Controlling Congestion when Consumers Choose their Service Times

Pnina Feldman and Ella Segev *

October, 2018

Abstract

Problem Definition: Two main challenges that service providers face when managing service systems is how to generate value and control congestion at the same time. **Academic/Practical Relevance:** To this end, classical queueing models suggest managers charge a per-use fee and invest in capacity to speed up the service. However, in discretionary services where customers value time in service and choose how long to stay, per-use fees result in suboptimal performance and speeding up does not apply. We consider two alternative mechanisms: price rates that are duration-based fees and time limits. We derive revenue and welfare implications and highlight operational consequences of these mechanisms. **Methodology:** We employ a queueing model of a service provider and rational consumers who are heterogeneous in their requirements for service durations, though they may be flexible in deviations from their time needs. Customers incur disutility from waiting and choose whether to join and how long to stay in service. Service providers can charge per-use fees or price rates and may also decide to limit customers' time in service. **Results:** Price rates and time limits benefit providers in different ways. Price rates are effective because of their superiority in extracting rents from heterogeneous customers and time limits successfully regulate congestion. Revenue maximizing firms benefit from implementing both. Social planners who seek to maximize consumer welfare, however, should focus on regulating congestion and should therefore offer the service for free, but implement time limits. **Managerial Implications:** Time limits reduce waiting times and benefit both firms and consumers, especially if congestion and consumer flexibility are high. Thus, we conclude that providers of discretionary services should not shy away from setting time limits.

1 Introduction

In many service systems customers choose the duration of their service, they value longer service, but also dislike waiting for the service to commence. Such services are often referred to as "discretionary services", because service time is not exogenous, as in classical service systems, but chosen by customers (Hopp *et al.* (2007)). For example, people who go to the gym to exercise choose how long they want to run on the treadmill, value longer time, but dislike waiting for the treadmill to become available. Drivers that secure a parking spot choose how long they want to park, value longer parking, but don't like circling around in

*Feldman: Questrom School of Business, Boston University, pninaf@bu.edu; Segev: Ben Gurion University of the Negev, ellasgv@bgu.ac.il

search of a parking spot; customers who come to a coffee shop to work enjoy using the wifi and choose how long they stay there, but dislike waiting for a table to become available; and visitors to a museum don't like waiting in line to view a popular exhibit, but would prefer spending more time at the museum once they gain admission. All of these examples illustrate that when customers value time in service and choose how long to stay in service, it can cause congestion and delay in the system.

How, then, can a service provider control congestion in this setting? The most natural way to control congestion in classical service systems is to invest in fast capacity (e.g., hire a more capable agent). Unfortunately, this solution does not apply to settings where customers choose how long to stay in service. Another way to regulate congestion is to charge a per-use fee or an entry toll for service. It has been shown to work well for both the firm and a social planner (Naor (1969)). However, other mechanisms may be more effective when consumers value the time spent in service and differ in their service needs. In particular, we examine two such mechanisms: charge customers based on how long they stay in service ("price rate") and impose time limits which put a cap on the maximum time spent in service. These mechanisms are often applied in practice separately or together. To return to our previous examples, some gyms impose limits on the use of exercise equipment (e.g., allow at most 30 minutes on the treadmill); parking meters limit parking time and charge a price per-unit of time; coffee shops limit the use of wifi or charge a fee per unit of time; and museums have recently begun limiting the time visitors can spend at an exhibition: The Guggenheim Museum limited visitors to the Doug Wheeler's "Synthetic Desert III" show to 10 minutes and the Hirshhorn Museum in DC (along with five other venues) took even more extreme measures and limited Kusama's "Infinity Mirrors" exhibition to only 30 seconds per visitor per room.

Despite their potential appeal to service providers, implementing price rates and time limits pose challenges. Consumers frequently complain about the restrictions that time limits impose. For example, Salty Running, an exercise and training blog, writes about running time restrictions on treadmills and includes suggestions on what runners could do to exercise efficiently given these "frustrating time maximums" (Cilantro (2013)). CBS reports on angry customers at Panera Bread following the bakery's decision to limit wifi use to 30 minutes at busy times (Ebben (2013)) and several bloggers even provide advise on how to circumvent these time limits (Dachis (2012)). Furthermore, these mechanisms may be difficult to implement. To implement both mechanisms, service providers need to keep track of the time individual customers spend in service, which requires more technological sophistication than charging a per-use fee and letting customers use the service for as long as they want. Even if systems are in place that can track time, certain customers may refuse to honor the limits. It is therefore imperative to understand the conditions under which it is useful to implement these mechanisms and how much benefit they provide compared to their costs.

In this paper we focus on exactly these issues. In particular, we examine four research questions. First,

should a service provider set a price rate and/or a time limit to control congestion when consumers choose not only whether they want to join, but also how long they want to stay in service? Second, can the firm and consumers benefit from price rates and time limits? Third, if they can, but implementation is costly, what is the cost of not implementing these mechanisms? Fourth, what are the operational implications of these levers? We address these issues for firms that aim to maximize revenue and for social planners who are interested in maximizing consumer surplus.

The optimal policies depend crucially on the objective that the service provider is trying to achieve and on the characteristics of the market–the potential for congestion and how flexible consumers are with their time needs. More specifically, we find that in the absence of implementation costs, to maximize revenue the firm should charge price rates that enable it to extract rents from heterogenous customers, but should restrict maximum usage only if the market size is large and customers have some flexibility in their need for time. By contrast, a social planner who wishes to maximize consumer surplus should let customers use the service for free, despite losing the ability to directly control congestion with prices. Distinct from prices, time limits play an important role and are able to substantially increase consumer surplus in markets where congestion is an issue. On a fundamental level, time limits have a non-standard effect on consumers' joining behavior. While prices work to eliminate low willingness to pay consumers from the market, time limits may cause both high and low value customers to balk. We find that to maximize revenue, firms should not set limits that are so strict that high value customers balk, but social planners, who have consumer welfare in mind, should. This behavior results in a more equitable use of common goods and has interesting operational implications, because waiting times and utilization in these settings depend not only on how many consumers join, but also on how long they spend in service.

2 Related Literature

Naor's seminal paper (Naor (1969)) started a very large stream of queueing game papers. (Extensive summaries of the literature can be found in Hassin & Haviv (2003) and Hassin (2016)). However, most of this literature is different from ours in at least two important ways. First, while the literature assumes that consumers value service, most of the papers assume that they do not value the time spent in service. In our model, consumers enjoy the time they spend in service and their value increases the longer they are in service. Second, most of the literature assumes that service times are exogenous (drawn from some known distribution) and can at best be controlled by the service provider through investment in capacity. In our model, service times are endogenous and chosen by consumers. These two features are central in our setting: customers choose not only whether to visit, but also how long to stay in service and have a (weakly) in-

creasing utility in service time. Interestingly, even though settings like these are abundant in practice, these features are mostly missing in the literature on queuing games.

A few streams within this literature are more related to ours. Like in our paper, the literature on diagnostic services assumes that consumers value time spent in service (e.g., Anand *et al.* (2011), Alizamir *et al.* (2013), Kostami & Rajagopalan (2013)), but they do not choose how long to spend there–the service provider determines the service speed. Therefore, the service provider encounters a trade-off between the speed and the quality of service he provides. A longer service is of a higher quality to consumers (e.g. a more accurate diagnostic), but it increases the congestion in the system. Therefore, these papers characterize optimal pricing and server-determined service rate policies in the presence of this trade-off. We also analyze alternative pricing strategies, but in our setting it is the customers who determine how long they spend in service. Conversely, Ha (2001) examines a model where customers choose their time in service, but customers do not enjoy longer times and therefore agree to pay a higher price for shorter service. By contrast, in our paper customers enjoy the time spent in service and will be willing to pay more for longer stays. Similar to our finding on the superiority of price rates, in a setting in which customers have time requirements but do not value this time (i.e., all customers obtain the same value for service) or choose how long to stay in service, Haviv (2014) shows that firms and consumers benefit from duration-based pricing.

