MANAGEMENT SCIENCE
Vol. 55, No. 4, April 2009, pp. 619–634
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5504 | 0619

DOI 10.1287/mnsc.1080.0946

Competition in Service Industries with Segmented Markets

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

Awi Federgruen
Columbia Business School, Columbia University, New York, New York 10027,
af7@columbia.edu

We develop a model for the competitive interactions in service industries where firms cater to multiple customer classes or market segments with the help of shared service facilities or processes so as to exploit pooling benefits. Different customer classes typically have distinct sensitivities to the price of service as well as the delays encountered. In such settings firms need to determine (i) the prices charged to all customer classes; (ii) the waiting time standards, i.e., expected steady state waiting time promised to all classes; (iii) the capacity level; and (iv) a priority discipline enabling the firm to meet the promised waiting time standards under the chosen capacity level, all in an integrated planning model that accounts for the impact of the strategic choices of all competing firms. We distinguish between three types of competition: depending on whether firms compete on the basis of their prices only, waiting time standards only, or on the basis of prices and waiting time standards. We establish in each of the three competition models that a Nash equilibrium exists under minor conditions regarding the demand volumes. We systematically compare the equilibria with those achieved when the firms service each market segment with a dedicated service process.

Key words: marketing; competitive strategy; noncooperative games; multichannel; queues
History: Received December 7, 2004; accepted December 29, 2007, by Paul H. Zipkin, operations and supply chain management. Published online in *Articles in Advance* January 5, 2009.

1. Introduction

We analyze the equilibrium behavior in service industries where firms cater to multiple customer classes or market segments with the help of shared service facilities or processes so as to exploit pooling benefits. Different customer classes typically have rather disparate sensitivities to the price of service as well as to the delays encountered. Conversely, from the firm's perspective it is vital to offer differentiated service charges and levels of service to different customer classes so as to maximize (long run) profits.

Examples of industries with the above characteristics are numerous. Banks and credit card companies segment their customers into regular and VIP or Gold and Platinum customers. Computer software and hardware firms often segment their customers, for example, into Home and Home Office users, Small Businesses, Large Businesses and the Government, or Education and Health Care sectors, using an integrated pool of technical support personnel to serve the different customer segments according to a specific priority discipline; each customer segment is associated with a specific price and waiting time expectation. Finally, overnight delivery services use their planes and trucks to deliver letters, boxes, and cargo, each with different prices and delivery time standards. In many service industries, waiting time standards are used as a primary advertised competitive instrument. For example, most major electronic brokerage firms (e.g., Ameritrade, Fidelity, E-trade) prominently feature the average or median execution speed per transaction, which is monitored by independent firms. Some firms go so far as to provide an individual execution time scorecard as part of the customer's personal account statements. As a second example, in the airline industry, independent government agencies (e.g., the Aviation Consumer Protection Division of the DOT), as well as internet travel services (e.g., Expedia) report the average delay on a flight by flight basis.

In this paper, we propose and analyze a model in which firms select all or part of the following: (i) the prices charged to all customer classes, (ii) the waiting time standards promised to all classes, (iii) the capacity level, and (iv) a priority discipline enabling the firm to meet the promised waiting time standards under the chosen capacity level. We define the *waiting time standard* offered by a given firm to a given market segment as the *maximum expected steady state waiting time in system* the firm guarantees. As to the priority discipline, modern call centers or computerized service processes allow for the easy adoption of very general priority schemes, whereas traditional "brick and mortar" service facilities may, for psychological or other reasons, be confined to simple priority rules such as FCFS or *absolute* priority schemes with an absolute priority ranking among the customer classes.

We distinguish between three types of competition: (i) *Price competition*—here, all waiting time standards are exogenously given and the firms compete on the basis of their prices only; (ii) *Waiting time* competition—here, all prices are exogenously given and the competition is in terms of waiting time standards; and (iii) *Simultaneous competition*—all prices and waiting time standards are selected simultaneously. Prices and waiting time standards are the only two essential strategic instruments. Once these are chosen by all service providers, each firm can determine a combined capacity level and priority scheme that minimizes its own cost without affecting the revenues or the costs of its competitors.

We first (§5) represent the demand rate faced by a given firm for a given market segment (customer class) as a separable function of *all* prices and waiting time standards offered to this segment in the industry, which in addition is linear in the price vector. This representation assumes that the customers are completely segmented. Each individual potential customer unambiguously belongs to one of the market segments without being able to switch between segments or to misrepresent his segment identity. In this context, a consumer is defined as an individual service-requiring unit, for instance, an individual box or letter rather than the household or firm that selects the service provider, possibly selecting one provider for its letters and a different one for its parcels. Complete segmentation is possible, for example, on the basis of (i) geographic differentiation (internet and mail delivery services or banking services); (ii) different product features (boxes versus letters, different financial products handled by electronic brokerage firms); (iii) age (senior citizens, children, and others); and (iv) the business sector (education, government, and the commercial sector).

In §6, we outline how our model and results can be generalized to settings where customers can select which class they wish to belong to, and the demand volumes are specified as functions of all prices and waiting time standards offered to *all* segments throughout the industry. The demand models allow us to represent *general* trade-offs between (i) prices, (ii) waiting time standards, and (iii) all other attributes. For example, for competing mail services, the "other attributes" include the convenience of the pick-up process, the ease with which deliveries can be traced, and the likelihood of the packages being damaged. For internet service providers, customers consider the frequency of service interruption and the quality of the support staff along with the price and waiting time. Electronic brokerage services monitor and advertise execution price, price improvement, and effective spread as "other attributes" along with the commission and execution speed (see, for instance, www.fidelity.com). We treat price and waiting time as truly independent attributes: in general, a change in a firm's waiting time (distribution) *cannot* be compensated for by a price change that will leave all firms' demand volumes unchanged.

Because the waiting time standard is a *guarantee*, the *actual* expected waiting time experienced by the customers may sometimes be lower—but never higher—than the waiting time standard. The actual expected waiting time must match the standard exactly if customers can apprise themselves of the *actual* expected waiting time their class experiences, e.g., if it is monitored (perhaps by an independent organization) or if one assumes that an individual customer has unbounded rationality and is able to compute the expected actual waiting times that arise in equilibrium under optimal capacity levels and optimal dynamic priority schemes. Note that our representation of the demand rates as being dependent on stated (or advertised) prices and waiting time standards imposes a weaker assumption on individual customers' ability or willingness to process competitive information. At the same time, the waiting time standards are believable when customers can apprise themselves of the actual average waiting time either by the aforementioned independent monitoring or when they can develop their own estimates via repeated usage of the service.

We model each service provider as an $M/M/1$ queueing facility. Each customer class generates an independent Poisson stream of customers to this service provider at the rate determined by the aforementioned demand functions. Its service times are independent and identically distributed (i.i.d.) with a firm and class dependent service rate proportional to the firm's capacity level. Each firm incurs a given class-dependent cost per customer as well as a cost per unit of time proportional to the adopted capacity level. (Generalizations to settings where the capacity cost depends on the capacity level according to a general convex function are straightforward.) Each firm attempts to maximize its own expected profits.

We derive an analytical expression of the capacity level each firm needs to adopt to accommodate a given vector of demand volumes and waiting time standards under an optimal associated dynamic priority rule. We show that this capacity level is the maximum of a number of closed form capacity bounds, one for each subset of the customer classes. Interestingly, for arbitrarily specified waiting time standards, the maximum may be achieved for a strict subset of the collection of all classes, the so-called *bottleneck set*; in this case, strategic idleness times, i.e., artificial after-service delays, may be adopted for the so-called *residual* classes outside the bottleneck set. The capacity function displays economies of scope; i.e., it is always beneficial for a firm to pool service processes of different collections of customer classes. The capacity function is always jointly convexly decreasing in all of the segments' waiting time standards. The capacity function exhibits economies of scale for the customer classes with relatively large waiting time standards, i.e., those receiving relatively low service. At the same time, it exhibits *diseconomies of scale* for the customer classes with relatively small (i.e., demanding) waiting time standards. More specifically, when we express a customer class's waiting time standard as a multiple of the expected amount of work per customer—the so-classed *normalized* waiting time—the marginal capacity cost decreases (increases) with a segment's demand volume, if the segment receives worse (better) than average service, i.e., if the segment's normalized waiting time is above (below) the firm's *waiting time benchmark*, a weighted average of the normalized waiting times among all classes. Thus, unless all normalized waiting times are identical (and there is no need to segment the classes), the capacity cost function is always concave in some of the segments' demand volumes and convex in the others.

The optimal capacity level is to be complemented with a randomized absolute priority rule. While residual customer classes may arise under arbitrary exogenous expected waiting time standards, they do not when these waiting times are endogenously determined by the firms in any of the competition models, below in which these waiting time standards are (part of) the strategic choices.

In each of the three competition models, we establish that a pure Nash equilibrium exists under minor conditions regarding the demand volumes, and we characterize how the equilibrium varies as a function of the cost parameters and other exogenously specified parameters. (Although of theoretical interest, randomized Nash equilibria are far more difficult to implement and hence less likely to be adopted.) These existence results are in stark contrast to the known behavior in existing service competition models. For example, the models of Levhari and Luski (1978) and Li and Lee (1994) both consider two service providers and a single class of customers and assume all customers choose their provider strictly on the basis of the full price, i.e., the price plus a cost rate times the waiting time. The former paper assumes the full price is based on the *steady state* expected waiting time, with customer-specific i.i.d. cost rates, whereas Li and Lee (1994) assume that each arriving customer considers his expected waiting time based on the prevailing queue sizes at both firms (under a uniform cost rate). With service rates exogenously given, the competition between the two firms is, in both models, confined to their price choices only, and a pure equilibrium often fails to exist. See Chen and Wan (2003) for the complete analysis of Levhari and Luski (1978).

We compare the equilibria with those achieved when the firms service each market segment with a dedicated service process, i.e., without pooling service resources. In the price competition model, for example, the equilibrium is obtained, under both service pooling and dedicated service facilities, when for each class the relative markup vis-a-vis the marginal cost equals the reciprocal of the demand elasticity. This generalizes the well known Lerner index rule, derived for simple price competition models (see, e.g., Vives 2000). The marginal cost per customer per unit of time always consists of the variable service cost plus the marginal capacity cost (per unit of time). When service is provided with dedicated facilities, the marginal increase in the required capacity equals the expected amount of work per customer of the considered class. Under service pooling it is zero for residual classes and less (more) than this benchmark value depending on whether the customer class receives worse (better) than average service.

Our numerical studies show that firms are always better off under service pooling. Do the consumers benefit as well? More specifically, are the members of a given customer class charged less *throughout the industry* when the firms service the various customer classes in dedicated facilities as opposed to pooling the service processes? The answer is affirmative if the given customer class is in the bottleneck set and receives better than average service, at all firms; i.e., its normalized waiting time is, at all firms, *lower* than the above waiting time benchmark. In all other cases, if the customer class receives worse than average service or is in a residual class, its members *benefit* from service pooling. In other words, VIP customer classes in the bottleneck set, demanding better than average service under service pooling, are made to pay for the additional capacity cost their relatively demanding service standards impose on the firms beyond what they would pay in the absence of service pooling. All other customer classes benefit from service pooling. The same conclusions apply if only *part* of the industry pools the service processes, at least as far as the equilibrium prices of the service pooling firms are considered.

