

# Competition in Large-Scale Service Systems: Do Waiting Time Standards Matter?

---

Customers often select service providers by trading three categories of service attributes: prices, service levels as measured by the response time to a service request, and all other intangible attributes. The purpose of this paper is to study the equilibrium behavior in a market with both small- and large-scale service providers that compete on both the price and service-level attributes. We find that, while price-differentiation plays a key role in determining the competition outcomes for both large and small firms, service-level differentiation is of greater importance for the small firms. The large firms will provide high service levels in equilibrium, but might choose to set these according to an industry standard rather than actively differentiate themselves from the competition on this attribute. With additional assumptions on the demand model we are able to identify the Quality-and-Efficiency-Driven regime, which was known to be optimal in many monopolistic settings, as the equilibrium outcome in competition. We are also able to identify in relative precision the equilibrium prices and relate these to the equilibrium in a pure price-competition. To establish these results we introduce a novel framework that combines  $\epsilon$ -Nash equilibria and heavy-traffic queueing analysis and apply these to a sequence of replicated markets with growing aggregate demand.

---

## 1. Introduction

While analogies can be made between manufacturing of goods and service delivery, there is one aspect of services which has no parallel in the manufacturing process – the customer’s role and experience. Indeed, while goods are very tangible entities, “Service is a time perishable, intangible experience performed for a customer acting in a role of a co-producer”; see Zeithaml et al. (2005).

As the customer plays an active role in the service-delivery process, his subjective experience from this process becomes an important attribute of the “product” offered by the provider, and his satisfaction is crucial for the firms’ success. Upset customers can replace the service provider with a competitor while satisfied customers are amenable for up-selling and cross-selling attempts and are a good source for extra revenues. In response to the customers’ sensitivity to multiple attributes of the service “product”, service providers try to find the right tradeoff between prices and service levels that will maximize their profits in equilibrium.

In many service industries, the most important aspect of the service experience is the response time or the delay experienced by customers waiting to be served. There is an extensive literature (which we survey in §2) that studies competition between service providers. While these models

predict that firms would differentiate themselves in terms of both their prices and their service levels, we can observe that in certain markets the large firms do not openly compete on the latter dimension of their service and don't use the service levels as a strategic lever. That is, in many industries, the firms do not differentiate themselves along the service-level dimension and choose to set it to an industry standard. One of the main questions this paper tries to answer is whether there is an economically rational explanation to that behavior.

In the call center industry, for example, it is quite prevalent to use an industry benchmark to determine the speed of answer instead of using the customer's sensitivity to this dimension. In an article from CRMToday (Spera (2003)), a call center expert states: "Most service-level targets are set without truly listening to the customer. Instead, they are often set by relying on myths of 'best practice' or 'industry standard.'" In this example, then, the firms do not compete on the response time dimension but rather set the same industry benchmark as a service-level target. The model developed in this paper tries to provide one possible answer for setting industry standards even in competitive environments, and show that such seemingly irrational behavior can be justified economically.

In the online brokerage industry, on the other hand, smaller firms, such as Ameritrade and E-Trade promise their customers a service-level guarantee on the speed with which their transactions are processed. While this trend started quite a few years ago, none of the larger firms have made an attempt to compete on this dimension by explicitly advertising their response times. It is plausible that these larger firms do use their relative economies of scale to provide high service levels but for some reason choose not to advertise them. We observe, then, that in some markets in which both large and small providers operate, small firms do compete actively on the response time dimension while large providers choose often not to. While there may be several explanations for this phenomenon, we will try to provide a reasoning that relates the strategic side to the operational one through the analysis of equilibria in a market in which both small- and large-scale service providers operate.

Two fundamental questions are central to the study of equilibria: (a) *existence*: do Nash equilibria exist in the examined markets, and (b) *characterization*: given some sort of existence, is it possible to characterize the set of equilibria in order to obtain qualitative insights into the market outcomes. In this paper we address both questions with regard to markets with competing service-providers. As a first step towards this goal, we identify an immanent gap between real market behavior and the clear limitations of the Nash-equilibria notion. Specifically, whereas casual observation of actual service markets suggests that they are relatively stable, it is known that Nash

equilibria need not exist even under the most plausible demand functions, such as Multinomial Logit, and the simplest supply systems, such as the  $M/M/1$  queue (see e.g. Cachon and Harker (2002)). Even in cases in which Nash equilibria are guaranteed to exist their analysis often fails to provide qualitative insights into the market behavior. Indeed, while previous work was able to characterize the equilibrium behavior in limited subset of these markets (i.e. under  $M/M/1$  and restrictive demand models), the characterization was usually fairly complex and did not lend itself to simple qualitative insights that could explain some of the motivating examples mentioned above.

To address these problems we introduce an analysis framework that stands on three pillars: (i)  $\epsilon$ -Nash (or approximate) equilibria (ii) heavy-traffic queueing analysis, and (iii) market replication. The introduction of approximate equilibrium is aimed, initially, to overcome the non-existence of Nash equilibria. Its eventual benefits, however, go significantly beyond this initial objective when combined with market replication and heavy-traffic analysis. We examine the behavior of equilibria, not on a single market, but rather on a sequence of markets with increasing aggregate demand—these are referred to as replicated markets. Using heavy-traffic analysis we characterize the way in which the equilibrium capacity-decisions (and in turn the service-level decisions) scale along the sequence of markets. The economies-of-scale characterization allows us to identify potential  $\epsilon$ -Nash equilibria points. With this we achieve two objectives: first, we bridge the gap between the market stability and the non-existence of Nash equilibria by showing that the market exhibits approximate equilibrium. Second, we provide a simple characterization of the equilibrium prices and service-levels (see more below). Thus, while heavy-traffic is traditionally applied to sequences of queueing systems operating in isolation, its application to the study of a sequence of markets with growing demand is far-reaching. It allows us to gain insights into the way in which firms make their strategic price and service-level decisions.

The insights we obtain correspond to two types of markets: *homogenous markets* and *mixed markets*. These two markets differ by the scale of the firms that operate in the market. We measure the scale of a firm by the demand-volume that it faces through one point of contact. Consequently we treat each outlet of McDonald's as a point-of-contact, and thus as a firm. A call center has one point of contact even if its agents are dispersed as long as all calls are directed to agents through a central call-distributor. We then say that a firm is large if the demand volume it faces (and consequently, the capacity it assigns to serve it) is large. The above will be precisely defined in §3.

We then say that a market is a **homogenous market** if all firms competing in the market are of a similar scale. We show that if the overall demand volume in the market is large, the competition will lead to price levels that are very close to the those arising under pure price-competition game

in which the service levels are fixed at the best possible service level. This implies, in turn, that the actual equilibrium service levels will be high. Beyond that, however, they will have only little impact on the pricing decision. In particular, we show that firms might, in equilibrium, choose to set their service levels in accordance with an “industry standard” rather than actively competing on it. We establish the above for a very large family of demand models and service provision systems. We regard the above result as a **first-order** result in the sense that it identifies the relative importance of price and service-level and characterizes the equilibrium prices as being “close” to the prices in a pure price-competition game and the equilibrium service-levels as being “close” to the maximal service-level possible.

While the first-order analysis gives guidelines on the importance of the different instruments, firms should be interested in **fine-tuning** the above in order to achieve better market outcomes. We thus proceed to the **second-order** analysis that aims to refine the understanding of how close the equilibrium prices and service levels shall be to the benchmark of pure-price competition and high service level. Specifically, we aim to better understand the optimal competitive choices of price and capacity for the firms competing in the market. This analysis yields two important results: (a) In terms of **service-level**, we show that under certain demand models the competition will lead the firms to operate in a regime in which extremely high service levels and high efficiency co-exist; this is known as the *Quality and Efficiency Driven (QED)* regime. While this operational regime was known to be optimal for a monopoly, we are the first to show that this regime arises in a competitive setting where the firms compete on prices and service levels and provide economic justification to its existence; and (b) in terms of **prices** we identify exactly how “close” the equilibrium prices would be to the pure price-competition prices. These **fine-tuning** results are driven by the ability to translate bounds on the profits to bounds on the action space using the specific structure of the demand models. The game-theoretic results on approximate Nash are always given in terms on the bound on the payoffs obtained by the firms. To our knowledge, there are no general results that identify that maximum that the firms can deviate from their actions without compromising the approximate equilibrium. Our result that the bounds on the equilibrium service-levels become increasingly tighter are negatively proportional to the square-root of the demand volume, is thus unique, and is obtained through the developed framework employing the concepts of replicated markets in conjunction with heavy-traffic queueing theory.

We next extend our market replication framework in order to analyze **mixed markets** that constitute of both large and small firms. In this extended framework, some firms will grow—these will represent the large firms—and some firms will be kept at a constant size but will be duplicated—

these will represent the small firms. We show that while the large firms will use high service levels they will not use them as a competitive dimension (i.e., firms will not differentiate themselves along this dimension) and will compete only on prices. The small firms, in contrast, will compete among themselves on the service-level dimension. Using the operational regime terminology - one may say that, while large firms operate optimally in the QED regime and hence have simultaneously high efficiency with high quality of service, the small firms need to carefully balance the efficiency and quality of service, some choose to focus on quality while other on efficiency (allowing them to compete on prices).

In contrast to the homogeneous market case, in the mixed market setting we focus on a simple model of demand and supply for which the *existence* of Nash equilibrium is guaranteed. This allows us, through the replication process, to focus on the *characterization* of the existing equilibria and in particular on the coexistence of the two operational regimes.

In addition to the findings above, we make a contribution by being the first that combine the notions of  $\epsilon$ -Nash equilibrium, market replication and heavy-traffic to study market equilibria. Each of these has been used in isolation. The  $\epsilon$ -Nash framework allows us go beyond the limited scope of Nash equilibrium and use general demand and capacity models. The notion of market replication allows us to discuss trends in sequences of markets to identify key characteristics for large and small firms. Combined with the notion of heavy-traffic, which is well studied for monopolists, this framework allows to characterize the equilibria behavior and obtain insights that are usually lost in traditional Nash equilibria analysis.

Previous work in game theory that studied sequences of games focused on three types of sequences: (i) sequences of games in which the action space is getting finer and finer, and while each game has discrete action space, the limiting game has continuous action space and (ii) sequences of games in which the number of agents grows, and (iii) a sequence of replicated markets with growing market size. In analyzing homogeneous markets we use the third framework, whereas to study the problem of mixed markets we develop a new framework that combines (ii) and (iii) above.

Our contribution to the existing literature is then threefold: first, on the *existence* front we show that an approximate equilibrium exists under very weak conditions both on the demand model and the service supply model, when the aggregate demand is large enough. Thus, helping to bridge the gap between the practice and the theory of service competition. Second, in the *characterization* front, we identify key characteristics of competition in markets with large-scale providers that were overlooked in the current literature. For example, while the optimality of high service levels

has been established for single, isolated, facilities, we are the first to show this in a competitive setting. In particular, we are the first to show that the QED regime is the optimal operational regime in a competitive setting. Third, we use a novel approach to the above, in using the ideas of replicated markets, and developing the notion of replication and duplication in order to study various phenomena in such markets.

## 2. Literature Review

Our work draws both on game theory and its application to the analysis of competition as well as on the analysis of service systems and in particular large-scale ones through queueing theory. These two streams of literature are not disjoint, and some recent work lies at the intersection of the two streams.

The literature on competition in service industries dates back to the late 1970s, but mostly focuses on competition on a single attribute – price *or* service level – rather than multiple attributes. Luski (1976) and Levhari and Luski (1978) confine themselves to two service providers, assuming all customers choose their provider strictly on the basis of the full price, i.e., the price plus the expected waiting time multiplied by a customer-specific cost rate. Customers' cost rates are independent and identically distributed (i.i.d). With service rates exogenously given, the competition between the two firms is confined to their price choices only. Chen and Wan (2003) showed that Nash equilibrium exists for the above model when the firms' service rates are identical, while demonstrating that the above may not hold under non-identical service rates.

Kalai et al. (1992) were the first to study a model in which service firms compete in terms of their capacity choices with exogenously given prices. Customers are served on a FIFO basis from a single queue by one of two service facilities. The authors show that asymmetric Nash equilibria of service rate pairs may arise, sometimes associated with infinite waiting times.

De Vany and Saving (1983) are the first to address a model in which firms compete with several rather than a single strategic instrument. This paper addresses a model with an arbitrary number of identical firms who simultaneously choose a price and service rate. The authors establish the existence of a symmetric equilibrium in a model where all customers share the same waiting cost rate and the total demand volume in the industry is given by a general function of the lowest full price.

Johari et al. (2007) consider investment and market structure in a model of congestion-sensitive service provision. They find that returns to investment and the timing of strategic decisions are crit-

ical determinants of the outcome of the game. They show that if the decisions are made simultaneously and firms exhibit decreasing return to investment, then if a pure strategy Nash equilibrium exists, it is unique, symmetric and efficient. They also establish conditions for existence of pure strategy Nash equilibrium in special cases.

All of the above papers assume that customers either have no choice in selecting their service provider or make the selection on the basis of the *full price* only. So (2000) and Cachon and Harker (2002) study competition in the presence of economics of scale and are the first to consider *differentiated* services, i.e., to analyze a model in which other service attributes matter along with the full price. Cachon and Harker (2002), for the case of two service providers, allow each firm's demand rate to be specified as a function of both firms' full-price values; When the demand rate functions are linear, the known equilibrium results merely exclude the existence of multiple equilibria, and this only when the demand rates are sufficiently large. In contrast to the latter, So (2000) establishes the existence of a unique equilibrium with an arbitrary number of competing M/M/1 service firms, when the demand rate functions are specified as a special type of attraction model. Afeche and Mendelson (2004) address a single firm model in which customers aggregate price and waiting time via a full price measure, now specified as a function of the price and two characteristics of the waiting time distribution.

Allon and Federgruen (2007) appear to provide the first competition model to address differentiated services while treating the prices and waiting time standards as fully independent attributes. This allows for different customers to exercise different explicit or implicit tradeoffs. Different types of competition and resulting equilibrium behavior are discussed, depending on the industry dynamics through which the firms select their strategic choices. Allon and Federgruen (2007) prove the existence of Nash equilibrium in the different types of competition. Moreover, they show that under certain conditions if firms make their strategic decisions sequentially, selecting service levels, hence waiting time standards, first, this results in an equilibrium with higher service levels, prices and demand volumes, as compared to the equilibrium reached in a simultaneous competition model. We extend Allon and Federgruen (2007) in multiple directions: (i) we extend the supply side to allow the service provider to adjust its capacity by increasing or decreasing the number of customer service representatives (giving rise to an M/M/N queue, where  $N$  is a decision made by the firm). This is a prevalent method of capacity management of service providers; (ii) on the demand side we allow for full generality of the customers' sensitivity to service level (and putting very mild conditions on their sensitivity to price changes), and (iii) while Allon and Federgruen (2007) focus on providing full analytical characterization of the equilibria that arise in the differ-

ent games, they fail to explain some key practical observations, such as the existence of industry standard and different levels of service-level differentiation in different markets. The methodology in the current paper, while appropriate only for markets with high volumes of demand, allows us to identify key characteristics of the market behavior that could not be identified in the framework of Allon and Federgruen (2007).

We refer the reader to Allon and Federgruen (2007) for a systematic discussion of several variants and extensions thereof and to Hassin and Haviv (2003) for a general survey of queueing models with competition.

Most of the analysis of service systems focuses on settings with a single supplier and no competition. Within this body of literature, the analysis of large-scale systems has gained increasing attention in the literature, starting with the seminal paper by Halfin and Whitt (1981). This paper was the first to formally identify a regime in which high service level and high efficiency coexist. Specifically, consider a sequence of  $M/M/N$  queues, all with the same service rate  $\mu$ . Then, Halfin and Whitt show that as the demand,  $\Lambda$ , grows, the probability of delay  $P\{W > 0\}$  will converge to a number strictly between 0 and 1 if and only if the number of agents satisfies a *square-root rule*, i.e,

$$N = R + \beta\sqrt{R} + o(\sqrt{R}), \quad (1)$$

where  $R := \Lambda/\mu$  is the offered load and  $\beta$  is a strictly positive constant. In particular, a square root staffing rule as proposed in (1) leads to a situation in which non-trivial fraction of the customers do not wait at all before entering service and at the same time the utilization  $\rho := \Lambda/N\mu$  is very close to 1. It is from this combination of high efficiency and high service level that the alternative name *Quality and Efficiency Driven* (QED) regime emerges. The use of the square-root safety staffing rule dates back to Erlang in his 1923 paper (that appeared in [12]), and has been used in different applications before its formalization by Halfin and Whitt (1981); see e.g. Kolesar and Blum (1973).

