Quality-Speed Conundrum: Tradeoffs in Customer-Intensive Services

Krishnan S. Anand
David Eccles School of Business, University of Utah, Salt Lake City, UT 84112. k.anand@utah.edu
M. Fazıl Paç
Wharton School, University of Pennsylvania, Philadelphia, PA 19104. mpac@wharton.upenn.edu
Senthil Veeraraghavan
Wharton School, University of Pennsylvania, Philadelphia, PA 19104. senthilv@wharton.upenn.edu

August 2010

Forthcoming in Management Science

Abstract

In many services, the quality or value provided by the service increases with the time the service-provider spends with the customer. However, longer service times also result in longer waits for customers. We term such services, in which the interaction between quality and speed is critical, as *customer-intensive services*. In a queueing framework, we parameterize the degree of customer-intensity of the service. The service speed chosen by the service-provider affects the quality of the service through its customer-intensity. Customers queue for the service based on service quality, delay costs and price. We study how a service provider facing such customers makes the optimal "quality-speed tradeoff". Our results demonstrate that the customer-intensity of the service is a critical driver of equilibrium price, service speed, demand, congestion in queues and service provider revenues. Customer-intensity leads to outcomes very different from those of traditional models of service rate competition. For instance, as the number of competing servers increases, the price increases and the servers become slower.
Keywords: *Customer-Intensity, Service Operations, Strategic Customers, Queues, Cost Disease.*

1. Introduction

'Festina Lente' [*Make haste slowly*]

– motto of *Aldus Manutius* (1449 – 1515).

In a wide variety of service industries, providing good customer service requires a high level of diligence and attention. We refer to such services as *customer-intensive services*. Examples of such services are health care, legal and financial consulting, and personal care (such as spas, hair-dressing, beauty care and cosmetics). Economists have noted that some industries in the service sector, including health services and education, have lagged significantly in their productivity growth, despite rapid productivity improvements overall, in the last few decades (Triplett and Bosworth 2004, Varian 2004). For example, in the last decade, the health care industry displayed a negative annual growth of -0.4% (Triplett and Bosworth, 2004. pp. 262-263). We note that low-productivity

industries[1] are predominantly customer-intensive.[2]

A major difficulty in improving productivity in such customer-intensive services is the sensitivity of the *service quality* provided to the *speed of service*: as the service speed increases, the quality of service inevitably declines. Often, the only way to increase productivity without sacrificing quality is to increase capacity investments, which increases costs. This phenomenon has been termed *Baumol's cost disease* (Baumol 1993). Primary health care practice in the United States epitomizes this problem. Due to high levels of demand, doctors need to rush between patients,[3] spending most of their time treating acute illnesses - a process that is also dissatisfying to patients (Yarnall *et al* 2003). As Surowiecki (2003) notes, *"Cost disease isn't anyone's fault. (That's why it's called a disease.)* [...] *you can control drug costs and limit expensive new procedures, but, when it comes to, say, hospital care and doctor visits, the only way to improve productivity is to shrink the size of the staff and have doctors spend less time with patients (or treat several patients at once). Thus the Hobson's choice: to lower prices you have to lower quality."* Thus, primary health care services provide a clear context for the quality degradation associated with a service system stretched to work at a fast pace while trying to serve a large number of patients.

The above examples suggest that focusing exclusively on improving productivity by increasing the speed of service leads to a reduction in the value of the service provided. On the other hand, increasing the service value by increasing the time spent serving each customer has its pitfalls. First, it increases customers' waiting times due to congestion effects from the slower service times. Second, it increases the cost of the service, as the productivity (number of customers served) falls. The first effect leads to lower customer value; the second, to higher prices.

In this paper, we study how a service provider can make the optimal "quality-speed" tradeoff in the face of strategic customers– customers who join the service only if the utility (the value of the service net of congestion costs) exceeds the price charged by the service provider. Congestion costs are an outcome of the *aggregate* procurement decisions of all consumers in the market, since every customer who joins the service imposes a negative externality (in the form of additional expected waiting time) on all other customers. In turn, the tradeoff faced by the provider of a customer-intensive service between *quality* (service value) and *service speed* forms the crux of our model.

The extant academic research has not addressed the interaction between service value and service speed, or its consequences. In general, the extant literature treats service value and service times as independent performance metrics, despite the fact that their interaction is critical for

[1] It is difficult to compare productivity *per se* between different industries; what can be compared is their productivity growths over time. The literature (*cf* Triplett and Bosworth 2004) describes industries with low-productivity *growth* as "low-productivity" industries. Of course, a sustained period of low productivity growth in an industry would lead to low productivity relative to other industries. We thank the Departmental Editor for suggesting this distinction.

[2] Customer-intensive services are generally characterized by high labor content, but high labor content need not imply high customer-intensity (e.g., construction services).

[3] "*I was seeing 30 people a day and always rushing. Patients were dissatisfied.... I was dissatisfied.*" Dr. Bernard Kaminetsky, M.D., F.A.C.P., (formerly with New York University, currently with MDVIP) in his testimony to the Joint Economic Committee of the United States Congress, April 28, 2004.

customer-intensive services. In our queueing model, "customer-intensity" is indexed explicitly by the parameter α. The greater the customer-intensity of the service, the higher the value of α. (The special case of $\alpha = 0$ corresponds to the traditional queuing model, in which the service value is independent of the service speed.)

We find that modeling customer intensity leads to outcomes very different from those of traditional queuing models. To give a flavor of these differences, we mention two such insights: *(i)* We find that the service provider slows down (i.e., increases its service-time) as the customer-intensity of the service increases. Thus, the equilibrium value of the service provided to customers is *always* increasing in customer-intensity. As a consequence, such services are likely to have partial market coverage; *(ii)* We find that competition in service rates does *not* dampen prices – in fact, the price charged by the service provider *increases* as the number of competing servers increases. Furthermore, the equilibrium waiting costs are invariant with respect to the number of competing servers (even as the price increases).

Related Literature:

The existing research in Service Operations treats quality and speed as *independent* performance metrics. To our knowledge, there is no precedent in the queueing literature that models the customer-intensity of a service or studies the interactions between service quality and service speed arising from customer-intensity.

A number of papers address the decision-making of customers who choose whether or not to join a queue based on rational self-interest, as in our model. Our paper differs from all of the extant literature in that we explicitly model the dependence of service quality on service duration, and explore the resulting equilibrium behavior of customers as well as the service provider's service rate and pricing decisions.

Admission fees have long been considered an important tool to control congestion in service queues, dating back to the seminal paper by Naor (1969). Edelson and Hildebrand (1975) extend Naor's (1969) model by analyzing unobservable service queues. Following Mendelson and Whang (1990), papers that explore equilibrium queue joining, pricing and/or service rate decisions include Afeche (2006), Armony and Haviv (2000), Cachon and Harker (2002), Chen and Frank (2004), Chen and Wan (2003), Gilbert and Weng (1998), Kalai *et al* (1992), Lederer and Li (1997), Li (1992) and Li and Lee (1994). We refer the reader to Hassin and Haviv (2003)'s excellent review of this literature. Other notable papers that explore the interaction between service quality and congestion include Allon and Federgruen (2007), Chase and Tansik (1983), Gans (2002), Hopp *et al* (2007), Lovejoy and Sethuraman (2000), Oliva and Sterman (2001), Png and Reitman (1994), Ren and Wang (2008), Veeraraghavan and Debo (2009) and Wang *et al* (2010).

Research articles that acknowledge the existence of interactions between service duration and quality in different domains include Kostami and Rajagopalan (2009) (dynamic decisions), de Vericourt and Zhou (2005) (routing unresolved call-backs), Lu *et al* (2008) (manufacturing rework), Hasija *et al* (2009) (an empirical study of call centers), de Vericourt and Sun (2009) (judgement

accuracy), and Wang *et al* (2010) (medical diagnostic services). In these papers, the customer demand is assumed to be exogenous and/or pricing decisions are absent.

2. A Model of Customer-Intensive Service Provision

We consider a monopolist providing a customer-intensive service to a market of homogenous, rational consumers. We model the monopolist service setting using an unobservable $M/M/1$ queueing regime.[4] We use the $M/M/1$ model in the interests of expositional simplicity; however, we can show that all our analytical results extend to general service distributions. Customers arrive at the market according to a Poisson process at an exogenous mean rate Λ. We shall refer to Λ as the *potential demand* for the service. We assume that customers are homogenous in their valuations of the service, and incur a waiting cost of c per unit of time spent in the system. Upon arrival, every customer decides whether to procure the service (join the queue) or quit (balk from the service) based on the value of the service, the expected waiting cost and the price.

The service rate μ of the service provider is assumed to be common knowledge. The effective demand for the service (i.e., the effective arrival rate), λ, is the aggregate outcome of all customers' decisions (joining or balking). For any customer, the expected waiting time in an M/M/1 system is as follows:[5]

$$W(\mu, \lambda) = \left\{ \frac{1}{\mu - \lambda} \text{ (if } 0 \leq \lambda < \mu), \ \infty \text{ (otherwise).} \right\} \tag{1}$$

Before we formalize our model of customer-intensive services, we discuss the classical queueing model, which will serve as a useful benchmark.

2.1 The Classical Queueing Model

The *classical queueing model* (e.g. Naor 1969, Edelson and Hildebrand 1975) assumes that customers receive a service value V_b, that is independent of the service rate μ_b (or equivalently, of the service time $\tau_b = 1/\mu_b$). This will serve as a useful *benchmark* for our analysis of customer-intensive service queues, and is indexed throughout this paper by the subscript b.)

In the classical queueing model, increasing the service rate (i.e., reducing the service time spent with each customer), always results in higher revenues, as it allows the firm to serve more customers and/or lower their expected waiting time. In this paper, we depart from the classical assumption that the service value remains unaffected by changes in the service rate.

2.2 Modeling Value in Customer-Intensive Services

In customer-intensive services, the quality of the service provided to a customer (and hence, service value) increases with the time spent in serving the customer. In our model, service quality

[4]The $M/M/1$ queueing approximation has of course been applied to a large variety of settings– too numerous to be listed here. See Green and Savin (2008) for an application to primary health care, and Brahimi and Worthington (1991) on outpatient appointment systems.

[5]For an $M/G/1$ system, the mean waiting times can be calculated by the Pollaczek-Khinchin formula (Ross 2006).

is reflected in the service value function $V(\tau)$ which increases with the mean service time τ. Furthermore, in most situations, the marginal value to customers from an increase in service time are diminishing. Therefore, we model customer-intensive services by constructing the service value function $V(\tau)$ as a non-decreasing and concave function of the mean service time τ.[6] Specifically, we let $V(\tau) = (V_b + \alpha/\tau_b - \alpha/\tau)^+$ or simply expressed in service rates as,

$$V(\mu) = (V_b + \alpha\mu_b - \alpha\mu)^+ \tag{2}$$

where $x^+ = max(x, 0)$.[7] The parameter $\alpha \geq 0$ captures the *customer-intensity* of the service provided. It determines the sensitivity of the service value to the service speed, and is a descriptor of the "nature" of the service. Clearly, higher values of α suggest a stronger dependence of the service value on the service time (highly customer-intensive tasks).

When α is zero, the value of the service provided equals V_b; this case is equivalent to the classical queueing model. Thus, as discussed previously, V_b serves as a benchmark service value. Secondly, for all α, when the service rate is $\mu_b = 1/\tau_b$, the value of the service provided is V_b. Therefore, $\mu_b(\tau_b)$ could be considered a benchmark service rate (time), providing a service value V_b to customers.

2.3 Customers' Queue Joining Decision

Rational customers arrive to the system according to a Poisson process at rate Λ, and decide whether to join the (unobservable) service queue. The potential demand (market size), Λ, price, p, service rate, μ, waiting cost per unit time, c, and the resulting service value, $V(\mu)$, are common knowledge to all arriving customers. We model the queue-joining decisions of customers as in Hassin and Haviv (2003), and focus on symmetric equilibrium queue-joining strategies since all customers are homogenous. Let $\gamma_e(\mu, p)$ denote the *equilibrium* probability that a customer would join the queue at a server whose service rate is μ and admission price is p.[8] Thus, the equilibrium decision of customers $\gamma_e(\mu, p)$ is based on the value of the service, the price and the expected cost of waiting.