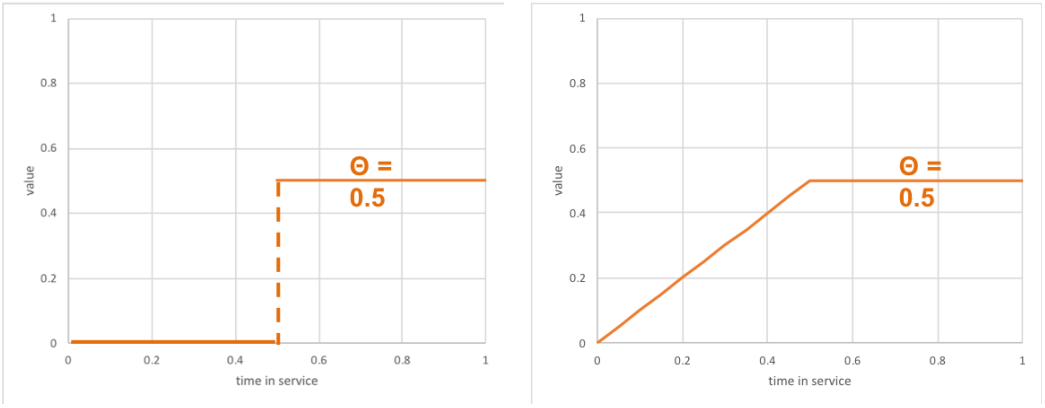
The literature on discretionary services is most similar to our paper in that customers value service and choose how long they want to stay there. As in our paper, Tong & Rajagopalan (2014) study a service system in which consumers value service time and analyze a time-based pricing scheme that allows consumers to choose how long to stay in service. However, distinct from our model, in their model consumers do not observe their own types before entering service (in our model consumers know their need for time and only the firm does not). Hence, in Tong & Rajagopalan (2014)'s model consumers are ex-ante homogeneous and their entry decision is based only on the parameters of the system and not on their individual characteristics. Only after consumers join they observe their type and choose how long to stay and consequently how much to pay. Thus, unlike in our model, the service provider cannot a-priori "select" which type of consumers to serve by choosing a policy and this has implications on his ability to manage system congestion. Debo & Li (2017) consider a service system with consumers who have an increasing value from time in service and choose from a service line that offer a menu of service times, but their objective is different–they find the optimal pricing and service line design and focus on the comparison between queuing disciplines (FIFO vs SPT). Similarly to Tong & Rajagopalan (2014), in their model higher type consumers reap higher value from service than lower type consumers for any time spent in service. This feature of the value functions, which is different from our model allows Tong & Rajagopalan (2014) and Debo & Li (2017) to characterize the optimal pricing policy of the firm, but also implies that high type consumers always want to use service, so

that the firm cannot influence joining decisions of consumers who have high service needs. To the best of our knowledge, our model is unique in that the firm can affect joining decisions not only of low type consumers, but also of high types. This is an important distinction if firms can extract higher rents from high-type consumers or if high type consumers more negatively affect congestion.

While we focus on consumer heterogeneity in service time needs, several papers consider heterogeneity in demand rates (as in the frequency of service requests) (e.g., Masuda & Whang (2006), Randhawa & Kumar (2008), Van Mieghem (2000), Cachon & Feldman (2011), Plambeck & Wang (2013), Afèche *et al.* (2018)). These papers also illustrate that pricing schemes other than per-use fees (e.g., subscriptions) can benefit the service provider. Service times in these models are determined exogenously as in the rest of the literature, but in these papers the firm can control the frequency of usage by implementing an appropriate mechanism. In addition to price rate and time limits, other papers have recently studied innovative pricing strategies in services prone to congestion with heterogenous consumers. For example Yang & Wu (2018) study bundle pricing and Wang *et al.* (2018) study auxiliary service subscriptions. Again, service times are exogenous in these models.

The investigation of consumer surplus rather than revenue or social welfare in services subject to congestion is relatively rare, but growing. Also for discretionary services, Feldman *et al.* (2018) estimate the efficacy of time limits as opposed to congestion pricing using urban parking data and find that while congestion pricing is better for social welfare, the effect on consumer welfare is ambiguous. While congestion pricing is a richer pricing strategy compared to the ones we propose in this paper, this finding supports our conclusion that prices are an ineffective way to maximize consumer welfare. In classical service systems with exogenous service times, Cui *et al.* (2016) investigate a setting where consumers can come back if the queue is too long and find that retrials can harm consumer surplus and Cui *et al.* (2018) analyze a business model of line-sitters and show that this practice can increase consumer surplus.

Finally, there is a large literature in economics that discusses the over-use of common goods known as the "tragedy of the commons" and ways to address it (e.g., Stavins (2011)). Many services in which customers choose how long they stay in service and enjoy their time there fall in this category (e.g., public parking, city managed pools and gyms) and hence this body of work relates to our analysis of a social planner's objective to maximize consumer surplus. It is similar in flavor to our result on the benefit of time limits, but does not include time considerations. This literature argues that governmental regulations should limit the amount of a common good available for use by any individual. For example, the government establishes permit systems for activities such as mining, fishing, hunting, livestock raising and timber extraction. (For an overview see Bowles (2009)).



(a) Strict service needs ($\beta = 1$) **(b)** Flexible service needs ($\beta = 0$)

Figure 1. Service time value functions for extreme levels of β.

3 Model Description

We study a model in which a single provider offers service to self-interested consumers with service needs that arise according to a Poisson process with rate Λ. Consumers enjoy longer service and have heterogeneous service time needs. A consumer's type θ denotes her optimal time for service, which is observed by the consumer at the moment the service opportunity arises, but is unobserved by the firm. Let $F(\cdot)$ be the distribution function and $f(\cdot)$ be the density function of the consumer types defined on the interval $\left[0, \bar{\theta}\right]$. We assume that F is twice differentiable, $F(0) = 0$, and F exhibits an increasing generalized failure rate (IGFR), i.e., $xf(x)/\bar{F}(x)$ is increasing in x, where $\bar{F}(x) = 1 - F(x)$. When necessary, we further assume that consumers' types are uniformly distributed. Consumers' service needs are heterogeneous in that they receive different value from the time in service. Specifically, a consumer of type θ gets value $v_\theta(t)$ from spending t units of time in service:

$$v_\theta(t) = \begin{cases} 0 & 0 \leq t < \beta\theta \\ \frac{t-\beta\theta}{1-\beta} & \beta\theta \leq t < \theta \\ \theta & \theta \leq t \end{cases},$$

where the parameter $\beta \in [0, 1]$ measures the flexibility in service time needs relative to θ. A higher β implies stricter time needs. Figure 1 plots examples of the consumers' value functions for the extreme values $\beta = 0$ and $\beta = 1$. To explain, $\beta = 1$ describes markets where time is critical in that consumers have strict service needs. A type-θ consumer will obtain no value for the service if she cannot use the service for (at least) θ units of time. Conversely, $\beta = 0$ describes markets where consumers have loose service needs in that they

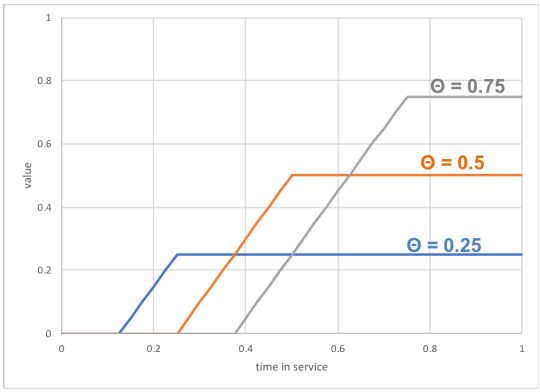


Figure 2. Service time value functions (parameters: $\beta = 0.5, \theta = 1$).

obtain (weakly) higher value from longer service, but will gain a positive value from service as long as they can spend some time in it. We assume that all consumers share the service flexibility parameter β, but that they are heterogeneous with the parameter θ. That is, consumers optimal needs for time in service, θ, are different and these optimal needs correspond to different times, $\beta\theta$, starting from which consumers begin to reap value from service. For example, imagine going to a science museum. A consumer may be interested in spending the morning at the museum. Her optimal time to spend there is 2 hours, which is represented by her type $\theta = 2$. Spending more than 2 hours at the museum will not add to her perceived value from service, but spending less than that (linearly) decreases her value. Finally, if she cannot spend at least 1 hour in service ($\beta = 1/2$) she would not reap any value from going to the museum. Figure 2 illustrates the service time value functions for different values of θ. We assume that consumers have different service needs. That is, while one consumer would like to spend 2 hours at the museum, another may only be interested in visiting the new space exhibition which should take around 30 minutes. Her value also increases with time and she would deem the service not valuable if she cannot spend at least 15 minutes there. Importantly, high type consumers do not value the service more than low type consumers in all cases. For example, 1 hour and 10 minutes are enough for visiting the space exhibition and would yield maximum value for that customer, but this amount of time is barely sufficient for the first customer and she would then obtain limited value from a visit.

While consumers enjoy longer service, they dislike waiting for service to commence and incur a waiting cost c per unit of time spent waiting. The service provider serves customers on a first come first served basis (FIFO). He considers price and time-based mechanisms for controlling congestion. We examine two pricing schemes: per-use fees and price rates. A per-use fee is a fixed fee per service. This is a commonly used and probably the most widely examined tool for controlling congestion in service systems with exogenous service times. We use the term price rate to denote a price that is charged per unit of time a customer

spends in service. In addition to the different pricing schemes, we also consider time limits as a lever to control congestion. A time limit, T, is a restriction on the total amount of time that a customer can spend in service, but customers who have shorter service needs can leave sooner. We analyze the decisions of a service provider who maximizes one of two objective functions, revenue and consumer surplus.