Similar conclusions prevail under waiting time competition. Under this type of competition, we show that *all* customer classes are in the bottleneck set. (Thus, the necessity to introduce strategic delays is, in our setting, confined to the case of price competition with *exogenously* specified waiting time standards.) Those receiving worse than average service at a given firm under service pooling can be consoled by the fact that their waiting time standard, although worse than the weighted average among all customer classes, is still better than what they would receive in the absence of service pooling. Conversely, if a customer class receives *better* than average service at a given firm under service pooling, its equilibrium waiting time standard would be even better if the firms employed dedicated facilities for the different customer classes. These results can be guaranteed when the normalized waiting time of a customer class is, percentage-wise, not too far from the firm's benchmark value; our numerical study shows that the results hold throughout. More strongly than the results under price competition, to guarantee a particular ranking of a customer class's waiting time at a *specific* firm, with and without service pooling, it suffices to know whether at this firm (class) the customer class enjoys better or worse than average service. Under simultaneous price and waiting competition, all customer classes are in the bottleneck set at all firms, as is the case under strict waiting time competition. Numerical examples show that the above comparisons between service pooling and service in dedicated facilities may fail to apply: even when a customer class receives better than average service at all firms, its equilibrium waiting time standards may be *smaller* under service pooling as compared to service with dedicated facilities. The reason is that under smaller simultaneous competition, such a customer class may may be charged considerably *more* under service pooling. Finally, one might conjecture that higher paying customer classes are always compensated by receiving better service, but this conclusion may fail to hold under both price and waiting time competition.

Section 2 provides a review of the relevant literature. Section 3 introduces the model and notation. The capacity choice and associated priority rules are discussed in §4. For the case of completely segmented markets, the equilibrium behavior in the competition models is characterized in §5. In §6, we outline how our results can be extended to the general model, in which customers can choose which class they want to join (or which firm to patronize). Section 7 provides additional insights obtained through numerical examples. Section 8 summarizes our major conclusions and outlines possible generalizations of the model.[1]

[1] Proofs of Theorems 4.1 and 5.1 are provided in the appendix, and the remaining proofs are provided in the online appendix, which can be found in the e-companion (http://mansci.journal.informs.org/).

2. Literature Review

In this section, we provide a brief review of the relevant literature on models with multiple customer classes.

Mendelson and Whang (1990) addressed the problem of how an $M/M/1$ service provider with a given capacity or service rate should select service charges and an optimal priority rule so as to maximize the expected social welfare, defined as the firm's revenues plus the consumer welfare minus the customers' waiting cost for multiple customer classes. The demand rate of each class is given by a decreasing function of the full price defined as the service charge plus a class-specific multiple of the expected waiting time. The optimal priority rule is a simple $c\mu$ rule, and the solution is shown to be incentive-compatible in settings where customers are able to misrepresent their class identity. (A solution is incentive-compatible if no individual customer has an incentive to feign a class identity different from his own.) Recently Afeche (2004), dealing with the case of *two* customer classes, has shown that the *unrestricted* optimal policy may fail to be incentive-compatible when the firm's revenues rather than social welfare are maximized. (This *unrestricted* policy continues to employ the above $c\mu$ rule.) Conversely, no absolute priority rule may be used as part of an optimal incentive compatible policy; in addition, such a policy may require the use of the aforementioned strategic idle times.

In the economics literature Gal-Or (1983), Champsaur and Rochet (1989), and Johnson and Myatt (2003) deal with price competition among oligopolists offering a menu of related products or services with different quality levels. As in our §6 model, these papers assume that the market cannot be segmented at all. However, in contrast to our model, they assume no interdependencies among the costs incurred for the different quality variants. We refer to Hassin and Haviv (2003) and Allon and Federgruen (2007) for a review of the literature on oligopolistic competition models in which the firms' demand rates depend on the customer expected steady state waiting times in system. The papers reviewed there, and here, all assume that customers aggregate the price and the waiting time standard into a single full price measure; most papers assume in addition that all customers select the service provider with the *lowest full price*, disregarding any other service attributes. Allon and Federgruen (2007) deal with the special case of our model in which all customers belong to a *single* customer class, with each firm offering a uniform price and waiting time standard to all.

To our knowledge, Loch (1991), Lederer and Li (1997), and Armony and Haviv (2001) are the only papers that have addressed competition models in which waiting-time-sensitive customers are segmented into multiple classes. When considering market segmentation, Loch (1991) considers an industry with two $M/M/1$ service providers and two customer classes, each with a given waiting cost rate and average service time. All customers within a class select the firm that offers the lowest full price, where the total demand volume in the class is given by a known function of this full price value. Under quantity competition, the author establishes the existence of a Nash equilibrium under which the customers are prioritized according to the $c\mu$ rule. Lederer and Li (1997) generalize this model to allow for an arbitrary number of nonidentical $M/G/1$ firms and an arbitrary number of customer classes. Assuming the firms engage in price competition, the authors establish the existence of a Nash equilibrium under which each firm, once again, prioritizes customers according to the $c\mu$ rule. The existence result is based on the assumption that each class's expected waiting time at a given firm is a convex function of all of the firm's demand rates for the different customer classes. Note also that whereas in the Afeche (2004) *monopoly* model, the incentive compatible optimal policy frequently cannot be based on the $c\mu$ priority rule, $c\mu$ priority rules *are* part of the Nash equilibrium in the Lederer and Li (1997) *perfect price competition* model (provided the above convexity assumption is satisfied).

In the above oligopoly models with multiple customer classes, prices or demand volumes are selected by the service providers. Lee and Cohen (1985) consider a model with exogenous prices, in which each of the customer classes decides, as a single entity, what fraction of its collective business to assign to each of the service providers. The total demand rate of each customer class is exogenously given, as are the service rates of the $M/M/1$ (or $M/M/c$) service providers who serve all customers on a FCFS basis, irrespective of their class identity. The authors establish the existence of a Nash equilibrium for the allocation decisions of the different customer classes. To relax the assumption of the customer classes' total demand rate being independent of service charges and waiting times, Armony and Haviv (2001) analyze a two-stage competition model with two $M/M/1$ service providers and two customer classes, each again acting as a single entity in deciding what fraction of its business to assign to each of the providers. In the first stage, the two providers compete with each other by announcing their service charges. In the second stage, the customer classes compete with their allocation decisions. A pure price equilibrium may fail to exist in this two-stage game.

3. Model and Notation

We consider a service industry with N competing service providers in a market which is segmented into J segments or customer classes. Let $E = \{1, \ldots, J\}$. Each firm i positions itself in the market by selecting a vector of prices for the different customer classes as well as an associated vector of expected steady state waiting times. More specifically,

$p_i^l =$ firm i's (service) charge for customers in class l, $i = 1, \ldots, N$; $l \in E$,

$w_i^l =$ firm i's expected steady state waiting time for customers in class l, $i = 1, \ldots, N$; $l = 1, \ldots, J$.

Let $p = \{p_i^l\colon\ i, l\}$, $w = \{w_i^l\colon\ i, l\}$, and for each $l \in E$, $p^l = (p_1^l, p_2^l, \ldots, p_N^l)$, and $w^l = (w_1^1, w_2^1, \ldots, w_N^l)$ denote the vectors of price and waiting time standards offered to class l. As illustrated in the introduction, in many service industries, the waiting time standards are explicitly advertised by the service providers themselves; in others, they are reported by independent organizations. The standard should be viewed as a (collective) guarantee allowing for the possibility that the actual expected waiting time is lower than the stated value. For each firm $i = 1, \ldots, N$ and customer class $l \in E$, the price p_i^l and waiting time standard w_i^l are chosen from given closed intervals $[p_i^{l,\max}, p_i^{l,\min}]$, $[w_i^{l,\max}, w_i^{l,\min}]$.

Each firm i faces a demand stream of customers of class l, generated by a Poisson process with rate λ_i^l. In the most general model, the rates $\{\lambda_i^l\}$ depend on all prices and waiting time standards offered by the various firms to all market segment, i.e., $\lambda_i^l = f_i^l(p, w)$, $i = 1, \ldots, N$ and $l = 1, \ldots, J$.

The amounts of work associated with customers of class l are i.i.d. with rate ν^l. $1/\nu^l$ is thus the average amount of work each class l customer brings. Each firm i selects a capacity level μ_i, where capacity is defined as the number of units of work that can be processed per unit of time. Thus, customers of class l who opt for service provider i experience service times that are exponentially distributed with rate $\mu_i \nu^l$. Each firm i selects its capacity level μ_i in conjunction with a priority rule so as to be able to service each customer class l with an expected steady state sojourn time no larger than w_i^l, given demand rates $\{\lambda_i^k\}_{k=1}^J$. γ_i denotes the per unit capacity cost rate of firm i. The only other cost component is a variable service cost c_i^l per customer of class l served by firm $i = 1, \ldots, N$.

As far as the priority rules are concerned, we consider the complete class Π of all rules with steady state waiting time distributions that are *non-anticipative*, i.e., under which priorities are assigned, with possible service preemption, on the basis of any part of the history of the process. Note that priorities cannot be assigned on the basis of the remaining service times of the customers in service because this information does not become available to firms until the actual service completions. At the same time, the priority rule may proscribe that a server be idle while customers are in the system or that customers' sojourn times are to be extended with post-service *strategic delays*, a term coined by Afeche (2004).

When discussing priority rules and their associated vectors of expected waiting time standards, we invoke the following properties of set functions $f: 2^E \to \mathbb{R}$. A set function $f(\cdot)$ is called *monotone* if $f(S) \le f(T)$, $\forall S \subseteq T$. It is called *submodular* [supermodular, modular] if $f(T \cup \{j\}) - f(T) \le [\ge, =]$ $f(S \cup \{j\}) - f(S)$, $\forall j \notin T \supseteq S$; i.e., the increment in the set function value due to the addition of a new element $\{j\}$ is smaller [bigger, identical], if this element is added to a larger set T as compared to a smaller set S (see, e.g., Nemhauser and Wolsey 1989 for equivalent definitions). A polyhedron in $\mathbb{R}^J$ is a *polymatroid* if it can be represented by the following set of constraints

$$\sum_{l\in S} X^l \le f(S), \quad \forall S \subseteq E \tag{1}$$
$$X \ge 0,$$

where the set function f is monotone and submodular with $f(\varnothing) = 0$. The *base* of this polymatroid is the polyhedron described by (1) with the constraint for $S = E$ specified as an *equality*.

4. Capacity Choice and Associated Priority Rules

Because a firm's capacity choice only affects its own cost and profits, it is clearly optimal for each firm to adopt the minimal capacity level that allows for a priority rule under which the waiting time standards $\{w_i^l: l \in E\}$ can be accommodated under the projected demand rates $\{\lambda_i^l: l \in E\}$. To characterize this minimum feasible capacity level μ_i for a given firm i, we first address the inverse question of which set of vectors of waiting time standards $\{W_i^l: l \in E\}$ is achievable under some priority rule in Π for a given capacity level μ_i^0.