Three market outcomes – full, zero or partial market coverage – are possible, depending on the market size Λ and other parameters. These outcomes are: 1. Full coverage: If the net utility is non-negative for a customer even when all the other potential customers join (i.e., $V(\mu) - (p + cW(\mu, \Lambda)) \geq 0$), then every customer will join the queue in equilibrium (i.e. $\gamma_e(\mu, p) = 1$). 2. No coverage: If the net utility is not positive for a customer joining the queue even when no other customer joins the queue, (i.e., $V(\mu) - (p + c/\mu) \leq 0$), then no one joins the queue (i.e., $\gamma_e(\mu, p) = 0$). 3. Partial coverage: When $p + c/\mu < V(\mu) < p + cW(\mu, \Lambda)$, each customer plays a mixed strategy in equilibrium, meaning that each customer joins the queue with the same probability

[6] Customer-intensity depends only on the relationship between the service time and the service value for a customer. Thus, a highly customer-intensive service need not be a high-contact service (Lovelock 2001).

[7] We can generalize $V(\mu)$ to be convex and decreasing in μ. Similarly, we can generalize $V(\tau)$ to be an increasing and concave function of τ. While this leads to more analytical complexity in the model, our conclusions remain identical.

[8] We indicate the *equilibrium* values of the various model variables by the subscript e.

$\gamma_e(\mu, p) \in (0, 1)$ and balks with probability $1 - \gamma_e(\mu, p) \in (0, 1)$. Therefore, the equilibrium arrival rate is $\lambda_e(\mu, p) = \gamma_e(\mu, p)\Lambda$ and satisfies the condition $V(\mu) - p = cW(\mu, \lambda_e(\mu, p))$.

2.4 Characterization of the Service Rate Decision Space

Clearly, the interaction between the service speed and the service value imposes a constraint on the service provider's feasible operating region (i.e. the range of service rates and prices he can choose from, while still drawing customers). In this Section, we characterize the *feasible* range of service-rates for the service provider, which is maximized at $p = 0$. We do this for two reasons: *(i)* This characterization will be useful when we formulate and solve the service provider's revenue maximization problem, with the price determined endogenously, in the next Section; and *(ii)* The characterization of the feasible space itself illustrates the impact of customer-intensity on business and customer outcomes.

A service should be at least valuable enough that a customer does not mind waiting *during* the process of service provision. Therefore, the value $V(\mu)$ must exceed $\frac{c}{\mu}$, the expected waiting costs during the service; i.e., $V(\mu) - c/\mu \geq 0$. This condition ensures that a customer can expect non-negative net value from the service (at $p = 0$), at least when no other customer precedes him in the queue. Note that a customer's service procurement imposes negative externalities on others, as the expected waiting cost, $\frac{c}{\mu-\lambda}$, increases with the effective demand, λ.

Rewriting $V(\mu) - c/\mu \geq 0$, we have $V_b + \alpha\mu_b - \alpha\mu \geq c/\mu$, or equivalently, $A_1(\alpha) \leq \mu \leq A_2(\alpha)$, where $A_1(\alpha), A_2(\alpha)$ are the solutions for μ to the quadratic $V_b + \alpha\mu_b - \alpha\mu = c/\mu$. Thus,

• $\mu \geq A_1(\alpha) = \frac{V_b+\alpha\mu_b-\sqrt{(V_b+\alpha\mu_b)^2-4\alpha c}}{2\alpha}$. The service has to be fast enough. No one will wait forever *even* if the service value is high. (Note that $A_1(0) = \lim_{\alpha\to 0} A_1(\alpha) = \frac{c}{V_b}$.)

• $\mu \leq A_2(\alpha) = \frac{V_b+\alpha\mu_b+\sqrt{(V_b+\alpha\mu_b)^2-4\alpha c}}{2\alpha}$. The service cannot be *too fast*. It is not possible to provide valuable service at very high service speeds. This additional constraint is unique to customer-intensive services (Observe that $A_2(0) = \lim_{\alpha\to 0} A_2(\alpha) = \infty$.).

For a customer-intensive service of type α, we denote this operating service-rate region by $\mathcal{F}(\alpha) = [A_1(\alpha), A_2(\alpha)]$.[9] Figure 1 shows the operating region and the associated net service value for any service rate in the operating region, for various α. Figure 1 shows that the service provider can choose from a larger range of service rates when the service is not very customer-intensive (i.e., when α is small). When $\alpha = 0$, the net service value is *increasing* in the service rate μ, in the entire range $\mathcal{F}(0) = [A_1(0), \infty)$. When $\alpha > 0$, the net service value is *unimodal* in the region $\mathcal{F}(\alpha)$, and thus our results are applicable to services in which customers' *net value decreases* after a service time threshold.

3. Service Provider's Revenue Maximization

The service provider's objective is to maximize his revenues with respect to the service rate, μ and the price, p. The service provider's revenue function is given by $R(\mu, p) = p\lambda_e(\mu, p)$, where $\lambda_e(\mu, p)$

[9] As long as $V_b \geq c/\mu_b$, $\mathcal{F}(\alpha)$ is non-empty, $\forall\alpha \geq 0$.

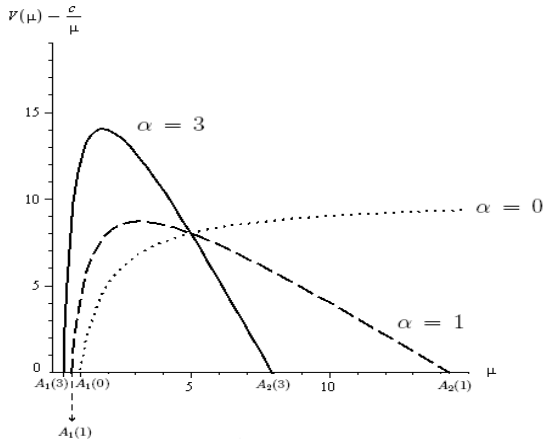

Figure 1: The net service value $(V(\mu) - c/\mu)$ and the operating region $\mathcal{F}(\alpha)$ shown for $\alpha = 0$ (dotted curve), $\alpha = 1$ (dashed curve) and $\alpha = 3$ (thick curve). However, for $\alpha > 0$ the service rates that provide non-negative net value are bounded in the interval $[A_1(\alpha), A_2(\alpha)]$. This implies that for a customer-intensive service of type α, the customer experiences a decrease in net value, if the service time exceeds a threshold.

is the equilibrium demand induced at the setting (μ, p). Therefore, the objective function of the service provider is given by:

$$max_{\{p\geq 0, \mu\in\mathcal{F}(\alpha)\}} \ \{R(\mu,p) = p\lambda_e(\mu,p)\} \equiv max_{\{\mu\in\mathcal{F}(\alpha)\}} \left\{max_{\{0\leq p\leq V(\mu)\}}\{p\lambda_e(\mu,p)\}\right\}. \qquad (3)$$

Thus, we solve the service provider's revenue maximization problem in two steps. First, we find the optimal price $p(\mu)$ for a given service rate, μ. Then, using $p(\mu)$, we find the revenue maximizing service rate in the operating region $\mathcal{F}(\alpha)$.

Recall that for any $\mu \notin \mathcal{F}(\alpha)$, the net service value derived by a customer is negative, and hence no customer will join the service. Conversely, for each $\mu \in \mathcal{F}(\alpha)$, there exists a non-negative price at which the service provider can attract customers. Hence we focus on $\mu \in \mathcal{F}(\alpha)$. Also recall that $W(\mu, \lambda) = \frac{1}{\mu-\lambda}$ (by equation (1); hence the value derived by any customer at the arrival rate λ is $V(\mu) - p - \frac{c}{\mu-\lambda}$. Customers will join the service until this value is driven to zero. The equilibrium demand, $\lambda_e(\mu, p)$, as a function of the price is given as follows:

$$\lambda_e(\mu,p) = \begin{cases} \Lambda & if\ 0 \leq p \leq V(\mu) - \frac{c}{\mu-\Lambda} \\ \mu - \frac{c}{V(\mu)-p} & if\ V(\mu) - \frac{c}{\mu-\Lambda} < p \leq V(\mu) - \frac{c}{\mu} \\ 0 & if\ V(\mu) - \frac{c}{\mu} < p. \end{cases} \qquad (4)$$

Using (4), it is easy to verify that for a given μ, the equilibrium demand, $\lambda_e(\mu, p)$, is a non-increasing function of the price. The following proposition derives the service provider's optimal pricing policy for a given service rate μ. (The proofs of all results are provided in the Appendix.)

Proposition 1. *Consider a customer-intensive service of type α. For any service rate $\mu \in \mathcal{F}(\alpha)$,*

the optimal price equals:

$$p^*(\mu) = \begin{cases} V(\mu) - \frac{c}{\mu - \Lambda} & if\ 0 \le \Lambda \le \hat{\lambda}(\mu) \\ V(\mu) - \sqrt{cV(\mu)/\mu} & if\ \hat{\lambda}(\mu) < \Lambda \end{cases}$$

where $\hat{\lambda}(\mu) = \mu - \sqrt{\frac{c\mu}{V(\mu)}}$. *The resulting equilibrium arrival rate is equal to:*

$$\lambda_e(\mu, p^*(\mu)) = \begin{cases} \Lambda & if\ 0 \le \Lambda \le \hat{\lambda}(\mu) \\ \hat{\lambda}(\mu) & if\ \hat{\lambda}(\mu) < \Lambda. \end{cases}$$

The corresponding equilibrium revenues are $R(\mu, p^*(\mu)) = p^*(\mu)\lambda_e(\mu, p^*(\mu))$.

Proposition 1 derives the optimal price and the equilibrium demand (arrival rate) for any arbitrary service rate μ. We find a threshold $\hat{\lambda}(\mu)$ that defines the maximum number of customers the service provider would serve at a given service speed μ. When $\Lambda \le \hat{\lambda}(\mu)$, the service provider clears the market. However, when the potential demand is higher (i.e., for all $\Lambda > \hat{\lambda}(\mu)$), the service provider serves exactly $\hat{\lambda}(\mu)$ customers and repels the rest, by making adjustments to the admission price $p^*(\mu)$. In each case, the service provider extracts all the consumer surplus.

This result is driven by the negative externality that each customer imposes on all other customers– in the form of an increase in their waiting costs. Thus, to accommodate an additional customer, the service provider has to compensate *all* of its current customers for the additional waiting costs they incur, by decreasing the price. As the arrivals to the system increase, serving every additional customer requires an additional reduction in price, which eventually leads to the scenario (at $\lambda = \hat{\lambda}(\mu)$) in which the increase in demand does not make up for the revenues lost due to the corresponding price reduction. Hence, for large Λ, the service provider limits the number of customers admitted to the system to $\hat{\lambda}(\mu)$, by charging a suitable admission price. Therefore, as long as Λ remains higher than the threshold $\hat{\lambda}(\mu)$, small fluctuations in potential demand do not affect the optimal price, and hence, revenues.

Proposition 1 showed that for any $\mu \in \mathcal{F}(\alpha)$, there exists a price $p^*(\mu)$ that maximizes the service provider's revenues. Having derived the optimal price for each service rate μ, we now analyze the service provider's optimal service rate decision. In the next Section (3.1), we analyze the case of partial market ceverage ($\Lambda > \hat{\lambda}(\mu)$). We analyze the case of full market ceverage ($\Lambda \le \hat{\lambda}(\mu)$) in Section 3.2.

3.1 Partial Market Coverage

In this section, we assume that the potential demand Λ is "high enough" that the service provider's optimal price and service rate decisions are not constrained by the availability of potential customers (We can show that this condition translates, mathematically, to $\Lambda > \lambda^*_\alpha \triangleq \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.). To derive the optimal service rate under partial coverage, we first establish that the equilibrium demand and price curves (as a function of the service rate) are unimodal (details in the Appendix).

The unimodality property of both demand and price are outcomes of the tension between service value and waiting costs, as follows: Focusing exclusively on delivering a high value service requires setting a slow service rate. This leads to high customer waiting costs and low demand. High waiting costs also translate to a low price, since the maximum price the service-provider can charge is the value of the service *net* of waiting costs. On the other hand, increasing the service rate to minimize waiting costs leads to a low service value (and hence, low demand and a low price). Thus, both demand and price are maximized at some intermediate service rates in $\mathcal{F}(\alpha)$. Further, since the service provider's revenues are a product of the equilibrium demand and the price, the revenue-maximizing service rate is an interior point in $\mathcal{F}(\alpha)$.