When a service need arises, a consumer makes two decisions. First, she decides whether or not to seek service (i.e., whether to join the service system). Then, conditional on seeking service, she decides how long to stay there. Both of these decisions depend on the consumer's value for service and the service provider's pricing and timing decisions. However, while the joining decision also depends on the cost associated with the expected waiting time, the consumer's decision of how long to stay in service does not. Conditional on choosing to seek service, the time spent waiting and the cost associated with it are sunk and have no effect on the length-of-stay decision. There may be behavioral effects to time limits that result in longer stays than in the absence of limits (e.g., customers choose to stay longer than they need just to "use up" the allowed time). In this model, we do not consider irrational consumer behavior.

We assume that although consumers observe their value for service, they do not observe the current queue length when making a decision. Rather, they form expectations regarding the expected waiting time $W(\cdot)$ and these expectations are rational (i.e., correct in equilibrium). In equilibrium, the waiting time function W depends not only on how many consumers join (which is common in most queueing games models), but also on how long each customer spends in service. The dependence on the length of service, which is endogenously determined, is a key feature in our model. To be more precise, let λ denote the effective arrival rate of consumers and S denote the random variable of the length of service for a single consumer, conditional on consumers' joining behavior. The expected waiting time, $W(\lambda, S)$ can then be written according to the Pollaczek-Khinchine formula for M/G/1 queues with a FIFO discipline as

$$W(\lambda, S) = \frac{\lambda \mathbb{E}\left[S^2\right]}{2\left(1 - \lambda \mathbb{E}\left[S\right]\right)},$$

where $\mathbb{E}[\cdot]$ is the expectation operator. Naturally, the pricing strategy and time limits influence consumers' decisions and the waiting time function and consumers account for them when choosing whether to join and how long to stay. Therefore, we derive the relevant expected waiting time functions and the consumer equilibrium behavior for each policy separately.

4 Benchmark: Per-Use Fees

In this section, we analyze a version of the model where the service provider sets a per-use fee for service. Charging a fixed fee for the use of a service is the most commonly used pricing mechanism and has been analyzed extensively as means to control congestion in classical service systems where service times are exogenous. We therefore use it as the benchmark mechanism in this paper.

A consumer observes her type θ and requests service only if the net utility of doing so is non-negative. With per-use pricing, to join, the value from service should be greater than or equal to $p+cW(\lambda, S)$. Because the value from service (weakly) increases in the time in service and the fee is constant, a consumer of type θ that chooses to join always prefers to stay her optimal service time need, θ: staying for a shorter time strictly decreases her utility and staying for a longer time does not benefit her, but harms other consumers. We assume that in case of indifference a customer chooses the shorter service time. Given that p, c, λ and S are common to all consumers, in equilibrium there exists a threshold type θ^0 such that a consumer seeks service when her type is θ^0 or greater, and otherwise the consumer does not seek service. That is, the threshold type θ^0 is such that: $\theta^0 - p - cW(\lambda, S) = 0$.

To be consistent with actual operational conditions, in the consumer behavior equilibrium the actual arrival rate to the service λ is then $\Lambda\bar{F}\left(\theta^0\right)$, S is distributed on $\left[\theta^0, \bar{\theta}\right]$, and the expected service time is $\mathbb{E}[S] = \int_{\theta^0}^{\bar{\theta}} xf(x)\,dx/\bar{F}\left(\theta^0\right)$. In this case both λ and S are functions of θ^0 and we therefore write $W\left(\theta^0\right)$. Lemma 1 establishes that the threshold θ^0 is unique. (Proofs of all results are provided in the appendix.)

Lemma 1. *There exists a unique threshold type θ^0 such that in equilibrium all consumers with $\theta \geq \theta^0$ seek service and use it for θ units of time and all other consumers balk. θ^0 satisfies*

$$\theta^0 - p = cW\left(\theta^0\right). \tag{1}$$

Furthermore, θ^0 is increasing in the per-use price p.

With per-use pricing, the firm's revenue is $\pi_f = \lambda p$, which can be expressed in terms of the threshold θ^0, because of the monotonicity of θ^0 in p. That is,

$$\pi_f\left(\theta^0\right) = \Lambda\bar{F}\left(\theta^0\right)\left(\theta^0 - cW\left(\theta^0\right)\right).$$

The following theorem establishes that an optimal threshold, θ_p, exists and is unique.

Theorem 1. *There exists a unique threshold θ_f that satisfies $\theta_f = \arg\max_{\theta^0} \pi_f\left(\theta^0\right)$ and a corresponding positive optimal per-use fee $p_f = \theta_f - cW(\theta_f)$.*

The consumer surplus function is $CS_f = \Lambda \int_{\theta^0}^{\bar{\theta}} xf(x)\,dx - \lambda(p + cW)$, where θ^0 satisfies equation (1). As with the revenue function, the consumer surplus function can also be expressed in terms of θ^0 alone:

$$CS_f\left(\theta^0\right) = \Lambda \int_{\theta^0}^{\bar{\theta}} xf(x)\,dx - \lambda\theta^0,$$

where the first term is the value each consumer gets from the use of the service and the second term is the expected total cost (price and expected waiting cost) expressed in terms of the threshold, θ^0.

Theorem 2. *To maximize CS_f it is optimal to set the per-use fee to 0. There is a unique threshold $\hat{\theta}_f$ that solves $\hat{\theta}_f = cW\left(\hat{\theta}_f\right)$. Therefore, $\hat{\theta}_f < \theta_f$.*

The comparison of Theorems 1 and 2 illustrates the stark contrast between a revenue maximizing firm and a social planner whose aim is to maximize consumer surplus. While firms must charge usage fees to extract revenue, to maximize consumer surplus a social planner chooses to offer the service for free. But fees do not only contribute to revenues. They also serve as a lever to control congestion, because consumers with low values would rather balk than pay the fee. By offering the service for free, a social planner loses the ability to control congestion. Congestion is still subject to some self-regulation by consumers since they only request service if their value exceeds their expected congestion cost. That is, in the absence of a fee, consumers select whether to join service purely on the basis of their value for service compared to the expected cost of waiting. Even though congestion is higher when setting the price to zero ($\hat{\theta}_f < \theta_f$) consumers are on aggregate better off with a zero per-use fee.

5 Price Rates

When consumers value the time spent in service, an alternative to per-use fees is to charge a price rate, p, which is a fixed price per-unit of time spent in service. A survey of professional service firms finds that along with per-use fees, price rates are the most commonly use pricing scheme (Lowenhahl (2005)). With a price rate scheme customers pay different prices if they stay in service for different lengths of time: customers who use the service more, pay more. Specifically, a consumer of type θ who uses the service for t units of time, gets utility $v_\theta(t) - p \cdot t - cW(\lambda, S)$. The decision of how long to stay in service depends on the price: if $p > 1$, the price is so high that no consumer gets a positive utility from joining. Otherwise, a consumer of type θ who joins, uses the service for her optimal amount of time, θ. Again, in equilibrium, there exists a threshold type θ^0 such that a consumer seeks service when her type is θ^0 or greater, and otherwise the consumer does not seek service. That is, the threshold type θ^0 is such that: $\theta^0(1-p) - cW(\lambda, S) = 0$.

Lemma 2. *There exists a unique threshold type θ^0 such that in equilibrium all consumers with $\theta \geq \theta^0$ seek service and use it for θ units of time and all other consumers balk. θ^0 satisfies*

$$\theta^0 (1-p) = cW\left(\theta^0\right). \tag{2}$$

Furthermore, θ^0 is increasing in the per-use price p.

With a price rate, the firm's revenue is $\pi_r = \Lambda p \int_{\theta^0}^{\bar{\theta}} x f(x)\, dx$, which as with per-use pricing can be expressed in terms of the threshold θ^0:

$$\pi_r\left(\theta^0\right) = \Lambda\left(1 - \frac{cW}{\theta^0}\right)\int_{\theta^0}^{\bar{\theta}} x f(x)\, dx.$$

Price rates are intuitively appealing, because they enable the firm to price discriminate between consumers (by charging them different prices depending on the time spent in service) and thereby extract more rents.

Consumer surplus is $CS_r = \Lambda(1-p)\int_{\theta^0}^{\bar{\theta}} x f(x)\, dx - \lambda cW$, or

$$CS_r\left(\theta^0\right) = \Lambda cW\left(\theta^0\right)\left(\frac{1}{\theta^0}\int_{\theta^0}^{\bar{\theta}} x f(x)\, dx - \bar{F}\left(\theta^0\right)\right).$$

The following theorem establishes the optimal thresholds and corresponding price rates for the firm and a social planner.