The QED regime has been shown to rise as an outcome of congestion-dependent demand and as the optimal regime in many capacity optimization settings. Borst et al. (2004) show that the QED regime is the optimal regime when minimizing linear staffing and waiting-time costs in an  $M/M/N$  setting. To the best of our knowledge, Armony and Maglaras (2004) are the first to show that the optimal staffing in a system with congestion-dependent demand leads to the QED regime. They do so in a setting where customers can choose between real-time service and a call-back option. Motivated by this work, Whitt (2003) analyzes the steady state behavior in an  $M/M/N$  setting where the demand is congestion-dependent. He identifies cases under which the

resulting equilibrium places the system in the QED regime. Maglaras and Zeevi (2003, 2005) show that the QED is the optimal regime in a setting with best effort and guaranteed performance customers. Finally, Green and Kolesar (2004) provide a summary of the use of the square root rule in emergency services.

Beyond the identification of this regime, Halfin and Whitt (1981) underline the significance of economies of scale. Their result motivated a large body of literature that exploits these economies of scale to derive tractable solutions for rather complex problems. Examples are Armony and Maglaras (2004), Armony (2005), Atar et al. (2004), Atar (2004), Gurvich et al. (2006), Gurvich and Whitt (2007), Tezcan (2006) and Tezcan and Dai (2006).

It is noteworthy that all of the above results consider a single facility with demand that at most depends on the congestion and price in this single facility. Our paper is the first to show that QED emerges as the optimal operational regime for firms in a competitive setting with certain linear demand models. In doing so we relate the competition setting with linear demand models to that monopolist setting with linear waiting-time costs.

Finally, the notion of  $\epsilon$ -Nash equilibria that we use in this paper has been used extensively in the economics literature. For the basic definition we rely on Tijms (1981). Dixon (1987) uses the idea of market replication in the context of price competition. While our form of replication is different, our analysis is inspired by his concept. The application of  $\epsilon$ -Nash in the operations literature is rare. Lu et al. (2007) use this concept in a setting where Nash equilibrium is shown to exist but the  $\epsilon$ -Nash helps in characterizing the equilibrium in a game with a large number of players approaching a continuum of players. Dasci (2003) uses this concept in the context of  $\epsilon$ -subgame-perfect equilibrium. We are the first to combine the concepts of  $\epsilon$ -Nash, market replication and heavy-traffic in the context of operational settings. As stated above, this allows us to discuss both stability and trends in markets of competing service providers.

### 3. The Model

We consider a market with a set  $\mathcal{I} = \{1, \dots, I\}$  of competing service firms, each operating as an  $M/M/N$  facility. Firm  $i$  positions itself in the market by selecting a price  $p_i$  and a service level  $\theta_i$ . Initially, we focus our attention on service-level guarantees that are given in terms of the customers' waiting time rather than their whole sojourn time in the system. In terms of customer sensitivity, this choice is supported by the subjective sense that "time crawls" while a customer is passively waiting in line, whereas it "flies" while the customer is actively talking to the agent. It is

also the prevalent form of guarantees in the call-center and the banking industry. To be consistent with the industry practice, the service level is assumed to be defined through service-level (SL) targets of the form

$$\mathbb{P}\{W_i > T_i\} \leq \phi, \quad i \in \mathcal{I},$$

where  $W_i$  is the steady-state waiting time,  $T_i > 0$  is the target response times and  $0 < \phi < 1$  is the satisfaction probability. This is consistent with industry practice that commonly uses  $\phi = 0.2$  (corresponding to 80% of the service requests being answered within target).<sup>1</sup> We further assume that customers in each facility are served FCFS. Service rates are assumed to be fixed and equal to  $\mu_i$  for firm  $i$ , and the capacities are adjusted through the choice of the number of agents (or service representatives), as given by the integral decision variable  $N_i$ . With the above restrictions, the service level  $\theta_i$  is defined as the difference between a *benchmark* upper bound  $\bar{\theta}$  and the actual target time  $T_i$ , i.e.  $\theta_i = \bar{\theta} - T_i$ . Consequently, the higher the service level, the lower the waiting time target. The maximal service level is  $\bar{\theta}$ , corresponding to  $T_i = 0$ . Each firm  $i$  is able to select its capacity or service rate so as to guarantee any given service-level target in  $[0, \bar{\theta}]$ . Thus,  $\theta_i \in [0, \bar{\theta}]$ . We let  $\Theta := \times_{i=1}^I [0, \bar{\theta}]$ .

It is well known that an  $M/M/N$  queue is stable if and only if the capacity is greater than the demand. Formally, service provider  $i \in \mathcal{I}$  will be stable if and only if

$$N_i > R_i, \quad (2)$$

where  $R_i := \lambda_i / \mu_i$  is the offered load given the demand  $\lambda_i$  faces by firm  $i$ . The required capacity for firm  $i$  can be then expressed in the form

$$N_i = R_i + \hat{e}_i(\lambda_i, \theta_i), \quad (3)$$

where  $\hat{e}_i(\lambda_i, \theta_i)$  is the excess capacity required to satisfy the service-level target and such that  $N_i$  is an integer number. Naturally, we define  $\hat{e}_i(\lambda_i, \theta_i) = 0$  whenever  $\lambda_i = 0$ . In the  $M/M/N$  setting the actual value of  $\hat{e}_i(\lambda_i, \theta_i)$  can be easily calculated using any Erlang-C calculator<sup>2</sup>. The two terms in (3) represent the two components of the required capacity: the first, *volume-based capacity*, is the base capacity ensuring that the service process is stable; the second component ensures that the desired service levels are achieved and is referred to as the *service-based capacity*.

Firm  $i$  incurs a cost  $c_i$  per customer served and a cost  $\gamma_i$  per agent, per unit of time. This corresponds to the cost of capacity being linear in the number of agents. We hasten to say that

<sup>1</sup>Our results are easily extended to the case where  $\phi$  is allowed to vary between different firms.

<sup>2</sup>Freeware calculators can be found, for example, at <http://iew3.technion.ac.il/~serveng/4CallCenters/Downloads.htm> or <http://www.cs.vu.nl/~koole/ccmath/ErlangC>.

some of our results hold for more general increasing capacity cost functions; see Remarks 4.4 and 4.6. The price  $p_i$  may be chosen from a compact interval  $[p_i^{min}, p_i^{max}]$ ,  $i = 1, \dots, I$ . Clearly, firm  $i$  selects a price  $p_i$  which results in a non-negative gross profit margin  $p_i - c_i - \gamma_i/\mu_i$ . Thus, without loss of generality, we select

$$p_i^{min} = c_i + \frac{\gamma_i}{\mu_i}, \quad i = 1, \dots, I. \quad (4)$$

As to  $p_i^{max}$ , it is allowed to obtain any value in  $[p_i^{min}, \infty)$ . We set  $\mathcal{P}_i := [p_i^{min}, p_i^{max}]$  and  $\mathcal{P} := \times_{i=1}^I \mathcal{P}_i$ .

In full generality, the demand rates are specified as general functions of *all* prices and waiting time standards (i.e.  $\lambda_i \equiv \lambda_i(p, \theta)$ ) that obey obvious monotonicity properties. The following Assumption is assumed to hold throughout the rest of the paper.

**Assumption 3.1 (regularity assumptions on the demand functions for differentiated services)**

For each  $i \in \mathcal{I}$ , the function

$$\lambda_i(\cdot, \cdot) : \mathcal{P} \times \Theta \mapsto \mathbb{R}_+$$

is strictly positive, continuous in all arguments, strictly decreasing in  $p_i$  and strictly increasing in  $\theta_i$ .

Firm- $i$ 's long-run-average profit  $\Pi_i$  is given by

$$\Pi_i(p, \theta) = \lambda_i \cdot \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\lambda_i, \theta_i). \quad (5)$$

We let

$$M_i = \max_{\theta \in \Theta, p \in \mathcal{P}} \lambda_i(p, \theta) \quad (6)$$

be the potential market size for firm  $i$ <sup>3</sup>. Note that the maximizer above is well defined by the compactness of  $\mathcal{P}$  and the assumed continuity of the demand functions. We consider the normalized profit functions  $i$

$$\bar{\Pi}_i(p, \theta) := \frac{\Pi_i(p, \theta)}{M_i}, \quad i \in \mathcal{I}. \quad (7)$$

The scaling is required to “level the playing-field” for all markets. Indeed, as we will be considering a sequence of markets with growing demand, the profits might diverge to infinity without the proper normalization and prevent any form of analysis. In practical terms, this scaling reflects the fact that companies measure their profits in terms of the market potential. Airlines, for example,

---

<sup>3</sup>For example, we note that in the popular MultiNomial Logit demand model, the constant  $M$  is defined as one of the model primitives.

measure their profits (or losses) in billions of dollars rather than in hundreds of thousands. Since  $M_i$  is a constant, the directional effect of price or service-level adjustments on the profit remains the same. Consequently, the analysis of the game with scaled payoffs has direct implications on the game with unscaled profits; see Remark 4.3.

The assumption of large-scale service systems is formally introduced by considering a sequence of replicated markets indexed by a market-scale multiplier  $\Lambda$  and letting  $\Lambda$  grow. The demand is assumed to scale with the market-scale multiplier in a natural way. Specifically, we let

$$\Lambda_i(p, \theta) := \Lambda \cdot \lambda_i(p, \theta), \quad (8)$$

be the demand facing firm  $i$  in the  $\Lambda^{th}$  market. The corresponding sequence of profit functions is then given by

$$\Pi_i^\Lambda(p, \theta) = \Lambda_i \cdot \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \hat{e}_i(\Lambda_i, \theta_i), \quad i \in \mathcal{I}, \quad (9)$$

and the sequence of normalized profits is defined to be

$$\bar{\Pi}_i^\Lambda(p, \theta) := \frac{\Pi_i^\Lambda(p, \theta)}{M_i^\Lambda}, \quad i \in \mathcal{I}, \quad (10)$$

where  $M_i^\Lambda$  is the scaled market potential for firm  $i$  and is given by  $M_i^\Lambda = M_i \Lambda$  with  $M_i$  as defined in (6). For future reference we make the following formal definition

**Definition 3.1** The  $\Lambda$ -game with normalized profits is the  $I$ -player game with profit functions  $\bar{\Pi}_i^\Lambda$ ,  $i \in \mathcal{I}$  and strategy space  $\mathcal{P} \times \Theta$ . The **un-normalized  $\Lambda$ -game** is the  $I$ -player game with profit functions  $\Pi_i^\Lambda$  and strategy space  $\mathcal{P} \times \Theta$ .

Note that considering a sequence of markets with growing demand is not equivalent to considering a single market with infinite demand. Rather, the key idea in market replication is to embed the real market (with fixed market size) into a sequence of markets with growing demand. If one is able to get meaningful results for the sequence of markets, these can be applied to the market with fixed size, as long as the size is large enough. The main idea behind looking at a sequence of markets, rather than on one, is to examine how the stability of the market and the market outcomes change with the increase in market size.

We now turn to discuss some preliminaries towards the statement of our main results. These are: (i) the  $\epsilon$ -Nash equilibrium (ii) the pure price-competition game, and (iii) the many-server heavy-traffic regime.

### 3.1 $\epsilon$ -Nash equilibria

We will be considering the limiting behavior of the equilibria as the market size  $\Lambda$  grows indefinitely. As the existence of equilibrium is not guaranteed, we cannot explicitly characterize the equilibria for a given value of  $\Lambda$  and take the formal limit of these. Rather, we use the more general notion of  $\epsilon$ -Nash, equilibria which we adopt from Tijms (1981). Rather than giving it in general notation, we give the definition as it would apply to our setting. Towards this end, we let

$$(p_i, \theta_i) \uparrow (\tilde{p}, \tilde{\theta})_{-i} = ((\tilde{p}_1, \tilde{\theta}_1), \dots, (\tilde{p}_{i-1}, \tilde{\theta}_{i-1}), (p_i, \theta_i), (\tilde{p}_{i-1}, \tilde{\theta}_{i-1}), \dots, (\tilde{p}_I, \tilde{\theta}_I))$$

**Definition 3.2 ( $\epsilon$ -Nash equilibrium)** *Let  $\epsilon_1, \dots, \epsilon_I$  be non-negative real numbers and consider the  $I$  players game with utility functions  $\bar{\Pi}_i(p, \theta)$ ,  $i \in \mathcal{I}$  and strategy space  $\mathbb{X} := \mathcal{P} \times \Theta$ . We say that  $\hat{x} = ((\hat{p}_1, \theta_1), \dots, (\hat{p}_I, \theta_I)) \in \mathbb{X}$ , is an  $(\epsilon_1, \dots, \epsilon_I)$ -Nash equilibrium point of the game, if for each  $i \in \mathcal{I}$  and for each  $\tilde{x} \in \mathbb{X}$ , we have*

$$\bar{\Pi}_i(\tilde{x}_i, \hat{x}_{-i}) \leq \bar{\Pi}_i(\hat{x}) + \epsilon_i. \quad (11)$$

The abbreviated term  $\epsilon$ -Nash equilibrium refers to an  $(\epsilon, \dots, \epsilon)$ -Nash equilibrium.

Note that the regular Nash equilibrium is a special case of  $\epsilon$ -Nash equilibrium in which  $\epsilon = 0$ . The importance of  $\epsilon$ -Nash equilibrium was discussed in the introduction. Not only it relaxes the requirement for stability, but it also allows us to obtain key insights about the market behavior. More specifically, it allows us to relate our market game to the pure price-competition game that is defined in the next section.

### 3.2 Pure price-competition

This paper is concerned with competition on both the price and service-level dimensions. In the current section, however, we introduce a somewhat relaxed market in which firms compete only on price. We refer to this competition as the **pure price-competition**. Our main results will relate the competition in the original framework with this pure price-competition.

Towards that end, we consider a market with the same set  $\mathcal{I}$  of firms, having the same capacity costs and same price sets  $\mathcal{P}$ , but with demand functions given by

$$\Lambda_i^P(p) := \Lambda_i(p, \vec{\theta}), \quad (12)$$

where  $\vec{\theta} := (\bar{\theta}, \dots, \bar{\theta})$ . In other words,  $\Lambda_i^P(p)$  is the demand function obtained when all firms in the market are providing the best possible service level and using the price vector  $p$ . To define the pure price-competition we also define the profit functions

$$\bar{\Pi}_i^{\Lambda, P}(p) := \frac{\Lambda_i^P \cdot \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right)}{M_i^\Lambda} = \frac{1}{M_i} \lambda_i(p, \vec{\theta}) \cdot \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right), \quad i \in \mathcal{I}. \quad (13)$$

This profit function is obtained, then, by removing the service-based capacity and setting the service level at its maximal value. Note that  $\bar{\Pi}^{\Lambda, P}$  is invariant to the scale  $\Lambda$ , so we may without loss of generality use  $\bar{\Pi}_i^P(\cdot) := \bar{\Pi}_i^{\Lambda, P}(\cdot)$ . The *Pure Price-Competition Game* is then defined as follows:

**Definition 3.3 (The pure price-competition game)** *The pure price-competition game is the  $I$ -player game with profit functions  $\bar{\Pi}_i^P(\cdot)$  and strategy space  $\mathcal{P}$ .*

Our  $\epsilon$ -Nash equilibrium framework allows us to avoid issues of existence of real Nash equilibrium for the game with strategy space  $\mathcal{P} \times \Theta$ , and thus allows us to use very general demand functions. Indeed, existence of real Nash equilibrium is not guaranteed in the game with strategy space  $\mathcal{P} \times \Theta$ . The pure price-competition model is a simpler, one dimensional, game with numerous sufficient conditions for the existence of equilibria. For example, it suffices to have that  $\bar{\Pi}_i^P(\cdot)$  be continuous and quasi-concave with respect to  $p_i$  (see §2.3 of Cachon and Netessine (2004)). The sufficient conditions are guaranteed, for example, for attraction models, a specific instance of which is the Multinomial Logit demand model, and the Cobb Douglas demand model. We emphasize that the uniqueness is not essential and our subsequent results are easily extended to the case of multiple price equilibria. The uniqueness is imposed for simplicity of presentation. We formally state these requirements in the following assumption, which will be assumed to hold throughout the rest of the paper.