Lemma 4.1. *Fix $i = 1, \ldots, N$. Assume firm i adopts a capacity level μ_i^0. The space of achievable vectors of waiting time standards $\{W_i^l: l \in E\}$ is a polyhedron $\mathcal{W}$, described by*

$$\sum_{l\in S} \rho_i^l W_i^l \ge b_i(S), \quad \forall S \subset E, \tag{2}$$

where $\rho_i^l = \lambda_i^l/(\mu_i^0 \nu^l)$, and

$$b_i(S) = \left(\sum_{l\in S} \frac{\lambda_i^l}{(\mu_i^0)^2(\nu^l)^2}\right)\frac{1}{1-\sum_{l\in S}(\lambda_i^l/(\mu_i^0\nu^l))}$$
$$= \frac{1}{\mu_i^0}\left(\sum_{l\in S}\frac{\lambda_i^l}{(\nu^l)^2}\right)\frac{1}{\mu_i^0 - \sum_{l\in S}(\lambda_i^l/\nu^l)}.$$

Lemma 4.1 immediately identifies what capacity level μ_i allows firm i to offer a given vector of waiting time standards $\{w_i^l, l \in E\}$ under a given vector of demand rates $\{\lambda_i^l: l \in E\}$: in (2), replace μ_i^0 by the variable μ_i, and the variables $\{W_i^l: l \in E\}$ by the specific vector w, to obtain that the latter is achievable, under some priority rule in Π, if and only if

$$\sum_{l\in S} \frac{\lambda_i^l}{\mu_i \nu^l} w_i^l \ge \frac{1}{\mu_i}\left(\sum_{l\in S}\frac{\lambda_i^l}{(\nu^l)^2}\right)\frac{1}{\mu_i - \sum_{l\in S}(\lambda_i^l/\nu^l)}, \quad \forall S \subset E.$$

Multiplying both sides of the inequality by $\mu_i(\mu_i - \sum_{l\in S}(\lambda_i^l/\nu^l))$, we obtain, after some algebra, that a capacity level μ_i is feasible if and only if

$$\mu_i \ge \sum_{l\in S}\frac{\lambda_i^l}{\nu^l} + \frac{\sum_{l\in S}(\lambda_i^l/(\nu^l)^2)}{\sum_{l\in S}(\lambda_i^l/\nu^l)w_i^l}, \quad \forall S \subset E. \tag{3}$$

Corollary 4.2. (a) *Fix $i = 1, \ldots, N$, given vectors of waiting time standards $\{w_i^l: l \in E\}$ and arrival rates $\{\lambda_i^l: l \in E\}$. The minimum feasible capacity level is given by*

$$\mu_i^* = \max_{S\subset E}\left\{\sum_{l\in S}\frac{\lambda_i^l}{\nu^l} + \frac{1}{W_i(S)}\right\}, \tag{4}$$

where $W_i(S) = \sum_{l\in S}(\lambda_i^l/(\nu^l)^2)(w_i^l\nu^l)/\sum_{l\in S}(\lambda_i^l/(\nu^l)^2)$.

(b) *There exists a largest set S_i^* that achieves the maximum in* (4). *We refer to this set as the bottleneck set* (*of customer classes*).

(c) *If $S_i^* = E$, the capacity choice μ_i^* can be optimally combined with a* (*possible randomization of*) *absolute priority rule*(*s*).

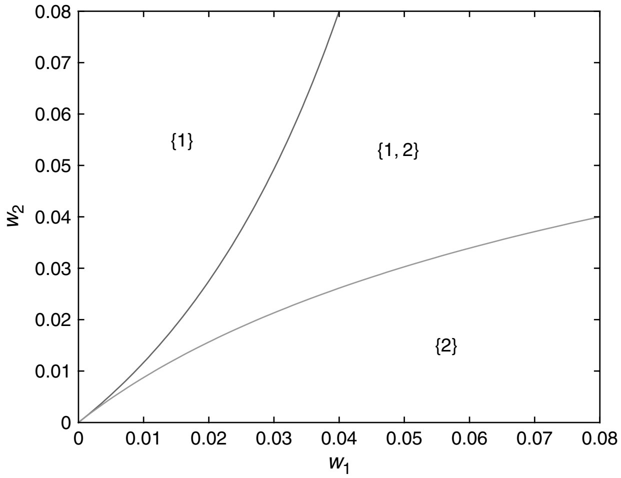
(d) *If $S_i^* \ne E$, the capacity choice μ_i^* can be optimally combined with one of the following two priority rules*:

r_1: *A* (*possible randomization of*) *at most $J+1$ absolute priority rule*(*s*) *combined with strategic delays $\{x^l: l \in E \setminus S^*\}$ for the classes in $E \setminus S_i^*$;*

r_2: *A* (*possible randomization of*) *at most $J+1$ absolute priority rule*(*s*) *under which the actual expected sojourn time for classes $l \in S^*$ is given by w_i^l and for classes $l \in E \setminus S^*$ by $w_i^l - x^l$.*

The maximand in (4) represents a *lower bound* for the capacity level required to meet the waiting time standards for the classes in the set S, under the projected demand rates. This lower bound consists of two terms: the first is $\sum_{l\in S}(\lambda_i^l)/\nu^l$, the total workload demanded by customer classes in the set S per unit of time, a base capacity required to ensure stability of the system irrespective of what waiting time standards are offered. The second term $1/W_i(S)$ represents a *safety margin* given by the reciprocal of a weighted average of the so-called *normalized waiting time standards*, $\{w_i^l\nu^l\}$, the waiting time experienced by a customer in class l expressed as a multiple of the expected amount of work demanded by the customer. The safety margin thus decreases with any of the waiting time standards. However, it may fail to be monotone in the set S, and the same may be true for the complete lower bound, even though its first term, the offered

Figure 1 Capacity Cost Function Regions for Two Classes



load, *does* increase as more customer classes are considered in the bound. Consider, for example, the case of two customer classes with fixed arrival rates λ_i^1 and λ_i^2, and $\nu^1 = \nu^2 = 1$. The capacity bound for the single class 1 dominates over that for the set E whenever the waiting time standard for class 2 is chosen to be in excess of a threshold value which increases with w_i^1, i.e., whenever $w_i^2 \geq w_i^1(\lambda_i^1 w_i^1 + 1)/(1-\lambda_i^2 w_i^1)$ if $w_i^1 \leq 1/\lambda_i^2$. By symmetry, the bound for class 2 dominates if w_i^1 is chosen to be in excess of a threshold value that increases with w_i^2 and has a horizontal asymptote at $w_i^2 = 1/\lambda_i^1$. Figure 1 thus exhibits that the positive quadrant of the (w_i^1, w_i^2) pairs can be partitioned into three regions, with one of the possible sets of classes representing the bottleneck in each. It is easily verified that the two switching curves intersect only in the origin. We conclude that the bottleneck set S_i^* may be a strict subset of E. In this case, it appears in general to be preferable for all parties concerned to employ rule r_2 as opposed to rule r_1. However, if the customers can apprise themselves of the actual average sojourn times, either because they are monitored and reported by independent firms (see the Introduction for examples) or because they are able to compute them by themselves, customers become aware of the fact that their *actual* expected sojourn time is lower than the guaranteed value w_i^l. This will result in increased demand for the relevant customer classes and hence increased congestion in the service facility, necessitating an increase in the capacity level. In this case, the firm may need to opt for rule r_1. This rule is easily implemented without any adverse effects if the customer is physically separated from the actual service process, e.g., when service is provided via the internet or in remote facilities. However, when able to observe progress in the actual service process, customers may resent their strategic delays. Strategic or intentional delays were first introduced by Afeche (2004) and have been used as an essential component of priority schemes by Maglaras and Zeevi (2003) and Yahalom et al. (2005). Because these papers address industries with asymmetric information, i.e., the service provider is unable to observe the class identity of its customers, the essential use of strategic delays appears to be the consequence of this asymmetry. We show that strategic delays are a required mechanism when selecting capacity levels and priority schemes, even in a setting with *symmetric* information, assuming rules of type r_2 are either infeasible or not desired. In the price competition model, strategic delays may be a part of the equilibrium strategy of the firm, as shown below. Note that in the Afeche (2004) model, rules of type r_2 are not an option because firms are assumed to announce their complete scheduling policies, and customers are capable of computing the resulting expected sojourn times for all customer classes.

The following proposition identifies a number of structural properties of the capacity function: We say that at a given firm i, class l receives better (worse) than average service if and only if its normalized waiting time $(w_i^l \nu^l) \leq (\geq) W_i(S_i^*)$, the weighted average of these normalized waiting times.

Proposition 4.1. *Fix $i = 1, \ldots, N$.*

(a) *Let E^1, E^2 denote two disjoint sets of customer classes with given demand rates and waiting time standards, $\{(\lambda_i^l, w_i^l)\colon l \in E^1\}$ and $\{(\lambda_i^l, w_i^l)\colon l \in E^2\}$. Let μ_i^{*c} denote the capacity in a single facility that provides* combined *service to E^1 and E^2, and μ_i^{*1} (μ_i^{*2}) the capacity of a facility that provides service to E^1 (E^2) only. Then $\mu_i^{*c} \leq \mu_i^{*1} + \mu_i^{*2}$; i.e., the capacity function always exhibits economies of scope.*

(b) $\mu_i^* \leq \sum_{l=1}^{J}(\lambda_i^l/\nu^l + 1/(\nu^l w_i^l))$.

(c) *μ_i^* is decreasing and jointly convex in $\{w_i^l\colon l \in E\}$.*

(d) *μ_i^* is increasing in the demand rates $\{\lambda_i^l\colon l \in E\}$. If class l is residual at firm i, the marginal capacity requirement is $\partial\mu_i^*/\partial\lambda_i^l = 0$. If class l is in the bottleneck set S_i^*, the marginal capacity requirement $\partial\mu_i^*/\partial\lambda_i^l$ exists (assuming S_i^* is the unique maximand in (7)), and*

$$\frac{\partial \mu_i^*}{\partial \lambda_i^l} = \frac{1}{\nu^l}\left\{1+\frac{1}{\nu^l}\frac{\sum_{m\in S_i^*}(\lambda_i^m/(\nu^m)^2)w_i^m\nu^m - w_i^l\nu^l\sum_{m\in S_i^*}(\lambda_i^m/(\nu^m)^2)}{(\sum_{m\in S_i^*}(\lambda_i^m w_i^m/\nu^m))^2}\right\} \qquad (5)$$

Thus, the marginal capacity requirement for a bottleneck class is larger (smaller) than the expected amount of work a marginal customer in the class adds if and only if the class receives better (worse) than average service.

