heavy traffic is highly relevant in our context. Most call centers, for example, aim at a utilization or occupancy rate in the 85%–90% range; see, e.g., Reynolds (2005) and Rosenberg (2005). Mandelbaum (2004) shows a table with utilization rates for the regional call centers of a nationwide bank. All utilization rates vary between 82% and 95%.
2. We assume that the set of competing firms is given and do not model potential entry or exit from the industry.

Acknowledgments

The authors thank Ward Whitt and Assaf Zeevi for many helpful comments regarding an earlier version of this paper.

References

- Abate, J., G. L. Choudhury, W. Whitt. 1995. Exponential approximations for tail probabilities in queues, I: Waiting times. *Oper. Res.* **43**(5) 885–901.
- Abate, J., G. L. Choudhury, W. Whitt. 1996. Exponential approximations for tail probabilities in queues, II: Sojourn time and workload. *Oper. Res.* **44**(5) 758–763.
- Allon, G., A. Federgruen. 2007. Competition in service industries. *Oper. Res.* **55**(1) 37–55.
- Anderson, S. P., A. de Palma, J. F. Thisse. 1992. *Discrete Choice Theory and Product Differentiation*. MIT Press, Cambridge, MA.
- Armony, M., M. Haviv. 2003. Price and delay competition between two service providers. *Eur. J. Oper. Res.* **147**(1) 32–50.
- Asmussen, S. 1987. *Applied Probability and Queues*. John Wiley, New York.
- Banker, R. D., I. Khosla, K. K. Sinha. 1998. Quality and competition. *Management Sci.* **44**(9) 1179–1192.
- Bell, D., R. Keeny, J. Little. 1975. A market share theorem. *J. Marketing Res.* **12** 136–141.
- Bernstein, F., A. Federgruen. 2004a. A general equilibrium model for industries with price and service competition. *Oper. Res.* **52**(6) 868–886.
- Bernstein, F., A. Federgruen. 2004b. Comparative statics, strategic complements and substitutes in oligopolies. *J. Math. Econom.* **40**(6) 713–746.
- Borovkov, A. A. 1976. *Stochastic Processes in Priority Queueing Theory*. Springer-Verlag, New York.
- Cachon, G. P., P. T. Harker. 2002. Competition and outsourcing with scale economics. *Management Sci.* **48**(10) 1314–1333.
- Cachon, G. P., F. Zhang. 2003. Procuring fast delivery, part I: Multi-sourcing and scorecard allocation of demand via past performance. Working paper, The Wharton School, University of Pennsylvania, Philadelphia, PA.
- Chen, H., Y.-W. Wan. 2003. Price competition of make-to-order firms. *IIE Trans.* **35**(9) 817–832.
- Chen, H., Y.-W. Wan. 2005. Capacity competition of make-to-order firms. *Oper. Res. Lett.* **33**(2) 187–194.
- Christ, D., B. Avi-Itzhak. 2002. Strategic equilibrium for a pair of competing servers with convex cost and balking. *Management Sci.* **48**(6) 813–820.
- De Vany, A., T. Saving. 1983. The economics of quality. *J. Political Econom.* **91**(6) 979–1000.
- Feller, W. 1971. *An Introduction to Probability Theory and Its Applications*, Vol. II, 2nd ed. John Wiley, New York.
- Gallego, G., W. T. Huh, W. Kang, R. Phillips. 2006. Price competition with the attraction demand model: Existence of unique equilibrium and its stability. *Manufacturing Service Oper. Management* **8**(4) 359–375.
- Gilbert, S., Z. K. Weng. 1997. Incentive effects favor non-competing queues in a service system: The principal-agent perspective. *Management Sci.* **44**(12) 1662–1669.
- Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.
- Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston.
- Kalai, E., M. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Sci.* **38**(8) 1154–1163.
- Kingman, J. F. C. 1962. On queues in heavy traffic. *J. Roy. Statist. Soc.* **B25** 383–392.
- Kleinrock, L. 1975. *Queueing Systems, Vol. I: Theory*. John Wiley, New York.
- Kleinrock, L. 1976. *Queueing Systems, Vol. II: Computer Application*. John Wiley, New York.
- Lederer, P. J., L. Li. 1997. Pricing, production, scheduling, and delivery-time competition. *Oper. Res.* **45**(3) 407–420.
- Leeflang, P., D. Wittink, M. Wedel, P. Naert. 2000. *Building Models for Marketing Decisions*. Kluwer Academic Publishers, Dordrecht/Boston/London.
- Levhari, D., I. Luski. 1978. Duopoly pricing and waiting lines. *Eur. Econom. Rev.* **11** 17–35.
- Li, L., Y. S. Lee. 1994. Pricing and delivery-time performance in a competitive environment. *Management Sci.* **40**(5) 633–646.
- Loch, C. 1991. Pricing in markets sensitive to delay. Ph.D. dissertation, Stanford University, Stanford, CA.
- Luski, I. 1976. On partial equilibrium in a queueing system with two servers. *Rev. Econom. Stud.* **43** 519–525.
- Mandelbaum, A. 2004. Introduction to service engineering. Lecture notes, Technion, Israel Institute of Technology, Haifa, Israel.
- Mendelson, H., S. Shneerson. 2003. Internet peering, capacity and pricing. Working paper, Stanford University, Stanford, CA.
- Milgrom, P., J. Roberts. 1990. Rationality, learning and equilibrium in games with strategic complementarities. *Econometrica* **58** 1255–1277.
- Newell, G. F. 1973. Approximate stochastic behavior of n -server service systems with large n . *Lecture Notes in Economics and Mathematical Systems*, Vol. 87. Springer, New York.
- Reitman, M. 1998. Endogenous quality differentiation in congested markets. *J. Indust. Econom.* **39** 621–647.
- Reynolds, P. 2005. Understanding agent occupancy. www.thecallcenterschool.com.
- Rosenberg, A. 2005. Best practices in workforce management. www.callcentermagazine.com.
- Seelen, L. P., H. C. Tijms, M. H. Van Hoorn. 1985. *Tables for Multi-Server Queues*. North-Holland, Amsterdam.
- Smith, W. L. 1953. On the distribution of queueing times. *Proc. Cambridge Philos. Soc.* **49** 449–461.
- So, K. C. 2000. Price and time competition for service delivery. *Manufacturing Service Oper. Management* **2**(4) 392–409.
- Tijms, H. C. 1986. *Stochastic Modelling and Analysis: A Computational Approach*. John Wiley, New York.
- Tsay, A. A., N. Agrawal. 2000. Channel dynamics under price and service competition. *Manufacturing Service Oper. Management* **2**(4) 372–391.
- Vives, X. 2000. *Oligopoly Pricing: Old Ideas and New Tools*. MIT Press, Cambridge, MA.
- Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Sci.* **38**(5) 708–723.
- Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.
- Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.