**Assumption 3.2 (existence and uniqueness of equilibrium for the pure price-competition model)** *The pure price-competition game has a unique equilibrium price vector  $p^* = (p_1^*, \dots, p_I^*)$ .*

Henceforth, we will use the notation  $p^*$  when referring to this unique price equilibrium in the pure price-competition game. The precise characterization of this equilibrium is of no importance for our analysis beyond the basic definition of Nash equilibrium, which guarantees that

$$p_i^* = \operatorname{argmax}_{p_i \in \mathcal{P}_i} \bar{\Pi}_i^P(p_i, p_{-i}^*). \quad (14)$$

We emphasize that the modified payoff functions  $\bar{\Pi}_i^P$  are used only to define  $p^*$  properly. From here onward we return to the original, combined service and price competition model and, in particular, to the payoff functions  $\Pi_i^\Lambda$  and  $\bar{\Pi}_i^\Lambda$  as defined in (9) and (10) respectively.

### 3.3 Economies of scale and the QED regime

Our equilibria analysis relies on three pillars:  $\epsilon$ -Nash equilibria, market replication and heavy-traffic analysis. The main result that we borrow from the heavy-traffic literature concerns the economies of scale in the many-server heavy-traffic regime. The relevant economies of scale are summarized in Lemma 3.4 below. Towards the statement of the lemma, let  $M(\Lambda)/M(\mu)/N^\Lambda$  stand for an  $M/M/N$  queue with some arrival rate  $\Lambda$ , service rate  $\mu$  and  $N^\Lambda$  agents, and let  $W^\Lambda$  be the corresponding steady-state waiting time. We set  $W^\Lambda = \infty$  when steady state does not exist. Finally, we let

$$N^{*,\Lambda}(T) := \min \{ N \in \mathbb{Z}_+ : \mathbb{P} \{ W^\Lambda > T \} \leq \phi. \},$$

that is  $N^{*,\Lambda}(T)$  is the minimal number of agents required to satisfy the service level with target time  $T$ .

**Lemma 3.4** ( *$M/M/N$  economies of scale*) *Fix  $0 < \phi < 1$ . Consider a sequence of  $M(\Lambda)/M(\mu)/N^\Lambda$  queues and a sequence  $T^\Lambda$  of service-level targets with  $T^\Lambda \rightarrow 0$ . Then,*

$$\mathbb{P} \{ W^\Lambda > T^\Lambda \} \leq \phi, \text{ for all } \Lambda,$$

*if and only if*

$$N^{*,\Lambda}(T^\Lambda) = \frac{\Lambda}{\mu} + e^\Lambda(T^\Lambda),$$

*where  $e^\Lambda(T^\Lambda) \rightarrow \infty$  but  $e^\Lambda(T^\Lambda)/\Lambda \rightarrow 0$  as  $\Lambda \rightarrow \infty$ . Moreover, if  $T^\Lambda = \bar{T}/\sqrt{\Lambda}$ , then,*

$$\mathbb{P} \{ W^\Lambda > T^\Lambda \} \leq \phi, \text{ for all } \Lambda,$$

*if and only if*

$$N^{*,\Lambda}(T^\Lambda) = \frac{\Lambda}{\mu} + \beta^*(\bar{T})\sqrt{\Lambda} + o(\sqrt{\Lambda}), \tag{15}$$

*for some  $\beta^*(\bar{T}) > 0$ .*

Note that in this Lemma we can have  $T^\Lambda = 0$  for all  $\Lambda$ . This, by the second part of the Lemma would dictate staffing according to the square-root safety staffing rule. The proof of Lemma 3.4 is omitted. The interested reader is referred to §9 of Borst et al. (2004) or to the Lemma 4.1 of Gurvich and Whitt (2007).

The most important implication of Lemma 3.4, and one that will play a key role in our results, is the fact that even with extremely high service level the service-based capacity consists of only a small fraction of the overall required capacity. Moreover, in the case where  $T^\Lambda$  decreases at

rate  $1/\sqrt{\Lambda}$ , the cost of the service-based capacity is of the order of  $\sqrt{\Lambda}$  which corresponds to order of  $1/\sqrt{\Lambda}$  percentage of the overall capacity cost. Hence, the marginal investment in service level becomes negligible as the demand volume grows. These economies of scale, however, do not trivialize the question of determining the right service levels for large-scale service providers. Rather, it implies that to understand the system behavior one should focus on the second order terms and characterize their behavior and in particular the order of magnitudes of the equilibrium values of the service levels. This is the path we take in the following section.

## 4. Homogeneous Markets

In this section we analyze homogeneous markets. That is, markets in which all firms share the same scale. We start with the **first-order** results that identify the relative importance of price and service-level for the general demand models defined in Assumption 3.1 and provide a rough (first-order) approximation for the equilibrium price and service-levels. In §4.2 we proceed to further refine these results and identify the operational and pricing regimes that emerge when one further restricts the demand model.

### 4.1 Existence and first-order characterization

We answer the two fundamental equilibria questions of existence and characterization for the sequence of  $\Lambda$ -game with normalized profits. This existence and characterization have some direct implications on the market behavior that are discussed right after the statement of the two results.

**Theorem 4.1 (existence)** *Fix a sequence  $\epsilon^\Lambda$  such that, as  $\Lambda \rightarrow \infty$ ,  $\epsilon^\Lambda \rightarrow 0$  but  $\Lambda\epsilon^\Lambda \rightarrow \infty$ . Then, there exists a sequence  $\theta^\Lambda \rightarrow \bar{\theta}$  as  $\Lambda \rightarrow \infty$  such that, for each  $\Lambda$ , the vector*

$$(p^*, \theta^\Lambda) = ((p_1^*, \theta^\Lambda), \dots, (p_I^*, \theta^\Lambda))$$

*is an  $\epsilon^\Lambda$ -Nash equilibrium for the  $\Lambda$ -game with normalized profits.*

**Theorem 4.2 (first-order characterization)** *Let  $(p^\Lambda, \theta^\Lambda)$  be an  $\epsilon^\Lambda$ -Nash equilibrium for each  $\Lambda$ -game with normalized profits, where  $\epsilon^\Lambda \rightarrow 0$  and  $\Lambda\epsilon^\Lambda \rightarrow \infty$ . Then, there exists a sequence  $\delta^\Lambda \rightarrow 0$  such that*

$$\theta_i^\Lambda \in [\bar{\theta} - \delta^\Lambda, \bar{\theta}], \quad i \in \mathcal{I}, \quad (16)$$

*and*

$$p_i^\Lambda \in [p_i^* - \delta^\Lambda, p_i^* + \delta^\Lambda], \quad i \in \mathcal{I}. \quad (17)$$

We now turn to discuss the implications of the above existence and characterization results.

**Service-level differentiation:** A consequence of Theorem 4.1 is that the firms do not need to differentiate themselves in terms of service-level. Specifically, the sequence  $(p^*, \vec{\theta}^\Lambda)$  will be an  $\epsilon$ -Nash equilibrium for all  $\Lambda$  large enough, even if for each  $\Lambda$ ,  $\theta^\Lambda$  is a vector with all identical elements  $(\theta^\Lambda, \dots, \theta^\Lambda)$ . This implies a very strong decoupling result between prices and service levels. The companies may set their prices to  $p^*$ . Once the price is fixed, and as a consequence of the large firms' economies-of-scale, they can match the service level of the competitor without moving much away from the equilibrium. Hence, as long as the firms provide high service levels they can disregard the competition on the service-level dimension. This is consistent with firms in certain industries aligning themselves in accordance with industry standards, such as the well known 80 – 20 rule that stipulates serving 80% of the customers within 20 seconds.

Theorem 4.2 strengthens the existence result by showing that, for large enough systems, any equilibrium point must have extremely good service levels, and in particular response-time targets that converge to zero as the market size scale grows.

**The equilibrium prices:** Theorem 4.2 implies that, in equilibrium, the market will, in some sense, behave according to the price-competition equilibrium. The underlying reason is the economies of scale allows (and actually forces) the firms to provide extremely high service levels. In turn, the equilibrium service levels bring the market very close to the pure price-competition market in which the demand function are defined by setting the service-level to their maximum possible value; see §3.2.

**Remark 4.3 (Nash equilibria for the un-normalized  $\Lambda$ -game)** *To emphasize the implications of the corollary, assume that, for each  $\Lambda$ , a Nash equilibrium exists. Under this assumption the games with normalized and un-normalized profit function are equivalent in terms of their equilibria (as the difference is only scaling by a constant). Consequently, Theorem 4.2 implies that, provided that Nash equilibria exists for each  $\Lambda$ , the sequence of Nash equilibria  $(p^\Lambda, \theta^\Lambda)$  must satisfy that*

$$p_i^\Lambda = p_i^* \pm o(1), \quad i \in \mathcal{I},$$

and

$$\theta_i^\Lambda = \bar{\theta} - o(1), \quad i \in \mathcal{I},$$

where we say that a sequence  $a^\Lambda$  is  $o(1)$  if  $a^\Lambda \rightarrow 0$  as  $\Lambda \rightarrow \infty$ . This consequence is hard to obtain through direct analysis of the original markets. The ability to extract such results illustrates the benefits of our approximate equilibria framework.

**Remark 4.4 (general increasing capacity costs)** *It should be emphasized that all of our results above still hold if the linear capacity cost  $C_i(N_i) = \gamma_i N_i$  is replaced with an arbitrary strictly increasing function  $C_i(\cdot)$  with  $C_i(0) = 0$ . ■*

To summarize we have established the existence of  $\epsilon^\Lambda$ -Nash equilibria. More importantly, we have shown so far that the prices in equilibrium will be close to the pure price-competition prices and that the service-levels will be close to the ideal service levels. The next question is exactly how close will these be. This is the subject of the next section.

## 4.2 Second-order characterization

Our first-order characterization proved the existence of an approximate equilibrium in a sequence of markets and identified the regions in which the firms should price and set their service levels to achieve this type of stability. Specifically, the first-order analysis shows that if  $(p^\Lambda, \theta^\Lambda)$  is a corresponding sequence of  $\epsilon^\Lambda$ -Nash equilibria, then  $p_i^\Lambda \in [p_i^* - \delta^\Lambda, p_i^* + \delta^\Lambda]$ , and  $\theta_i^\Lambda \in [\bar{\theta} - \delta^\Lambda, \bar{\theta}]$ ,  $i \in \mathcal{I}$ .

We now proceed to further refine our understanding of the decisions made by the firms. In particular we are interested in characterizing the relationship between the size of the market and the service level and pricing decisions. The key question is how a firm should fine-tune its pricing and service level decision with growing demand and in the presence of economies of scale. We will refine our first-order results, by translating the bounds  $\epsilon^\Lambda$  on the profit-deviations to bounds on the actions  $p^\Lambda$  and  $\theta^\Lambda$ .

In absence of general results on such translations, and without additional restrictions on the demand mode, beyond continuity, further refinements are not possible. They are achievable, however, with stronger assumptions on the underlying demand model. Indeed, we will show that if one restricts the model to a linear demand model one can identify these convergence rates. Specifically, we assume that

$$\lambda_i(p, \theta) = \left[ a_i(\theta_i) - b_i p_i - \sum_{j \neq i} \alpha_{ij}(\theta_j) + \sum_{j \neq i} \beta_{ij} p_j \right]^+. \quad (18)$$

The functions  $a_i(\cdot)$  are assumed to be differentiable and strictly increasing in the service level  $\theta_i$ , with

$$a'_i(\bar{\theta}) > 0. \quad (19)$$

If  $a_i(\cdot)$  is defined only on  $[0, \bar{\theta}]$  then this should be interpreted as the left-derivative. The cross term functions  $a_{ij}(\cdot)$  are assumed to be non-decreasing. Beyond these basic requirements, the precise characteristics of the functions  $a_i(\cdot)$  and  $a_{ij}(\cdot)$  are immaterial for our results. We assume that a *uniform* price increase by all  $I$  firms cannot result in an increase in any firm's demand volume and that a price increase by a given firm cannot result in an increase of the industry's aggregate demand volume, i.e.,

$$(D) \quad b_i > \sum_{j \neq i} \beta_{ij}, i = 1, \dots, I; \quad (D') \quad b_i > \sum_{j \neq i} \beta_{ji}, i = 1, \dots, I. \quad (20)$$

This condition is usually referred to as the "Dominant Diagonal" condition. We will assume that

$$\lambda_i(p, \theta) > 0, \forall (p, \theta) \in \mathcal{P} \times \Theta. \quad (21)$$

Equipped with this more refined demand model, we are able to show next the optimality of the QED regime in a competition setting. Hereafter, we say that  $b^\Lambda = O(a^\Lambda)$  if  $\limsup_{\Lambda \rightarrow \infty} b^\Lambda / a^\Lambda < \infty$ .

**Theorem 4.5 (optimality of the QED regime)** *Suppose the demand functions  $\lambda_i(\cdot, \cdot)$  satisfy (18) and (19). Fix a sequence  $\epsilon^\Lambda$  such that  $\epsilon^\Lambda = O(1/\sqrt{\Lambda})$ , and suppose that  $(p^\Lambda, \theta^\Lambda)$  is a sequence of  $\epsilon^\Lambda$ -Nash equilibria. Then,*

$$p_i^\Lambda = p_i^* \pm O\left(\frac{1}{\sqrt{\Lambda}}\right), i \in \mathcal{I}, \quad (22)$$

and

$$\theta_i^\Lambda = \bar{\theta} - O\left(\frac{1}{\sqrt{\Lambda}}\right), i \in \mathcal{I}. \quad (23)$$

Observe that, by Lemma 3.4, equation (23) implies that the sequence of staffing level vectors  $N^\Lambda = (N_1^\Lambda, \dots, N_I^\Lambda)$ , corresponding to the sequence  $(p^\Lambda, \theta^\Lambda)$  of  $\epsilon^\Lambda$ -Nash equilibria, must satisfy

$$\liminf_{\Lambda \rightarrow \infty} \frac{N_i^\Lambda - R_i}{\sqrt{\Lambda}} > 0, \text{ and } \limsup_{\Lambda \rightarrow \infty} \frac{N_i^\Lambda - R_i}{\sqrt{\Lambda}} < \infty,$$

placing all service providers in the QED regime. Indeed, all service-levels such that  $\theta_i^\Lambda - \bar{\theta} = O(\sqrt{\Lambda})$  (including  $\theta_i^\Lambda = \bar{\theta}$ ) are satisfied only with a square-root safety staffing. We also note that

the while (22) does depend on the linearity of the demand with respect to prices, the optimality of the QED regime, as given by (23) actually requires only the condition on the derivative of  $a_i(\cdot)$  in the point  $\bar{\theta}$ . In particular, it would hold under general continuous demand models as long as the derivative of  $\lambda_i(p, \theta)$  with respect to  $\theta_i$  is uniformly bounded from below by a positive constant, where the uniformity is with respect to  $p$  and  $\theta_j$  for  $j \neq i$ . Furthermore, the assumption that the capacity cost is linear not necessary for Theorem 4.5 to hold. The following remark addresses this issue.

**Remark 4.6 (concave capacity-cost functions)** The above result is proved for linear capacity cost functions  $C_i(x) = \gamma_i x$ . The proof reveals, however, that the result holds as long as the functions  $C_i(\cdot)$  are increasing, continuously differentiable and have a bounded first derivative, that is  $\sup_{x \geq 0} C_i'(x) < \infty$  and  $\inf_{x \geq 0} C_i'(x) > 0$  for all  $i \in \mathcal{I}$ . The result holds, in particular, for arbitrary concave capacity-cost functions with a first derivative that is bounded from below. As agent salaries are usually bounded from below by some minimal wage that is independent of the firm's size, such concave cost functions seem to be general enough for all practical purposes. Convex functions, on the other hand, present some problems in this framework. To wit, consider the cost function  $C_i(x) = x^2$ . Then, as  $\Lambda \rightarrow \infty$ , the revenue is at most  $M_i \Lambda$  but the cost grows as  $\Lambda^2$ , leading to divergence to  $-\infty$  of the profit functions. Hence, firms with such convex capacity cost functions will not survive in the large markets.