(e) *Fix $\{\lambda_i^r, r \neq l\}$ at a given firm i and a given customer class l. Assume the same bottleneck set S_i^* applies for all demand volumes λ_i^l. Thus, the optimal capacity level*

$$\mu_i^* \text{ is } \begin{cases} \text{independent of } \lambda_i^l & \text{if class } l \notin S_i^*,\\ \text{concave in } \lambda_i^l & \text{if class } l \in S_i^* \text{ and receives worse than average service at firm } i,\\ \text{convex in } \lambda_i^l & \text{if class } l \in S_i^* \text{ and receives better than average service at firm } i. \end{cases}$$

The condition in part (d) is satisfied everywhere except for a set of measure zero, in which several capacity bounds for different subsets of customer classes are exactly equal and maximal among all capacity bounds for *all* possible sets of E; see (10). The condition in part (e) is satisfied wherever the waiting time standards are endogenously determined as part of a competitive model (for example, the waiting time and simultaneous competition models, as we will show, in these cases $S_i^* = E$ throughout). We conclude from part (d) that under service pooling, the marginal capacity cost at a given firm for a given bottleneck customer class is lower (higher) than its value with dedicated service facilities if the class receives worse (better) than average service at this firm. Part (e) shows that under service pooling, the required capacity level exhibits decreasing (increasing) marginal cost to scale with respect to the demand volume of a bottleneck set if and only if this class receives worse (better) than average service at the firm. When service is provided with dedicated facilities, the required capacity level is always *affine* in any of the demand volumes. Because by definition some classes receive better than average and others worse than average service, the capacity function always fails to be convex in all of the demand volumes separately, let alone to be jointly convex; the only exception is the trivial case where all classes receive the same service (i.e., have the same normalized waiting time), in which case no differentiation between customer classes is required.

5. Competition Model: The Case of Completely Segmented Markets

In this section, we analyze the competition models under the assumption that the market is completely segmented; i.e., each customer is unambiguously assigned to a specific customer class. See the introduction for a discussion of this assumption. The demand rates for a given class are therefore entirely independent of the prices and waiting time standards offered to other customer classes and the interdependence between the customer classes stems from the structure of the joint capacity cost described above. More specifically, we consider the following demand functions:

$$\lambda_i^l(p^l, w^l) = a_i^l(w_i^l) - \sum_{j\neq i} \alpha_{ij}^l(w_j^l) - b_i^l p_i^l + \sum_{j\neq i} \beta_{ij}^l p_j^l; \quad i = 1, \ldots, N. \quad (6)$$

Here a_i^l is a decreasing concave function reflecting the fact that a waiting time reduction by a firm results in an increase in its demand volume, however, with nonincreasing marginal returns to scale. The functions α_{ij}^l are general decreasing functions, representing the fact that firm i's demand volume can only increase in response to an increase in the waiting time standard of any of its competitors.

Several relationships may be assumed regarding the magnitude of b_i^l compared with other parameters in (6). First, prices may be scaled in units such that

$$(S) \quad b_i^l > \max_{w_i^{l,\min} \le w_i^l \le w_i^{l,\max}} \left| \frac{da_i^l(w_i^l)}{dw_i^l} \right| = \left| \frac{da_i^l(w_i^{l,\max})}{dw_i^l} \right| \quad i = 1, \ldots, N,\ l \in E.$$

Also, 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 and a price increase by a given firm cannot result in an increase of the industry's aggregate demand; i.e.,

$$(\text{D}) \quad b_i^l > \sum_{j\neq i} \beta_{ij}^l, \quad i = 1, \ldots, N,\ l \in E;$$

$$(\text{D}') \quad b_i^l > \sum_{j\neq i} \beta_{ji}^l, \quad i = 1, \ldots, N,\ l \in E.$$

The demand function (6), may, e.g., be derived from a representative consumer model with utility function $U^l(\lambda^l, w^l) \equiv C + \frac{1}{2}\lambda^{lT}(B^l)^{-1}\lambda^l - \lambda^{lT}(B^l)^{-1}\bar{a}(w)$, where the $N \times N$ matrix B^l has $B_{ii}^l = -b_i^l$ and $B_{ij}^l = \beta_{ij}^l$, $i \neq j$, $\bar{a}^l(w) \equiv a_i^l(w_i^l) - \sum_{j\neq i} \alpha_{ij}^l(w_j^l)$, and $C > 0$. ((D) ensures that $(B^l)^{-1}$ exists and is negative semi-definite, giving rise to a jointly concave utility function.) The demand functions (6) arise by optimizing the utility function subject to a budget constraint.

The expected profit for firm i is, by Corollary 4.2, given by

$$\begin{aligned}\pi_i(p, w) &= \sum_{l\in E} (p_i^l - c_i^l)\lambda_i^l(p^l, w^l) - \gamma_i(\mu_i(\lambda, w)) \\ &= \sum_{l\in E} (p_i^l - c_i^l)\lambda_i^l(p^l, w^l) \\ &\quad - \gamma_i\left(\max_{S\subset E}\left\{\sum_{l\in S} \frac{\lambda_i^l(p^l, w^l)}{\nu^l} + \frac{\sum_{l\in S}(\lambda_i^l(p^l, w^l)/(\nu^l)^2)}{\sum_{l\in S}(\lambda_i^l(p^l, w^l)/\nu^l)w_i^l}\right\}\right). \quad (7)\end{aligned}$$

Even though the firms make selections for four types of strategic decisions, i.e., prices, waiting time standards, the capacity level, and the priority rule, the closed form expected profit function in (7) allows us to represent each firm's profit as a function of the price vector p and waiting time standards vector w only. Let $\Delta_i \triangleq \max_{l\in E}(w_i^l \nu^l) - \min_{l\in E}(w_i^l \nu^l)$, $i = 1, \ldots, N$, the *span* of the vector of normalized waiting time standards, denote the *degree of service differentiation* for firm i. Note that the measure is dimensionless; it is, in particular, invariant with respected to the chosen time unit. Finally, to allow for comparisons with systems without service pooling, we assume that the minimum prices are set to ensure a positive variable profit margin under dedicated service; i.e.,

$$p_i^{l,\min} > c_i^l + \frac{\gamma}{\nu^l}. \tag{8}$$

5.1. Price Competition

In the price competition (PC) model, all waiting time standards are exogenously given. Firms compete by choosing a price list for the different customer classes along with a capacity level and associated priority rule. This type of competition arises when waiting time standards are either chosen in a way different from the way they are chosen in noncooperative competition, or they are selected with lower frequency than the prices. The PC model differs fundamentally from earlier price competition models addressing segmented markets, which assume that the firms' cost can be represented as a separable (linear or convex) function of the demand rates. We have argued that, without loss of generality (w.l.o.g.), $p_i^{l,\min} > c_i^l + \gamma_i/\nu^l$; see (8). The derivations of our results for the price competition model are, however, simplified when expanding the feasible region by specifying $p_i^{l,\min} = c_i^l$.

Theorem 5.1. *There exist $B_i > 0$ such that if for all i, l the demand volumes $\lambda_i^l \geq B_i\sqrt{\Delta_i}$ on the entire feasible price region, the following results hold:*[2]

(a) *A price equilibrium p^* exists and any such equilibrium is in the interior of the price region.*

(b) *For any price equilibrium p^* and corresponding demand vector $\lambda(p^* \mid w)$, assume that firm i's optimal capacity level μ_i^* is achieved for a unique set S_i^* in (4), $i = 1, \ldots, N$. Then p^* and $\lambda(p^* \mid w)$ satisfy the system of equations $\lambda_i^l = b_i^l(p_i^l - c_i^l - \gamma_i(\partial\mu_i/\partial\lambda_i^l))$.*

(c) *Any price equilibrium p^* is component-wise increasing in each of the cost parameters $\{c_i^l; \gamma_i\}$.*

Thus, a price equilibrium exists, provided the demand volumes are not too small. The theorem states specific lower bounds as sufficient conditions derived from (highly generous) bounding arguments. The lower bounds are proportional to the square roots of the degrees of service differentiation $\{\Delta_i\}$. The closer the normalized waiting time standards for the different customer classes are to each other, the smaller the lower bounds are. Also, the bounds decrease to zero in the case of a single class or when the normalized waiting time standards are identical for all customer classes. Theorem 5.1 thus provides a full generalization for the equilibrium existence result in Allon and Federgruen (2007). The condition in part (b) is satisfied almost everywhere on the feasible price region $\times_{i,l}[p_i^{l,\min}, p_i^{l,\max}]$. Whereas equilibrium prices are monotone in each of the cost parameters, no such monotonicity can be expected with respect to the waiting time standards (even for sufficiently large demand volumes). Allon and Federgruen (2007) established this, even for the case where all customers belong to a *single* segment, identifying a sufficient condition with respect to the derivatives of the functions $\{a_i^l\}$ and $\{\alpha_{ij}^l\}$, under which prices decrease with waiting time standards.

The equilibrium conditions are thus structurally identical to those under dedicated service. In the latter case, the marginal capacity requirement $\partial\mu_i^*/\partial\lambda_i^l = 1/\nu^l$ for *all* customer classes. As shown in Proposition 4.1(b), under pooled service the marginal capacity requirement is zero for residual class, and for bottleneck classes it is either lower or higher than the benchmark value $(\nu^l)^{-1}$ depending on whether the class receives worse or better than average service at firm i. The equilibrium conditions state that at each firm and for each customer class the variable profit margin equals the reciprocal of the demand elasticity. This generalizes the so-called Lerner index condition, derived for basic price competition models with linear costs.

These observations give rise to the following proposition, which compares the price equilibrium achieved under pooled service with that arising when each firm serves every class with a dedicated service facility.

Proposition 5.1. *Let p^D denote the price equilibrium that arises when each of the firms serves every class with a dedicated service facility. Let S_i^* be the bottleneck set of customer classes for firm i under a price equilibrium p^* for the model with pooled service. Fix $l \in E$.*

(a) *Assume class $l \in S_i^*$, $\forall i = 1, \ldots, N$, and receives* better than average *service: $\nu^l w_i^l \leq W_i(S_i^*)$; i.e., its normalized waiting time is less than or equal to the weighted average of normalized waiting times in S_i^*. Then $p_i^{Dl} \leq p_i^{*l}$, $\forall i = 1, \ldots, N$.*

(b) *Assume that for all $i = 1, \ldots, N$, either $l \notin S_i^*$ or $l \in S_i^*$ and receives* worse than average *service: $\nu^l w_i^l \geq W_i(S_i^*)$; i.e., its normalized waiting time is greater than or equal to the weighted average of normalized waiting times in S_i^*. Then $p_i^{Dl} \geq p_i^{*l}$, $\forall i = 1, \ldots, N$.*

(c) *Assume only one of the firms, w.l.o.g. firm 1, pools service for the J customer classes, and all other firms serve their customers in dedicated facilities. Let $\hat{p}$ denote a price equilibrium and $\hat{S}$ an associated bottleneck set for firm 1. If $l \in \hat{S}$ and $\nu^l w_1^l \leq W_1(\hat{S})$, then $p_i^{Dl} \leq \hat{p}_i$, $\forall i = 1, \ldots, N$. If $l \notin \hat{S}$ or $l \in \hat{S}$, but $\nu^l w_1^l \geq W_1(\hat{S})$, then $p_i^{Dl} \geq \hat{p}_i$, $\forall i = 1, \ldots, N$.*

Proposition 4.1 shows that all firms reduce their cost structure by switching from dedicated to pooled service. Proposition 5.1 shows, however, that these cost savings do not necessarily result in price reductions for all customer classes. Indeed, if a customer class gets better than average service (and belongs to

[2] The following conditions are easily verified to guarantee that the condition in Theorem 5.1 is satisfied: $a_i^l(w_i^l) - \sum_{j\neq i}\alpha_{ij}^l(w_j^l) - b_i^l p_i^{l,\min} + \sum_{j\neq i}\beta_{ij}^l p_j^{l,\max} \geq B_i$.

the bottleneck set) at all firms, it is charged a higher price under service pooling than under dedicated service. The following provides some intuition behind this result: assume all classes initially get identical normalized waiting time standards; if class 1, say, subsequently bargains for a lower waiting time standard, the cost for the other customer classes increases, for which externalities class 1 is made to pay.