Theorem 3. *Revenue: there exists a unique threshold θ_r that satisfies $\theta_r = \arg\max_{\theta^0} \pi_r\left(\theta^0\right)$ and a corresponding positive optimal price rate $p_r = 1 - cW(\theta_r)/\theta_r$. Consumer surplus: the optimal price rate is 0 and the unique threshold $\hat{\theta}_r$ solves $\hat{\theta}_r = cW\left(\hat{\theta}_r\right)$. Therefore, $\hat{\theta}_r < \theta_r$ and $\hat{\theta}_r = \hat{\theta}_f$.*

Theorem 3 illustrates once again that while a firm sets a positive price rate to maximize revenue and by doing so is also able to control congestion, a social planner does not. In other words, a social planner that maximizes consumer surplus does not use prices (per-use fees or price rates) despite their ability to control congestion.

6 Time Limit

In addition to the which pricing scheme to use, in markets where consumers choose how long to stay in service, providers can control congestion by setting limits on the time spent in service. The use of time limits is distinct to these markets and is not suitable in settings where the time spent in service is determined by the service provider and is independent of the consumers' identities. In this section we analyze a version of the

model where the service provider can institute a time limit in conjunction with a pricing scheme (a per-use fee or a price rate).

6.1 Per-Use Pricing

We characterize the consumer equilibrium behavior for a given per-use price p and time limit T. Since the value from service (weakly) increases in the time in service and the per-use price is constant, a consumer that chooses to join always prefers to stay her optimal service time need. Only a time limit may restrict a customer's time in service. Conditional on joining, the time a customer chooses to spend in service is $\min\{\theta, T\}$. A consumer observes her type θ and requests service only if the net utility of doing so is non-negative. That is, a type $\theta \leq T$ consumer chooses to join and spends θ time units in service if $\theta - p - cW(\lambda, S) \geq 0$ and a consumer of type $\theta > T$ chooses to join and spend T time units in service if

$$\frac{T - \beta\theta}{1 - \beta} - p - cW(\lambda, S) \geq 0.$$

The waiting time function $W(\lambda, S)$ is as before a function of the actual arrival rate λ and the random variable that represents the service time S, both of which may become more complex in the presence of a time limit. Lemma 3 characterizes the consumer behavior equilibrium.

Lemma 3. *For every per-use fee, p, and time limit, T, there exist two unique thresholds in equilibrium, $\theta_l(p,T)$ and $\theta_h(p,T)$, such that $0 \leq \theta_l(p,T) \leq T$ and $T \leq \theta_h(p,T) \leq \bar{\theta}$. Customers with $\theta \in [\theta_l, T]$ stay for θ units of time and customers with $\theta \in [T, \theta_h]$ use the service for T units of time. All other customers do not join. The effective arrival rate is $\lambda = \Lambda(F(\theta_h) - F(\theta_l))$ and the expected service time is*

$$\mathbb{E}[S] = \frac{\int_{\theta_l}^{T} x f(x)\, dx + T(F(\theta_h) - F(T))}{F(\theta_h) - F(\theta_l)}.$$

Lemma 3 describes a consumer behavior equilibrium that is unique to our model with time limits. In classical queueing games with heterogeneous customers and an unobservable queue, each customer trades off her value from service with the expected total cost from service. This results in a threshold equilibrium under which low type consumers choose not to join. This is captured in Lemmas 1 and 2 by the threshold θ^0. The same behavior happens when the service provider sets a time limit and is captured by the threshold θ_l. However, a provider that chooses to limit the time in service may also impact the behavior of high type consumers. This is captured by the threshold θ_h. To explain, a time limit lowers the value that high type consumers receive from service. High type consumers with $\theta > T$, would like to stay in service for $t = \theta$, but are restricted to use it for $t = T$. Some of these consumers, with $T < \theta \leq \theta_h$, would choose to enter

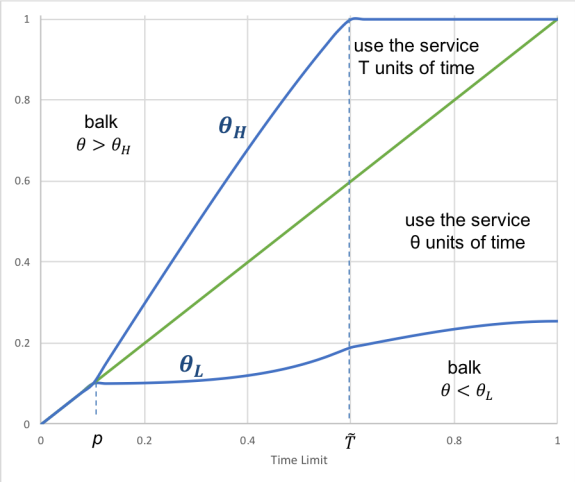


Figure 3. Consumer behavior regions as a function of the time limit, T. (Parameters used in the example: $\Lambda = 1, c = 0.5, \bar{\theta} = 1, \beta = 0.5, p = 0.1$)

the system despite the time limit, but consumers with very high types, $\theta > \theta_h$, may elect not to visit at all (because the time limit is too restrictive for them). For example, suppose that a gym restricts the use of treadmills to 30 minutes. A runner training for a marathon may choose not to use it and run outdoors instead, because 30 minutes may be insufficient according to her exercise plan.

The structure of the value function $v_\theta(t)$ that describes situations in which keeping everything else equal it is possible that given a time t, a high type consumer obtains lower value from service relative to a low type consumer is important for the existence of θ_h. Models that describe situations in which the value of a high type consumer is higher than that of a low type for every t will not result in high type consumers balking (For examples of such models, see Tong & Rajagopalan (2014) and Debo & Li (2017)).

The thresholds θ_l and θ_h are functions of the per-use fee, p, and the time limit, T. Figure 3 plots the thresholds as a function of the time limit. If the time limit is high ($T \geq \tilde{T}$), all high type consumers join. Customers with $\theta \in [T, \bar{\theta}]$ use the service for T units of time and customers with $\theta \in [\theta_l, T]$ use the service for θ units of time. Below $\tilde{T}$ some high type consumers choose to balk. As the time limit becomes more strict, fewer high type consumers join. Specifically, for a moderate time limit ($T \in \left[p, \tilde{T}\right)$), customers with $\theta \in [T, \theta_h]$ join and use the service for T units of time and customers with $\theta \in [\theta_l, T]$ use the service for θ units of time. All other consumers do not join. Finally, if the time limit is lower than the per-use fee p, no consumer joins.

The effective arrival rate λ and the expected service time $\mathbb{E}[S]$ can both be uniquely expressed using the thresholds θ_l and θ_h. Therefore, when convenient, we write the expected waiting time, given a time limit T as a function of the two thresholds: $W_T(\theta_l, \theta_h)$. The firm sets a per-use fee, p, and a time limit, T, to

maximize revenue

$$\pi_f(p,T) = \Lambda p\left(F\left(\theta_h(p,T)\right) - F\left(\theta_l(p,T)\right)\right).$$

Theorem 4 establishes that despite the objective to extract revenue, a firm willingly chooses to restrict customers' time in service.

Theorem 4. *For every $\beta < 1$, the firm optimally sets a strict time limit, i.e., $T_f < \bar{\theta}$.*

Theorem 4 highlights that the flexibility parameter β plays an important role in whether the firm sets a time limit. It does not affect pricing decisions in the absence of a time limit mechanism because customers' values depend on their flexibility of time needs only if they do not spend their optimal time, θ, in service. In particular, we show that for every $\beta \neq 1$, despite its objective to extract revenue and the ability to control congestion by choosing the optimal per-use fee which influences *how many* consumers join, the firm uses a second lever and limits the maximum allowable service time thereby influencing not only how many consumers join, but also *how long* each customer spends in service. In fact, we show that it is optimal to set a strict time limit for every per-use fee, p, and therefore also for the optimal per-use fee, p_f.

Characterizing the optimal time limit analytically for a general distribution is challenging. The difficulty stems mainly from the waiting time function, $W(\cdot)$, which may be neither convex nor concave in the time limit T, a complexity that does not arise when the firm optimizes on price alone. Nevertheless, we are able to characterize the equilibrium fully for $\beta = 0$ and $\beta = 1$ for a uniform distribution.