Thus we see that even in markets where the potential demand is very large (i.e., $\Lambda \to \infty$), increasing the service speed does not lead to an increase in effective demand for customer-intensive services, because of the drop in service quality. Thus, partial market coverage is a by-product of the customer-intensity of services. Building on these observations, Proposition 2 provides the equilibrium outcomes from the maximization of (3), the service-provider's objective function.

Proposition 2. *For a customer-intensive service of type $\alpha > 0$, and when $\Lambda > \lambda^*_\alpha$,*

1. *The optimal service rate is equal to $\mu^* = \frac{V_b + \alpha\mu_b}{2\alpha}$.*
2. *The corresponding optimal price is equal to $p^*(\mu^*) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2}$.*
3. *The demand at the optimal price and service rate equals $\lambda_e(\mu^*, p^*(\mu^*)) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha} = \lambda^*_\alpha$.*

Therefore, the optimal revenue for the service is equal to $R(\mu^*, p^*(\mu^*)) = \frac{(V_b + \alpha\mu_b - 2\sqrt{c\alpha})^2}{4\alpha}$.

Proposition 2 shows that there exists a unique, interior service rate μ^* in $\mathcal{F}(\alpha)$ that maximizes revenues. Proposition 2.1 shows that the optimal service rate, μ^*, is decreasing in α: as the service becomes more customer-intensive, the service provider has a greater incentive to slow down and spend more time on each customer. We also see this in the expression for the equilibrium service value. From equation (2), the service value provided to customers in equilibrium is $V(\mu^*) = (V_b + \alpha\mu_b)/2$, which is increasing in α.

From Proposition 2.2, we note that the optimal price, $p^*(\mu^*)$, is unimodal in α – decreasing for $\alpha < c/\mu_b^2$ and increasing for $\alpha > c/\mu_b^2$. We saw that as the service becomes more customer-intensive, the optimal service time increases. However, this does not imply that the *net* value of the service provided also increases with customer-intensity. This is demonstrated by Proposition 2.2, since the optimal price tracks the net value of the service. For low α (i.e., $\alpha < c/\mu_b^2$), congestion effects dominate the increase in service value as α increases. Hence, as the task becomes more customer-intensive (i.e., α increases), the price *falls*. However, for high α values ($\alpha > c/\mu_b^2$), the optimal price is increasing in α: The increased service value from a longer service time dominates any increase in the equilibrium waiting cost.

The equilibrium demand $\lambda_e(\mu^*, p^*(\mu^*))$ is also determined by the tradeoff between waiting costs and the service value, and behaves similarly to the optimal price. At low values of α, waiting costs

are more sensitive to small increases in α than is the service value. Hence, congestion considerations dominate in this range. For higher values of α, the reverse is true – the service value is more sensitive to increases in α than the waiting cost. The net effect is that the equilibrium demand is unimodal – decreasing in α for $\alpha < V_b^2/c$, and increasing in α for $\alpha > V_b^2/c$ (Proposition 2.3).

Finally, Proposition 2 captures the effect of the delay parameter c on service outcomes. Interestingly, the optimal service rate, μ^*, is independent of the waiting cost, c; i.e., if customers are more impatient, the additional waiting cost does not result in a faster service. As one might expect, higher waiting costs lead to both lower prices, $p^*(\mu^*)$ and lower equilibrium demand $\lambda_e(\mu^*, p^*(\mu^*))$. Consequently, the optimal revenues, $R(\mu^*, p^*(\mu^*))$, decrease with increased waiting costs.

3.1.1 Analysis of Value-Price-Demand Interactions

We shed further light on the subtle interactions among price, demand and service value as the service rate changes in customer-intensive services. The equilibrium price, equilibrium demand, waiting costs and the service value to customers are outcomes of these complex interactions. Lemma 1 studies the relationship between the equilibrium price and the equilibrium demand at any service rate μ.

Lemma 1. ***[Property of α-symmetry:]*** *For a customer-intensive service of type α, $p^*(\mu)$ and $\lambda_e(\mu, p^*(\mu))$ have the following symmetric relationship around the optimal service rate μ^* for any given $\mu \in \mathcal{F}(\alpha)$: $p^*(\mu^* + \epsilon) = \alpha\lambda_e(\mu^* - \epsilon, p^*(\mu^* - \epsilon))$, where $\epsilon = (\mu - \mu^*)$.*

Lemma 1 clearly demonstrates that prices and effective demand are two levers related to each other by the customer-intensity parameter α. To better understand the implications of α-symmetry between price and demand, we divide the operating region $\mathcal{F}(\alpha)$ into 3 sub-regions as shown in Figure 2. Region 1 corresponds to low service rates, Region 2 corresponds to intermediate service rates, and Region 3 corresponds to high service rates.

When the service rate is low (Region 1), there is an *over-investment* in time of service, from both the customers' and the service provider's perspectives. Although the service provided is of high value, the cost of waiting is also high. In other words, increasing the service rate would improve each customer's service value (*net* of waiting costs), as well as the service provider's total revenues. In Region 1, increasing the service rate will lead to some loss of service value; however, the gains from the waiting cost reduction dominate the loss in service value (At low μ, waiting costs drop precipitously as μ increases.). Hence, the net service value provided to a customer is increasing in the service rate. This allows the service provider to charge customers a higher price. Furthermore, this service rate increase leads to higher throughput. By increasing the service rate, the service provider therefore has the opportunity to simultaneously increase the price and the number of customers served, thus increasing his revenues.

For intermediate service rates (Region 2), increasing the service rate no longer increases the net value of the service for the customer, because the reduction in the service value is greater than the reduction in waiting costs. Therefore, at any given price in this region, increasing the service rate

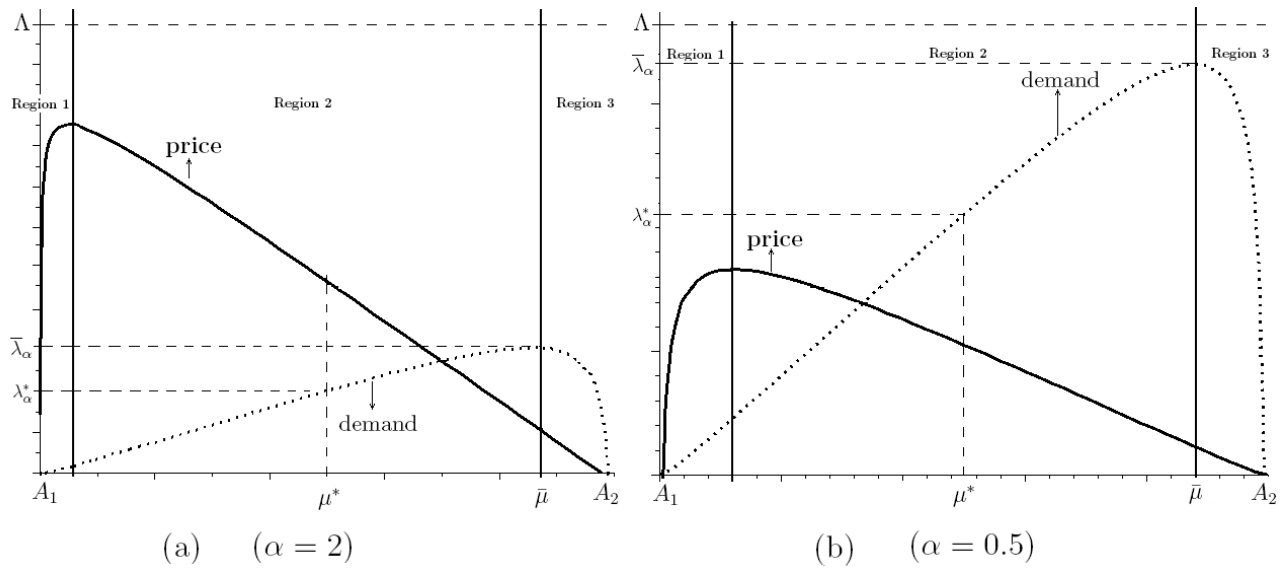

Figure 2: The figure illustrates the symmetry of $p^*(\mu)$ (denoted by the thick line) and $\lambda_e(\mu, p^*(\mu))$ (denoted by the dotted curve) around μ^* for customer-intensive services of types $\alpha = 2$ and $\alpha = 0.5$ (for $V_b = 10$ and $\mu_b = 5$). The optimal service rate μ^* and the corresponding equilibrium demand is $\lambda^*_\alpha = \lambda_e(\mu^*, p^*(\mu^*))$. The maximum throughput, induced by the service rate $\overline{\mu}$, is $\overline{\lambda}_\alpha$.

leads to lower equilibrium demand (and consequently, lower revenues). However, by *simultaneously* increasing the service rate *and* lowering the price, the service provider can increase the demand. As the service rate is increased in Region 2, the net effect of lower price and higher demand is to increase revenues up to the point μ^* (see Figure 2). Beyond this point, revenues start falling.

When the service rate provided is in Region 3, decreasing the service rate is desirable, as it leads to a higher price *and* higher equilibrium demand. In this region, the gain in service value from decreasing the service rate is greater than the losses accrued from the increase in customer waiting costs. Thus, as the service rate is reduced in Region 3, the equilibrium demand increases in spite of the increase in the equilibrium price.

To summarize, service rates in both Region 1 and Region 3 are untenable in equilibrium. The optimal service rate μ^* must lie in the intermediate service rate region (i.e., Region 2).

Figure 2 also illustrates the potential for service systems with very different service value propositions to earn identical revenues. A service provider may choose to provide high quality service at a high price to a limited number of customers, or it may provide lower quality service at a lower price to a large number of customers. Comparable revenues may be attained through either of these service strategies. Modeling customer-intensity through α allows us to capture the presence of such options in service provision.

3.2 Full Market Coverage

In Section 3.1, we analyzed the equilibrium in markets with partial coverage. Proposition 3 below derives the equilibrium in markets in which full coverage is possible. We find that customer intensity plays a similar, important role in these markets–thus, the insights for partial coverage derived in Section 3.1 continue to hold under full coverage.

Proposition 3. *For a customer-intensive service of type $\alpha > 0$, and when $\Lambda \leq \lambda^*_\alpha$,*

1. *The optimal service rate is equal to $\mu^* = \Lambda + \sqrt{c/\alpha}$*
2. *The corresponding optimal price is equal to $p^*(\mu^*) = V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c}$.*
3. *The equilibrium demand $\lambda_e(\mu^*, p^*(\mu^*))$ at the optimal price and service rate is Λ.*

Proposition 3.1 shows that, just as in the case of partial market coverage, the optimal service rate μ^* decreases in α. The service provider spends more time on each customer as the service becomes more customer-intensive. As one would expect for full coverage, $\lambda_e(\mu^*, p^*(\mu^*)) = \Lambda$; i.e., the service provider serves all customers in equilibrium (Proposition 3.3). Thus, as the customer-intensity α increases, the optimal service rate μ^* falls, while the equilibrium arrival rate remains unchanged at Λ. This leads to increased waiting costs as α increases. In fact, the expected waiting cost is $\sqrt{c\alpha}$. In equilibrium, customers wait longer (i.e., the congestion increases) as the service becomes more customer-intensive.

From Proposition 3.2, we see that the optimal price is convex in α. We first focus on the case of $\Lambda < \mu_b$. In this range, when $\alpha < \frac{c}{(\mu_b - \Lambda)^2}$, the optimal price is *decreasing* in α. Yet we saw that when α increases, the service provider increases the service time with every customer, which would increase the service value provided (Proposition 3.1). As α increases in this range of parameter values, the higher waiting cost (due to the increased service time) dominates the increase in service value, leading to a degradation in the net value of the service provided to customers. To accommodate this loss, the service provider has to cut the price as α increases. Thus, if we compare two services of low customer-intensity (with $\alpha < \frac{c}{(\mu_b - \Lambda)^2}$), the more customer-intensive service will be more congested but less expensive than the other.

However, when α is high ($> \frac{c}{(\mu_b - \Lambda)^2}$), the gains in service value are significant enough to offset the increase in waiting costs as α increases. Therefore, both the optimal price and the service time (or, service value) increase in α. Comparing two services that are both highly customer-intensive ($\alpha > \frac{c}{(\mu_b - \Lambda)^2}$), the service with higher α is both more expensive and more congested than the other.