Example. Let $N = J = 3$ and $w_i^{l,\min} = 10^{-3}$, $\overline{w} \equiv w_i^{l,\max} = 4\cdot 10^{-3}$, $p_i^{l,\min} = 70$, $p_i^{l,\max} = 105$. Let $a_i^l(w_i^l) = a_i^0 + \sigma_w^l a_i \log(\overline{w} - w_i^l)$ and $\alpha_{ij}^l(w_j^l) = \sigma_w^l \alpha_{ij} \log(\overline{w} - w_j^l)$, while $b_i^l = 10\sigma_p^l$, $\beta_{ij}^l = 4.5\sigma_p^l$. Thus, all classes share the same intercepts a_i^0 in the demand functions. Also, all functions $a_i^l(w_i^l)$ and α_{ij}^l are proportional to the common function $\log(\overline{w} - w_i^l)$ and $\log(\overline{w} - w_j^l)$ respectively, with proportionality factors that are identical across classes up to a class specific factor σ_w^l. The same applies to the price sensitivity coefficients $[b_i^l; \beta_{ij}^l]$, which, in addition, are identical across firms. We consider the parameter values: $\sigma_w^1 = 2$; $\sigma_w^2 = 1.5$; $\sigma_w^3 = 1$; $\sigma_p^1 = 1$; $\sigma_p^2 = \sigma_p^3 = 1$. $(a_1^0, a_2^0, a_3^0) = (435, 435, 705)$; $(a_1, a_2, a_3) = (100, 100, 100)$; $\alpha_{12} = \alpha_{21} = \alpha_{31} = \alpha_{32} = 40$ while $\alpha_{13} = \alpha_{23} = 50$. As to the cost parameters $(\gamma_1, \gamma_2, \gamma_3) = (35, 35, 50)$, $c_1^1 = c_2^1 = 40$ and $c_3^1 = 25$ while $c_1^2 = c_2^2 = c_1^3 = c_2^3 = 20$ and $c_3^3 = c_3^2 = 5$. Finally, $\nu^1 = 4$, $\nu^2 = 2$, $\nu^3 = 1$. Thus, the classes are ranked in decreasing order of their prices and waiting time sensitivities and in increasing order of their expected service times. The instance may reflect an industry with an established domestic firm and two entrant oversees competitors. The domestic firm 3 enjoys a larger brand recognition, as reflected by larger intercepts of the demand functions, and operates with a higher capacity cost rate but lower per customer variable service cost. The two oversees competitors have identical characteristics. Finally, variable service costs are incurred for class 1 customers.

Table 1 exhibits the price equilibrium under both dedicated and pooled service when all firms are offered an identical waiting time standard of $3\cdot 10^{-3}$ time units. Under pooled service, classes 1 and 2 experience, at all firms, higher than average normalized waiting times, which equal $4\cdot 10^{-3}$, $4\cdot 10^{-3}$, $4.6\cdot 10^{-3}$ for firms 1, 2, 3. Class 3 experiences a lower than average normalized waiting time at all firms. It is a VIP class in spite of its *absolute* waiting time standard being identical to those offered to the other classes. Consistent with Proposition 5.1, classes 1 and 2 benefit under pooled service, but class 3 does not.

Table 1 Price Competition Under Pooled and Dedicated Service

	Firms 1 and 2			Firm 3		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Pooled	71	65	80	72	69	71
Dedicated	97.37	82.9	79.70	102.75	89.69	70.58

5.2. Waiting Time Competition

In some settings, *prices* are chosen exogenously, in a manner different from how they are chosen in noncooperative competition. Alternatively, prices may exhibit significantly larger stickiness than service levels. See Allon and Federgruen (2007) for a detailed discussion. In the waiting time (WT) competition model, we thus assume that prices $\{p_i^l, i, l\}$ are exogenously given and firms compete by selecting waiting time standards.

In the following theorem, we establish the existence of an equilibrium in the WT competition model, assuming that the minimum acceptable waiting time standards $\{w_i^{l,\min}\}$ are not chosen to be excessively small. In particular, we assume

$$w_i^{l,\min} \geq \sqrt[3]{\frac{\gamma_i}{4(p_i^l - c_i^l - \gamma_i/\nu^l)a_i^{l,(2)}\underline{\nu}_i}}, \tag{9}$$

where $a_i^{l,(2)} \equiv \min_{w_i^{l,\min} \leq w_i^l \leq w_i^{l,\max}} |d^2 a_i^l(w_i^l)/d(w_i^l)^2|$. (Note the minimum acceptable waiting times decrease to zero as the exogenously given prices increase.)

Theorem 5.2. *Assume* (9) *holds for a given vector of prices* $\{p_i^l\}$. *There exist lower bounds* $B_i \geq 0$ *such that if demand rates* $\lambda_i^l \geq B_i$ *throughout the feasible waiting time region, a Nash equilibrium exists.*

As in the case of the PC model, a simple condition may be established to ensure that any Nash equilibrium of the WT model is an interior point of the feasible region and therefore satisfies the following system of first-order conditions (see the proof of Theorem 5.2):

$$w_i^l = a_i^{l'-1}\left[\frac{-\gamma_i \lambda_i^l / \sum_{m\in S} \lambda_i^m/(\nu_m)^2}{(p_i - c_i^l - \gamma_i(\partial\mu_i^*/\partial\lambda_i^l))\nu^l}\right], \tag{10}$$

where $a_i^{l'-1}(\cdot)$ denotes the inverse of the decreasing function $a_i^{l'}(\cdot)$. These equilibrium conditions generate the following insights, assuming all classes have the same marginal waiting time sensitivity functions $a_i'(\cdot)$: if two customer classes offer identical demand volumes, the lowest waiting time is offered to the class for which the profit margins per unit of work per customer, i.e., $[p_i - c_i^l - \gamma_i(\partial\mu_i^*/\partial\lambda_i^l)]\nu^l$ are highest. At the same time, if two classes show the same profit margins per unit of work per customer, the class generating the higher volume of customers is associated with a lower equilibrium waiting time standard. In general, the equilibrium waiting times standards are ranked in the same order as the ratios of the demand volumes and the profit margins per customer per unit of work $\lambda_i^l/(p_i - c_i^l - \gamma_i(\partial\mu_i^*/\partial\lambda_i^l))\nu^l$. Conversely, if this ratio is identical for a given pair of classes $\{k, l\}$, but class k has a point-wise larger waiting time sensitivity, i.e., $|a_i^{k'}(\cdot)| \geq |a_i^{l'}(\cdot)|$, then class k receives a lower waiting time standard than class l. Finally, if firm i's capacity cost rate γ_i goes up, the firm compensates by increasing the waiting time standards for *all* classes, as opposed to only some.

In contrast to the PC model, in equilibrium all customer classes belong to the bottleneck set, and as a consequence, no strategic delays need to be imposed on any of the classes.

Proposition 5.2. *Let w^* denote an interior point equilibrium in the WT competition model. Then, $S_i^* = E$, $\forall\, i = 1, \ldots, N$; i.e., all customer classes are part of each firm's bottleneck set and the vector of waiting time standards w^* can be achieved without imposing strategic delays on any of the customer classes.*

We conclude this subsection, again, with a comparison between the equilibrium under pooled versus dedicated service. In the PC model, Proposition 5.1 showed that a customer class with better (worse) than average service experiences a lower (higher) equilibrium price under dedicated versus pooled service. The following proposition shows that for a customer class with better (worse) than average service under pooling, a move to dedicated service is, again, beneficial (detrimental) but only if its normalized waiting time is not too far below (above) the weighted average.

Proposition 5.3. *Let w^D denote the waiting time equilibrium that arises when each of the firms serves every class with a dedicated service facility, and assume it is an interior point of the feasible waiting time space. Let w^* denote an interior point equilibrium under pooled service. Let $\tilde{\lambda} = \sum_{m\in E}(\lambda_i^m/(\nu^m)^2)$.*

(a) *Assume class $l \in E$ receives moderately better than average service under service pooling* at a given firm i, *i.e.*, $\sqrt{\nu^l}\sqrt{\tilde{\lambda}/(\lambda_i^l/(\nu^l)^2)} \leq w_i^{*l}\nu^l/W^*(E) \leq 1$, *then* $w_i^{Dl} \leq w_i^{*l}$.

(b) *Assume class $l \in E$ receives moderately worse than average service under service pooling* at a given firm i, *i.e.*, $\sqrt{\nu^l}\sqrt{\tilde{\lambda}/(\lambda_i^l/(\nu^l)^2)} \geq w_i^{*l}\nu^l/W^*(E) \geq 1$, *then* $w_i^{Dl} \geq w_i^{*l}$.

(c) *If firm i serves its customers with dedicated facilities, its equilibrium waiting time standards are independent of any of the competitors' characteristics. In particular, a firm with dedicated service is unaffected by the choice of any of its competitors whether to adopt pooled or dedicated service.*

No specific ranking of class l's equilibrium waiting times under dedicated versus pooled service can be guaranteed when class l receives extremely better [worse] than average service, i.e., $w_i^{*l}\nu^l/W^*(E) \leq [\geq]$ $\min\{\sqrt{\nu^l}\sqrt{\tilde{\lambda}/(\lambda_i^l/(\nu^l)^2)}, 1\}$ $[\max\{\sqrt{\nu^l}\sqrt{\tilde{\lambda}/(\lambda_i^l/(\nu^l)^2)}, 1\}]$. In this respect, the ranking result is more limited than its counterpart in Proposition 5.1; at the same time, to guarantee a specific ranking for a given class at a given firm, it suffices to compare this class' normalized waiting time with the average value at *this* firm only. We expect that the results of parts (a) and (b) continue to apply under more general demand functions and queueing models for the firms' facilities. In contrast, the independence of each firm's equilibrium waiting time standards with respect to any of the competitors' characteristics is a consequence of three specific assumptions: (i) the demand function is separable from the firms' waiting time standards; (ii) each firm services the different customer classes in a dedicated facility; and (iii) the safety margin of a firm's capacity level is a function of its own waiting time standard only.