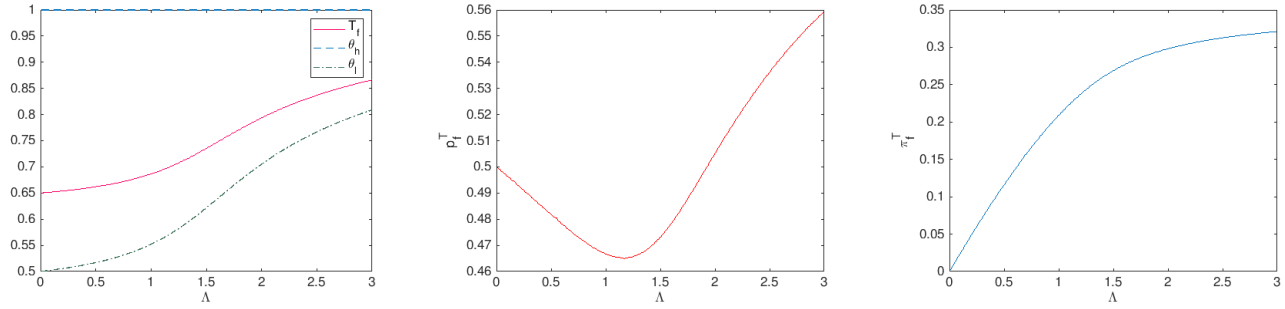
Theorem 5. *Assume that the distribution of consumer types follows a uniform distribution on $\left[0, \bar{\theta}\right]$.*

1. *If $\beta = 1$, the firm does not set a time limit, i.e., $T_f = \bar{\theta}$. Therefore, the optimal thresholds are $\theta_l = \theta_f$ and $\theta_h = \bar{\theta}$ and $p_f^T = p_f$.*
2. *If $\beta = 0$, the firm always sets a time limit. Specifically, let $\Lambda_f = 4\left(1 - \sqrt{c/(c+2)}\right)/\bar{\theta}$.*
   (a) *If $\Lambda \leq \Lambda_f$, the optimal time limit is unique and equals $T_f = \bar{\theta}/2$. The optimal thresholds are $\theta_l = T_f$ and $\theta_h = \bar{\theta}$ and the optimal per-use fee is $p_f^T = \frac{c\bar{\theta}}{4}\left(\frac{2}{c} - \frac{\Lambda\bar{\theta}}{4-\Lambda\bar{\theta}}\right)$. The optimal revenue is $\pi_f^T = \frac{1}{8}\Lambda\bar{\theta}c\left(\frac{2}{c} - \frac{\Lambda\bar{\theta}}{4-\Lambda\bar{\theta}}\right)$.*
   (b) *Otherwise, there exist a continuum of optimal time limits. Let $T^1 = \frac{1}{2}\bar{\theta}\left(1 - \sqrt{1-\Lambda_f/\Lambda}\right)$ and $T_2 = \frac{1}{2}\bar{\theta}\left(1 + \sqrt{1-\Lambda_f/\Lambda}\right)$. Any time limit $T \in \left[T^1, T^2\right]$ with the corresponding per-use fee $p(T) = \frac{1}{2}T\left(c + 2 - \sqrt{c(c+2)}\right)$ is optimal. The optimal thresholds are $\theta_l = T_f$ and $\theta_h = T + \bar{\theta}^2\Lambda_f/4T\Lambda$. The optimal revenue is $\pi_f^T = c + 1 - \sqrt{c(c+2)}$ and is independent of Λ. In these equilibria the firm extracts all surplus from consumers.*

Theorem 5 stresses the importance of the parameter β even further. While the firm always chooses to restrict service times when $\beta < 1$ (Theorem 4), we show that if $\beta = 1$, which implies that time needs are so strict that consumers balk if their optimal service need is not met, the firm does set a time limit and the results mimic those of Theorem 1. By contrast, when consumers have very flexible time needs ($\beta = 0$), the optimal time limit is set so that all consumers with types greater than the time limit join (high type consumers do not balk, $\theta_h = \bar{\theta}$) and all customers spend exactly T_f units of time in service. That is, the firm sets a time limit that makes the heterogenous consumers homogenous in their use of the service. The same happens for all values of the potential arrival rate, Λ, but when $\Lambda > \Lambda_f$ the same type of behavior can be achieved in (infinitely) many different ways: charge a high per-use fee with a high time limit and get low demand, or lower the per-use fee along with a stricter time limit and get greater demand. With all the different time limit / per-use fee combinations, the firm is able to extract all rents from consumers and leave them with zero utility, even though consumer behavior may be significantly different: when $T = T^1$ or $T = T^2$ all high type consumers join, but when $T \in \left(T^1, T^2\right)$, some high type consumers choose to balk (i.e., $\theta_h < \bar{\theta}$).

Figure 4 graphs the equilibrium thresholds, time limits, per-use fees and optimal revenue as a function of Λ ($\beta = 0.3$, $c = 0.5$, $\bar{\theta} = 1$) . It confirms that the firm sets a time limit for all potential arrival rates. In fact, the firm sets the smallest time limit that still ensures that all high type consumers join, i.e. $\theta_h = \bar{\theta}$. In addition, the figures highlight two interesting observations: The optimal per-use fee is non-monotone in the arrival rate (decreases and then increases) and the optimal time limit increases with the potential arrival rate (i.e., the firm allows longer stays as the potential for congestion increases). To explain, with per-use pricing there are two opposing revenue generating effects: the number of customers who join and the price they pay. When Λ is low, congestion is not an issue and the firm cares more about increasing demand so it lowers the price to mitigate low-type consumers' resistance to join. In other words, in this equilibrium the choice of a time limit affects the high-type consumers' joining decision so that $\theta_h = \bar{\theta}$, and the price affects the low type consumers' joining decision, so that θ_l does not increase too much. Conversely, when Λ is high, the firm can focus on selling to high type consumers because even selling to only a subset of consumers guarantees a large customer base. Therefore, in that range, the firm's main objective is to extract sufficient revenue from every sale and it increases the price.

We solve for the equilibrium numerically for $0 < \beta < 1$. Figure 5 illustrates the equilibrium thresholds and optimal time limit (Figure 5(a)), per-use fee (Figure 5(b)) and resulting revenue (Figure 5(c)) as a function of β for $\Lambda = 1, c = 0.6, \bar{\theta} = 1$. Figure 5 highlights three main insights: First, irrespective of how flexible consumers are with their service needs, the firm never prefers to set a time limit that results in high-type consumers balking (i.e., in equilibrium $\theta_h = \bar{\theta}$). We prove that this is the case for $\beta = 0$ and $\beta = 1$



(a) Equilibrium thresholds, θ_l and θ_h, and the optimal time limit, T_f

(b) Equilibrium per-use fee, p_f^T

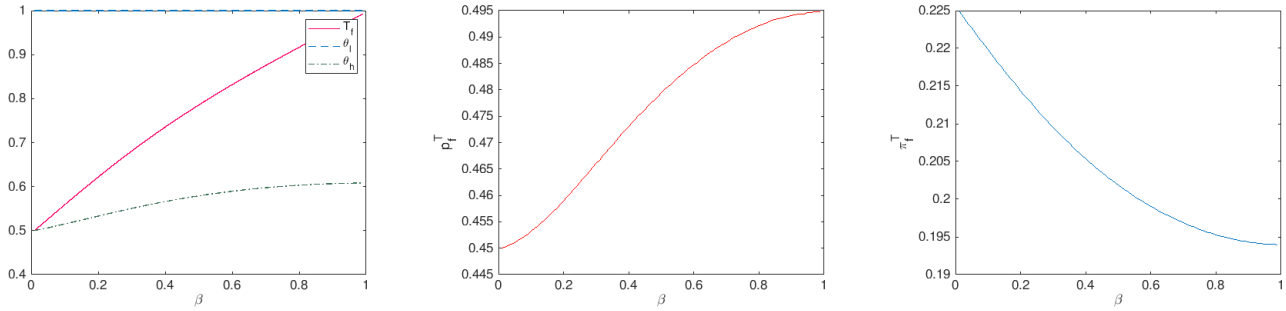
(c) Optimal revenue, π_f^T

Figure 4. Optimal time limit, T_f, joining thresholds, θ_l^f and θ_h^f, per-use fees, p_f^T, and revenue, π_f^T as a function of Λ. (Parameters used in the example: $\beta = 0.3, c = 0.6, \bar{\theta} = 1$).

and show it numerically for all other β. Second, both the optimal time limit and the per-use fee increase with the flexibility parameter β: when customers are very strict with their time needs, the firm chooses not to restrict time much–if it does, high-type consumers choose not to join. Rather, the firm charges a high per-use fee and serves fewer consumers that have high service needs. With strict service needs, price is the more effective lever to control congestion. Alternatively, when consumers have looser service needs, the firm sets a strict time limit but a low price - more consumers choose to join, but some use the service for shorter time. Here, consumers are hurt less by time restrictions, and limit is the main lever to control congestion. Third, the firm extracts higher revenue as consumers have more flexible needs with highest revenue achieved at $\beta = 0$. The flexibility in service needs enables the firm to set a time limit that makes consumers more homogenous. To do that, the firm has to charge a lower per-use fee, but the low fee allows it to serve a larger market that benefits the firm overall. This result resembles the idea behind advance selling and subscription pricing schemes when consumers buy in advance of the time of service (e.g., DeGraba (1995), Xie & Shugan (2001), Cachon & Feldman (2011), Cachon & Feldman (2017) and Cachon & Feldman (2018)). When consumers buy in advance, they are more homogeneous compared to the spot market, and the firm is able to extract more revenue because of that. In our model, consumers do not purchase over time, so the homogeneity does not stem from consumers' uncertain information. Instead, when the firm sets a time limit it makes customers' value from service more homogenous which enables the firm to extract more rents by selling to more consumers.