**Remark 4.7 (implication for Nash equilibria)** Theorem 4.5 implies, in particular, that if existence of an actual Nash equilibrium can be established for each value of  $\Lambda$ , then the resulting sequence of Nash equilibria must satisfy (22) and (23). In particular, following the logic in Remark 4.3 we obtain some properties of the sequence of competition in the sequence of markets with unscaled profits. ■

**Remark 4.8 (connection to the monopolist setting)** Borst et al. (2004) show that a monopolist that seeks to minimize the sum of linear waiting-time and capacity costs will optimally operate in the QED regime. A firm that operates in a competitive environment but chooses its service level as a monopolist, by following the recommendations in Borst et al. (2004), is ignoring important aspects of competition. Still, our results show that as long as the demand model satisfies certain conditions, this firm will be actually doing “the right thing for the wrong reasons” as it will be operating in the right operational regime. This reasoning also suggests that in terms of choosing the right service-levels, a firm that faces a demand model with the right condition on the derivative

of  $\lambda_i(p, \theta)$  should choose their operational regime as if they were monopolists facing linear holding costs. Clearly, other conditions on the demand functions should correspond to different analogies and lead to different operational regimes.

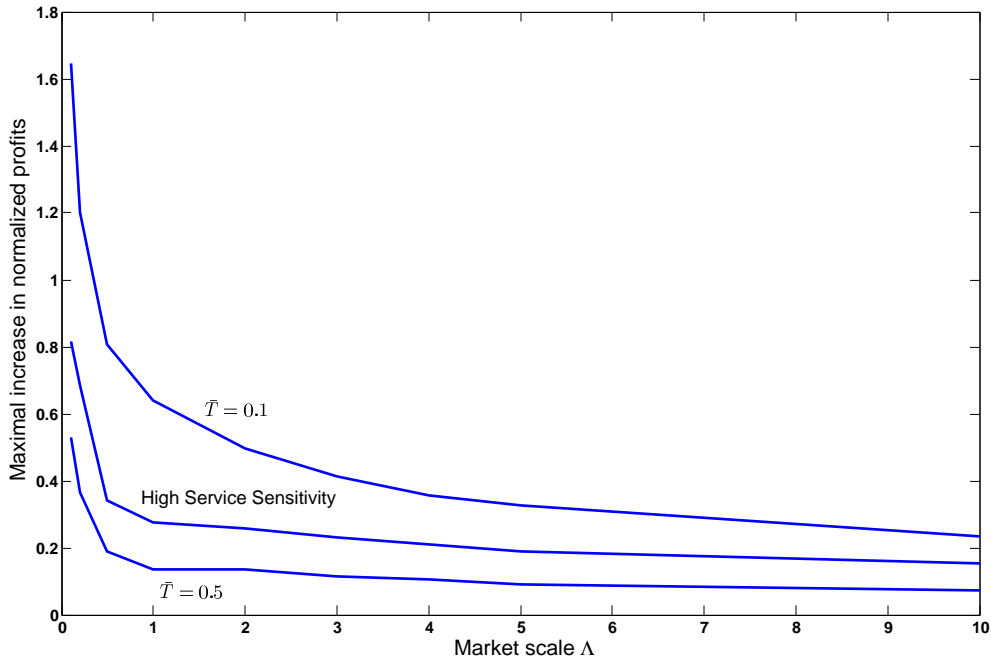
We end this section with an example that illustrate the concept of  $\epsilon^\Lambda$ -Nash equilibria.

**Example 4.1** To demonstrate the impact of the size of the market on the equilibrium behavior, we consider an industry with  $N = 3$  firms,  $\bar{\theta} = 1$ , and cost parameters  $c_1 = c_2 = 20, c_3 = 5$ , while  $\gamma_1 = \gamma_2 = 20, \gamma_3 = 35$ . The example may, therefore, apply to a setting with firm 3 an established local service provider and firms 1 and 2 competitors that have entered the local market more recently from a foreign or remote location, where capacity costs ( $\gamma$ ) are lower but the per customer access costs ( $c$ ) are higher. In *this* example firms experience identical price sensitivities, i.e.,  $b_i = 10$  and  $\beta_{ij} = 4.75, \forall i \neq j$ . Finally,  $a_1(\theta) = 205 + 0.1\theta_1, a_2(\theta) = 205 + 0.1\theta_2, a_3(\theta) = 295 + 0.1\theta_3. \alpha_{ij} = -0.01\theta_j, \forall i \neq j$ .

We study a sequence of games indexed by  $\Lambda$ , as defined above. First, to illustrate the importance of the  $\epsilon$ -Nash framework, we verified that the game and any of its  $\Lambda^{th}$  scaled games do not have a Nash equilibrium. Indeed, with  $\Lambda = 1$ , for example, the game oscillates between the following price vectors (73.1, 73.5, 77.5), (73.6, 73.6, 76.8), and (73.3, 74.1, 76.90) and service levels (0.02, 0.01, 0.02), (0.01, 0.02, 0.02), and (0.03 0.03 0.03) when starting the search with different starting points.

We now proceed to the actual characterization of the sequence of  $\epsilon$ -Nash equilibria. We first compute the *pure price*  $p^* = (73.54, 73.54, 77.18)$ . We fix  $T > 0$  and show that the sequence  $(p^*, \theta^\Lambda)$  with  $p^*$  as above and  $\theta_i^\Lambda = \bar{\theta} - \frac{T}{\sqrt{\Lambda}}$  is an  $\epsilon$ -Nash equilibria for all  $\Lambda$  large enough. We do this by evaluating the maximum difference in scaled profit that can be gained by a firm unilaterally deviating from the equilibrium in which all firms set their prices and service levels according to  $(p^*, \theta^\Lambda)$ . In Figure 1 we depict these differences for different values of  $\Lambda$ . The upper graph depicts the differences for  $T = 0.1$ , while the bottom graph depicts these differences for  $T = 0.5$ . For the middle graph we increased the sensitivity of the firms' demand to service levels a thousandfold and ran for different values of  $\Lambda$  with  $T = 0.1$ . Note that regardless of the  $T$  level and the service level elasticity, as the systems become larger, unilateral deviations become less and less beneficial, so that  $(p^*, \theta^\Lambda)$  is indeed an  $\epsilon$ -Nash equilibrium for large enough  $\Lambda$ . We can observe that if the demand sensitivity to service level is high, deviations from high level of service are discouraged at a very rapid rate.

Figure 1: The benefit of deviation from the proposed  $\epsilon$ -Nash equilibria



### 4.3 Non-Simultaneous Competition

In the previous sections, we considered simultaneous competition on both service level and price. In some service systems, however, capacity is not easily adjusted and it is often set in advance. It is of interest, then, to examine what happens when the competition is non-simultaneous, i.e., when firms make their decision sequentially. When the firms first choose their service level, the resulting competition is referred to as *service-level first* (SF) competition, whereas it is referred to as *price first* (PF) competition if they choose their prices first. Allon and Federgruen (2007) show that while the price-first and simultaneous competition models share the same set of equilibria, the service-level-first mode of competition can result in higher prices, higher service levels and higher demand volumes for all firms.

We show, however, that in a market consisting of large service providers, the order of competition has a diminishing significance as the market scale is larger. First, we need to define the notion of  $\epsilon$ -Nash equilibria for the non-simultaneous game. Towards that end, we say for the two-stage SF

game that a point  $(p, \theta)$  is an  $\epsilon$ -Nash equilibrium if (i) given a price vector  $p$ , the service-level vector  $\theta$  is an  $\epsilon$ -Nash equilibrium for the residual game on the strategy space  $\Theta$  and (ii) the price  $p$  is an  $\epsilon$ -Nash equilibrium for the first stage price game. This notion is defined with the corresponding modifications for the PF game. We then have the following result:

**Theorem 4.9 (non-simultaneous competition)** *The results of Theorems 4.1 and 4.2 continue to hold when the simultaneous competition is replaced with the SF (PF) competition.*

Intuitively, the result that the order of competition has a diminishing influence is a direct consequence of the previous section. There, we established that large firms compete only on price and align themselves to some industry standard in terms of the service level – an alignment that has negligible influence on their costs. The diminishing significance of the order of the competition is then clear. Indeed, since the firms will, in any case, compete only on the price, the order in which they determine the service level and price should be of no practical importance.

Before proceeding to the analysis of mixed markets, we devote the next section to the connection between the many-server setting we are considering in this paper and the single-server setting that has been considered in the competition-in-service-industries literature. We show that our results hold under this type of service provision as well, generalizing some previous results (when existence was already shown) and extending other (where examples were provided for non existence of Nash equilibrium).

#### 4.4 The single-server service-supply model

The current literature on competition in services firms typically assumes that service is provided through single-server facilities. In those cases, capacity is chosen by adjusting the service rate. We, on the other hand, chose to consider the case where the service rates are fixed and capacity is adjusted through the choice of the number of servers. These two settings are essentially distinct. While the multiple-server framework seems to be more realistic for most settings with human service representatives, the single-server assumptions have several advantages. First, the single-server setting seems to be more adequate for settings in which capacity is fluid. More importantly, though, the single-server setting is tractable in terms of existence and uniqueness of Nash equilibria due to the concavity of the cost function; see, for example, Allon and Federgruen (2007). In contrast, in the multiple-server model, the mere existence of a Nash equilibrium is questionable.

The aim of this section is, therefore, to connect these two distinct frameworks with respect to the equilibria results. Specifically, we show that our results for the multi-server setting are easily

applied to the single-server setting. In this respect, then, our multi-server setting is much more general than the single-server setting.

In order to present the results for the single-server case, consider the same set  $\mathcal{I} = \{1, \dots, I\}$  of service firms, each now providing service through an  $M/M/1$  facility. For  $M/M/1$  queues we will consider the sojourn time rather than the waiting time in queue. An alternative is to stay with the waiting time in queue but use a high-load approximation (see for example Allon and Federgruen (2006)). Hence, we redefine  $W_i$  to be the steady-state sojourn time for firm  $i$ . For a given service rate  $\mu_i$  and a given demand volume  $\lambda_i$ , it is well known that  $\mathbb{P}(W_i > T_i) = e^{-(\mu_i - \lambda_i)T_i}$ , from which we can deduce that the service-based capacity for a target time  $T_i$  and satisfaction probability  $\phi$  is given by  $\mu_i - \lambda_i = \frac{\ln(\frac{1}{\phi})}{T_i}$ . Defining, as before, the service level  $\theta_i$  to be the difference between a *benchmark* upper bound  $\bar{\theta}$  and the actual waiting time target  $T_i$ , i.e.,  $\theta_i = \bar{\theta} - T_i$ , we determine that the required capacity (service-rate) is defined by the equation

$$\mu_i = \lambda_i + \frac{\ln(\frac{1}{\phi})}{\bar{\theta} - \theta},$$

where  $\theta \in [0, \bar{\theta}]$ . The cost per served customer is  $c_i$  and the cost of capacity is proportional to the service-rate and given by  $\gamma_i \mu_i$ . We repeat the market replication procedure that we used before, using the market scaler  $\Lambda$  and considering a sequence of markets; see §3. The corresponding sequence of profit functions is obtained from (9) by calculating the service-based capacity using the explicit expressions for the  $M/M/1$  queue. We then have

$$\Pi_i^\Lambda(p, \theta) = \Lambda_i \cdot \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i \frac{\ln(\frac{1}{\phi})}{\bar{\theta} - \theta}, \quad i \in \mathcal{I}, \quad (24)$$

and the normalized version is defined similarly to (10). We then re-define the normalized and un-normalized games with respect to these profit functions. We refer to the market with  $M/M/1$  providers and the linear demand model given in (18) as *the single-server linear-demand market*.

Theorem (4.5) and Remark 4.7 allow us to deduce that if equilibrium exists for each  $\Lambda$ , it will place the system in the QED regime. In contrast to the multi-server setting, in the single-server linear-demand setting we can guarantee this existence. This allows to focus on the game with un-normalized profits. Specifically, we have the following result, the first part of which is itself a non-direct corollary of Theorem 3 in Allon and Federgruen (2007). Note that we have removed the assumption on the magnitude of  $\bar{\theta}$  that is required in Allon and Federgruen (2007) (there  $\bar{\theta}$  is denoted by  $\bar{w}$ ).

**Theorem 4.10 (existence and second-order characterization of equilibria)** *For all  $\Lambda$  large enough there exists a Nash equilibria  $(p^\Lambda, \theta^\Lambda)$  in the  $\Lambda$ -game with un-normalized profits. The*

equilibria  $(p^\Lambda, \theta^\Lambda)$  satisfy the following system of equations for all  $i \in \mathcal{I}$ :

$$\frac{\partial \Pi_i^\Lambda}{\partial p_i^\Lambda} = -\Lambda b_i(p_i^\Lambda - c_i - \gamma_i) + \Lambda_i = 0, \quad (25)$$

$$\theta_i^\Lambda(p_i^\Lambda) = \begin{cases} \text{the unique root of } \Lambda a'_i(\theta_i^\Lambda)(p_i^\Lambda - c_i - \gamma_i) = \frac{\gamma_i}{(\bar{\theta} - \theta_i^\Lambda)^2} & , \text{ if } p_i^\Lambda \geq c_i + \gamma_i \left(1 + \frac{1}{\bar{\theta}^2 \Lambda a'_i(0)}\right) \\ 0 & \text{otherwise .} \end{cases} \quad (26)$$

Moreover,

$$p_i^\Lambda = p_i^* \pm O\left(\frac{1}{\sqrt{\Lambda}}\right), \quad i \in \mathcal{I}, \quad (27)$$

and

$$\theta_i^\Lambda = \bar{\theta} - O\left(\frac{1}{\sqrt{\Lambda}}\right), \quad i \in \mathcal{I}. \quad (28)$$

We emphasize that result above is stated for the game with un-normalized profits. Of course, the normalized and un-normalized games are equivalent give the existence of equilibria as the difference is only the multiplication of the profit functions by a constant.

To summarize this section, we have shown that the results that obtained in the previous section for the multi-server settings are easily transferred to the single-server setting with adjustable service rate. We will use this extension in the next section where we consider mixed markets.

To prepare the ground for the next section, we recall that the results throughout this section apply to a setting with the market consisting of only substantially large firms. We know that in a market that consists of smaller firms, but still all of the same order of magnitude, not only is the competition on service levels significant, but also the order of competition has substantial importance. The question we handle in the next section is one of reconciling the two results by examining a mixed market with both large and small companies. We show that such a market exhibits simultaneously characteristics identified in Allon and Federgruen (2007) and those identified so far in the current paper. Moreover, we show that such a market exhibits some sort of one-sided decoupling.

## 5. A Mixed Market With Proportional Influence

We now turn our attention to competition in mixed markets, consisting of both large and small firms, and characterize its outcomes. While the size of a firm is clearly a function of its pricing and service-level strategies, other attributes, such as branding and location, create a natural distinction

between large and small firms. Indeed, in most demand models, firms, unless completely symmetric, experience different demand volumes even when employing the same policies. Accordingly, we denote firms as large or small based on their demand volume under identical prices and service levels. The current section is devoted, then, to the impact of these natural size differences on the competition outcome.

In the context of mixed markets, we address two main questions: (a) Do the large firms continue, as before, to ignore the service level as a competitive attribute? (b) Will the small firms, unlike the large ones, compete on the service levels to improve their revenues? Our analysis in this section answers both questions in the affirmative. The formal statement of this result is given in Theorem 5.1 in the end of this section followed by numerical results that illustrate the strength of this result.

We start by properly constructing a sequence of replicated mixed markets. In replicating the initial market, we choose to allow the number of small firms to grow with the market size. Consequently we will simultaneously replicate the market and duplicate some firms. Alternatively, one might consider keeping the number of small and large firms fixed as the market size grows. This, however, would trivialize the result as the aggregate volume of all the small firms combined would then be negligible and, consequently, their impact on the market outcome would be negligible.

Our replication and duplication construction will correspond to a market with **proportional influence**. Each firm has an influence on the firms that is proportional to its maximum potential. This is very natural as in most markets a firm that serves only a tiny proportion of the customers will have negligible influence on the competitors' decision. The aggregate of all the small firms will have, however, an influence on the large firms.