5.3. Simultaneous Competition

When firms simultaneously compete in terms of their prices and waiting time standards, the existence of a Nash equilibrium can be guaranteed under conditions very similar to those required in the waiting time competition model. It suffices to replace (9) by

$$w_i^{l,\min} \geq \sqrt[3]{\frac{\gamma_i}{4[(p_i^l - c_i^l - \gamma_i/\nu^l)a_i^{(2),l} + da_i(w_i^{l,\max})/dw_i^l]\underline{\nu}_i}}. \tag{11}$$

Once again, the larger the minimum markups $(p_i^l - c_i^l - \gamma_i/\nu^l)$, the lower the minimum waiting time standard that may be chosen, ensuring that a Nash equilibrium exists.

Theorem 5.3. (a) *Assume* (11). *There exist lower bounds $B_i \geq 0$ such that if demand rates $\lambda_i^l \geq B_i$ throughout the feasible price-waiting time standard region, a Nash equilibrium exists.*

(b) *If the equilibrium is an interior point, the bottleneck sets $S_i^* = E$, $\forall\, i = 1, \ldots, N$.*

6. Competition Models for Unsegmented Models

In this section, we discuss generalizations of the models in §5, to allow for settings where the market fails to be presegmented; i.e., individual customers have the option to select a service class along with the firm they wish to patronize. The following is the natural extension of the demand model (6):

$$\lambda_i^l(p, w) = a_i^l(w_i^l) - \sum_{j\neq i}\alpha_{ij}^l(w_j^l) - \sum_{k\neq l}\sum_{m=1}^{N}\kappa_{im}^{lk}(w_m^k) - b_i^l p_i^l + \sum_{j\neq l}\beta_j^l p_j^l + \sum_{k\neq l}\sum_{m=1}^{N}\varphi_{im}^{lk}p_m^k;$$
$$i = 1, \ldots, N,\ l = 1, \ldots, J, \tag{12}$$

when the functions $\kappa_{im}^{lk}(\cdot)$ are again general decreasing functions and the parameters $\varphi_{im}^{lk} \geq 0$, to reflect the fact that any increase of the price or waiting time standard for a customer class $k \neq l$ at firm i or any of its competitors can only result in an increase of the expected demand volume for service class l at firm i. The demand model (6) clearly arises as a special case of (12). Analogous to (D) and (D′), we assume again, without loss of practical generality, that a uniform price increase by all firms and for all types of customers (for a given firm and customer class) cannot result in an increase of the demand volume for any given customer class and any given firm (the total demand volume).

$$\text{(D)}\quad b_i^l > \sum_{j\neq i} \beta_{ij}^l + \sum_{k\neq l}\sum_{m=1}^{N} \varphi_{im}^{lk}, \quad i=1,\ldots,N,\ l=1\ldots,J;$$

$$\text{(D}'\text{)}\quad b_i^l > \sum_{j\neq i} \beta_{ij}^l + \sum_{k\neq l}\sum_{m=1}^{N} \varphi_{mi}^{lk}$$

Theorem 6.1. *There exists minimal demand threshold $\underline{\lambda}_i^l$ such that if for all i, l the demand volumes $\lambda_i^l \geq \underline{\lambda}_i^l$ on the entire feasible price region, the following results hold:*

(a) *A price equilibrium p^* exists and any such equilibrium satisfies the first-order conditions*

$$0 = \lambda_i^l - b_i^l\left(p_i^l - c_i^l - \gamma_i \frac{\partial \mu_i^*}{\partial \lambda_i^l}\right) + \sum_{m\neq l} \varphi_{ii}^{ml}\left(p_i^m - c_i^m - \gamma_i \frac{\partial \mu_i^*}{\lambda_i^m}\right). \tag{13}$$

(*b*) *Any price equilibrium p^* is component-wise increasing in each of the cost parameters $\{c_i^l, \gamma_i\}$.*

The equilibrium no longer specifies that the percentage profit margin should equal the reciprocal of the demand elasticity, the generalization of the Lerner index rule discussed in §5. Similarly, the condition under which a given customer class is charged more or less under pooled service as compared to service with dedicated facilities is no longer as simple as the condition in Proposition 5.1.

7. Examples

In this section, we illustrate our results and identify some important qualitative observations regarding the equilibria in the three competition models. These observations complement our theoretical results and stem from extensive numerical experiments. For the sake of brevity, we report here on the results of one instance obtained from the example by modifying the following parameters: $\sigma_p^1 = 1.5$, $\nu^1 = \nu^2 = \nu^3 = 1$. Table 2 displays the equilibrium under pooled and dedicated service for the simultaneous competition model. Unlike firm 3, firms 1 and 2 select, under simultaneous competition, different service levels for the three customer classes. The total variable cost per customer is identical at all firms and all classes. The greater brand recognition of firm 3 (intercepts in the demand functions) permits it to charge classes 1 and 2 a higher price while providing inferior service. Nevertheless, to increase its market share and revenues, it offers class 3 a lower price along with superior service. Table 3 (4) displays the price (waiting time) equilibria for the price (waiting time) competition model.

First, one might conjecture that, under price competition, the ranking of the equilibrium prices across different classes is the reverse of the ranking of the basic or the normalized waiting time standards. Conversely, one might expect that if class l is charged a higher price than class l', it is rewarded with a lower waiting time in the waiting time competition model. The results for firm 3 in Tables 3 and 4 disprove both conjectures. For example, in the first (second) instance of Table 3 (4) class 3 is charged the lowest price while receiving the highest service. Proposition 5.1 shows that a class with better (worse) than average service by *all* providers is better (worse) off at *all* firms under dedicated as opposed to pooled service. This leaves open the question whether providing worse (better) than average service to a specific customer class at a *specific* firm ensures that this class has a lower (higher) equilibrium price at this specific firm under pooled versus dedicated service. The results for class 1, at firms 1 and 2, in both instances of Table 3 disprove this localized version of Proposition 5.1.

Next, Proposition 5.3 provides conditions under which a given class at a given firm benefits or suffers from service pooling under waiting time competition. These conditions fail to exhaust the spectrum of possibilities, but our numerical experiments have shown that, invariably, *all classes* benefit from pooling. The same applies to all *firms*, under all three

Table 2 Simultaneous Competition Under Pooled and Dedicated Service

	Firm 1			Firm 3		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Prices—Pooled	87.5	87.5	77	91	91	73.5
Waiting times—Pooled	$25\cdot 10^{-4}$	$19\cdot 10^{-4}$	$28\cdot 10^{-4}$	$25\cdot 10^{-4}$	$25\cdot 10^{-4}$	$25\cdot 10^{-4}$
Prices—Dedicated	87.5	83.3	80.15	89.6	88.8	70
Waiting times—Dedicated	$32\cdot 10^{-4}$	$29\cdot 10^{-4}$	$33\cdot 10^{-4}$	$33\cdot 10^{-4}$	$30\cdot 10^{-4}$	$34\cdot 10^{-4}$

Table 3 Price Competition Under Pooled and Dedicated Service

	Firm 1			Firm 3		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Waiting times (exogenous)	$25 \cdot 10^{-4}$	$30 \cdot 10^{-4}$	$35 \cdot 10^{-4}$	$35 \cdot 10^{-4}$	$30 \cdot 10^{-4}$	$25 \cdot 10^{-4}$
Prices—Pooled	90	83	76	84	90	77
Prices—Dedicated	91.21	82.9	77.12	86.1	89.7	75.8
Waiting times (exogenous)	$22 \cdot 10^{-4}$	$30 \cdot 10^{-4}$	$37 \cdot 10^{-4}$	$37 \cdot 10^{-4}$	$30 \cdot 10^{-4}$	$22 \cdot 10^{-4}$
Prices—Pooled	95	82	73	79	89	81
Prices—Dedicated	93.7	82.9	74.79	81.7	89.7	78.8

Table 4 Waiting Time Competition Under Pooled and Dedicated Service

	Firm 1			Firm 3		
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
Prices (exogenous)	85	85	85	85	85	85
Waiting times—Pooled	$26 \cdot 10^{-4}$	$22 \cdot 10^{-4}$	$25 \cdot 10^{-4}$	$29 \cdot 10^{-4}$	$27 \cdot 10^{-4}$	$17 \cdot 10^{-4}$
Waiting times—Dedicated	$34 \cdot 10^{-4}$	$29 \cdot 10^{-4}$	$31 \cdot 10^{-4}$	$35 \cdot 10^{-4}$	$31 \cdot 10^{-4}$	$31 \cdot 10^{-4}$
Prices (exogenous)	85	90	90	90	90	85
Waiting times—Pooled	$26 \cdot 10^{-4}$	$20 \cdot 10^{-4}$	$24 \cdot 10^{-4}$	$26 \cdot 10^{-4}$	$26 \cdot 10^{-4}$	$20 \cdot 10^{-4}$
Waiting times—Dedicated	$34 \cdot 10^{-4}$	$28 \cdot 10^{-4}$	$31 \cdot 10^{-4}$	$33 \cdot 10^{-4}$	$30 \cdot 10^{-4}$	$31 \cdot 10^{-4}$

types of competition. Although Proposition 4.1 shows that a firm's profit function under service pooling is point-wise larger than under dedicated service, this by itself does not guarantee that the same ranking applies to the equilibrium profits. Invariably, all three classes belong to the bottleneck set. (Recall that Proposition 5.2 and Theorem 5.3(b) show that this must hold for any interior point equilibrium in the waiting time and simultaneous competition models.) Also, invariably, a mixture of absolute priority rules is required to meet the offered waiting time standards. For example, in the first instance of Table 4, firm 1 [firm 3] needs to mix the absolute priority rules $(3-1-2)$, $(2-3-1)$, $(1-2-3)$; $[(3-2-1), (2-1-3), (1-3-2)]$ with close to equal probabilities. (Under absolute priority rule (A-B-C), class A receives absolute priority over class B, and B over C.)

8. Conclusions and Extensions

We have developed a general model for the competitive interactions between providers in a service industry that cater to multiple customer segments with the help of shared service facilities. Under mild regularity conditions, we have established that a Nash equilibrium exists in each of the three competition models considered, i.e., the price competition (PC), the waiting time (WT) competition, and simultaneous competition (SC) models. The existence conditions merely preclude that demand volumes or minimum waiting time standards are excessively low. We systematically compare the equilibria with those arising under *dedicated* service: all firms always benefit from service pooling, usually with major profit increases. In the PC model, a class always pays a lower (higher) price under dedicated service if, under pooled service, it receives a better (lower) than average normalized waiting time at all firms. In the WT model, for a class to be better (worse) off under dedicated versus pooled service at a given firm, it suffices that, under pooled service, it receives better (worse) than average service at *this* firm only.

We have also investigated various comparative statics results for the equilibria. For example, we have proved that, under price competition, each firm's equilibrium prices are monotone in each of its cost parameters as well as those pertaining to its competitors. However, equilibrium prices (waiting time standards) may under the PC model (WT model) fail to be be monotone with respect to the exogenous waiting times (prices). Moreover, equilibrium prices may fail to be ranked in accordance with the waiting time standards the classes receive, and this is true in each of the competitive models.