Finally, the case when $\Lambda \geq \mu_b$ is similar to the first case above: The price is decreasing in α, while the service value is increasing in α, for the entire range of α. In this case, the greater the customer-intensity, the more congested but cheaper the service will be.

4. Model with Service Rate Competition

In this section, we consider the effect of multiple competing servers owned by a single service provider (firm) that provides a service of customer-intensity α. Although the service provider sets an admission price to maximize total revenues, the individual servers have the flexibility to set their own service speed / quality (for example, consider primary care physicians who belong to the same health network or the same hospital that, in turn, determines the admission price for patient visits). For ease of exposition, we initially restrict our attention to a firm with two servers and then show how our results extend to multiple servers. The firm sets the admission price p to maximize its total revenues. Each server individually decides its service rate to maximize its own revenues under the admission price p set by the firm. Arriving customers decide whether to join the system, and if they join, which server to go to, based on the service value offered by the servers, waiting costs at the servers, and the price. We focus on the Nash equilibrium of the system comprised of the firm, the servers and customers making all these decisions.

We model each server as an $M/M/1$ queueing regime.The queue joining decision of a customer is given by $\gamma_j(\mu_1, \mu_2, p, \Lambda)$, for $j = 0, 1, 2$, where $\gamma_0(\cdot)$ denotes the probability of balking, and $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ denote the probability of joining queue 1 and queue 2, respectively. Under pure strategies, i.e. $\gamma_i = 1$ for some i, either one server gains all the customers (Λ), or none of the servers serves any customers. We prove that none of these outcomes are possible in equilibrium. We thus focus on mixed strategies. Again, as in Section 3, we divide our analysis into two cases based on market coverage. When the market is sufficiently large, i.e. $\Lambda \geq 2 \cdot \lambda_\alpha^* = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{\alpha}$, we show that, in equilibrium, both servers choose their service rates as if they were monopolies, and the firm chooses the single-server monopoly price. When the market is small, i.e. when $\Lambda < \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{\alpha}$, we show that the firm chooses a price such that all of the consumer surplus is extracted, and the market is fully covered by the firm.

Proposition 4. *When the market is sufficiently large, given by the condition $\Lambda \geq 2\lambda_\alpha^* = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{\alpha}$, the servers act as monopolists. The optimal service rate set by server i is given by: $\mu_i^* = \frac{V_b+\alpha\mu_b}{2\alpha}$ for $i = 1, 2$. The firm's optimal price is $p^* = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2}$.*

Proposition 4 simply states that in a large enough market, the price charged by the service provider remains unaffected by competition within his network. All our insights on customer-intensive services, derived for the single-server monopoly in Section 3, continue to hold.

When the market is smaller, i.e. when $\Lambda < \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{\alpha}$, competition affects the servers' and the firm's strategies. The servers compete by adjusting their service rates, while the firm adjusts its admission price for the service. We find that the net values ($V(\mu_i) - cW(\mu_i, \lambda_i)$, for $i = 1, 2$) provided by the servers are equal and positive in equilibrium. Server i's equilibrium demand is $\lambda_i = \Lambda/2$; thus, the entire market is covered by the two servers (i.e., $\lambda_1 + \lambda_2 = \Lambda$). The service provider extracts the entire consumer surplus by charging an appropriate price p^*.

Proposition 5. *When the potential demand for the service is low, i.e., $\Lambda < 2 \cdot \lambda_\alpha^* = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{\alpha}$, the two servers share the market demand equally in equilibrium by setting their service rates to*

$\mu_i^e = \frac{\Lambda}{2} + \sqrt{c/\alpha}$, *for* $i = 1, 2$.

Proposition 5 shows that the equilibrium service rate μ_i^e is less than μ^*, the optimal service rate under monopoly (recall Proposition 3.1). Thus, under service-rate competition, the firm provides a higher service value at a slower rate through its servers than it would if there were only one server. Moreover, we find that customers' expected waiting costs in the multi-server case ($cW(\mu_i^e, \Lambda/2) = \sqrt{c\alpha}$) is *identical* to that under monopoly ($cW(\mu^*, \Lambda) = \sqrt{c\alpha}$). Therefore, the service value net of waiting costs increases under server-competition for market share.

Our structural results continue to hold when there are n (> 2) servers competing on service rates. We find that (i) In small markets ($\Lambda < n \cdot \lambda_\alpha^* = n\frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$), each additional server induces every server to slow down further; (ii) Otherwise, the market is large enough that each server acts as a local monopolist.

Proposition 6. *When there are n servers, full market coverage is assured for $\Lambda < n \cdot \lambda_\alpha^* = n\frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$. The service provider charges an admission price $p_n^* = V_b + \alpha\mu_b - \alpha\Lambda/n - 2\sqrt{c\alpha}$, which is strictly* increasing *and concave in the number of servers, n.*

Proposition 6 shows that under full market coverage, the firm's price (and therefore, the total revenues) *increases* with the number of competing servers. As a special case, we see that the equilibrium price under competing servers is greater than the monopoly price.

The results of Proposition 6 are driven by the impact of customer-intensity on service rates. For customer-intensive services, the greater the competition, the slower the equilibrium service rates chosen by the servers. Although the servers compete amongst themselves for customers, they choose to provide higher service value over faster service rates, which in turn allows the firm to charge higher admission prices.

In practice, there may be investment costs to hire and maintain servers. In such cases, the service provider needs to calculate the optimal number of servers as a trade-off between the additional revenues earned by adding servers and the incremental investment costs. Suppose the cost of additional servers is increasing and convex. Since the price p_n^* is increasing and concave in the number of competing servers (as established in Proposition 6), the service provider will increase the number of servers until the marginal revenue from adding one more server is exceeded by the marginal cost. Since p_n^* is increasing in α, the optimal number of servers is *ceteris paribus* increasing in customer-intensity.

5. Summary, Insights and Future Directions

We have argued that the results from traditional queueing models are not applicable to customer-intensive services, wherein the service quality is sensitive to the time spent with the customer. The tradeoff between *quality* and *speed* is at the crux of the service-provider's problem, and his choice of an *intermediate* service rate in the face of rational customers reflects this tradeoff.

Thus our model provides fundamentally new insights into the nature of customer-intensive services. In the discussion below, we focus on a couple of these insights and examine the related

empirical evidence in the context of primary care services. However, the conclusions from our model are applicable across a wide variety of industries.

Service Speed and Market Coverage:

An implication of service degradation with speed is that service-level (quality) targets are met only at slow service times, necessitating a larger investment in capacity/servers. This increases the costs of providing the service. Thus, *Baumol's cost disease* (discussed in Section 1) is a consequence of the customer-intensity of the service. What could exacerbate this disease is our analytical result that, as the service gets more customer-intensive, the service provider *slows down*, and increases the time spent with each customer.

For a highly customer-intensive service such as primary care, the service provider gains by focusing on service quality and spending adequate time with each patient, rather than by increasing throughput by speeding up the service. Paradoxically, this approach leads to greater revenues and service value. Recent empirical research findings in primary care services confirm our conclusions. Chen *et al* (2009) and Mechanic *et al* (2001) examine primary care visit data in the United States between 1989 and 2005, and show that primary care visit durations have increased (i.e., the average service rate is slower) with an accompanied increase in service value. It is optimal for firms providing primary care services to invest in high-quality, slower service, and therefore, partial market coverage is likely to be observed. It has been documented that an increasing fraction of the U.S. population resorts to emergency room visits due to the lack of adequate access to primary care (Pitts et al 2008).

Slowdown and longer services have also created a new primary care model, termed "concierge medicine". Concierge primary care practices announce and limit the number of patients they accept, and offer them highly customized primary care, spending as much time as needed with each patient, with minimal delays. In return, concierge physicians charge higher fees, that also have the effect of limiting the demand for the service, thus reducing congestion. For example, MDVIP (http://www.mdvip.com/) founded in 2000, is a national network of 250+ physicians who provide preventive and personalized health care. Concierge doctors affiliated to MDVIP care for a maximum of 600 patients each. MDVIP, MD^2 International (http://www.md2.com), Current Health (http://www.currenthealth.com), and Qliance Primary Care (http://www.qliance.com) are some leading concierge primary care firms in the market. The longer duration of patient care in concierge practice leads to the service provider providing more valuable service to a limited number of customers.

Partial market coverage is likely to be observed in other highly customer-intensive services, such as legal/financial consulting, educational services and other healthcare related services.

Pricing under Service Rate Competition:

We find that as the number of competing servers increases, every server slows down further. As a consequence, server competition enhances the service value delivered in equilibrium, while holding the equilibrium congestion (waiting) costs constant, and exerts an upward pressure on the price charged by the provider of the customer-intensive service. These results, which are in sharp contrast

to previous queuing research, are driven by customer-intensity.

For customer-intensive services such as primary care, adding service agents improves quality but may not reduce congestion. Service rate competition among the agents leads to the desirable outcome of higher quality but also to a higher admission price. In primary care settings, there is strong empirical evidence of higher prices when the number of primary care physicians in a market increases. This empirical finding in the seminal paper by Pauly and Satterthwaite (1981) has subsequently found support in several studies that confirm price increases due to competition in primary care service provision (see Gaynor and Haas-Wilson (1999) and references therein). The theoretical explanations offered for such observations of price increases have been based on tacit collusion and informational inefficiencies. In contrast to these explanations, our paper posits that such increases in price can emerge naturally under service competition, when the value of the service increases with the time spent in serving the customer.

Future Directions:

Several directions seem promising for future research. One extension could be to model and study the effects of different kinds of market heterogeneity. Competing servers could vary in their customer-intensities (e.g. based on their choices of patient-care models or investments in training agents). Whether *customer-intensity differentiation* is a viable competitive strategy or not is an interesting research question. A second extension would be to model multiple service providers that independently set their prices and service rates, which would require a model of full-price competition. Simplifying such a model in other ways (such as eliminating customer choice by assuming an exogenously specified joining rate) and/or employing other methodologies, such as computational approaches, might be required. Another interesting extension of this research would be to model information asymmetry – especially in customer-intensity. Presumably, there are occasions when customers do not know the exact content of the service offered. Debo *et al* (2008) model incentive effects in the context of such 'credence' services; similar issues are pertinent to customer-intensive services.

Acknowledgements: A previous version of this paper was a Finalist in the 2009 INFORMS Junior Faculty paper competition. The authors would like to thank the anonymous reviewers, an associate editor and the departmental editor for their valuable suggestions. Special thanks go to Philipp Afeche, Mor Armony, Baris Ata, John Birge, Gérard Cachon, Francis de Véricourt, Jack Hershey, Ananth Iyer, Ashish Jha, Hsiao-hui Lee, Raj Rajagopalan, Alan Scheller-Wolf, Robert Shumsky, Anita Tucker, Ludo van der Heyden, Eitan Zemel and Yong-Pin Zhou for their thoughts and discussions during various stages of the paper. We also thank seminar participants at Carnegie Mellon University, Cornell University, Northwestern University, University of Chicago, University of Maryland, University of Pennsylvania, Utah Winter Conference 2010, University of Rochester's Workshop on Information Intensive Services, the reviewers and participants of the 2009 Service SIG Conference at MIT, and the judges of the 2009 INFORMS JFIG Competition. We acknowledge financial support from the Fishman-Davidson Center at the Wharton School, University of Pennsylvania.