To maximize consumer surplus, a social planner sets p and T to maximize

$$CS_f(p,T) = \Lambda\left(\int_{\theta_l}^{T} xf(x)\,dx + \frac{T-\beta\theta_h}{1-\beta}\left(F(\theta_h) - F(T)\right)\right) - \lambda(p+cW).$$



(a) Equilibrium thresholds, θ_l and θ_h, and the optimal time limit, T_f

(b) Equilibrium per-use fee, p_f^T

(c) Optimal revenue, π_f^T

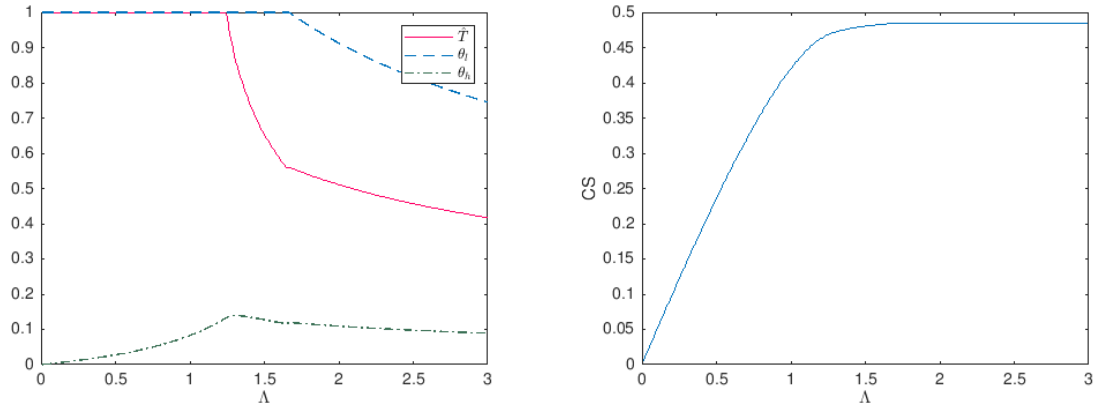
Figure 5. Optimal time limit, T_f, joining thresholds, θ_l^f and θ_h^f, per-use fees, p_f^T, and revenue, π_f^T as a function of β. (Parameters used in the example: $\Lambda = 1, c = 0.6, \bar{\theta} = 1$).

Theorem 6. *To maximize consumer surplus the social planner sets $\hat{p}_f = 0$. Furthermore, assume that θ is uniformly distributed on $[0, \bar{\theta}]$. Then the following hold:*

1. *If $\beta = 1$, there exists $\Lambda_1 > 0$ such that if $\Lambda \leq \Lambda_1$, consumer surplus is maximized when $\hat{T} = \bar{\theta}$ and otherwise the social planner sets a time limit, i.e., $\hat{T} < \bar{\theta}$.*

2. *If $\beta = 0$, there exist Λ_2, Λ_3 and Λ_4 with $0 < \Lambda_2 < \Lambda_3 < \Lambda_4$ such that if $\Lambda \leq \Lambda_2$, consumer surplus is maximized when $\hat{T} = \bar{\theta}$ and if $\Lambda \in [\Lambda_3, \Lambda_4]$, then $\hat{T} < \bar{\theta}$.*

Theorem 6 establishes that as in the case without a time limit, to maximize consumer surplus, a social planner allows customers to use the service for free. However, even though the social planner chooses not to use per-use fees to control congestion, he may choose to set a time limit. Specifically, we find that to maximize consumer surplus, a social planner should set a time limit if the potential arrival rate is sufficiently large. We observe that the consumer surplus function, $CS_f(T; \Lambda)$, takes on different shapes depending on the potential arrival rate, Λ. It is increasing in T for small Λ, increasing-decreasing for $\Lambda \in [\Lambda_3, \Lambda_4]$, and increasing-decreasing-increasing for $\Lambda > \Lambda_4$. Because of the complex structure of the consumer surplus function, the analytical result for the existence of a strict time limit is limited to values of $\Lambda \in [\Lambda_3, \Lambda_4]$. However, as is intuitively appealing, we confirm numerically that the social planner will set a time limit for any $\Lambda \geq \Lambda_3$. The equilibrium time limit is illustrated in Figure 6(a). Observe that there are three regions in the figure: When Λ is low the potential for congestion is low and the social planner does not set a time limit. Consequently, in that region $\theta_h = \bar{\theta}$ and θ_l increases with Λ so that fewer consumers join. When Λ is intermediate, the potential for congestion increases and the social planner responds by setting a time limit which becomes more strict as Λ increases (i.e., $\hat{T}$ decreases). For this intermediate range, the time limit is set such that all high type consumers join ($\theta_h = \bar{\theta}$), but at the same time, the market size increases because the



(a) Equilibrium thresholds, θ_l and θ_h, and the optimal time limit, $\hat{T}$

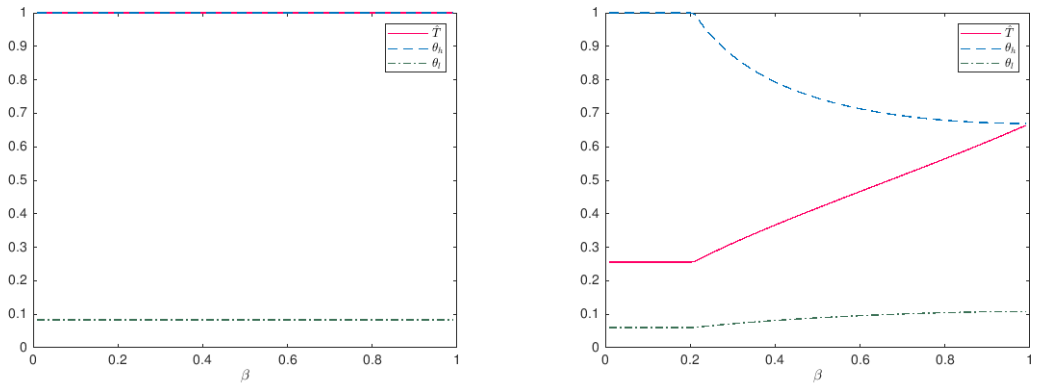
(b) Optimal Consumer Surplus, CS^T

Figure 6. Optimal time limit, $\hat{T}$, joining thresholds, θ_l and θ_h, and consumer surplus, CS^T as a function of Λ. (Parameters used in the example: $\beta = 0.5, c = 0.25, \bar{\theta} = 1$).

strict time limit encourages more low type consumers to request service. Finally, when Λ is large, the service provider continues to reduce the time limit, but the optimal time limit is too strict for high type consumers who choose to balk, i.e., $\theta_h < \bar{\theta}$. Figure 6(b) graphs the optimal consumer surplus for the same parameter values. Observe that consumer surplus strictly increases as long as all high type consumers join and remains constant for larger Λ despite the increase in potential congestion, the social planner cannot increase consumer surplus further and his best bet is to set a strict time limit that makes high type consumers balk, but keeps surplus at a constant rate. Overall, by using a time limit the social planner is able to overcome the negative implications of congestion. We will return to this point in Section 7.

Interestingly, we find that to maximize consumer surplus when the potential for congestion is high, the social planner chooses to set a time limit that causes some high type consumers to balk. This result comes in stark contrast to the revenue maximizing equilibrium that manages not to lose high type consumers even when it is optimal to set a time limit (except for the special case when $\beta = 0$ and large Λ, where it may choose to do so). With revenue maximization the firm sets a time limit, but also a price and both help regulate congestion. Without price as a lever (which reduces consumer surplus), the optimal time limit may become too strict for inflexible high type consumers when congestion is an issue.