To achieve the above results, we have characterized how a firm's capacity level and associated priority rule depend on the demand volumes it faces and the waiting time standards it offers to the various customer classes. The capacity level, for example, can be expressed as a closed form function of the vector of demand volumes and waiting time standards. The capacity function, of importance in its own right, is monotone and jointly convex in the waiting time standards, exhibiting economies of scope but not necessarily of scale.

An important assumption in our model is that customers are completely segmented and that their class identity is given. In some settings, customers may be able to choose a class identity. To model this variant, the demand rate for a given customer class at a given firm would need to be specified as a function of *all* prices and *all* waiting time standards offered to *all* classes (and by all firms) rather than just the class under consideration. This generalization imposes no additional difficulties on the characterization of the required capacity level or priority schemes. The above analysis methods can continue to be employed to establish the existence of a Nash equilibrium in the various competition models and to study its qualitative properties. Only the existence conditions for the Nash equilibria become more complex.

Future work will extend the above results to settings where customers are primarily sensitive to the *delay* they experience rather than to the full sojourn time, those where service is best characterized as a fractile of the waiting time distribution rather than its expected value, and those where the service facilities need to be described by more general queueing models. For example, in the former case, it is possible to derive a capacity cost function analogous to (8), i.e., where the capacity is the maximum of $(2^J - 1)$ closed form functions of the demand volumes and the expected delays, one for each subset of the classes of customers. (This characterization requires a restriction to *nonpreemptive* priority rules but allows for a *general* service time distribution scaled down in proportion to the invested capacity.) The structure of the closed form capacity bounds μ_i^S (such that $\mu_i = \max_{S\subseteq E} \mu_i^S$) is more complex than that of the maximand in (8).

9. Electronic Companion

An electronic companion to this paper is available as part of the online version that can be found at http://mansci.journal.informs.org/.

Appendix. Proofs

Proof of Theorem 4.1. Consider an arbitrary priority rule $r \in \Pi$ and let $W_i^l(r)$ denote the expected steady state sojourn time for class l customers under rule r at firm i. We first verify that the vector $\{W_i^l(r)\}$ satisfies the constraints (2). Thus, select an arbitrary subset $S \subset E$. Note that $\sum_{l\in S}(\lambda_i^l/(\mu_i^0 \nu^l))W_i^l(r)$ denotes the aggregate expected steady state amount of work for customers belonging to one of the classes in S under rule r. The right-hand side of (2) denotes the aggregate expected amount of work for classes in S, under *any* rule which is nonidling and gives preemptive priority to customers belonging to S over all others (see, e.g., Federgruen and Groenevelt 1988). It therefore also denotes the expected steady state amount of work in the *single* class $M/G/1$ system that arises when all classes $l \in S$ are merged into a single class, no other customer classes are admitted, and the server operates with no idling. Similarly, $\sum_{l\in S}(\lambda_i^l)/(\mu_i^0 \nu^l)W_i^l(r)$ denotes the expected amount of work in the same single class $M/G/1$ system under a rule that forces the server to idle while customers are waiting whenever, in the original system, rule r assigns the server to a customer *not* belonging to one of the classes in S or prescribes him to be idle. Because in the single class $M/G/1$ system the amount of work is minimized by any nonidling rule, the vector $\{W_i^l(r)\colon l \in E\}$ satisfies (2) for this set S.

Conversely, consider an arbitrary vector $w \triangleq \{w_i^l : l \in E\}$ in the polyhedron described by (2). We show that a rule $r \in \Pi$ exists such that $W_i^l(r) = w_i^l$, for all $l = 1, \ldots, J$. Let $\overline{\mathcal{W}}_i \subset \mathcal{W}$ denote the base polyhedron described by (2), however, with the constraint for $S = E$ specified as an *equality*. If $w \in \overline{\mathcal{W}}_i$, it is well known from Coffman and Mitrani (1980) and Federgruen and Groenevelt (1988) that w is the vector of expected sojourn times under a simple absolute priority rule or a randomization of such rules. If $w \notin \overline{\mathcal{W}}_i$, there exists a vector $x \triangleq \{x^1, \ldots, x^J\} \geq 0$ such that $w' \triangleq w - x \in \overline{\mathcal{W}}_i$. To verify this, note that $w - x \in \overline{\mathcal{W}}_i$ iff

$$\sum_{l\in S} \rho_i^l (w^l - x^l) \geq b_i(S), \quad S \subsetneq E;$$

$$\sum_{l=1}^{J} \rho_i^l (w^l - x^l) = b_i(E). \tag{14}$$

Let $X^l \triangleq \rho_i^l x^l$, $l \in E$. Thus, $x \geq 0$ and $w - x \in \overline{\mathcal{W}}_i$, iff

$$\sum_{l\in S} X^l \leq \hat{b}_i(S), \quad S \subsetneq E; \qquad \sum_{l=1}^{J} X^l = \hat{b}_i(E); \quad X \geq 0, \tag{15}$$

where $\hat{b}_i(S) \triangleq \sum_{l\in S} \rho_i^l w_i - b(S)$, $S \subset E$. Theorem 2 in Federgruen and Groenevelt (1988) shows that the set function $b(\cdot)$ is supermodular, so that the set function $\hat{b}_i(\cdot)$, as the difference between a modular function and a supermodular function, is submodular. Moreover, $\hat{b}_i(S) \geq 0$ for all $S \subset E$ because $w \in \mathcal{W}_i$. The set function $\hat{b}_i(\cdot)$ may fail to be monotone; i.e., $\hat{b}_i(S) > \hat{b}_i(T)$ may arise for some pair of sets $S \subset T$. At the same time, it is easily verified that the polyhedron described by (15) remains unaltered when replacing the right-hand side $\hat{b}_i(S)$ by $\bar{b}_i(S) \triangleq \min_{T\supset S} \hat{b}_i(T)$, $S \subset E$:

$$\sum_{l\in S} X^l \leq \bar{b}_i(S), \quad S \subsetneq E; \qquad \sum_{l=1}^{J} X^l = \bar{b}_i(E); \quad X \geq 0. \tag{16}$$

The set function $\bar{b}_i(\cdot)$ is clearly monotone and non-negative because $\hat{b}_i(\cdot) \geq 0$; it is also submodular (see, for instance, Theorem 135 in Edmonds 2003). This implies that the polyhedron described by (16) is the base of a polymatroid which is always nonempty. For example, the vector $(\bar{b}_1(\{1\}), \ldots, \bar{b}_1(\{1,\ldots,l\}) - \bar{b}_1(\{1,\ldots,l-1\}), \ldots, \bar{b}_1(\{1,\ldots,J\}) - \bar{b}_1(\{1,\ldots,J-1\}))$ satisfies (16). This shows the existence of a vector $x \geq 0$ such that $w' = w - x \in \overline{\mathcal{W}}_i$ for which we have pointed out that a (possible randomization of) absolute priority rule(s) $r \in \Pi$ exists such that $W(r) = w'$. Let $\tilde{r}$ denote the rule obtained from r by extending the sojourn time of any customer in class l by a post-service (strategic) delay x^l. Clearly, $w = w' + x = W(\tilde{r})$. $\square$

Proof of Theorem 5.1. (a) The profit function π_i can be written as $\pi_i(p) = \min_{S\subset E} \pi_i^S(p)$, where

$$\pi_i^S(p) = \sum_{l\in E}(p_i^l - c_i^l)\lambda_i^l(p^l) - \gamma_i\left(\sum_{l\in S}\frac{\lambda_i^l(p^l)}{\nu^l} + \frac{\sum_{l\in S}(\lambda_i^l(p^l)/(\nu^l)^2)}{\sum_{l\in S}(\lambda_i^l(p^l)/\nu^l)w_i^l}\right).$$

In view of the Nash-Debreu Theorem, to show the existence of an equilibrium p^*, it suffices to verify that each of the functions $\{\pi_i^S: S \subset E\}$ is jointly concave in $(p_i^1, \ldots, p_i^J)$ because in that case π_i, as the minimum of $2^J - 1$ jointly concave function, is jointly concave itself. Let $\bar{w}_i \triangleq \max_{m\in E} w_i^m$, $\underline{\nu}_i \triangleq \min_{m\in E} \nu_i^m$, $\bar{b}_i = \max_{m\in S} b_i^l$, $\underline{b}_i = \min_{m\in S} b_i^l$, and $(\underline{w\nu})_i = \min_m w_i^m \nu^m$. Also let

$$B_i = \sqrt{\gamma_i \bar{b}_i} \cdot \sqrt{J\bar{b}_i/\underline{b}_i} \max\left\{1/(\underline{w\nu})_i, \sqrt{(1/2\underline{\nu}_i^3)[1/\underline{\nu} + 2\bar{w}_i/(\underline{w\nu})_i]}\right\}.$$

Note that

$$\frac{\partial \pi_i^S}{\partial p_i^l} = \lambda_i^l - b_i^l(p_i^l - c_i^l) + \mathbb{1}_{\{l\in S\}} \gamma_i b_i^l \frac{\partial \mu_i^S}{\partial \lambda_i^l}. \tag{17}$$

Thus, for $l \notin S$, $\partial^2 \pi_i^* / \partial (p_i^l)^2 = -2b_i^l$, and $\partial^2 \pi_i^* / (\partial p_i^l \partial p_i^k) = 0$, for $k \neq l$. Let $\tilde{\lambda}_s^W = \sum_{m\in S} (\lambda_i^m w_i^m / \nu^m)$.

For all $l \in S$, $\partial^2 \pi_i^* / \partial (p_i^l)^2 = -2b_i^l + \delta_{ii}^{ll}(p)$, where by (7), $\delta_{ii}^{ll} = \gamma_i (b_i^l)^2 (\partial^2 \mu_i^S / \partial (\lambda_i^l)^2) = 2\gamma_i w_i^l (b_i^l)^2 / ((\nu^l)^3 (\tilde{\lambda}_S^w)^3)(\sum_{m\in S} (\lambda_i^m / (\nu^m)^2) \cdot (\nu^m w_i^m - \nu^l w_i^l)) \leq \Delta_i / (\underline{w\nu})_i (2\gamma_i w_i^l (b_i^l)^2) / ((\nu^l)^3 (\tilde{\lambda}_S^w)^2) \leq \epsilon_i$, if

$$\begin{aligned}
\lambda_i^m &\geq \sqrt{\frac{2\gamma_i (\bar{b}_i)^2 \Delta_i}{\epsilon_i (\underline{w\nu}_i)^2}} \geq \sqrt{\frac{2\gamma_i (\bar{b}_i)^2 (w_l^l/\nu^l)^2 \Delta_i}{\epsilon_i (\underline{w\nu}_i)^2 (\sum_{m\in S} (w_i^m/\nu^m))^2}} \\
&\geq \sqrt{\frac{2\gamma_i (\bar{b}_i)^2 (w_l^l/\nu^l)^2 \Delta_i}{\epsilon_i (\underline{w\nu}_i) w_i^l \nu^l (\sum_{m\in S} (w_i^m/\nu^m))^2}} \\
&\geq \sqrt{\frac{2\gamma_i (\bar{b}_i)^2 w_i^l \Delta_i}{\epsilon_i (\underline{w\nu}_i) (\nu^l)^3 (\sum_{m\in S} (w_i^m/\nu^m))^2}},
\end{aligned} \tag{18}$$

where the inequality follows from the bound $(\tilde{\lambda} / \sum_m (\lambda_i^m/(\nu^m)^2) \nu^m w_i^m \geq 1/(\underline{w\nu})_i)$, where $\tilde{\lambda} = \sum_m (\lambda_i^m/(\nu^m)^2)$, because the left-hand side of this inequality is the reciprocal of a weighted average of the normalized waiting time standards.