Appendix

Proof of Proposition 1: We begin by showing the optimal price, $p^*(\mu)$ for $\Lambda > A_2(\alpha)$. In this case, the service provider cannot serve all potential customers even when the price is equal to zero. The equilibrium arrival rate, $\lambda_e(\mu, p)$, is determined by the following equation in this case:

$$V(\mu) - p = cW(\mu, \lambda_e(\mu, p)). \tag{5}$$

The revenue of the service provider, $R(\mu, p)$, is given by:

$$R(\mu, p) = p\left(\mu - \frac{c}{V(\mu) - p}\right). \tag{6}$$

Recall that the service value is the upper-bound for the price, i.e. $V(\mu) \geq p$. Therefore, the revenue function is concave in the price, p, for the set of admissible prices (for $0 \leq p \leq V(\mu)$), as the second order condition is negative:

$$0 > \frac{\delta^2 R(\mu, p)}{\delta p^2} = -\frac{2c}{(V(\mu) - p)^2} - \frac{2pc}{(V(\mu) - p)^3}$$

The optimal price, maximizing the service provider's revenues for a given service rate μ, is found using the first order condition:

$$0 = \frac{\delta R(\mu, p)}{\delta p} = \mu - \frac{c}{V - p} - \frac{pc}{(V - p)^2}.$$

$p^*(\mu) = V(\mu) - \sqrt{cV(\mu)/\mu}$ is the unique solution of the first order condition within the set of admissible prices, $p \in [0, V(\mu)]$. Plugging $p^*(\mu)$ into equation (5) we find the resulting equilibrium arrival rate:

$$\lambda_e(\mu, p^*(\mu)) = \mu - \sqrt{\frac{c\mu}{V(\mu)}}.$$

The equilibrium arrival rate, $\lambda_e(\mu, p^*(\mu))$, is independent of the potential demand, Λ. This shows that the optimal price given service rate μ is equal to $V(\mu) - \sqrt{cV(\mu)/\mu}$ for all $\Lambda \geq \mu - \sqrt{\frac{c\mu}{V(\mu)}}$. So far, we have derived the optimal price p^* for all $\Lambda \geq \mu - \sqrt{\frac{c\mu}{V(\mu)}}$.

To complete the proof, we need to derive the optimal price for $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$. Note that the service provider can serve all potential customers at a non-negative price for $\Lambda \leq \mu - \sqrt{\frac{c\mu}{V(\mu)}}$. For a given service rate μ, the equilibrium demand, $\lambda_e(\mu, p)$, is decreasing in price. Therefore, the maximum number of customers that can be served (maximum throughput) at rate μ, $\bar{\Lambda}(\mu)$, is found by setting the price equal to zero. Using the following equation we find $\bar{\Lambda}(\mu)$.

$$V(\mu) = \frac{c}{\mu - \bar{\Lambda}(\mu)} \Rightarrow \bar{\Lambda}(\mu) = \mu - \frac{c}{V(\mu)}$$

If $\bar{\Lambda}(\mu)$ is greater than the potential demand Λ, then the service provider can serve all potential

customers, charging a price greater than zero. For $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$, $\bar{\Lambda}(\mu) > \Lambda$:
$\bar{\Lambda}(\mu) = \mu - \frac{c}{V(\mu)} \geq \mu - \sqrt{\frac{c\mu}{V(\mu)}} > \Lambda$, since $V(\mu) \geq \frac{c}{\mu}$ for all $\mu \in \mathcal{F}(\alpha)$.

For $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$, all arriving customers join the queue, if the net utility of joining when all others join the queue at price p is non-negative, i.e. $V(\mu) - p - cW(\mu, \Lambda) \geq 0$. The net utility decreases in price, and it is non-negative for $p \leq V(\mu) - \frac{c}{\mu - \Lambda}$. Thus the service provider's revenue as a function of price can be written as:

$$R(\mu, p) = \begin{cases} p\Lambda & if\ 0 \leq p \leq V(\mu) - cW(\mu, \Lambda) \\ p\left(\mu - \frac{c}{V(\mu) - p}\right) & if\ V(\mu) - cW(\mu, \Lambda) < p \leq V(\mu) - \frac{c}{\mu} \\ 0 & if\ p > V(\mu) - \frac{c}{\mu}, \end{cases} \tag{7}$$

Differentiating the revenue function with respect to price we get:

$$\frac{\delta R(\mu, p)}{\delta p} = \begin{cases} \Lambda & if\ 0 \leq p \leq V(\mu) - \frac{c}{\mu - \Lambda} \\ \mu - \frac{c}{V(\mu) - p} - \frac{pc}{(V(\mu) - p)^2} & if\ V(\mu) - \frac{c}{\mu - \Lambda} < p \leq V(\mu) - \frac{c}{\mu} \\ 0 & if\ p > V(\mu) - \frac{c}{\mu}, \end{cases} \tag{8}$$

The revenue, $R(\mu, p)$ is clearly increasing in price for $p \leq V(\mu) - \frac{c}{\mu - \Lambda}$. Increasing the price further at $p = V(\mu) - \frac{c}{\mu - \Lambda}$ will decrease the demand (throughput) but it may still increase the revenues. Note that the revenue function for $p > V(\mu) - \frac{c}{\mu - \Lambda}$ is equivalent to the revenue function given by equation (6), which is maximized at $p = V(\mu) - \sqrt{\frac{cV(\mu)}{\mu}}$. The revenues decrease in price at $p = V(\mu) - \frac{c}{\mu - \Lambda}$ because $V(\mu) - \frac{c}{\mu - \Lambda} > V(\mu) - \sqrt{\frac{cV(\mu)}{\mu}}$ for $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$:

$$\sqrt{\frac{cV(\mu)}{\mu}} = \frac{c}{\mu - (\mu - \sqrt{\frac{c\mu}{V(\mu)}})} > \frac{c}{\mu - \Lambda}.$$

As a result, the optimal price at service rate μ, for $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$ is $p^*(\mu) = V(\mu) - \frac{c}{\mu - \Lambda}$. The resulting equilibrium arrival rate is equal to $\lambda_e(\mu, p^*(\mu)) = \Lambda$.

$$p^*(\mu) = \begin{cases} V(\mu) - cW(\mu, \Lambda) & if\ 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\ V(\mu) - \sqrt{cV(\mu)/\mu} & if\ \hat{\lambda}(\mu) < \Lambda. \end{cases} \tag{9}$$

Thus we have derived p^* for all Λ.

■

Preparatory Results for Lemmas 1, 2 and Proposition 3: Before we prove the lemmas, we prove two main preparatory results.

Result 1: Service provider's revenue function $R(\mu, p)$ is non-decreasing in demand, Λ.

Proof: For a given service rate μ and price p, the service provider's revenue as a function of the potential demand, Λ, is given as follows:

$$R(\mu,p) \;=\; \begin{cases} p\Lambda & if\ cW(\mu,\Lambda) \le V(\mu)-p \\ p\left(\mu-\frac{c}{V(\mu)-p}\right) & if\ cW(\mu,0) < V(\mu)-p < cW(\mu,\Lambda) \\ 0 & if\ V(\mu)-p < cW(\mu,0). \end{cases} \tag{10}$$

$R(\mu,p)$ is continuous in Λ. To show this, we only need to show that the function is continuous at the transition point $cW(\mu,\Lambda) = V(\mu)-p$. $cW(\mu,\Lambda) = \frac{c}{\mu-\Lambda}$ is increasing in Λ for $\Lambda \le \mu$. Rewriting the transition point, $cW(\mu,\Lambda) = V(\mu)-p$, and solving for Λ we get: $\frac{c}{\mu-\Lambda} = V(\mu)-p \Rightarrow \Lambda = \mu - \frac{c}{V(\mu)-p}$. Which shows that $R(\mu,p)$ is continuous in Λ for $\Lambda \ge 0$.

Clearly, $R(\mu,p)$ is increasing in Λ for $cW(\mu,\Lambda) \le V(\mu)-p$ and constant in Λ for $cW(\mu,\Lambda) > V(\mu)-p$. This proves that $R(\mu,p)$ is non-decreasing in Λ for $\Lambda \ge 0$. ∎

Result 2: For $\Lambda \ge \bar{\lambda}_\alpha = max_{\{\mu\in\mathcal{F}(\alpha)\}}\{\lambda_e(\mu,p^*(\mu))\}$, the optimal price, $p^*(\mu)$, and the resulting equilibrium arrival rate, $\lambda_e(\mu,p^*(\mu))$, have the following symmetric relationship around $\beta = \left(\frac{V_b+\alpha\mu_b}{2\alpha}\right)$ for any $\mu \in \mathcal{F}(\alpha)$.

$$p^*(\beta+\epsilon) = \alpha\lambda_e(\beta-\epsilon, p^*(\beta-\epsilon)),$$

where $\epsilon = \mu - \frac{V_b+\alpha\mu_b}{2\alpha}$. The desired result is obtained by plugging in $\mu = (\beta+\epsilon)$ and $\mu = (\beta-\epsilon)$ and into $p^*(\mu)$ and $\lambda_e(\mu,p^*(\mu))$ (derived in Proposition 1) respectively.

Remark 1. *As long as the typical service value V_b is greater than the expected waiting cost during a typical service, $\frac{c}{\mu_b}$, i.e. $V_b > \frac{c}{\mu_b}$, we have a non-empty operating region, $\mathcal{F}(\alpha)$ for a customer-intensive service of type α.*

Having proven Results 1 and 2, we are now ready to prove the lemmas.

Proof of Proposition 2:

1. For $\Lambda > A_2(\alpha)$, the objective function in equation 3 is unimodal in the service rate, μ. The revenue function $R(\mu,p^*(\mu)) = \mu(V_b+\alpha\mu_b-\alpha\mu) - 2\sqrt{c\mu(V_b+\alpha\mu_b-\alpha\mu)} + c$ is continuous in μ for $\mu \in [A_1(\alpha), A_2(\alpha)]$.

The revenue function is differentiable in μ for $\mu \in [A_1(\alpha), A_2(\alpha)]$:

The first derivative $\frac{\delta R(\mu,p^*(\mu))}{\delta\mu} = V_b+\alpha\mu_b-2\alpha\mu - \frac{c(V_b+\alpha\mu_b)-2c\alpha\mu}{\sqrt{c\mu(V_b+\alpha\mu_b-\alpha\mu)}}$, exists and is continuous for $\mu \in [A_1(\alpha), A_2(\alpha)]$.

The first derivative, $\frac{\delta R(\mu,p^*(\mu))}{\delta\mu} = 0$, crosses 0 at three points; $A_1(\alpha)$, $\mu^* = \frac{V_b+\alpha\mu_b}{2\alpha}$ and $A_2(\alpha)$.

$\mu^* \in (A_1(\alpha), A_2(\alpha))$ for $\alpha, c > 0$. As a result, μ^* is either the unique maximizer or the unique minimizer of $R(\mu,p^*(\mu))$ for $\mu \in [A_1(\alpha), A_2(\alpha)]$. We show that μ^* maximizes the revenue function: $R(A_1(\alpha), p^*(A_1(\alpha)))$ and $R(A_2(\alpha), p^*(A_2(\alpha)))$ are equal to zero because both the optimal price, $p^*(A_i(\alpha))$, and the resulting equilibrium arrival rate, $\lambda_e(A_i(\alpha), p^*(A_i(\alpha)))$, are clearly equal to zero, since the service value, $V(\mu)$, is equal to the waiting cost during the service, c/μ, at these points.

Now, we show that $R(\mu^*, p^*(\mu^*))$ is greater than zero. $R(\mu^*, p^*(\mu^*)) = \frac{(V_b+\alpha\mu_b-2\sqrt{c\alpha})^2}{4\alpha} > 0$ for all $\alpha > 0$, since $V_b + \alpha\mu_b - 2\sqrt{c\alpha} > 0$ from Remark 1.

This shows that $R(\mu, p^*(\mu))$ is increasing in μ for $\mu \in (A_1(\alpha), \mu^*)$ and decreasing in μ for $\mu \in (\mu^*, A_2(\alpha))$, which proves that the optimal service rate is $\mu^* = \frac{V_b+\alpha\mu_b}{2\alpha}$.

2. From Proposition 1 $p^*(\mu^*) = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2}$.

3. From Proposition 1 $\lambda_e(\mu^*, p^*(\mu^*)) = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha}$.

This proves the optimality of the above service setting for $\Lambda > A_2(\alpha)$.

We now show that the above operating setting is optimal, even for all $\Lambda \geq \lambda_e^*(\mu^*, p^*(\mu^*)) = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha}$:

The revenue function $R(\mu, p)$ is non-decreasing in the potential demand, Λ, for all $\mu, p, \Lambda \geq 0$ from Result 1. Therefore, the optimal revenue $R(\mu^*, p^*(\mu^*))$ is non-decreasing in Λ for all $\Lambda \geq 0$. The service provider can achieve $R(\frac{V_b+\alpha\mu_b}{2\alpha}, \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2})$ by serving $\frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha}$ customers at rate $\mu = \frac{V_b+\alpha\mu_b}{2\alpha}$, charging price $p = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2}$ for all $\Lambda \geq \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha}$. In other words, the optimal revenue for $\Lambda > A_2(\alpha)$ can be achieved by the identical operating setting (price and service rate) when the potential demand is lower than $A_2(\alpha)$ ($\frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha} \leq \Lambda \leq A_2(\alpha)$).