Finally, it is instructive to understand how consumers' flexibility to time measured by the parameter β affects the optimal time limit and consumer surplus. Figure 7 graphs the joining behavior and optimal time limit as a function of β for different levels of Λ (low: $\Lambda = 1$ and high: $\Lambda = 3$ and parameters $\bar{\theta} = 1$, $c = 0.25$) and Figure 8 graphs the corresponding optimal consumer surplus. The figures illustrate that the level of flexibility matters when the potential for congestion is high, but not when it is low. In particular,



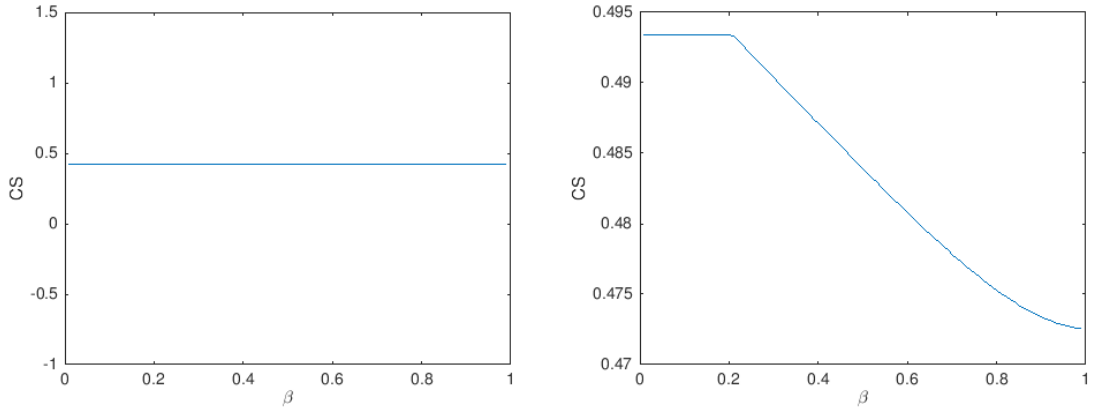
(a) $\Lambda = 1$ **(b)** $\Lambda = 3$

Figure 7. Optimal time limit, $\hat{T}$, joining thresholds, θ_l and θ_h, as a function of β. (Parameters used in the example: $c = 0.25, \bar{\theta} = 1$).

flexibility (or lack thereof) only matters when it is optimal for the social planner to set a time limit that causes some high type consumers to balk. Consumer surplus does not depend on β when there is no limit. And indeed, when Λ is small the social planner does not set a time limit and the consumer surplus is constant, independent of β. However, when Λ is high, the level of flexibility affects equilibrium decisions. Specifically, Figure 7(b) shows that there exists a β' below which a time limit does not cause high type consumers to balk ($\theta_h = \bar{\theta}$), which implies that in this region the equilibrium outcomes will not depend on β. Higher flexibility levels ($\beta > \beta'$) induce the social planner to choose a stricter time limit that discourages some high type consumers from joining ($\theta_h < \bar{\theta}$) and will therefore result in equilibrium outcomes that depend on β. Specifically, as consumers' time flexibility decreases, it is optimal to set a less stringent time limit, but even though the limit is less strict, it still makes some high type consumers stay out of the market. Consequently, as is illustrated by Figure 8(b), optimal consumer surplus decreases with β - time limits are especially helpful when they do not discourage high type consumers from joining, but it may be optimal to choose to do that to avoid congestion.

6.2 Price Rates

As with per-use fees, we can characterize the consumer equilibrium behavior for a given price rate p and time limit T. Again, only a time limit may restrict a customer's time in service such that conditional on joining, the time a customer chooses to spend in service is $\min\{\theta, T\}$. A consumer observes her type θ and requests service only if the net utility of doing so is non-negative. That is, a type $\theta \leq T$ consumer chooses to join and spends θ time units in service if $(1-p)\,\theta - cW(\lambda, S) \geq 0$ and a consumer of type $\theta > T$ chooses to join and spend T time units in service if $\frac{T-\beta\theta}{1-\beta} - pT - cW(\lambda, S) \geq 0$. Analogously to the model with per-use pricing



(a) $\Lambda = 1$ **(b)** $\Lambda = 3$

Figure 8. Optimal consumer surplus as a function of β. (Parameters used in the example: $c = 0.25, \bar{\theta} = 1$).

and time limits, the structure of the equilibrium involves two thresholds θ_l and θ_h that determine consumer behavior. Since the behavior is qualitatively similar to the model with per-use pricing and time limits (of course, the actual thresholds depend on the pricing scheme) and follows Lemma 3, for brevity we do not repeat it here. (The proof of the equilibrium behavior with the price rate scheme is significantly different than that with per-use pricing and is available from the authors.)

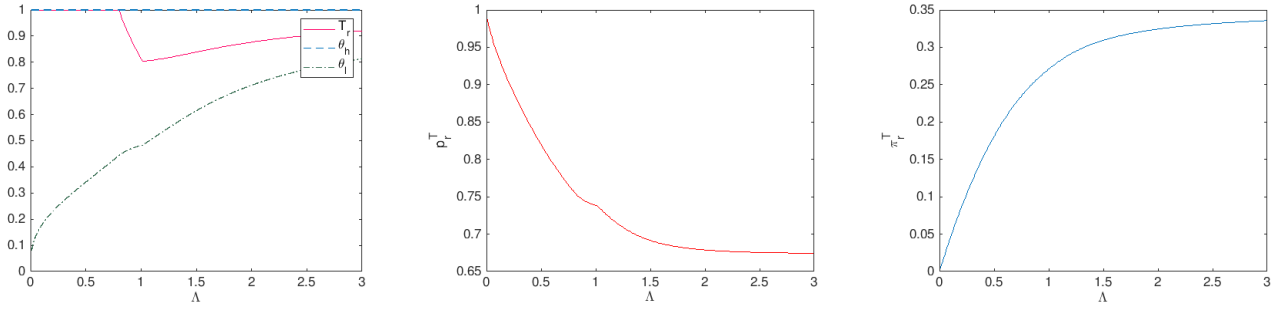
A firm that wishes to maximize revenue can set a price rate, p, and a time limit, T and obtain

$$\pi_r(p, T) = \Lambda p \left(\int_{\theta_l}^{T} x f(x)\, dx + T\left(F(\theta_h) - F(T)\right) \right).$$

The next result establishes the optimal time limit.

Theorem 7. *Assume that $\theta \sim U\left[0, \bar{\theta}\right]$. With a price rate scheme, if $\beta = 1$, the firm does not set a time limit, i.e. $T_r = \bar{\theta}$. If $\beta = 0$, there exists a threshold Λ_r such that the firm sets a time limit, $T_r < \bar{\theta}$, if $\Lambda > \Lambda_r$ and sets $T_r = \bar{\theta}$, otherwise.*

Theorem 7 establishes that when consumers have flexible needs, a firm that charges a price rate chooses to set a strict time limit only when the potential congestion is sufficiently high. This comes in contrast to per-use pricing, where as long as consumers have some time flexibility ($\beta \neq 1$) to maximize revenue the firm always limits the time spent in service. Recall that with per-use pricing and $\beta = 0$ the firm sets a time limit so that all customers who join spend exactly T_f units of time in service. All consumers, though heterogenous, essentially behave as homogenous–they pay the same price and spend the same time in service. Here, the same does not always happen. A price rate strategy is good at extracting rents from heterogeneous consumers and the firm prefers to price discriminate than to charge them the same fee when Λ is relatively



(a) Equilibrium thresholds, θ_l and θ_h, and the optimal time limit, T_r

(b) Equilibrium price rate, p_r^T

(c) Optimal revenue, π_r^T

Figure 9. Optimal time limit, T_r, joining thresholds, θ_l^r and θ_h^r, per-use fees, p_r^T, and revenue, π_r^T as a function of Λ. (Parameters used in the example: $\beta = 0.3, c = 0.6, \bar{\theta} = 1$).

low. We will elaborate on the comparison and the operational implications of that result in Section 7. Only when Λ is high, so that a time limit is important to control congestion will the firm abandon the ability to price discriminate consumers and also set a time limit that makes consumers homogenous. Naturally, for those high values of Λ the revenue with the two price schemes is the same (collecting \$1 per minute of a 1-hour service from each customer is the same as collecting \$60 for the full service.) By contrast, when $\beta = 1$, the firm never sets a time limit whether it charges a price rate or a per-use fee. When time needs are very rigid, a time limit will necessarily result in high-type consumers balking which harms the firm.

Though the proof focuses on $\beta = 0$ and $\beta = 1$, the equilibrium for a general β resembles that of $\beta = 0$ qualitatively. To see this, Figure 9 illustrates the equilibrium quantities and revenue as a function of Λ for $\beta = 0.3$. With a price rate scheme, the factors affecting revenue are the price charged per unit of time and the total time consumers spend in the system (rather than the number of consumers). In equilibrium, the optimal price decreases with Λ (as it did with no time limit), but the time limit changes in a non-monotone way. First, when Λ is low, congestion is not an issue and the firm does not set a time limit at all. Then, as Λ increases, congestion becomes more of an issue and the firm sets a time limit. This limit allows it to charge a relatively higher price rate. Finally, when Λ is high, a further reduction in time limit would induce some high type consumers to balk and hurt revenue. Instead, the firm raises the time limit so that all high type consumers request service.

To maximize consumer surplus, a social planner sets p and T to maximize

$$CS_r(p,T) = \Lambda(1-p)\left(\int_{\theta_l}^{T} xf(x)\,dx + \frac{T-\beta\theta_h}{1-\beta}\left(F(\theta_h) - F(T)\right)\right) - \lambda cW.$$

We prove that for $\beta = 0$ and $\beta = 1$ the social planner does not charge a price rate (i.e., $p_r = 0$). (The proof

Table 1. Parameter values used in the numerical study.