Similarly, for $k, l \in S$

$$\frac{\partial^2 \pi_i^S}{\partial p_i^l \partial p_i^k} = -\frac{\gamma_i b_i^l b_i^k}{\nu^k \nu^l} \frac{1}{(\tilde{\lambda}_S^w)^3} \left[\left(\frac{w_i^k}{\nu^l} - \frac{w_i^l}{\nu^k} \right) \tilde{\lambda}_S^w - 2w_i^k \left(\sum_{m\in S} \frac{\lambda_i^m}{\nu^m} \left(\frac{w_i^m}{\nu^l} - \frac{w_i^l}{\nu^m} \right) \right) \right]. \tag{19}$$

Then

$$\begin{aligned}
\left| \frac{\partial^2 \pi_i^S}{\partial p_i^l \partial p_i^k} \right| &\leq \frac{\gamma_i b_i^l b_i^k |w_i^k \nu^k - w_i^l \nu^l|}{(\nu^l \nu^k)^2 (\tilde{\lambda}_S^w)^2} + \frac{2\gamma_i b_i^l b_i^k w_i^k \sum_{m\in S} (\lambda_i^m/(\nu^m)^2) |w_i^m \nu^m - w_i^l \nu^l|}{(\nu^l)^2 \nu^k (\tilde{\lambda}_S^w)^3} \\
&\leq \frac{\Delta_i \gamma_i b_i^l b_i^k}{(\nu^l)^2 \nu^k (\tilde{\lambda}_S^w)^2} \left[\frac{1}{\nu^k} + \frac{2w_i^k}{(\underline{w\nu})_i} \right] \leq \epsilon_i, \quad \text{if}
\end{aligned}$$

$$\begin{aligned}
\lambda_i^m &\geq \sqrt{\frac{\Delta_i (\bar{b}_i)^2 \gamma_i}{(\sum_{m\in S} (w_i^m/\nu^m))^2 \underline{\nu}^3 \epsilon_i} \left[\frac{1}{\underline{\nu}} + \frac{2\bar{w}_i}{(\underline{w\nu})_i} \right]} \\
&\geq \sqrt{\frac{\Delta_i b_i^l b_i^k \gamma_i}{(\nu^l)^2 (\sum_{m\in S} (w_i^m/\nu^m))^2 \nu^k \epsilon_i} \left[\frac{1}{\underline{\nu}} + \frac{2\bar{w}_i}{(\underline{w\nu})_i} \right]}.
\end{aligned} \tag{20}$$

We conclude that for the chosen coefficients B_i, (18) and (20) hold for $\epsilon_i = \underline{2b}_i/J$. In this case, the Hessian of π_i^S with respect to the vector $(p_i^1, \ldots, p_i^J)$ has negative diagonal elements and is *diagonally dominant*; i.e., the absolute value of each diagonal element is larger than the sum of the absolute values of the off-diagonal elements in its row. This implies that the Hessian is negative-semidefinite, so that π_i^S is jointly concave in $(p_i^1, \ldots, p_i^J)$.

It remains to establish that any equilibrium p^* must be in the interior of the feasible price range. Given the choice of $p^{\max}$, it suffices to show that $p_i^{*l} > c_i^l = p_i^{l,\min}$ for all i and l. Assume to the contrary that for some pair (i, l) $p_i^{*,l} = c_i^l$. The profit function π_i is only piece-wise smooth and may fail to be differentiable in p^*. We show that for *any* subgradient $g = (g^1, \ldots, g^J)$ of π_i in the point p^*, $g^l > 0$, thus contradicting, by Proposition 5.1.2 in Mäkelä and Neittaanmäki (1992), the fact that p_i^{*l} is an optimal price for firm i, when all competitors charge according to the vector p^*. Because π_i is piece-wise smooth, any of its subgradients is a convex combination of the $2^J - 1$ gradients of π_i^S, $S \subset E$; see, e.g., Lemarechal and Mifflin (1978). It thus suffices to show that $\partial \pi_i^S / \partial p_i^l > 0$ for any $S \subset E$. By (17), if $l \notin S$, then $\partial \pi_i^S / \partial p_i^l = \lambda_i^l > 0$. If $l \in S$, by (17)

$$\begin{aligned}
\frac{\partial \pi_i^S}{\partial p_i^l} &> \lambda_i^l + \frac{\gamma_i b_i^l}{\nu^l} \left(\frac{\sum_{m\in S} (\lambda_i^m w_i^m / (\nu^m \nu^l)) - w_i^l \tilde{\lambda}_s}{(\tilde{\lambda}_S^w)^2} \right) \\
&\geq \lambda_i^l - \frac{\gamma_i b_i^l}{(\nu^l)^2} \left(\frac{\sum_{m\in S} (\lambda_i^m / (\nu^m)^2) |w_i^m \nu^m - w_i^l \nu^l|}{(\tilde{\lambda}_S^w)(\sum_{m\in S} (\lambda_i^m \nu^m w_i^m / (\nu^m)^2))} \right) \\
&\geq \lambda_i^l - \frac{\gamma_i b_i^l \Delta_i}{(\nu^l)^2 (\underline{w\nu})_i (w_i^l/\nu^l) \lambda_i^l} \\
&\geq \lambda_i^l - \left(\frac{\gamma_i \bar{b}_i \Delta_i}{(\underline{w\nu})_i^2} \right) \frac{1}{\lambda_i^l} \geq 0,
\end{aligned}$$

where $\tilde{\lambda}_s = \sum_{m\in S} \lambda_i^m / (\nu^m)^2$ and where the third inequality follows from the reciprocal of a weighted average of normalized waiting times being smaller than the reciprocal of the minimum value and $\sum_{m\in S} \lambda_i^m w_i^m / \nu^m \geq \lambda_i^l w_i^l / \nu^l$. The last inequality holds because $\lambda_i^l \geq (\sqrt{\gamma_i \bar{b}_i} / (\underline{w\nu})_i) \sqrt{\Delta_i}$, by the definition of B_i.

(b) Because μ_i is achieved for a *single* set S_i^* of customer classes, $\pi_i(\cdot)$ is differentiable in p^*; and because p^* is an interior point of the feasible price region, $\partial \pi_i(p^*) / \partial p_i^l = \partial \pi_i^{S_i^*}(p^*) / \partial p_i^l = 0$ for all i, l. Thus p^* satisfies (17).

(c) Let H denote the $NJ \times NJ$ matrix $(\partial^2 \pi_i^{S_i^*} / (\partial p_i^l \partial p_i^k))$ and G denote the matrix $G = \operatorname{diag}(b_1^1, \ldots, b_1^J; b_2^1, \ldots, b_2^J; b_N^1, \ldots, b_N^J)$. Applying the Implicit Function theorem to (17), we obtain, for λ sufficiently large: $(\partial p_i^{*l} / \partial c_j^k) = (-H)^{-1} G \geq 0$ because $(-H = -H^0 + o(\lambda))$, where the row corresponding with (i, l) in $(-H^0)$ has $2b_i^l$ as its diagonal element, $-\beta_{ij}^l$ in the column corresponding with (j, l), and zeros elsewhere. Thus, $(-H^0)^{-1} > 0$; see, e.g., Bernstein and Federgruen (2002) and $(-H)^{-1} = (-H^0)^{-1} + o(\lambda)$. Similarly, $(\partial p_i^{*l} / \partial \gamma_j) = (-H)^{-1} \Gamma$, where the $NJ \times N$ matrix $\Gamma = \Gamma^0 + o(\lambda)$ and $\Gamma^0 \geq 0$. □

References

Afeche, P. 2004. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delaying tactics. Working paper, Kellogg School of Management, Northwestern University, Evanston, IL.

Allon, G., A. Federgruen. 2007. Competition in service industries. *Oper. Res.* **55**(1) 37–55.

Armony, M., M. Haviv. 2001. Price and delay competition between two service providers. Technical report, Stern School, New York University, New York.

Bernstein, F., A. Federgruen. 2002. Comparative statics, strategic complements and substitutes in oligopolies. *J. Math. Econom.* **40**(6) 713–746.

Champsaur, P., J. C. Rochet. 1989. Multiproduct oligopolists. *Econometrica* **57**(3) 533–557.

Chen, H., Y.-W. Wan. 2003. Price competition of make-to order firms. *IIE Trans.* **35**(9) 817–832.

Coffman, Jr., E. G., I. Mitrani. 1980. A characterization of waiting time performance by single-server queues. *Oper. Res.* **28**(3) 810–821.

Edmonds, J. 2003. *Submodular Functions, Matroids, and Certain Polyhedra*. Springer-Verlag, Berlin.

Federgruen, A., H. Groenevelt. 1988. Characterization and optimization of achievable performance in general queueing systems. *Oper. Res.* **36**(5) 733–741.

Gal-Or, E. 1983. Quality and quantity competition. *Bell J. Econom.* **14** 590–600.

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

Johnson, J., S. Myatt. 2003. Multiproduct quality competition: Fighting brands and product line pruning. *Amer. Econom. Rev.* **93** 748–774.

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

Lee, H. L., M. A. Cohen. 1985. Multi-agent customer allocation in a stochastic service system. *Management Sci.* **31**(6) 752–763.

Lemarechal, C., R. Mifflin. 1978. *Nonsmooth Optimization*. Pergamon Press, Oxford, UK.

Levhari, D., I. Luski. 1978. Duoply 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 delays. Unpublished Ph.D. dissertation, Stanford University, Stanford, CA.

Maglaras, C., A. Zeevi. 2003. Pricing and performance analysis for a system with differentiated services and customer choice. *Proc. 42th Allerton Conf. Comm., Control Comput., Allerton, IL.*

Mäkelä, M. M., P. Neittaanmäki. 1992. *Nonsmooth Optimization*. World Scientific, Hackensack, NJ.

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

Nemhauser, G. L., L. A. Wolsey. 1989. *Integer Programming*. Elsevier, North-Holland, New York.

Vives, X. 2000. *Oligopoly Pricing*: *Old Ideas and New Tools*. MIT Press, Cambridge, MA.

Yahalom, T., J. M. Harrison, S. Kumar. 2005. Designing and pricing incentive compatible grades of service in queueing systems. Working paper, Stanford University, Stanford, CA.