The optimal revenues are non-decreasing in the potential demand, Λ, and therefore the above setting is optimal for all $\Lambda \geq \lambda_\alpha^* = \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha}$, since it is optimal for $\Lambda > A_2(\alpha)$ and $A_2(\alpha) > \frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha}$. ∎

Proof of Lemma 1:[α-symmetry] Proposition 2 indicates that β defined in Result 2 is equal to the optimal service rate, μ^*, for $\Lambda > \bar{\lambda}_\alpha$. The result of the Lemma immediately follows from Result 2, plugging in μ^* for β. ∎

Lemma 2. *For any $\alpha > 0$, the optimal price for a given service rate, $p^*(\mu)$, and the resulting equilibrium demand, $\lambda_e(\mu, p^*(\mu))$, are unimodal in the service rate, μ.*

Proof of Lemma 2: For $\Lambda > \bar{\lambda}_\alpha$, $p^*(\mu)$ and $\lambda_e(\mu, p^*(\mu))$ are unimodal (increasing and then decreasing) in the service rate, μ. We will prove that $p^*(\mu)$ is unimodal in μ. Unimodality of $\lambda_e(\mu, p^*(\mu))$ follows from the α-symmetry property. For $\Lambda > \bar{\lambda}_\alpha$, the optimal price for service rate $\mu \in \mathcal{F}(\alpha)$ is $p^*(\mu) = V(\mu) - \sqrt{\frac{cV(\mu)}{\mu}}$.

$p^*(\mu)$ is equal to zero for $\mu = A_i(\alpha)$ where $i = 1, 2$. We pick an interior point $\beta = \frac{V_b+\alpha\mu_b}{2\alpha}$ in the operating region, $\mathcal{F}(\alpha)$. Clearly, $\beta \in (A_1, A_2)$ for $\alpha, c > 0$.

The optimal price for β, $p^*(\beta)$ is non-negative as long as the condition in Remark 1 holds. For $\alpha, c \geq 0$,

$$p^*(\beta) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2} > 0 \Leftrightarrow (V_b + \alpha\mu_b)^2 - 4c\alpha > 0.$$

If $V_b > c/\mu_b$, then $(V_b + \alpha\mu_b)^2 - 4c\alpha > (V_b - \alpha\mu_b)^2 > 0$, since $c < V_b\mu_b$ from Remark 1.

We will prove the unimodality of $p^*(\mu)$ by showing that the first derivative $\frac{\delta p^*(\mu)}{\delta\mu}$ crosses 0 only once in $(A_1(\alpha), A_2(\alpha))$ and this point is the maximizer of the price, $p^*(\mu)$, with respect to μ for $\mu \in (A_1(\alpha), A_2(\alpha))$. Hence the first order condition is satisfied at a unique, interior point.

$FOC:\ \frac{\delta p^*(\mu)}{\delta\mu} = -\alpha + \frac{c(V_b+\alpha\mu_b)}{2\mu^2\sqrt{\frac{cV(\mu)}{\mu}}} = 0$

The first derivative is continuous for $\mu \in (A_1(\alpha), A_2(\alpha))$. Reorganizing the terms, we can write the first order condition as: $2\alpha\mu^2\sqrt{\frac{cV(\mu)}{\mu}} = c(V_b + \alpha\mu_b)$.

For notational convenience, let $(V_b + \alpha\mu_b) = K$. Therefore $V(\mu) = K - \alpha\mu$. Plugging in K and squaring both sides of the equation, we get: $4\alpha^2\mu^4\frac{c(K-\alpha\mu)}{\mu} = c^2K^2$

$$\Rightarrow \frac{cK^2}{4\alpha^2} = \mu^3(K - \alpha\mu). \tag{11}$$

Note that the left hand side of equation (11) is constant with respect to μ. We show that the right hand side crosses this constant only once for $\mu \in [A_1(\alpha), A_2(\alpha)]$.

The right hand side, $\mu^3(K - \alpha\mu)$, is unimodal in the service rate μ: $\frac{\delta}{\delta\mu}\mu^3(K - \alpha\mu) = \mu^2(3K - 4\alpha\mu)$.

Clearly the right hand side is increasing in μ for $\mu < \frac{3K}{4\alpha}$ and decreasing for $\mu > \frac{3K}{4\alpha}$.

We now show that the RHS term in equation (11) is less than $\frac{cK^2}{4\alpha^2}$ when $\mu = A_1(\alpha)$ and greater than $\frac{cK^2}{4\alpha^2}$ when $\mu = A_2(\alpha)$, which proves that the right hand side crosses the left hand side only once in the operating region, $\mathcal{F}(\alpha)$, since the right hand side is unimodal in μ.

If we plug in K for the value of $A_i(\alpha)$, we get:

$$A_i(\alpha) = \frac{K \mp \sqrt{K^2 - 4\alpha c}}{2\alpha} \quad for \quad i = 1, 2.$$

Let $RHS(\mu) = \mu^3(K - \alpha\mu)$. Let $LHS = \frac{cK^2}{4\alpha^2}$. Then we have:
$RHS(A_1(\alpha)) = \frac{c(K-\sqrt{K^2-4\alpha c})^2}{4\alpha^2} \leq LHS$ and $RHS(A_2(\alpha)) = \frac{c(K+\sqrt{K^2-4\alpha c})^2}{4\alpha^2} \geq LHS$,
which yield the desired result. Note that the weak inequalities in the above equations are strict if $\alpha > 0$, and that they are equalities if $\alpha = 0$.

The point satisfying the first order condition in the operating region is a maximizer since $p^*(A_1(\alpha)) = p^*(A_2(\alpha)) = 0$. Furthermore, the price is positive in the operating region, $\mathcal{F}(\alpha)$, since the optimal price is positive at an interior point β, i.e. $p^*(\beta) > 0$, and the first derivative of $p^*(\mu)$ crosses zero only once. Thus we have shown that $p^*(\mu)$ is unimodal (increasing and then decreasing) in the service rate μ. ■

Result 3: For a customer-intensive service of type $\alpha > 0$, the revenue maximizing, equilibrium demand $\lambda_e(\mu^*, p^*(\mu))$ is strictly lower than $\overline{\lambda}_\alpha = max_{\{\mu\in\mathcal{F}\}}\{\lambda_e(\mu, p^*(\mu))\}$. The service rate leading to $\overline{\lambda}_\alpha$ is greater than the optimal service rate, i.e., $\bar{\mu} = argmax_{\{\mu\in\mathcal{F}\}}\{\lambda_e(\mu, p^*(\mu))\} > \mu^*$. The equilibrium demand, $\lambda_e(\mu, p^*(\mu))$, is still increasing in the service rate, and the optimal price, $p*(\mu)$, is decreasing in the service rate, at the optimal service rate, μ^*.

Proof: The revenue maximizing equilibrium demand, $\lambda_e(\mu^*, p^*(\mu^*))$ is smaller than the maximum equilibrium demand, $\bar{\lambda}_\alpha = max_{\{\mu\in\mathcal{F}(\alpha)\}}\{\lambda_e(\mu, p^*(\mu))\}$. By definition we have $\bar{\lambda}_\alpha > \lambda_e(\mu^*, p^*(\mu^*))$.

We know that $\lambda_e(\mu, p^*(\mu))$ is unimodal in μ from Lemma 2. We first show that $\lambda_e(\mu, p^*(\mu))$ is increasing in the service rate at $\mu = \mu^*$.

Differentiating the equilibrium demand $\lambda_e(\mu, p^*(\mu))$ with respect to μ we get:

$$\frac{\delta\lambda_e(\mu, p^*(\mu))}{\delta\mu} = 1 - \frac{cK}{(K-\alpha\mu)\sqrt{c\mu(K-\alpha\mu)}}.$$

Evaluating the first derivative at μ^* we get: $\frac{\delta\lambda_e(\mu,p^*(\mu))}{\delta\mu}\Big|_{\mu=\mu^*} = 1 - 2\sqrt{\alpha c}/K$.
To prove the result, we need to show that $1 - 2\sqrt{\alpha c}/K \geq 0$. Re-organizing the equation, we get: $K^2 \geq 4\alpha c$, which holds for all $K, \alpha, c \geq 0$, which follows from Remark 1: Remark 1 suggests that $V_b > c/\mu_b$. Recall that $K = V_b + \alpha\mu_b$. Therefore, the condition in Remark 1 is equivalent to $(K - \alpha\mu_b)(K - V_b) > \alpha c$. Both μ_b and V_b are non-negative. Therefore we can rewrite the condition of Remark 1 as follows:

$$(K - V_b)V_b > c\alpha \;\; for\; V_b \in [0, K]. \tag{12}$$

Inequality 12 holds for all $V_b \in [0, K]$, and therefore it holds for the value of V_b that maximizes $(K - V_b)V_b$ in $[0, K]$. $(K - V_b)V_b$ is maximized when $V_b = K/2$.

$$max_{\{0\leq V_b\leq K\}}\{(K - V_b)V_b\} = K^2/4 \Rightarrow \frac{K^2}{4} > \alpha c.$$

Therefore, Remark 1 implies the result $K^2 \geq 4\alpha c$, which in turn implies that

$$\frac{\delta\lambda_e(\mu, p^*(\mu))}{\delta\mu}\Big|_{\mu=\mu^*} \geq 0.$$

Thus, we have shown that $\lambda_e(\mu, p^*(\mu))$ is increasing in the service rate at $\mu = \mu^*$. Therefore, the throughput maximizing service rate is $\bar{\mu} \geq \mu^*$. ∎

We begin with the following Lemma, characterizing the optimal price that clears the market demand, which helps us build conditions for Proof of Proposition 3.

Lemma 3. *For any customer-intensive service of type α, when the potential demand is such that $\Lambda < \overline{\lambda}_\alpha$, there exists $\bar{\mu}_1(\Lambda)$ and $\bar{\mu}_2(\Lambda)$ in $\mathcal{F}(\alpha)$, such that the optimal price $p^*(\mu)$ clears the market for all $\mu \in [\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$.*

Proof of Lemma 3: When the potential demand $\Lambda < \bar{\lambda}_\alpha$, there exists $\mu_1(\Lambda)$ and $\mu_2(\Lambda)$ in $\mathcal{F}(\alpha)$ such that $\lambda_e(\mu, p^*(\mu)) \geq \Lambda$ for $\mu \in [\mu_1(\Lambda), \mu_2(\Lambda)]$. The result of the Lemma follows from the fact that $\lambda_e(A_i(\alpha), p^*(A_i(\alpha))) = 0$ for $i = 1, 2$ and from the unimodality of $\lambda_e(\mu, p^*(\mu))$ (Lemma 2). ∎

Lemma 3 shows that for low values of potential demand, there exists a closed interval of service rates $[\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)] \subset \mathcal{F}(\alpha)$, where it is optimal to clear the market. A corresponding price $p^*(\mu)$ can be chosen so that the market demand is cleared for any service rate μ in this interval. Note that $\Lambda = \lambda_e(\bar{\mu}_1(\Lambda), p^*(\bar{\mu}_1(\Lambda))) = \lambda_e(\bar{\mu}_2(\Lambda), p^*(\bar{\mu}_2(\Lambda)))$.