In this section we restrict our attention to the single-server linear-demand model from §4.4. We do this for concreteness and to avoid existence-of-equilibrium issues that are only secondary to the main point we want to emphasize in this section, which is the characterization of the market behavior, rather than the existence itself. The result will hold, however, in greater generality as long as the existence of a unique equilibrium can be established. Note that when using the terms *small* and *large* we merely mean that when all the firms position themselves with the same price levels and the same service level, some will experience higher demand than others.

The first element of the sequence of markets is a *basic market* from which we will create the sequence of markets through a process of duplication. The basic market consists of a set  $\mathcal{L} \subset \mathcal{I}$  of large firms and a set  $\mathcal{S} = \mathcal{I} \setminus \mathcal{L}$  of small firms. Here the small and big firms are not really distinguishable, but we use this name as they will become distinguishable through the duplication

procedure when we construct the sequence of markets. The general form of the demand function is given by equation (18).

To avoid problems of definition when we construct the sequence of markets, we assume that there are at least two firms of each type. Formally, we assume that for every  $i \in \mathcal{S}$  there exists  $j \in \mathcal{S}$  that shares the same parameters  $a_i(\cdot)$ ,  $b_i$  as well as  $a_{ik}(\cdot)$  and  $\beta_{ik}$  for  $k \neq i, j$  and such that  $a_{ij}(\cdot) = a_{ji}(\cdot)$ . For such  $i$  and  $j$  we define  $a_{ii}(\cdot) := a_{ij}(\cdot)$  and  $a_{jj}(\cdot) := a_{ji}(\cdot)$ .

To measure service-level differentiation for a given vector  $\theta$ , we define

$$\Delta(\theta) = \frac{\max_{i \in \mathcal{I}} \theta_i - \min_{i \in \mathcal{I}} \theta_i}{\min_{i \in \mathcal{I}} \theta_i},$$

and we say that a market admits service-level differentiation in an equilibrium  $(\theta, p)$  if  $\Delta(\theta) > 0$ . We allow  $\Delta(\theta)$  to obtain the value  $\infty$  if a firm uses the minimal service level 0. We define similarly the price-differential  $\Delta(p)$ . To guarantee the existence of equilibrium, we will assume that

$$\bar{\theta} \leq \sqrt[3]{\frac{4b_k \gamma_k}{(a'_k(0))^2}} \quad (29)$$

for all  $k \in \mathcal{S}$ . This is consistent with the assumption in Theorem 3 of Allon and Federgruen (2007)<sup>4</sup>. Also, we make a technical assumption about the equilibrium in the basic market:

**Assumption 5.1 (differentiation in the basic market)** *Fixing  $\theta_i = \bar{\theta}$  for all  $i \in \mathcal{L}$ , the  $I$  dimensional game with profit functions  $\bar{\Pi}_i(\theta, p) = \Pi_i(\theta, p)/M_i$  and strategy space  $[0, \bar{\theta}] \times \mathcal{P}_i$  for player  $i \in \mathcal{S}$  and  $\{\bar{\theta}\} \times \mathcal{P}_i$  for player  $i \in \mathcal{L}$  has at least one equilibrium, and for any equilibrium at least two firms exist  $i, k \in \mathcal{S}$  such that  $\theta_i \neq \theta_k$ .*

It is important to emphasize that Assumption 5.1 is merely a technical assumption on the equations that define the equilibrium (such as equations (25) and (26) in Corollary 4.10), and it will hold in great generality unless there are some pathologies in the underlying demand functions, such as completely identical demand functions for all small firms which can be rarely the case in practice.

The fact that in the *basic market*, whenever the firms in the set  $\mathcal{L}$  fix the service levels to  $\bar{\theta}$ , the firms in  $\mathcal{S}$  will differentiate themselves in terms of service levels, does not, a priori, imply anything about the nature of the Nash equilibria as the market size grows. On the contrary, one would actually expect that in mixed markets, as the small firms cannot use the high service levels that the large firms use, the large firms will not find it beneficial to invest in high service levels, like those we observed when dealing with a homogeneous market, in which all firms are large.

---

<sup>4</sup>For reasons that become clear in the proof of Corollary 4.10, we do not need to impose this for the large firms.

Moreover, the fact that the firms differentiate themselves in the basic market does not imply that this differentiation will not become negligible as the market size grows.

**The replication and duplication procedure:** We construct the  $\Lambda^{th}$  market by keeping the set of large firms fixed but creating  $\Lambda$  duplicates of each of the small firms. Implicitly this implies that  $\Lambda$  is taken to be an integer. Then in the  $\Lambda^{th}$  market we will have a set  $\mathcal{S}^\Lambda = \{L + 1, \dots, L + \Lambda S\}$  of small firms. We will say that a firm  $i$  is of type  $k$  whenever firm  $i \in \mathcal{I}^\Lambda$  is a duplicate of firm  $k$  in the basic market. We will then write  $c(i) = k$ . Clearly, as we duplicate only the small firms, we have that  $c(i) = i$  for  $i \in \mathcal{L}$ . Given the firm-types and the functions  $a_i(\cdot)$ ,  $a_{ij}(\cdot)$  and the constant  $b_i$  and  $\beta_{ij}$  for the basic market, we construct the demand for a large firm  $i \in \mathcal{L}$  as follows:

$$\Lambda_i(p, \theta) = \Lambda \left[ a_i(\theta_i) - b_i p_i - \sum_{j \neq i, j \in \mathcal{L}} \alpha_{ij}(\theta_j) - \sum_{j \neq i, j \in \mathcal{S}^\Lambda} \frac{a_{ij}(\theta_j)}{\Lambda} + \sum_{j \neq i, j \in \mathcal{L}} \beta_{ij} p_j + \sum_{j \neq i, j \in \mathcal{S}^\Lambda} \frac{\beta_{ij} p_j}{\Lambda} \right]^+, \quad (30)$$

and for a small firm  $i \in \mathcal{S}^\Lambda$  as follows:

$$\Lambda_i(p, \theta) = \left[ a_i(\theta_i) - b_i p_i - \sum_{j \neq i, j \in \mathcal{L}} \alpha_{ij}(\theta_j) - \sum_{j \neq i, j \in \mathcal{S}^\Lambda} \frac{a_{ij}(\theta_j)}{\Lambda} + \sum_{j \neq i, j \in \mathcal{L}} \beta_{ij} p_j + \sum_{j \neq i, j \in \mathcal{S}^\Lambda} \frac{\beta_{ij} p_j}{\Lambda} \right]^+. \quad (31)$$

Note that we slightly abuse notation above. The reader should have in mind that  $a_i(\theta_i)$  is actually equal to  $a_{c(i)}(\theta_i)$  which is defined through the parameters in the basic market. Similar interpretation applies to the functions  $a_{ij}(\cdot)$  and the constant  $b_i$  and  $\beta_{ij}$ . We illustrate the construction of the demand function in Example 5.1. We point out that this specific construction of the demand induces a notion of proportional influence. The relative demand of a given small firm is, by construction, of order  $1/\Lambda$  and so is its influence on the demand faced by another firms, through the cross terms  $a_{ij}(\cdot)/\Lambda$  and  $\beta_{ij}/\Lambda$ . As the number of small firms is, however, of order  $\Lambda$ , their aggregate influence on the large firms is not negligible.

All firms operate through single server facilities, and we fix The price interval for firm  $i \in \mathcal{I}^\Lambda$  equals  $\mathcal{P}_{c(i)}$ , and the service-level interval is  $[0, \bar{\theta}]$  as before. Accordingly we define  $\times^\Lambda = [0, \bar{\theta}]^{I^\Lambda}$  and  $\mathcal{P}^\Lambda = \times_{i=1}^{I^\Lambda} [p_{c(i)}^{min}, p_{c(i)}^{max}]$ .

In passing, observe that if Nash equilibrium exists then all small firms of the same type will necessarily make the same choice of service level and price. Consequently, the real potential market size for firm  $i$  is given by

$$M_i^\Lambda = \max_{\theta \in \bar{\Theta}, p \in \bar{\mathcal{P}}} \Lambda_i(p, \theta) \quad (32)$$

where  $\tilde{\mathcal{P}} = \{p \in \mathcal{P} : p_i = p_j \text{ whenever } c(i) = c(j)\}$  and  $\tilde{\Theta} := \{\theta \in \Theta : \theta_i = p_\theta \text{ whenever } c(i) = c(j)\}$ . Accordingly, we redefine  $\bar{\Pi}_i^\Lambda$  as before with  $M_i^\Lambda$  now defined according to (32).

The  $\Lambda^{th}$  simultaneous mixed-market game then is defined as the game with  $I^\Lambda$  players, strategy space  $\Theta^\Lambda \times \mathcal{P}^\Lambda$  and payoff function  $\bar{\Pi}_i^\Lambda$ .

Finally, for a vector of service levels and prices  $(\theta^\Lambda, p^\Lambda) \in \Theta \times \mathcal{P}$  we let  $(p_{\mathcal{L}}^\Lambda, \theta_{\mathcal{L}}^\Lambda)$  and  $(p_{\mathcal{S}^\Lambda}^\Lambda, \theta_{\mathcal{S}^\Lambda}^\Lambda)$  be, respectively, the sub-vectors corresponding to the large and small firms. The following is, then, the main result of this section.

**Theorem 5.1 (simultaneous competition in the mixed market)** *Consider the sequence of simultaneous mixed-market games. Then there exist a sequence of Nash equilibria  $(\theta_{\mathcal{L}}^\Lambda, \theta_{\mathcal{S}^\Lambda}^\Lambda, p_{\mathcal{L}}^\Lambda, p_{\mathcal{S}^\Lambda}^\Lambda)$  and it satisfies that*

$$\liminf_{\Lambda \rightarrow \infty} \Delta(\theta_{\mathcal{S}^\Lambda}^\Lambda) > 0$$

but

$$\lim_{\Lambda \rightarrow \infty} \Delta(\theta_{\mathcal{L}}^\Lambda) = 0.$$

We observe, then, that the firms of large scale will not differentiate themselves on the basis of the service level experienced by their customers. The small firms, however, which do not enjoy the advantages of economies of scale, will offer different service levels in the market. The result is illustrated in the following example.

**Example 5.1** In order to demonstrate the impact of the size of the service level differentiation of small and large firms, we conducted the following numerical study. Consider an “initial” industry of four firms: firms 1 and 2 are large (and will be scaled), while firms 3 and 4 are small. Note that initially the differences between the alleged small and large firms are marginal. We set the cost parameters  $c_1 = c_2 = 20, c_3 = c_4 = 5$ , while  $\gamma_1 = \gamma_2 = 20, \gamma_3 = \gamma_4 = 35$ . The demand for the services of the four firms follows the following system of equations:

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} = \begin{pmatrix} 255 \\ 275 \\ 175 \\ 185 \end{pmatrix} + \begin{pmatrix} -21 & 3 & 5 & 2 \\ 2 & -23 & 4 & 1 \\ 4 & 4 & -19 & 4 \\ 5 & 2 & 3 & -15 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} - \begin{pmatrix} -100 & 10 & 15 & 20 \\ 15 & -110 & 15 & 20 \\ 20 & 20 & -120 & 20 \\ 10 & 15 & 20 & -150 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix}$$

To illustrate the duplication and replication procedure, note that the above demand specification corresponds to firm 3, for example, having a linear demand function with  $a_3(\theta_3) = 175 + 120\theta_3$ ,  $a_{3j}(\theta_j) = 20\theta_j$ , for  $j = 1, 2, 4$ ,  $b_3 = 19$ , and  $\beta_{3j} = 4$  for all  $j = 1, 2, 4$ . In the  $\Lambda^{th}$  market we will

have  $2\Lambda$  small firms with the demand function for a small firm  $i$  that is a duplicate of firm 3 being given by

$$(175 + 120\theta_3 - 19p_i) + \left( 4p_1 + 4p_2 + \sum_{j \neq i: c(j)=3} \frac{4}{\Lambda} p_j + \sum_{j \neq i: c(j)=4} \frac{4}{\Lambda} p_j \right) - \left( 20\theta_1 + 20\theta_2 \sum_{j \neq i: c(j)=3} \frac{20}{\Lambda} + \sum_{j: c(j)=4} \frac{20}{\Lambda} \right),$$

where the first parenthesis corresponds to the impact of the firms own price and service level and the second and third parenthesis correspond to the cross-influences of prices and service-levels. Note that the number of firms with  $c(j) = 3$  and  $c(j) = 4$  is of the order of  $\Lambda$  so that these cross-influences are not negligible.

In Figure 2 we depict the service-level differentiation among the large and the small firms, for different levels of  $\Lambda$ . Note that we followed the duplication process described above. We measure the level of differentiation by the percentage change between the highest and lowest service level among large firms and small firms, separately. We can observe that, initially, the level of differentiation is of the same magnitude among the two types of firms. However, once we scale up the size of the large firms and increase the number of small firms, the levels of differentiation in both sub markets go in opposite directions. In particular, we observe that while the large firms provide practically an identical service level, the small firms differ in their service-level offering up to 58%. In Table 1 we present in parallel both the prices and service levels in our experiment. In this table  $\min(x)$  and  $\max(x)$  stands for the minimal and maximal value in the vector  $x$ . Table 2 depicts both the price and service-level differentials for both small and large firms. Finally, Figure 2 was generated from the service-differential column in Table 2. Note that as the service-level differentiation increases, so does the price differential among the small firms. While the large firms continue to improve their quality, some small firms maintain these high service levels as well (and to keep up with the large firms, the small firms have to increase their prices to compensate for the increase in costs). Some other small firms just reduce their quality level, admitting that they cannot compete with the large ones, and they can now reduce the prices as well (and be more attractive on the price side).

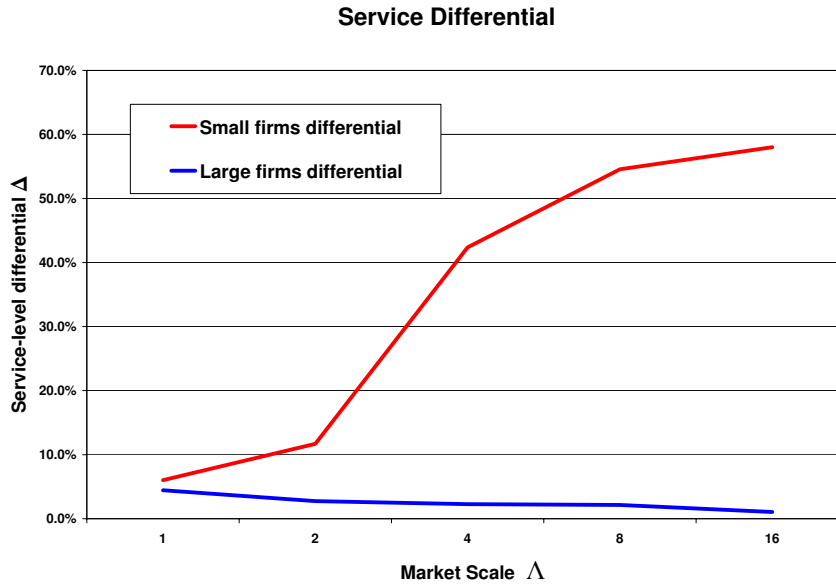
**Table 1: Mixed Markets**

$\Lambda$	$ \mathcal{L} $	$ S^\Lambda $	$\min(p_{\mathcal{L}}^\Lambda)$	$\max(p_{\mathcal{L}}^\Lambda)$	$\min(\theta_{\mathcal{L}}^\Lambda)$	$\max(\theta_{\mathcal{L}}^\Lambda)$	$\min(p_{S^\Lambda}^\Lambda)$	$\max(p_{S^\Lambda}^\Lambda)$	$\min(\theta_{S^\Lambda}^\Lambda)$	$\max(\theta_{S^\Lambda}^\Lambda)$
1	2	2	64.70	69.30	84.76	88.52	69.30	78.40	83.83	88.88
2	2	4	62.03	66.20	90.16	92.64	63.06	71.80	77.41	86.46
4	2	8	60.60	64.40	91.49	93.57	59.50	67.90	58.97	83.96
8	2	16	60.35	64.10	92.27	94.25	58.50	69.35	55.54	85.83
16	2	32	60.21	60.58	95.42	96.43	58.20	69.84	53.85	85.08

**Table 2: Mixed Markets**

$\Lambda$	$ \mathcal{L} $	$ S^\Lambda $	$\Delta(p_{\mathcal{L}}^\Lambda)$	$\Delta(\theta_{\mathcal{L}}^\Lambda)$	$\Delta(p_{S^\Lambda}^\Lambda)$	$\Delta(\theta_{S^\Lambda}^\Lambda)$
1	2	2	7.11%	4.44%	13.13%	6.02%
2	2	4	6.77%	2.75%	13.86%	11.69%
4	2	8	6.27%	2.27%	14.02%	42.37%
8	2	16	6.21%	2.15%	18.55%	54.55%
16	2	32	6.10%	1.06%	20.00%	58.00%

Figure 2: Service-level differentiation: Large Vs. Small firms



## 6. Conclusions and future research

In his seminal article, Porter (1996) states: “Differences in needs will not translate into meaningful positions unless the best set of activities to satisfy them also differs.” We show that regardless of

the cost structure of the firm, as long as all of the firms are of large scale, they do not have to make any tradeoffs and can support an extremely high service level. This indeed results in a market in which market positions are not distinguished based on the service level. This result, which is consistent with practices in many service industries, continues to hold in markets with different sizes of firms, but only for the large-scale firms.