If the potential demand, Λ, is higher than $\overline{\lambda}_\alpha$, then the service provider cannot clear the market at any service rate μ in the operating region $\mathcal{F}(\alpha)$. Using the results of Lemma 3 and Proposition

1, we can write the resulting equilibrium arrival rate as:

$$\lambda_e(\mu, p^*(\mu)) = \begin{cases} \hat{\lambda}(\mu) = \mu - \sqrt{\frac{c\mu}{V(\mu)}} & if\ A_1(\alpha) \le \mu < \bar{\mu}_1(\Lambda) \\ \Lambda & if\ \bar{\mu}_1(\Lambda) \le \mu \le \bar{\mu}_2(\Lambda) \\ \hat{\lambda}(\mu) = \mu - \sqrt{\frac{c\mu}{V(\mu)}} & if\ \bar{\mu}_2(\Lambda) \le \mu \le A_2(\alpha), \end{cases} \tag{13}$$

A representative example of the equilibrium demand leading to full market coverage can be seen in the right panel of Figure 2. Note that the service provider covers the entire demand Λ in the interval $[\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$. Again applying Proposition 1, we can rewrite the optimal price, $p^*(\mu)$, for a given service rate as follows:

$$p^*(\mu) = \begin{cases} V(\mu) - \sqrt{cV(\mu)/\mu} & if\ A_1(\alpha) \le \mu < \bar{\mu}_1(\Lambda) \\ V(\mu) - cW(\mu, \Lambda) & if\ \bar{\mu}_1(\Lambda) \le \mu \le \bar{\mu}_2(\Lambda) \\ V(\mu) - \sqrt{cV(\mu)/\mu} & if\ \bar{\mu}_2(\Lambda) \le \mu \le A_2(\alpha). \end{cases} \tag{14}$$

Therefore, the equilibrium revenue equals $R(\mu, p^*(\mu)) = p^*(\mu)\lambda_e(\mu, p^*(\mu))$. We can now derive the service provider's optimal service rate and price using the above revenue function.

Proof of Proposition 3: We will show that when the demand is low, ($\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$) the service provider's optimal service rate is $\mu^* = \Lambda + \sqrt{c/\alpha}$, the price is $p^*(\mu^*) = V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c}$, and the optimal equilibrium demand is equal to Λ (i.e. the market coverage is full).

1. For the small market scenario, the service provider's objective function is given by:

$$R(\mu, p^*(\mu)) = \begin{cases} \mu(K - \alpha\mu) - 2\sqrt{c\mu(K - \alpha\mu)} + c & if\ A_1(\alpha)\mu < \bar{\mu}_1(\Lambda) \\ (V(\mu) - \frac{c}{\mu - \Lambda})\Lambda & if\ \bar{\mu}_1(\Lambda) \le \mu \le \bar{\mu}_2(\Lambda) \\ \mu(K - \alpha\mu) - 2\sqrt{c\mu(K - \alpha\mu)} + c & if\ \bar{\mu}_2(\Lambda) \le \mu < A_2(\alpha). \end{cases} \tag{15}$$

The objective function is continuous in μ as $\Lambda = \lambda_e(\mu, p^*(\mu))$ and $V(\mu) - cW(\mu, \Lambda) = p^*(\mu)$ at $\bar{\mu}_1(\Lambda)$ and $\bar{\mu}_2(\Lambda)$, which implies that the revenues are equal at the transition points between regions.

Recall that Lemma 3 shows that there exists $\bar{\mu}_1(\Lambda)$ and $\bar{\mu}_2(\Lambda)$ such that all potential customers join the queue at the optimal price, $p^*(\mu)$, for all $\mu \in [\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$ when $\Lambda < \bar{\lambda}_\alpha$. $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$ and $\bar{\Lambda} > \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$ from Result 3. Therefore $A_1(\alpha) < \bar{\mu}_1(\Lambda) < \bar{\mu}_2(\Lambda) < A_2(\alpha)$ in the small market scenario.

Let Region A be $A_1(\alpha) < \mu \le \bar{\mu}_1(\Lambda)$, Region B be $\bar{\mu}_1(\Lambda) < \mu \le \bar{\mu}_2(\Lambda)$ and Region C be $\bar{\mu}_2(\Lambda) < \mu < A_2(\alpha)$. We will show that the optimal service rate is in Region B, for $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.

Note that in Region A and Region C the objective function is equivalent to that of the large market scenario ($R(\mu, p^*(\mu)) = \mu(K - \alpha\mu) - 2\sqrt{c\mu(K - \alpha\mu)} + c$), which is maximized at $\mu = \frac{K}{2\alpha}$ (Proposition 2).

$\bar{\mu}_2(\Lambda) = max\{\mu | \lambda_e(\mu, p^*(\mu)) = \Lambda\}$, is greater than $\frac{K}{2\alpha}$ since $\lambda_e(\mu, p^*(\mu))$ is unimodal by Lemma 2 and $\bar{\mu} = argmax_{\{\mu \in \mathcal{F}(\alpha)\}}\{\lambda_e(\mu, p^*(\mu))\} > \frac{K}{2\alpha}$ by Result 3. Therefore, the objective function in equation (15) is decreasing in μ in Region C.

We show that the objective function is increasing in μ in Region A, by showing that $\bar{\mu}_1(\Lambda) < \frac{K}{2\alpha}$. By definition $\lambda_e(\bar{\mu}_1(\Lambda), p^*(\bar{\mu}_1(\Lambda))) = \Lambda$. $\lambda_e(\mu, p^*(\mu))$ is unimodal in μ from Lemma 2 and $\arg\max_{\{\mu \in \mathcal{F}(\alpha)\}}\{\lambda_e(\mu, p^*(\mu))\} > \frac{K}{2\alpha}$ from Result 3. These facts imply that for $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha} = \lambda_e(\frac{K}{2\alpha}, p^*(\frac{K}{2\alpha}))$, $\bar{\mu}_1(\Lambda) < \frac{K}{2\alpha}$.

As a result, the service rate that maximizes the objective function is in Region B. Differentiating the objective function, $R(\mu, p^*(\mu))$, with respect to the service rate we get:

$$\frac{\delta R(\mu, p^*(\mu))}{\delta \mu} = \begin{cases} (K - 2\alpha\mu)\left[1 - \frac{c}{\sqrt{c\mu(K-\alpha\mu)}}\right] & if\ A_1(\alpha) < \mu < \bar{\mu}_1(\Lambda) \\ \Lambda\left(\frac{c}{(\mu-\Lambda)^2} - \alpha\right) & if\ \bar{\mu}_1(\Lambda) \le \mu \le \bar{\mu}_2(\Lambda) \\ (K - 2\alpha\mu)\left[1 - \frac{c}{\sqrt{c\mu(K-\alpha\mu)}}\right] & if\ \bar{\mu}_2(\Lambda) \le \mu < A_2(\alpha). \end{cases} \quad (16)$$

The first order condition is given by:

$$FOC: \quad 0 = \frac{\delta R(\mu, p^*(\mu))}{\delta \mu} = \Lambda\left(\frac{c}{(\mu-\Lambda)^2} - \alpha\right).$$

Recall that $\Lambda = \lambda_e(\mu, p^*(\mu)) = \mu - \sqrt{\frac{c\mu}{K-\alpha\mu}}$ at $\mu = \bar{\mu}_i(\Lambda)$ for $i = 1, 2$. Therefore we can write $\bar{\mu}_1(\Lambda) = \Lambda + \sqrt{\frac{c\bar{\mu}_1(\Lambda)}{K-\alpha\bar{\mu}_1(\Lambda)}}$. Plugging this value into $\frac{\delta R(\mu, p^*(\mu))}{\delta \mu}$ we get:

$$\left.\frac{\delta R(\mu, p^*(\mu))}{\delta \mu}\right|_{\mu=\bar{\mu}_1(\Lambda)} = \Lambda\left[\frac{K - 2\alpha\bar{\mu}_1(\Lambda)}{\bar{\mu}_1(\Lambda)}\right].$$

The value of $\frac{\delta R(\mu, p^*(\mu))}{\delta \mu}$ is positive for $\bar{\mu}_1(\Lambda) \le \frac{K}{2\alpha}$. $\bar{\mu}_1(\Lambda)$ is monotonically increasing in Λ for $0 \le \Lambda \le \bar{\Lambda}$ since $\lambda_e(\mu, p^*(\mu))$ is unimodal (increasing and then decreasing) in μ (Lemma 2). Therefore, the objective function is increasing in μ at $\mu = \bar{\mu}_1$, since $\bar{\mu}_1(\Lambda) \le \frac{K}{2\alpha}$. This proves that the optimal service rate, μ^*, is greater than $\bar{\mu}_1(\Lambda)$ for low potential demand ($\Lambda < \frac{K-\sqrt{c\alpha}}{2\alpha}$).

Similarly, we show that $R(\mu, p^*(\mu))$ is decreasing in μ at $\mu = \bar{\mu}_2(\Lambda)$ by plugging in $\Lambda + \sqrt{\frac{c\bar{\mu}_2(\Lambda)}{K-\alpha\bar{\mu}_2(\Lambda)}}$ for $\bar{\mu}_2(\Lambda)$:

$$\left.\frac{\delta R(\mu, p^*(\mu))}{\delta \mu}\right|_{\mu=\bar{\mu}_2(\Lambda)} = \Lambda\left[\frac{K - 2\alpha\bar{\mu}_2(\Lambda)}{\bar{\mu}_2(\Lambda)}\right] < 0.$$

The above value is negative since $\bar{\mu}_2(\Lambda) > \frac{K}{2\alpha}$. Therefore, an interior point of $[\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$ satisfies the first order condition for $\Lambda < \frac{K}{2\alpha} - \sqrt{c/\alpha}$. The unique solution of the first order condition for $\mu \in [\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$ is $\mu^* = \Lambda + \sqrt{c/\alpha}$, which proves the result of Proposition 3.

2. The optimal price, $p^*(\mu^*)$, is equal to $V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c}$, from Proposition 1.

3. The resulting equilibrium demand, $\lambda_e(\mu^*, p^*(\mu^*))$, is equal to the potential demand Λ by Proposition 1. ∎

Proof of Proposition 4: Proposition 4 shows that for $\Lambda \ge \frac{V_b + \alpha - 2\sqrt{c\alpha}}{\alpha}$, competing servers can achieve monopoly revenues, by using the optimal monopoly operating setting, μ^* and $p^*(\mu^*)$. Note that the potential demand Λ must be at least two times the optimal monopoly equilibrium demand,

$\lambda_e(\mu^*, p^*(\mu^*))$.

We begin by showing that for any potential demand $\Lambda \geq 0$, a server serving at rate μ and charging price p is better off in the single server setting than in the multi-server setting. Recall that the revenue of the server in the single server setting is given by $R(\mu, p)$. Let $R_1(\mu_1, \mu_2, p)$ be the revenue of server 1, providing service with rate μ_1, when server 2 is providing service with rate μ_2, at price p in the two server competition case. $R(\mu, p) \geq R_1(\mu, \mu_2, p)$ for all $\mu, \mu_2 \in \mathcal{F}(\alpha)$ and $p \geq 0$.

$$R_1(\mu, \mu_2, p) = \begin{cases} p\Lambda & if\ V(\mu) - p - cW(\mu, \Lambda) \geq max\{0, V(\mu_2) - p - cW(\mu_2, 0)\} \\ p\lambda_{1e}(\mu, \mu_2, p, \Lambda) & if\ cW(\mu, \Lambda) > V(\mu) - p \geq c/\mu \\ 0 & if\ V(\mu) - p - cW(\mu, 0) \leq max\{0, V(\mu_2) - p - cW(\mu, \Lambda)\}. \end{cases} \quad (17)$$

where $\lambda_{1e}(\mu_1, \mu_2, p, \Lambda)$ is the equilibrium arrival rate for server 1.

Let us examine the first line in equation (17). In the multi-server competition setting, server 1 can serve all potential customers if the net value of an arriving customer from joining server 1 when all other customers join server 1, is non-negative and greater than the net value of joining server 2 when no other customer joins server 2, i.e. $V(\mu) - p - cW(\mu, \Lambda) \geq max\{0, V(\mu_2) - p - cW(\mu_2, 0)\}$. Recall that in the single server setting, non-negativity of the net value ($V(\mu) - p - cW(\mu, \Lambda) \geq 0$) is sufficient to serve all potential customers. Clearly server 1 is better off in a single server setting than in a multi server setting when $V(\mu) - p - cW(\mu, \Lambda) \geq 0$.

Let us examine the case in which $cW(\mu, \Lambda) > V(\mu) - p \geq c/\mu$ (line 2 in Equation (17)). In the single server setting, customers join the queue until the net utility from joining equals zero. However in a multi-server setting, the net utility of joining server 2 may be positive when $\lambda_e(\mu, p)$ customers join server 1 and $\Lambda - \lambda_e(\mu, p)$ customers join server 2. Therefore, customers will deviate to server 2 until an equilibrium is reached, i.e. until the net utility from joining server 1 equals the net utility from joining server 2.

Hence, $R(\mu, p) \geq R_1(\mu, \mu_2, p)$ for all $\mu, \mu_2 \in \mathcal{F}(\alpha)$ and $p \geq 0$, which implies:

$$max_{\{\mu \in \mathcal{F}(\alpha), p \geq 0\}}\{R(\mu, p)\} \geq \max_{\{\mu \in \mathcal{F}(\alpha), p \geq 0\}}\{R_1(\mu, \mu_2, p)\} \quad (18)$$

for all $\Lambda \geq 0$.

Therefore, in the multi server setting when $cW(\mu, \Lambda) > V(\mu) - p \geq c/\mu$, the equilibrium arrival rate for server 1, $\lambda_{1e}(\mu, \mu_2, p, \Lambda)$, is less than or equal to the monopoly equilibrium arrival rate, $\lambda_e(\mu, p) = \left(\mu - \frac{c}{V(\mu) - p}\right)$.

For $\Lambda \geq \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, the optimal equilibrium demand in the single server setting is given by $\lambda_e(\mu^*, p^*(\mu^*)) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.

In the two server setting, the service provider can serve $2\lambda_e(\mu^*, p^*(\mu^*))$ customers, charging $p^*(\mu^*)$ when both servers serve at rate $\mu_i^* = \mu^*$, hence doubling the optimal monopoly revenues. In this case, the equilibrium demand at each server is equal to $\lambda_e(\mu^*, p^*(\mu^*))$. Clearly, the net value

of an arriving customer from joining either server is equal to zero, therefore an arriving customer is indifferent among the options to join server 1, join server 2 and to balk from the service. Hence this setting is an equilibrium for customers.

When the service provider charges price $p^*(\mu^*)$, (μ^*, μ^*) is a Nash Equilibrium for the servers, since they achieve the maximum revenue (shown in the LHS of the equation (12)) by choosing μ^*. (i.e. μ^* is the best response of a server to any service rate adopted by the other server).

Proof of Proposition 5: Proposition 5 indicates that for $\Lambda < \frac{V_b+\alpha-2\sqrt{c\alpha}}{\alpha}$, there exists a symmetric Nash equilibrium (μ^e, μ^e) such that all potential customers procure the service and agents equally share the potential demand. Further, we have $\mu^e = \frac{\Lambda}{2} + \sqrt{c/\alpha}$.

We prove the proposition by showing that none of the players (the servers and customers) have incentive to deviate from μ^e. An arriving customer's net value from joining server i when half of the customers join server i is given by: $NV_i(\mu_i, p, \Lambda/2) = V(\mu_i) - p - \frac{c}{\mu_i - \Lambda/2}$.

Let μ^e maximize $NV_i(\mu_i, p, \Lambda/2)$: Differentiating $NV_i(\mu_i, p, \Lambda/2)$ with respect to μ_i we get:

$$\frac{\delta NV_i(\mu_i, p, \Lambda/2)}{\delta \mu_i} = -\alpha + \frac{c}{(\mu_i - \Lambda/2)^2}$$

The second order condition indicates that $NV_i(\mu_i, p, \Lambda/2)$ is concave in μ_i for $\mu_i \geq \Lambda/2$:

$$\frac{\delta^2 NV_i(\mu_i, p, \Lambda/2)}{\delta \mu_i^2} = \frac{-2c}{(\mu_i - \Lambda/2)^3} < 0.$$

The first order condition is satisfied at $\mu^e = \frac{\Lambda}{2} + \sqrt{c/\alpha}$, hence μ^e maximizes $NV(\mu_i, p, \Lambda/2)$. The resulting net value is given by:

$$NV(\mu^e, p, \Lambda/2) = V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha} - p.$$

Clearly the net value for an arriving customer from joining server i is non-negative for $p \leq V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$, when agents serve at rate μ^e and customers mix equally between servers 1 and 2.

It can be shown that, server i, serving $\Lambda/2$ customers, has no incentive to deviate from μ^e, because doing so decreases the net value for an arriving customer joining server i, which will result in lower equilibrium demand for the server i. Given this, it follows that (μ^e, μ^e) is a Nash equilibrium for service agents for all $p \leq V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$.

The maximum price that can be charged by the service provider is $p_2^* = V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$. The service provider optimizes the revenues by charging p_2^* and fully extracting the consumer surplus. This result extends to n servers, and holds for all $\Lambda < n\frac{V_b+\alpha-2\sqrt{c\alpha}}{2\alpha}$.

■

Proof of Proposition 6: To prove the Proposition, we first derive the equilibrium service rate for n agents and the resulting optimal price, when $\Lambda < n\frac{V_b+\alpha\mu_b-2\sqrt{c\alpha}}{2\alpha}$.

We use the result of Proposition 5 to derive the service rates under the Nash equilibrium. Agents maximize the net value $NV(\mu_i, p, \Lambda/n)$. Therefore, the equilibrium service rate is $\mu^e(n) =$

$\frac{\Lambda}{n} + \sqrt{c/\alpha}$. The service provider then fully extracts the consumer surplus, and the resulting optimal price is thus given by:

$$p_n^* = V_b + \alpha\mu_b - \alpha\Lambda/n - 2\sqrt{c\alpha}.$$

Clearly p_n^* is increasing in the number of agents, n.

The marginal increase in price is given by:

$$p_{(n+1)}^* - p_n^* = \frac{\alpha\Lambda}{n(n+1)},$$

which is decreasing in the number of servers, n, and approaches zero as $n \to \infty$. ■

References

Afeche, Philipp, 2006. Incentive-compatible Revenue Management in Queueing systems: Optimal Strategic delay and other Delay Tactics. University of Toronto working paper.

Allon, G., A. Federgruen. 2007. Competition in Service Industries. *Opns. Res.*, **55**(1), 37–55.

Armony M., M. Haviv. 2000. Price and Delay Competition between Two Service Providers. *European Journal of Operational Research*, **147**(1), 32–50.

Baumol, W. J. 1993. Health Care, Education, and the Cost Disease: A Looming Crisis for Public Choice. *Public Choice*, **77**(1), 17–28.

Brahimi, M., D. J. Worthington. 1991. Queueing models for out-patient appointment systems – A case study. *Journal of Operational Research Society*, **42**(9), 733—746.

Cachon, G., P. Harker. 2002. Competition and Outsourcing with Scale Economies. *Management Science*, **48**(10), 1314–1333.

Chase, R.B., D.A. Tansik. 1983. The Customer Contact Model for Organization Design. *Management Science*, **29**(9), 1037–1050.

Chen, L. M., W. R. Farwell, A. K. Jha. 2009. Primary Care Visit Duration and Quality: Does Good Care Take Longer? *Archives of Internal Medicine*, **56**(6), 1866–1872.

Chen, H., M. Frank. 2004. Monopoly Pricing When Customers Queue, *IIE Transactions*, **36**(6), 569–581.

Chen, H., Y. Wan. 2003. Price competition of make-to-order firms, *IIE Trans.*, **35**(9), 871–832.

de Vericourt, F., P. Sun. 2009. Judgement Accuracy Under Congestion In Service Systems. Duke University working paper.

de Vericourt, F., Y. Zhou. 2005. Managing Response Time in a Call-Routing Problem with Service Failure. *Operations Research*, **53**(6), pp. 968-981

Debo, L., L.B. Toktay, L. N. Van Wassenhove. 2008. Queueing for Expert Services. *Management Science*, **54**(8), 1497–1512.

Edelson, N.M., D.K. Hildebrand. 1975. Congestion Tolls for Poisson Queuing Process. *Econometrica*, **43**(1), 81–92.

Gans, N. 2002. Customer Loyalty and Supplier Quality Competition. *Management Science*, **48**(2), 207–221.

Gaynor, M., D. Haas-Wilson. 1999. Change, Consolidation, and Competition in Health Care Markets. *The Journal of Economic Perspectives*, Vol. 13, No. 1, pp. 141–164.

Gilbert, S. M., Z. K. Weng. 1998. Incentive Effects Favor Non-Consolidating Queues in a Service System: The Principal Agent Perspective. *Management Science*, **44**(12), 1662–1669.

Green, L., S. Savin. 2008. Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research*, **56**(6), 1526–1538.

Hasija, S., E. Pinker, R.A. Shumsky. 2009. Work Expands to Fill the Time Available: Capacity Estimation and Staffing under Parkinson's Law. *M& SOM*, **12**(1), 1–18.

Hassin, R., M. Haviv. 2003. To Queue or not to Queue: Equilibrium behavior in queuing systems. Kluwer Academic Publishers, Norwell, MA.

Hopp, W.J., S. M. R. Iravani, G. Y. Yuen. 2007. Operations Systems with Discretionary Task Completion. *Management Science*, **53**(1), 61–77.

Kalai, E., M. Kamien, M. Rubinovitch. 1992. Optimal Service Speeds in a Competitive Environment. *Management Science*, **38**(8), 1154–1163.

Kaminetsky, B. 2004. Testimony to the Joint Economic Committee of the United States Congress. Consumer-Directed Doctoring: The Doctor is in, Even if Insurance is Out. Wednesday, April 28, 2004.

Kostami, V., S. Rajagopalan. 2009. Speed Quality Tradeoffs in a Dynamic Model. University of Southern California working paper.

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

Li, L. 1992. The role of inventory in delivery time-competition. *Management Science*, **38**(2) 182—197.

Li, L., Y. S. Lee. 1994. Pricing and Delivery-Time Performance in a Competitive Environment. *Management Science*, **40**(5), 633–646.

Lovejoy, W., K. Sethuraman. 2000. Congestion and Complexity Costs in a Plant with Fixed Resources that Strives to Make Schedule. *M & SOM* **2**(3), 221—239.

Lovelock, C. 2001. Service Marketing, People, Technology and Strategy. Prentice Hall, Upper Saddle River, NJ.

Lu, L., J. Van Mieghem, C. Savaskan. 2008. Incentives for Quality Through Endogenous Routing. *M & SOM*, **11**(2), 254–273.

Mechanic, D., D. McAlpine, M. A. Rosenthal. 2001. Are patients' visits with physicians getting shorter? *New England Journal of Medicine*, **344**(3), 198–204.

Mendelson, H., S. Whang. 1990. Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue. *Operations Research*, **38**(5), 870–893.

Naor, P. 1969. The Regulation of Queue Sizes by Levying Tolls. *Econometrica*, **37**(1), 15–24.

Oliva, R., R. J. Sterman. 2001. Cutting Corners and Working Overtime: Quality Erosion in the Service Industry. *Management Science*, **47**(7), 894–914.

Pauly, M. V., M. A. Satterthwaite. 1981. The Pricing of Primary Care Physicians Services: A Test of the Role of Consumer Information. *The Bell Journal of Economics*, **12**(2), 488–506.

Pitts, S. R., R. W. Niska, J. Xu, C. W. Burt. 2008. National Hospital Ambulatory Medical Care Survey: 2006 Emergency Department Summary. *National Health Statistics Reports*, **7**. Dated August 6, 2008. http://www.cdc.gov/nchs/data/nhsr/nhsr007.pdf

Png, I., D. Reitman. 1994. Service Time Competition. *RAND Journal of Economics*, **25**(4), 619–634.

Ren, Z. J., X. Wang. 2008. Should Patients be Steered to High Volume Hospitals? An Empirical Investigation of Hospital Volume and Operations Service Quality. Boston University working paper.

Ross, S.M. 2006. Introduction to Probability Models. Academic Press. Ninth Edition.

Surowiecki, J. 2003. What Ails us? *The New Yorker*, July 7, 2003.

Triplett, J. E., B. P. Bosworth. 2004. Productivity in the U.S. Services Sector, New Sources of Economic Growth. Brookings Institution Press, Washington, D.C.

Varian, H. 2004. Economic Scene; Information Technology May Have Been What Cured Low Service-Sector Productivity. *NY Times*, Published: February 12, 2004.

Veeraraghavan, S., L. Debo. 2009. Joining Longer Queues: Information Externalities in Queue Choice. *Manufacturing and Service Operations Management*, **11**(4), 543–562.

Wang, X., L. Debo, A. Scheller-Wolf, S. Smith. 2010. Design and Analysis of Diagnostic Service Centers. *Management Science*, Forthcoming.

Yarnall, K. S. H., K. I. Pollak, T. Ostbye, K. M. Krause and J. L. Michener, 2003. Primary Care: Is There Enough Time for Prevention? *American Journal of Public Health*, **93**(4), 635-641.