In order to obtain these results, we develop a novel framework that combines the notions of  $\epsilon$ -Nash equilibrium, market replication and heavy-traffic to study market equilibria. The  $\epsilon$ -Nash framework allows us go beyond the limited scope of Nash equilibrium and use general demand and capacity models. The notion of market replication allows us to discuss trends both in terms of stability and market outcomes in sequences of markets. Combined with the notion of heavy-traffic, which is well studied for monopolists, this framework allows to characterize the equilibria behavior and obtain insights that are usually lost in traditional Nash equilibria analysis.

Future work should consider settings of competition incorporated with learning. In most industries, competing firms cannot fully observe the demand characteristics of the other firms. We would like to study the problem of jointly competing and learning, utilizing the framework developed in this paper. One can extend this work to study outsourcing models in which competing firms outsource to one or more common suppliers. Previous outsourcing models assume that the demand characterization of the firms in the supply chain is known. The supplier or the retailer uses this information to coordinate the chain, by offering contracts to the different member of the chain that are geared to maximize the system-wide profits. However, in practice, demand information is rarely observed by competitors or other stakeholder in the industry.

## Appendix: Proofs

**Proof of Theorem 4.1:** The proof draws on Definition 3.2 of  $\epsilon$ -Nash equilibria, Assumption 3.2 on the uniqueness of the equilibrium  $p^*$  for the pure price-competition model, and the properties of the demand functions as listed in Assumption 3.1.

We fix a sequence  $\theta^\Lambda$  that satisfies the following three properties:

$$\max_{i \in \mathcal{I}} \sup_{p \in \mathcal{P}} |\lambda_i(p^\Lambda, \theta^\Lambda) - \lambda_i(p^\Lambda, \vec{\theta})| \leq \epsilon^\Lambda/8, \quad (33)$$

$$\frac{\hat{e}_i(\lambda_i, \theta_i^\Lambda)}{\Lambda} \leq \epsilon^\Lambda/8, \quad \text{and} \quad (34)$$

$$\theta^\Lambda \rightarrow 0, \quad \text{as } \Lambda \rightarrow \infty. \quad (35)$$

Such a sequence exists by the absolute continuity of the demand functions on the compact domain and by Lemma 3.4. Note that for (34) we are using the assumption that  $\Lambda\epsilon^\Lambda \rightarrow \infty$  as  $\Lambda \rightarrow \infty$ .

To show that  $(p^*, \theta^\Lambda)$  is a  $\epsilon^\Lambda$ -Nash equilibria, fix a firm  $i$  and  $(p_i'^\Lambda, \theta_i'^\Lambda) \in \mathcal{P}_i \times \Theta$  of prices and service levels for firm  $i$  such that  $(p_i'^\Lambda, \theta_i'^\Lambda) \neq (p_i^*, \theta_i^\Lambda)$ . Define

$$(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) := (p_i'^\Lambda, \theta_i'^\Lambda) \uparrow (p^*, \theta^\Lambda)_{-i}.$$

As  $e_i(\cdot, \cdot) \geq 0$ , we have that

$$\bar{\Pi}_i^\Lambda(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) \leq \frac{\lambda_i(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) \left( \tilde{p}_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right)}{M_i}.$$

By the choice of  $\theta^\Lambda$ , we have that

$$\left[ \lambda_i(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) \left( \tilde{p}_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) - \lambda_i(\tilde{p}^\Lambda, \bar{\theta}) \left( \tilde{p}_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) \right] \leq \epsilon^\Lambda/4. \quad (36)$$

Indeed, one writes  $\lambda_i(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) - \lambda_i(\tilde{p}^\Lambda, \bar{\theta}) = \lambda_i(\tilde{p}^\Lambda, \theta^\Lambda) - \lambda_i(\tilde{p}^\Lambda, \bar{\theta}) - \lambda_i(\tilde{p}^\Lambda, \theta^\Lambda) + \lambda_i(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda)$ . Then, by (33) we have that  $|\lambda_i(\tilde{p}^\Lambda, \theta^\Lambda) - \lambda_i(\tilde{p}^\Lambda, \bar{\theta})| \leq \epsilon^\Lambda/8$ . Now, there are two cases: if  $\theta_i'^\Lambda \leq \theta_i^\Lambda$  then we can apply (33) once again with  $\theta^\Lambda$  replaced with  $\tilde{\theta}^\Lambda$ . If, on the other hand,  $\theta_i'^\Lambda > \theta_i^\Lambda$ , then the monotonicity of the demand functions is invoked to get that  $\lambda_i(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) - \lambda_i(\tilde{p}^\Lambda, \theta^\Lambda) \leq 0$ .

Note that (36) is independent of the actual values of the sequence  $(p_i'^\Lambda, \theta_i'^\Lambda)$  and depends only on the values of  $(p^\Lambda, \theta^\Lambda)$ . By (33) we have that

$$\left| \lambda_i(p^*, \theta^\Lambda) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right) - \lambda_i(p^*, \bar{\theta}) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right) \right| \leq \epsilon^\Lambda/4. \quad (37)$$

By the definition of  $p^*$  as an equilibrium for the pure price-competition model we have that

$$\lambda_i(\tilde{p}^\Lambda, \bar{\theta}) \left( \tilde{p}_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) \leq \lambda_i(p^*, \bar{\theta}) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right). \quad (38)$$

Combining (36), (37) and (38) we readily have that,

$$\lambda_i(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) \left( \tilde{p}_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) \leq \lambda_i(p^*, \bar{\theta}) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right) + \frac{\epsilon^\Lambda}{2}.$$

Finally, using (34) we have that

$$\bar{\Pi}_i^\Lambda(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) \leq \frac{\lambda_i(p^*, \theta^\Lambda) \left( p_i^* - c_i - \frac{\gamma_i}{\mu_i} \right)}{M_i} + \epsilon^\Lambda.$$

We conclude that for each  $\Lambda$ ,  $(p^*, \bar{\theta}^\Lambda)$  is an  $\epsilon^\Lambda$ -Nash equilibrium. ■

**Proof of Theorem 4.2:** We divide the proof in two parts. First we prove the characterization for the equilibrium-service-level characterization in equation (16) and then proceed to prove the equilibrium-price characterization in equation (17).

**Proof of (16):** To reach a contradiction, assume there is no such sequence  $\delta^\Lambda$  for  $\theta^\Lambda$ . In particular, there exists  $i$  such that  $\limsup_{\Lambda \rightarrow \infty} \bar{\theta} - \theta_i^\Lambda \geq \delta$ , for some  $\delta > 0$ . Consequently, we may choose a subsequence  $\Lambda^j$  such that  $\limsup_{j \rightarrow \infty} \bar{\theta} - \theta_i^{\Lambda^j} \geq \delta$ .

Define,  $\bar{\theta}^\Lambda$  by setting  $\bar{\theta}_i^\Lambda = \bar{\theta} - \bar{T}/\sqrt{\Lambda}$  for this firm  $i$  and some  $\bar{T} > 0$  and  $\bar{\theta}_k^\Lambda = \theta_k^\Lambda$  for all  $k \neq i$ . Then we can re-choose  $j$  large enough so that  $\bar{\theta}_i^{\Lambda^j} \leq \tilde{\theta}_i^\Lambda - \eta$ , for some  $\eta > 0$  and since, by assumption,  $\lambda_i(p, \theta)$  is strictly increasing in  $\theta_i$ , we have that there exists  $\epsilon > 0$ , such that

$$\lambda_i(\tilde{p}^{\Lambda^j}, \bar{\theta}^{\Lambda^j}) \cdot \left( \tilde{p}_i^{\Lambda^j} - c_i - \frac{\gamma_i}{\mu_i} \right) - \lambda_i(\tilde{p}^{\Lambda^j}, \tilde{\theta}^{\Lambda^j}) \cdot \left( \tilde{p}_i^{\Lambda^j} - c_i - \frac{\gamma_i}{\mu_i} \right) \geq 4\epsilon.$$

Consequently, using the definition of the profit functions we have that

$$\begin{aligned} \Pi_i^{\Lambda^j}(\tilde{p}^{\Lambda^j}, \tilde{\theta}^{\Lambda^j}) - \Pi_i^{\Lambda^j}(\tilde{p}^{\Lambda^j}, \bar{\theta}^{\Lambda^j}) &\leq \Lambda^j \lambda_i(\tilde{p}^{\Lambda^j}, \tilde{\theta}^{\Lambda^j}) \left( p_i - c_i - \frac{\gamma_i}{\mu_i} \right) \\ &\quad - \left( \Lambda^j \lambda_i(\tilde{p}^{\Lambda^j}, \bar{\theta}^{\Lambda^j}) \cdot \left( \tilde{p}_i^{\Lambda^j} - c_i - \frac{\gamma_i}{\mu_i} \right) - \gamma_i e_i(\Lambda_i^j, \bar{T}/\sqrt{\Lambda}) \right) \\ &\quad - 4\epsilon \Lambda + \gamma_i e_i(\Lambda_i^j, \bar{T}/\sqrt{\Lambda}). \end{aligned}$$

Since by Lemma 3.4, for all  $\Lambda$  large enough,  $e_i(\Lambda_i, \bar{T}/\sqrt{\Lambda}) \leq K\sqrt{\Lambda}$  for some  $K > 0$ , we can re-choose  $j$ , so that

$$\Pi_i^{\Lambda^j}(\tilde{p}^{\Lambda^j}, \tilde{\theta}^{\Lambda^j}) - \Pi_i^{\Lambda^j}(\tilde{p}^{\Lambda^j}, \bar{\theta}^{\Lambda^j}) \leq -2\epsilon \Lambda^j.$$

Firm  $i$  can, then, improve its scaled profit,  $\bar{\Pi}_i^{\Lambda^j}$ , by more than  $\epsilon$ . Since  $\epsilon^\Lambda \rightarrow 0$ , there exists  $j_0$  such that  $\epsilon^{\Lambda^j} \leq \epsilon$  for all  $j \geq j_0$ . Consequently, for  $j$  large enough,  $(\tilde{p}^{\Lambda^j}, \tilde{\theta}^{\Lambda^j})$  can not be an  $\epsilon^\Lambda$ -Nash. Equation (16) is thus proved and we move to the price characterization.

**Proof of (17):** Fix the sequence  $(p^\Lambda, \theta^\Lambda)$  of  $\epsilon^\Lambda$ -Nash equilibria. To reach a contradiction assume that  $\limsup_{\Lambda \rightarrow \infty} \|p^\Lambda - p^*\| > 0$ . We then say that  $p^\Lambda$  is *asymptotically distinguishable from  $p^*$* . First, note that if  $\max_{i \in \mathcal{I}} \limsup_{\Lambda \rightarrow \infty} \bar{\theta} - \theta_i^\Lambda > 0$ , the result of the Theorem trivially follows from (16). Hence, we assume  $\bar{\theta} - \theta_i^\Lambda \rightarrow 0$  as  $\Lambda \rightarrow \infty$  for all  $i \in \mathcal{I}$ . We will show that under the assumption that  $p^\Lambda$  is distinguishable from  $p^*$ , every limit point  $p$  of  $p^\Lambda$  must be an equilibrium point for pure price-competition model. Such a limit point exists by the compactness of  $\times_{i=1}^I [p_i^{\min}, p_i^{\max}]$ . Since  $p^\Lambda$  is distinguishable from  $p^*$ , this will imply the existence of multiple equilibria for the pure price-competition model, contradicting Assumption 3.2. We proceed then to show that every limit point  $p$  is indeed an equilibrium point for the pure price-competition model. Towards that end, fix a limit point  $p$  of  $p^\Lambda$  and the corresponding convergent subsequence  $\Lambda^k$ ,  $k \geq 0$ . We will now show that for each  $\epsilon > 0$ ,  $p$  is an  $\epsilon$ -Nash equilibrium for the pure price-competition model and, in turn, a Nash equilibrium. Define  $\bar{p} := (\bar{p}_i, p_{-i})$ , for some price  $\bar{p}_i \in [p_i^{\min}, p_i^{\max}]$  with  $\bar{p}_i \neq p_i$ . Then,

since  $(p^\Lambda, \theta^\Lambda)$  is the assumed sequence of  $\epsilon^\Lambda$ -Nash equilibria, we have that for all  $k$  large enough,

$$\bar{\Pi}_i^{\Lambda^k}(\bar{p}^{\Lambda^k}, \theta^{\Lambda^k}) \leq \bar{\Pi}_i^{\Lambda^k}(p^{\Lambda^k}, \theta^{\Lambda^k}) + \epsilon/4,$$

for some  $\epsilon > 0$ . Observe that by Lemma 3.4,  $e_i(\Lambda_i, \theta_i^\Lambda)/\Lambda \rightarrow 0$  as  $\Lambda \rightarrow \infty$ . This, together with the continuity of the demand functions, implies that

$$\lim_{\Lambda \rightarrow \infty} \sum_{i \in \mathcal{I}} |\bar{\Pi}_i^\Lambda(p^\Lambda, \theta^\Lambda) - \bar{\Pi}_i^P(p)| = 0.$$

In particular,

$$\lim_{k \rightarrow \infty} \sum_{i \in \mathcal{I}} \left| \bar{\Pi}_i^{\Lambda^k}(p^{\Lambda^k}, \theta^{\Lambda^k}) - \bar{\Pi}_i^P(p) \right| = 0.$$

Hence,

$$\bar{\Pi}_i^{\Lambda^k}(\bar{p}^{\Lambda^k}, \theta^{\Lambda^k}) \rightarrow \bar{\Pi}_i^P(\bar{p}) \text{ and } \bar{\Pi}_i^{\Lambda^k}(\bar{p}^{\Lambda^k}, \theta^{\Lambda^k}) \rightarrow \bar{\Pi}_i^P(p), \text{ as } k \rightarrow \infty$$

where  $\bar{p} = (\bar{p}_i, p_{-i})$ , and we have that

$$\bar{\Pi}_i^P(\bar{p}) \leq \bar{\Pi}_i^P(p) + \epsilon.$$

In particular,  $p$  is an  $\epsilon$ -Nash equilibrium for the pure price-competition model. Since  $\epsilon$  was arbitrary, we have that  $p$  is a Nash equilibrium of the pure price-competition game. Since  $p \neq p^*$ , we have reached the desired contradiction to the uniqueness of  $p^*$ . This proves (17) and completes the proof of Theorem 4.2.  $\blacksquare$

**Proof of Theorem 4.5:** Fix a sequence  $(p^\Lambda, \theta^\Lambda)$  that satisfies the conditions of the theorem. We start by establishing (23). Assume, to reach a contradiction, that there exists a firm  $i$ , with  $\limsup_{\Lambda \rightarrow \infty} \sqrt{\Lambda} |\bar{\theta} - \theta_i^\Lambda| = \infty$ . By Taylor's expansion,

$$a_i(\theta_i^\Lambda) = a_i(\bar{\theta}) - a_i'(\bar{\theta})(\bar{\theta} - \theta_i^\Lambda) + o(\bar{\theta} - \theta_i^\Lambda),$$

and it is straightforward to show that  $\limsup_{\Lambda \rightarrow \infty} \sqrt{\Lambda} |\bar{\theta} - \theta_i^\Lambda| = \infty$  implies

$$\liminf_{\Lambda \rightarrow \infty} \frac{\Lambda_i(p^\Lambda, \theta^\Lambda) \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right) - \Lambda_i(p^\Lambda, \tilde{\theta}^\Lambda) \left( p_i^\Lambda - c_i - \frac{\gamma_i}{\mu_i} \right)}{\sqrt{\Lambda}} = \infty \quad (39)$$

where  $\tilde{\theta}^\Lambda$  is obtained from  $\theta^\Lambda$  by setting  $\tilde{\theta}_i^\Lambda = \bar{\theta} - c/\sqrt{\Lambda} + o(1/\sqrt{\Lambda})$ , and setting  $\tilde{\theta}_k^\Lambda = \theta_k^\Lambda$  for all  $k \neq i$ . By Lemma 3.4,  $e_i(\Lambda_i, T_i/\sqrt{\Lambda}) = O(\sqrt{\Lambda})$ , implying together with (39) that

$$\liminf_{\Lambda \rightarrow \infty} \sqrt{\Lambda} \left[ \bar{\Pi}_i(p^\Lambda, \theta^\Lambda) - \bar{\Pi}_i(p^\Lambda, \tilde{\theta}^\Lambda) \right] = \infty, \quad (40)$$

and in particular that for all  $\Lambda$  large enough  $(p^\Lambda, \theta^\Lambda)$  cannot be an  $\epsilon^\Lambda$ -Nash for any choice of the sequence  $\epsilon^\Lambda$ -Nash equilibrium such that  $\limsup_{\Lambda \rightarrow \infty} \sqrt{\Lambda} \epsilon^\Lambda < \infty$ . Consequently, we must have that  $\theta^\Lambda = \bar{\theta} - O(1/\sqrt{\Lambda})$ .

Having established (23) we turn now to prove (22). We may now assume, without loss of generality, that  $\theta_i^\Lambda = \bar{\theta} - O(1/\sqrt{\Lambda})$ , for all  $i \in \mathcal{I}$ . We define

$$\tilde{\Theta}^\mathcal{N} = \left\{ \{\theta^\Lambda\}_{\Lambda \geq 0} : \limsup_{\Lambda \rightarrow \infty} \sqrt{\Lambda} \|\theta^\Lambda - \bar{\theta}\| < \infty, \text{ as } \Lambda \rightarrow \infty \right\}.$$

Clearly,  $\tilde{\Theta}^\mathcal{N} \subset \mathcal{N}$ . The proof proceeds as follows: First, we show that, if  $(p^\Lambda, \theta^\Lambda)$  is a sequence of  $\epsilon^\Lambda$ -Nash equilibria, then  $p^\Lambda$  is an  $\tilde{\epsilon}^\Lambda$ -Nash equilibrium for the pure price-competition game, where  $\tilde{\epsilon}^\Lambda = \tilde{\epsilon}/\sqrt{\Lambda}$  for some fixed  $\tilde{\epsilon} > 0$ . Having shown this, we will use the pure price-competition model to establish equation (22).

First, you see that  $p^\Lambda$  is indeed the claimed  $\tilde{\epsilon}^\Lambda$  equilibrium for the pure price-competition game. Note that, since  $\theta^\Lambda \in \tilde{\Theta}^\mathcal{N}$ , we can use a Taylor expansion as above for both  $a_i(\cdot)$  and  $a_{ij}(\cdot)$ , to show that there exists  $\Lambda_0$  and  $K > 0$  so that

$$\sqrt{\Lambda} \sum_{i \in \mathcal{I}} |\bar{\Pi}_i^P(p^\Lambda) - \bar{\Pi}_i(p^\Lambda, \theta^\Lambda)| \leq K,$$

for all  $\Lambda \geq \Lambda_0$ . Note that  $\Lambda_0$  and  $K$  depend only on  $\theta^\Lambda$  and not on the price sequence  $p^\Lambda$ . Hence, recalling that  $(p^\Lambda, \theta^\Lambda)$  is a sequence of  $\epsilon^\Lambda$ -Nash equilibria with  $\epsilon^\Lambda = O(1/\sqrt{\Lambda})$ , we have that

$$\bar{\Pi}_i^P(\tilde{p}^\Lambda) \leq \bar{\Pi}_i^\Lambda(\tilde{p}^\Lambda, \theta^\Lambda) + K/\sqrt{\Lambda} \leq \bar{\Pi}_i^\Lambda(p^\Lambda, \theta^\Lambda) + (K+c)/\sqrt{\Lambda} \leq \bar{\Pi}_i^P(p^\Lambda) + (2K+c)/\sqrt{\Lambda},$$

for all  $\Lambda \geq \Lambda_0$  and for some  $c > 0$ . Setting  $\tilde{\epsilon} = 2K+c$  we have that  $p^\Lambda$  is a sequence of  $\tilde{\epsilon}^\Lambda$ -Nash equilibrium for the pure price-competition model.

Consider then the pure price-competition model. Using the definition of  $\bar{\Pi}_i^P(\cdot)$  we have that for any price  $\tilde{p}_i$

$$\bar{\Pi}_i^P(\tilde{p}_i, p_i^\Lambda) - \bar{\Pi}_i(p_i^\Lambda) = \frac{1}{M_i} \left[ \left( a_i(\bar{\theta}) - \sum_{j \neq i} \alpha_{ij}(\bar{\theta}) + \sum_{j \neq i} \beta_{ij} p_j^\Lambda - b_i \left( c_i + \frac{\gamma_i}{\mu_i} \right) \right) (\tilde{p}_i^\Lambda - p_i^\Lambda) + b_i (\tilde{p}_i^2 - (p_i^\Lambda)^2) \right],$$

and after some basic algebraic manipulations we write the inequalities

$$\begin{aligned} \frac{1}{M_i} [(C_1 + 2b_i p_i^\Lambda)(\tilde{p}_i - p_i^\Lambda) + b_i(\tilde{p}_i - p_i^\Lambda)^2] &\geq \bar{\Pi}_i^P(\tilde{p}_i, p_i^\Lambda) - \bar{\Pi}_i(p_i^\Lambda) \\ &\geq \frac{1}{M_i} [(C_2 + 2b_i p_i^\Lambda)(\tilde{p}_i - p_i^\Lambda) + b_i(\tilde{p}_i - p_i^\Lambda)^2], \end{aligned}$$

where

$$C_1 = \max_{p \in \times_{i=1}^I [p_i^{\min}, p_i^{\max}]} \left( a_i(\bar{\theta}) - \sum_{j \neq i} \alpha_{ij}(\bar{\theta}) + \sum_{j \neq i} \beta_{ij} p_j - b_i \left( c_i + \frac{\gamma_i}{\mu_i} \right) \right),$$

and

$$C_2 = \min_{p \in \times_{i=1}^I [p_i^{\min}, p_i^{\max}]} \left( a_i(\bar{\theta}) - \sum_{j \neq i} \alpha_{ij}(\bar{\theta}) + \sum_{j \neq i} \beta_{ij} p_j - b_i \left( c_i + \frac{\gamma_i}{\mu_i} \right) \right)$$

and  $C_2 > 0$  is strictly positive by our assumption that  $\lambda_i(p, \theta) > 0$  for all  $p \in \times_{i=1}^I [p_i^{\min}, p_i^{\max}]$  and  $\theta \in \times_{i=1}^I [0, \bar{\theta}]$ . In particular,

$$|\bar{\Pi}_i^P(\tilde{p}_i, p_{-i}^\Lambda) - \bar{\Pi}_i(p^\Lambda)| \geq \frac{1}{M_i} [(C_1 + 2b_i p_i^{\min}) |\tilde{p}_i - p_i^\Lambda| + b_i (\tilde{p}_i - p_i^\Lambda)^2].$$

Consequently, we have that there exists a constant  $M_i'$  such that

$$|p_i^\Lambda - \tilde{p}_i| \leq \frac{M_i' \tilde{\epsilon}}{\sqrt{\Lambda}} \quad (41)$$

whenever  $|\bar{\Pi}_i^P(\tilde{p}_i, p_{-i}^\Lambda) - \bar{\Pi}_i(p^\Lambda)| \leq \tilde{\epsilon}/\sqrt{\Lambda}$ . Set  $\bar{p}_i$  to be a best response of player  $i$  to the vector  $p_{-i}^\Lambda$  of prices of the competitors. Setting  $\tilde{p}_i = \bar{p}_i$  and using the fact that  $p^\Lambda$  is a sequence of  $\tilde{\epsilon}^\Lambda$ -Nash for the pure price-competition game, we have that  $|\bar{\Pi}_i^P(\bar{p}_i, p_{-i}^\Lambda) - \bar{\Pi}_i(p^\Lambda)| \leq \tilde{\epsilon}/\sqrt{\Lambda}$ , and consequently that

$$|p_i^\Lambda - \bar{p}_i| \leq \frac{M_i' \tilde{\epsilon}}{\sqrt{\Lambda}}. \quad (42)$$

We will now use the specific structure of the linear demand model and equation (42) to show that  $\|p^\Lambda - p^*\| = O(1/\sqrt{\Lambda})$ . First, note that the best response is the unique solution to the first order conditions

$$\left. \frac{\partial}{\partial p_i} \Pi_i^P(p^\Lambda) \right|_{\bar{p}_i} = 0;$$

hence, we may write  $\bar{p}_i = \bar{p}_i(p_{-i}^\Lambda)$ . In particular, using the specific structure of the model we have that

$$\bar{p}_i = \frac{a_i(\bar{\theta}) - \sum_{j \neq i} \alpha_{ij}(\bar{\theta}) + \sum_{j \neq i} \beta_{ij} p_j^\Lambda}{2b_i}.$$

By (41) we then have that

$$\left| p_i^\Lambda - \frac{\tilde{a}(\bar{\theta}) + \sum_{j \neq i} \beta_{ij} p_j}{2b_i} \right| \leq \frac{M_i' \tilde{\epsilon}}{\sqrt{\Lambda}}, \quad \forall i \in \mathcal{I}.$$

Repeating the same argument for all  $i \in \mathcal{I}$  and writing the result in a matrix form, as in equation (12) of Federgruen Allon and Federgruen (2007), we have that

$$\|Ap^\Lambda - \tilde{a}(\bar{\theta}) + k\| \leq \frac{IM' \tilde{\epsilon}}{\sqrt{\Lambda}}. \quad (43)$$

By Theorem 1 in Allon and Federgruen (2007),  $p^*$  is the unique solution,  $p$ , to

$$Ap - \tilde{a}(\bar{\theta}) - k = 0. \quad (44)$$

For simplicity of notation we let  $M_i'' = IM$  and let  $\mathbf{e}$  be the vector in  $\mathbb{R}^I$  with all components equal to 1. Combining (43) and (44), we then have that

$$-\frac{M_i''\tilde{\epsilon}}{\sqrt{\Lambda}}\mathbf{e} \leq A(p^\Lambda - p^*) \leq \frac{M_i''\tilde{\epsilon}}{\sqrt{\Lambda}}\mathbf{e}.$$

Since  $A$  is invertible, we have that

$$\|p^\Lambda - p^*\| \leq \left\| A^{-1} \left( \frac{M_i''\tilde{\epsilon}}{\sqrt{\Lambda}}\mathbf{e} \right) \right\| \leq \|A\| \frac{M_i''\tilde{\epsilon}}{\sqrt{\Lambda}}. \quad (45)$$

Here  $\|A\|$  is the matrix norm given by  $\|A\| = \sum_{i \in \mathcal{I}, k \in \mathcal{I}} |a_{ik}|$  where  $a_{ik}$  is the element of the matrix  $A$  in the  $i^{\text{th}}$  row and  $k^{\text{th}}$  column. Equation (22) now follows directly from equation (45). As we have already established (23), the proof is complete.  $\blacksquare$

**Proof of Theorem 4.9:** We consider the PF game and the proof for the SF game follows similarly. Towards this end, we first claim that given  $\epsilon^\Lambda$  with  $\epsilon^\Lambda \rightarrow 0$  as  $\Lambda \rightarrow \infty$  and such that  $\limsup_{\Lambda \rightarrow \infty} \Lambda \epsilon^\Lambda < \infty$ , we can construct  $\theta^\Lambda$  and  $\delta^\Lambda$  such that  $\lim_{\Lambda \rightarrow \infty} \delta^\Lambda = 0$  and  $\theta^\Lambda \in [\bar{\theta} - \delta^\Lambda, \bar{\theta}]$ . The proof of this observation proof of this claim is very similar to the proof of the first part of Theorem 4.2 and is omitted. Once this observation is made, the proof proceeds as follows: let the profit function of the first stage be given by  $\tilde{\Pi}_i^\Lambda(p)$  for a price vector  $p$ . Then,

$$\min_{\theta \in [\bar{\theta} - \delta^\Lambda, \bar{\theta}]} \bar{\Pi}_i^\Lambda(p, \theta) \leq \tilde{\Pi}_i^\Lambda(p) \leq \max_{\theta \in [\bar{\theta} - \delta^\Lambda, \bar{\theta}]} \bar{\Pi}_i^\Lambda(p, \theta). \quad (46)$$

By Lemma 3.4  $e_i(\Lambda_i, \theta_i^\Lambda)/\Lambda \rightarrow 0$ , as  $\Lambda \rightarrow \infty$ . This, together with the absolute continuity of the demand function, implies that

$$\Pi_i^P(p) - \epsilon^\Lambda \leq \tilde{\Pi}_i^\Lambda(p) \leq \Pi_i^P(p) + \epsilon^\Lambda, \quad (47)$$

for all  $\Lambda$  large enough and for any vector  $p \in \mathcal{P}$ . Here  $\Pi_i^P(\cdot)$  is the profit function the in the pure price-competition as defined in equation (13). Fix now a firm  $i$  and a price  $p_i \in [p_i^{\min}, p_i^{\max}]$ . Then,

$$\tilde{\Pi}_i^\Lambda(p_i \uparrow p^*) \leq \Pi_i^P(p_i \uparrow p^*) + \epsilon^\Lambda \leq \Pi_i^P(p^*) + \epsilon^\Lambda,$$

where the last inequality follows from  $p^*$  being the unique equilibrium of the pure price-competition game. In particular, using (47), we have that

$$\tilde{\Pi}_i^\Lambda(p_i \uparrow p^*) \leq \tilde{\Pi}_i^\Lambda(p^*) + 2\epsilon^\Lambda.$$

Hence  $(p^*, \theta^\Lambda)$  is an  $\epsilon$ -Nash equilibrium for the PF game for any  $\Lambda$  large enough. The proof of the characterization results is very similar to the proof of Theorem 4.2 using the bounding in (46). The complete argument is omitted.  $\blacksquare$

**Proof of Corollary 4.10:** The only part of the corollary that requires proof is the existence of an actual Nash equilibrium  $(p^\Lambda, \theta^\Lambda)$  for each  $\Lambda$ . Once this is established the second part of the corollary would follow from Theorem 4.5 and Remark 4.7. We first show that it is enough to consider a certain subspace of the strategy space to find the equilibria. Then we show that equilibrium exists for the sequence of games with these subspaces as strategy space.

First, therefore, we claim that if an actual Nash equilibrium  $(p^\Lambda, \theta^\Lambda)$  exists for all  $\Lambda$  large enough, the sequence of Nash equilibria must satisfy

$$\theta_i^\Lambda = \bar{\theta} - o\left(\frac{1}{\Lambda^{\frac{1}{2}-\epsilon}}\right),$$

for some  $\epsilon > 0$  and for all  $\Lambda$  large enough. To reach a contradiction, assume that this is not the case, i.e., that there exists a subsequence  $\Lambda_n$  such that  $\Lambda_n^{\frac{1}{2}-\epsilon}(\bar{\theta} - \theta_i^{\Lambda_n}) > \delta$  for some  $\delta > 0$  and all  $n$  large enough. But since  $(p^{\Lambda_n}, \theta^{\Lambda_n})$  is a Nash equilibrium for all  $n$  large enough, it is necessarily an  $\epsilon/\sqrt{\Lambda_n}$ -Nash equilibrium for all  $n$  large enough. By Theorem 4.5, then, this sequence must satisfy

$$\theta_i^{\Lambda_n} = \bar{\theta} - O\left(\frac{1}{\sqrt{\Lambda_n}}\right),$$

and this is a contradiction.

To establish the existence of equilibrium for all  $\Lambda$  large enough, we may focus on the sequence of games with strategy space in the  $\Lambda^{th}$  game given by  $\mathcal{P} \times [\bar{\theta} - C/\Lambda^{\frac{1}{2}-\epsilon}, \bar{\theta}]^I$ , for some constant  $C > 0$ . The proof of existence for the games with the truncated space relies now on the proof of Theorem 3 in Allon and Federgruen (2007). Specifically, the existence would be established if we show that due to our truncation the assumed bounds of  $\bar{\theta}$  in Theorem 3 in Allon and Federgruen (2007) may be removed. To show that this is indeed the case, note that the assumption on  $\bar{\theta}$  there is made to guarantee that

$$\frac{4b_i\gamma_i}{(\bar{\theta} - \theta_i)^3} \geq (a'_i(\theta_i))^2,$$

or with our scaling that

$$\frac{4b_i\Lambda_i\gamma_i}{(\bar{\theta} - \theta_i^\Lambda)^3} \geq (\Lambda a'_i(\theta_i))^2.$$

Letting  $\bar{a}'_i = \sup_{\theta \in [0, \bar{\theta}]} a'_i(\theta)$ , this will be satisfied for every  $\theta \in [\bar{\theta} - C/\Lambda^{\frac{1}{2}-\epsilon}, \bar{\theta}]$ , provided that

$$\frac{4b_i\Lambda_i\gamma_i}{C^3/\Lambda^{\frac{3}{2}-3\epsilon}} \geq (\Lambda\bar{a}'_i)^2. \quad (48)$$

Since  $\epsilon$  was arbitrary we may choose  $\epsilon < 1/3$  in which case (48) holds for all  $\Lambda$  large enough. Thus, equilibrium exists for all  $\Lambda$  large enough for the games with truncated strategy space. By Theorem 3 in Allon and Federgruen (2007) each equilibrium point must satisfy equations (25) and (26).  $\blacksquare$

**Proof of Theorem 5.1:** The proof is somewhat lengthy but it is based on a simple argument. First, we establish the existence of Nash equilibrium for each  $\Lambda$ . Then, observing that in equilibrium all firms that are of the same type will use the same price and service level, we get an explicit set of equations to determine the equilibrium. Finally, we show that the sequence of equilibria obtained from this set of equation converges in the limit to the set of equations that define the equilibrium in the *basic market*. The result of the theorem then follows from Assumption 5.1.

First, the existence of equilibrium for all  $\Lambda$  large enough is proved as in Corollary 4.10. In contrast to Corollary 4.10, here we do have to impose (29) to guarantee the existence of equilibrium in the basic market and also for the small firms as  $\Lambda$  grows large. The adjustment to the existence proof of Corollary 4.10 is, however, straightforward. As in Corollary 4.10, each equilibrium point  $(p^\Lambda, \theta^\Lambda)$  must satisfy, for all  $i \in \mathcal{I}^\Lambda$ , the equations

$$\frac{\partial \bar{\Pi}_i^\Lambda}{\partial p_i^\Lambda} = -\frac{\Lambda}{M_i^\Lambda} b_i^\Lambda (p_i^\Lambda - c_i - \gamma_i) + \frac{\Lambda_i}{M_i^\Lambda} = 0, \quad (49)$$

$$\theta_i^\Lambda(p_i^\Lambda) = \begin{cases} \text{the unique root of } a_i'^\Lambda(\theta_i^\Lambda)(p_i^\Lambda - c_i - \gamma_i) = \frac{\gamma_i}{(\bar{\theta} - \theta_i^\Lambda)^2} & , \text{ if } p_i^\Lambda \geq c_i + \gamma_i \left(1 + \frac{1}{\bar{\theta}^2 a_i'^\Lambda(0)}\right) \\ 0 & \text{otherwise.} \end{cases} \quad (50)$$

Equation (49) is equivalently written as

$$\Lambda(M^{-1})^\Lambda A^\Lambda p = \Lambda(M^{-1})^\Lambda (\bar{a}^\Lambda(\theta) + \kappa^\Lambda), \quad (51)$$

where the  $I^\Lambda \times I^\Lambda$  matrix  $A$  is specified by  $A_{ii}^\Lambda = b_i^\Lambda$ ,  $A_{ij}^\Lambda = -\beta_{ij}^\Lambda$  for  $i \neq j$  and where  $\kappa_i^\Lambda = b_i^\Lambda(c_i + \gamma_i)$ . Also,  $\bar{a}^\Lambda(\theta) = (a_i^\Lambda(\theta_i) - \sum_{j \neq i} a_{ij}^\Lambda(\theta_j))$ . Finally  $M_i^\Lambda$  is the  $I^\Lambda \times I^\Lambda$  diagonal matrix whose  $k^{\text{th}}$  diagonal element is  $M_i^\Lambda$  and  $(M^{-1})^\Lambda$  is its inverse. Clearly, all firms of the same type will use in Nash equilibrium exactly the same service-level and price. Formally,  $\theta_i^\Lambda = \theta_j^\Lambda$  and  $p_i^\Lambda = p_j^\Lambda$  whenever  $c(i) = c(j)$ . Consequently, to every solution  $(p^\Lambda, \theta^\Lambda) \in \mathcal{P}^\Lambda \times \Theta^\Lambda$  to (49) and (50) corresponds a vector  $(\tilde{p}^\Lambda, \tilde{\theta}^\Lambda) \in \mathcal{P} \times \Theta$  which is obtained from  $(p^\Lambda, \theta^\Lambda)$  by setting  $\tilde{\theta}_k^\Lambda = \theta_i^\Lambda$  and  $\tilde{p}_k^\Lambda = p_i$  for  $k = c(i)$ . We remind the reader that  $\mathcal{P} = \times_{k=1}^I [p_k^{\min}, p_k^{\max}]$  and  $\Theta = [0, \bar{\theta}]^I$ , i.e.,  $\mathcal{P} \times \Theta$  is the strategy space in the basic market.

It is easy to see that for every vector  $\tilde{\theta} \in \Theta$ , and every  $i \in \mathcal{S}^\Lambda$ ,

$$\bar{a}_i^\Lambda(\theta) = a_{c(i)}(\theta_{c(i)}^\Lambda) - \frac{\Lambda - 1}{\Lambda} a_{c(i)c(i)}(\theta_{c(i)}) - \sum_{c(i) \neq l} a_{c(i)l}(\theta_l).$$

Similarly for a firm  $j \in \mathcal{L}$  we have

$$\bar{a}_j^\Lambda(\theta) = \Lambda a_{c(j)}(\theta_{c(j)}) - \Lambda \sum_{c(j) \neq l} a_{c(j)l}(\theta_l).$$

Plugging these back into (51), we have that

$$M^{-1}Ap = M^{-1}(\bar{a}(\tilde{\theta}^\Lambda) + \kappa) + C^\Lambda, \quad (52)$$

where the  $I \times I$  matrix  $A$  is specified by  $A_{kk} = b_k$ ,  $A_{kl} = -\beta_{kl}$ , for  $k, l \in \mathcal{I}$ ,  $\kappa_k = b_k(c_k + \gamma_k)$  for  $k \in \mathcal{I}$  and for  $\theta \in \Theta$ ,  $\bar{a}_k(\theta) = (a_k(\theta_k) - \sum_{l \neq k} a_{kl}^\Lambda(\theta_l))$ . Also, for  $k \in \mathcal{I}$ ,

$$M_k = \max_{\theta \in \Theta, p \in \mathcal{P}} \left[ a_k(\theta_k) - b_k p_k - \sum_{k \neq l} \alpha_{kl}(\theta_l) + \sum_{k \neq l} \beta_{kl} p_l \right]^+.$$

Finally, for  $(p, \theta) \in \mathcal{P} \times \Theta$ ,

$$C^\Lambda(\theta, p) = \left( \Lambda(M^{-1})^\Lambda A^\Lambda p - M^{-1}Ap \right) + \left( \Lambda(M^{-1})^\Lambda (\bar{a}^\Lambda(\theta) + \kappa) - M^{-1}(\bar{a}(\theta) + \kappa) \right) - \frac{1}{\Lambda} \sum_{k \in \mathcal{I}} \frac{a_{kk}(\theta)}{M_k^\Lambda}.$$

It is lengthy but a matter of basic calculations to show that  $\sup_{\theta \in \Theta, p \in \mathcal{P}} C^\Lambda(\theta, p) \rightarrow 0$  as  $\Lambda \rightarrow \infty$ . Now, for every vector  $\theta \in \Theta$ , we define  $f(\theta) := M^{-1}(\bar{a}(\theta) - \bar{a}(\tilde{\theta}))$ , with  $\bar{a}(\cdot)$  as defined above. We claim that for every  $i \in \mathcal{L}$ ,  $\theta_i^\Lambda \rightarrow 0$  as  $\Lambda \rightarrow \infty$ . The argument is extremely similar to the proof of Theorem 4.2 and it is omitted. Hence, we will have that  $f(\theta^\Lambda) \rightarrow 0$  as  $\Lambda \rightarrow \infty$ .

To summarize, we can rewrite the set of equations (50) and (51) for  $\theta^\Lambda$  and  $p^\Lambda$  as the following equations for  $\tilde{\theta}^\Lambda$  and  $\tilde{p}^\Lambda$ :

$$M^{-1}A\tilde{p}^\Lambda = M^{-1}(\bar{a}(\tilde{\theta}^\Lambda) + \kappa) + f(\tilde{\theta}^\Lambda) + C^\Lambda,$$

and for  $k \in \mathcal{S}$ ,

$$\tilde{\theta}_k^\Lambda(\tilde{p}_k^\Lambda) = \begin{cases} \text{the unique root of } a'_k(\tilde{\theta}_k^\Lambda)(p_k^\Lambda - c_k - \gamma_k) = \frac{\gamma_k}{(\tilde{\theta}_k^\Lambda)^2} & , \text{ if } \tilde{p}_k^\Lambda \geq c_k + \gamma_k \left( 1 + \frac{1}{\tilde{\theta}_k^{\Lambda 2} a'_k(0)} \right) \\ 0 & \text{otherwise.} \end{cases} \quad (53)$$

Consider a convergent subsequence of  $\tilde{\theta}^\Lambda$ . Such a subsequence exists by the compactness of  $\Theta$ , and we already know that every limit point,  $\theta$ , must satisfy that  $\theta_i = \bar{\theta}$  for all  $i \in \mathcal{L}$ . Denote

the chosen subsequence by  $\{\Lambda_n\}_{n \in \mathbb{N}}$ . Then, by our previous arguments, both  $f(\tilde{\theta}^{\Lambda_n}) \rightarrow 0$  and  $C^{\Lambda_n} \rightarrow 0$  as  $n \rightarrow \infty$ . Since  $A$  is invertible (see Lemma 1 of Allon and Federgruen (2007)) we have the continuity of  $p^\Lambda$  as a function of  $\tilde{\theta}^\Lambda$ ,  $f(\theta^\Lambda)$  and  $C^\Lambda$ . Consequently,  $(p^{\Lambda_n}, \theta^{\Lambda_n})$  converges to a solution  $(p, \theta)$  in which  $\theta_k = \bar{\theta}$  for all  $k \in \mathcal{L}$  and one that satisfies the equations

$$M^{-1}Ap = M^{-1}(\bar{a}(\theta) + \kappa),$$

and for  $k \in \mathcal{S}$ ,

$$\theta_k(p_k) = \begin{cases} \text{the unique root of } a'_k(\theta_k)(p_k - c_k - \gamma_k) = \frac{\gamma_k}{(\bar{\theta} - \theta_k)^2} & , \text{ if } p_k \geq c_k + \gamma_k \left(1 + \frac{1}{\bar{\theta}^2 a'_k(0)}\right) \\ 0 & \text{otherwise.} \end{cases} \quad (54)$$

But these equations define the equilibrium in the game defined in Assumption 5.1, and by that assumption we must have that  $\theta_k \neq \theta_l$  for at least two types  $k$  and  $l$  in  $K^S$ . Consequently, we have that

$$\liminf_{n \rightarrow \infty} \Delta(\theta_{S^{\Lambda_n}}^{\Lambda_n}) > 0.$$

Since this holds for every convergent subsequence, the proof is complete. ■

## Acknowledgments

We are grateful to Robert Shumsky for helpful discussion. We also thank the referees for their valuable comments which substantially improved this manuscript.

## References

- Afeche, F., H. Mendelson. 2004. Pricing and priority auctions in queueing systems with generalized delay cost structure. *Management Science* **50**(7).
- Allon, G., A. Federgruen. 2006. Service competition with general queueing facilities. *Forthcoming in Operations Research*.
- Allon, G., A. Federgruen. 2007. Competition in service industries. *Operations Research* **55**(1).
- Armony, M. 2005. Dynamic routing in large-scale service systems with heterogenous servers. *Queueing Systems* **51** 287–329.
- Armony, M., C. Maglaras. 2004. On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Oper. Res.* **52**(2) 271–292.
- Atar, R. 2004. Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Prob.* **15**(4) 2606–2650.
- Atar, R., A. Mandelbaum, M. Reiman. 2004. Scheduling a multi-class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Prob.* **14**(3) 1084–1134.

- Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Oper. Res.* **52**(1) 17–34.
- Cachon, G., S. Netessine. 2004. Game theory in supply chain analysis. D. Simchi-Levi, S. D. Wu, Z. J Shen, eds., *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era*. Kluwer, 13–59.
- Cachon, P. C., P. T. Harker. 2002. Competition and outsourcing with scale economies. *Management Science* **48**(10) 1314–1333.
- Chen, H., Y-W. Wan. 2003. Price competition of make-to order firms. *IIE Transactions* **35**(9) 817–832.
- Dasci, A. 2003. Dynamic pricing of perishable assets under competition: a two-period model Working paper.
- De Vany, A.S., T.R. Saving. 1983. The economics of quality. *Journal of Political Economy* **91**(6) 979–1000.
- Dixon, H.D. 1987. Approximate bertrand equilibria in a replicated industry. *Review of Economic Studies* **54**(1) 47–62.
- Green, L. V., P. J. Kolesar. 2004. Improving emergency responsiveness with management science. *Management Science* **50**(8) 1001–1014.
- Gurvich, I., M. Armony, A. Mandelbaum. 2006. Service-level differentiation in call centers with fully flexible servers. *Management Science* .
- Gurvich, I., W. Whitt. 2007. Service-level differentiation in many-server service systems: A solution based on fixed-queue-ratio routing. *Working Paper, Columbia University, New York, NY* .
- Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29**(3) 567–587.
- Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.
- Johari, R., G. Weintraub, B. Van Roy. 2007. Investment and market structure in industries with congestion. *Working Paper, Columbia GSB* .
- Kalai, E., M. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Science* **38**(8) 1154–1163.
- Kolesar, P. J., E. H. Blum. 1973. Square root laws for fire engine response distances. *Management Science* **19**(12) 1368–1378.
- Levhari, D., I. Lusk. 1978. Duopoly pricing and waiting lines. *European Economic Review* **11** 17–35.
- Lu, L. X., J.A. Van Mieghem, R. C. Savaskan. 2007. Incentives for quality through endogenous routing Working paper.
- Lusk, I. 1976. On partial equilibrium in a queueing system with two servers. *The Review of Economic Studies* **43** 519–525.
- Maglaras, C., A. Zeevi. 2003. Pricing and capacity sizing for systems with shared resources: Scaling relations and approximate solutions. *Management Science* **49**(8) 1018–1038.
- Maglaras, C., A. Zeevi. 2005. Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* **53**(2) 242–262.
- Porter, M. 1996. What is strategy? *Harvard Business Review* 61–78.
- So, K. 2000. Price and time competition for service delivery. *Manufacturing Service and Operations Management* **2**(4) 392–409.
- Spera, S. 2003. Listening to customers is important; acting on the information is critical to business success. *CRMToday* .
- Tezcan, T. 2006. Optimal control of distributed parallel server systems under the Halfin and Whitt regime. *Working Paper, Georgia Institute of Technology, Atlanta, GA* .

- Tezcan, T., J.G. Dai. 2006. Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Working Paper. Georgia Institute of Technology, Atlanta, GA* .
- Tijms, S.H. 1981. Nash equilibria for noncooperative  $n$ -person games in normal form. *SIAM Review* **23**(2) 225–237.
- Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. *Oper. Res.* **51**(4) 531–542.
- Zeithaml, V., M. J. Bitner, D. D. Gremler. 2005. *Services Marketing*. McGraw-Hill Irwin.