Parameter	Values
Type Distribution	Uniform
$\bar{\theta}$	1
Λ	0.25,0.5,0.75,1,1.25,1.5,1.75,2,...,5
β	0,0.1,0.2,...,1
c	0.01,0.2,0.4,0.6,...,3

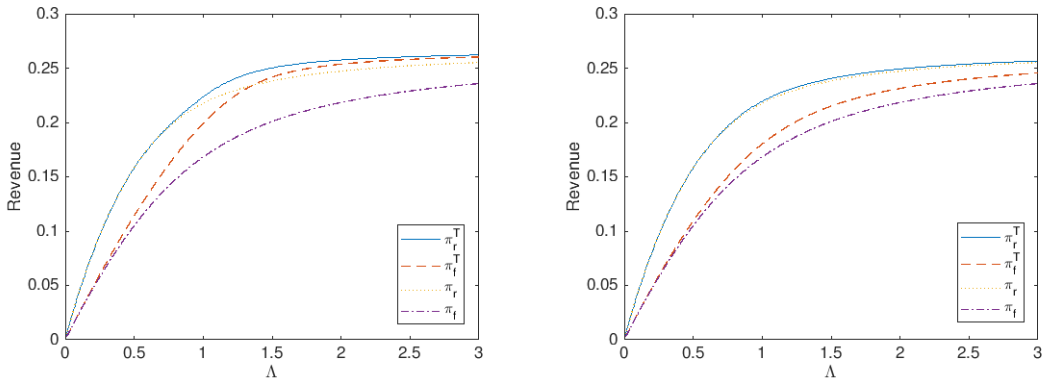
is given in the appendix.) Therefore, consumer surplus is only affected by the possibility to set a time limit, but not by the price scheme. Theorem 6, which finds the optimal time limit $\hat{T}$ that maximizes consumer surplus and the discussion that follows hold here too.

7 Discussion: Comparison and Operational Implications

In this section we compare the different mechanisms and discuss operational implications. Table 1 lists the parameters we use in a numerical study to compare the strategies discussed in the previous sections. We compare four strategies for revenue maximization - price rate with and without time limit and per-use fees with and without time limit and two strategies for consumer surplus maximization - with and without time limit (because a social planner always sets a zero price to maximize consumer surplus). This yields a total of 3,520 scenarios. However, if it is optimal for the service provider not to set a time limit, then the consumer surplus functions and the revenue functions for each pricing strategy coincide making the comparison less interesting. We therefore remove those instances from the analysis, leaving 2598 scenarios for revenue comparisons and 2640 scenarios for consumer surplus comparisons.

7.1 Revenue Comparison

Figure 10 illustrates the revenues achieved with each policy for two representative levels of β and a wide range of potential arrival rates. As expected revenue is maximized when the firm is able to charge a price rate as well as set a time limit. The ability to charge price rates enables the firm to discriminate between different types of consumers and allow even lower types to join because of the differentiated prices each customer pays. The ability to set a time limit improves revenue further and is important when the potential for congestion is high. Overall, the figure suggests that the firm sacrifices a considerable amount of revenue if it charges per-use fees and does not impose a time limit. This strategy, that works well in service systems where service times are exogenously set, may work poorly when customers choose their service times. There may be situations where the firm is unable to implement both price rates and time limits. In such situations, whether firms should focus on charging price rates or on restricting time spent in service is less clear and



(a) Revenue with flexible service needs ($\beta = 0.1$) **(b)** Revenue with less flexible needs ($\beta = 0.4$)

Figure 10. Revenue comparison with the four strategies. Parameters used in the example: $c = 1, \bar{\theta} = 1$.

depends on market conditions. The figures suggest that in many cases it is better to focus on charging price rates, but not always. A firm that charges per-use fees but is able to set a time limit will perform better when customers have relatively flexible service time needs and the potential congestion is high.

These findings extend to the larger sample. Figure 11 shows box-plots of the revenue ratios of the different strategies, where π_r stands for revenue with price rate, π_f stands for revenue with a per-use fee and the superscript T stands for a strategy that considers time limits. As observed in Figure 10, the price rate with time limit strategy performs better than all other strategies, often times substantially. Though all schemes can approach this strategy in some cases (with maximum ratios of $\pi_r/\pi_r^T = 1$ when the optimal time limit $T_r = \bar{\theta}$, $\pi_f^T/\pi_r^T = 1$ and $\pi_f/\pi_r^T = 0.975$), they often perform poorly in comparison: the ratio π_r/π_r^T can be 89.3% in the worst case with $\pi_f^T/\pi_r^T = 0.698$ and $\pi_f/\pi_r^T = 0.697$ at a minimum. Average ratios are: $\pi_r/\pi_r^T = 98\%$, $\pi_f^T/\pi_r^T = 93.4\%$ and $\pi_f/\pi_r^T = 89.1\%$.

The results in Figure 11 emphasize several points. First, the simple per-use fee mechanism that works well when service times are exogenously determined performs poorly when consumers have service time needs and can choose how long they want to spend in service. Failing to implement an appropriate mechanism may substantially harm revenue. Second, not all mechanisms perform equally. While the ability to implement both price rates and time limits is most desirable, in most cases the use of just one of these levers improves revenues significantly. This is important in settings where there are challenges to implement these more sophisticated mechanisms either because of lack of technological infrastructure and the costs associated with it (e.g., keeping track of the time customers use the service, enforcing limited use, collecting different fees from different customers) or due to consumer resistance to these levers. In fact, we show that moving from per-use fees to price rates alone can often achieve nearly optimal revenue if the potential for congestion is not too high. Otherwise, it is more important to focus on targeting congestion by restricting time in service.

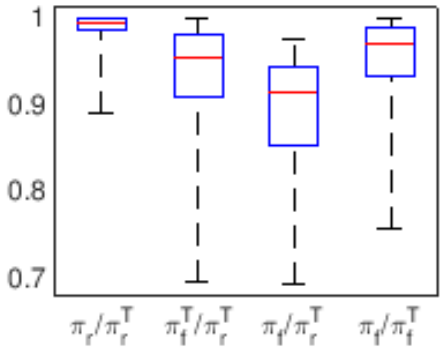


Figure 11. Revenue Ratios. The bottom, middle and upper horizontal lines of the box correspond to the 1st quartile, median and 3rd quartile, respectively; the bottom and the top lines at the ends of the vertical lines are the minimum and maximum, respectively.

Why do these mechanisms work? It is intuitive that faced with heterogenous consumers that value time in service differently, a firm can benefit from charging price rates. With price rates, different customers pay different prices for service and the firm can extract more rents. That is, charging a price rate is an effective revenue generating mechanism. In fact, if congestion is costless (i.e., $c = 0$), a firm that implements a price rate strategy extracts all consumer surplus and revenue with price rates is twice as high as with per-use fees. Time limits provide a different advantage. With time limits, the firm can keep the same price (whether it is a per-use fee or a rate) and increase demand while at the same time decrease expected waiting time: if high type consumers use the service for shorter time, more low type consumers are encouraged to join. With per-use fees this necessarily increases revenue - more consumers join that are willing to pay the same price. With price rates, the overall effect is less clear, because the firm cares about the total amount of time spent in service. While more consumers join, some consumers spend less time in service. And, if fact, we find that firms should only set time limits if the potential congestion is high, so that even though a time limit results in more customers, the overall time they spend in the system is controlled and waiting times are decreased.

7.2 Consumer Surplus Comparison

Without the ability to set time limits, consumer surplus, CS, is maximized for an intermediate level of potential congestion. Too low of a Λ and there are no consumers that generate value, too high of a Λ and customers waiting costs become so high that very few join (until in the limit, $\lim_{\Lambda\to\infty} CS = 0$). For a social planner it is undesirable to resolve this negative effect of congestion with prices which harm consumer surplus regardless of the level of congestion as opposed to a firm that would always set a positive price to earn a profit. Time limits can not only resolve this issue, but may even improve surplus further, as Figure 12

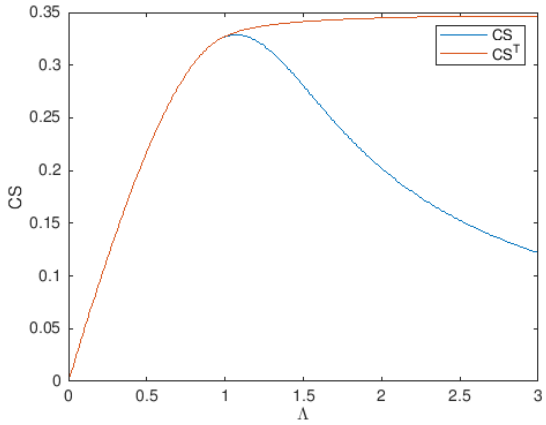


Figure 12. Comparison of consumer surplus with and without time limit. Parameters used: $\beta = 0.2$, $c = 0.6$, $\bar{\theta} = 1$.

illustrates. With time limits, consumer surplus increases in congested systems up to a point where high type consumers leave and surplus stabilizes, whereas in their absence it would decrease drastically. The figure emphasizes the erroneous views of consumers who raise objections to the use of time limits. Some consumers are certainly hurt by these limits - high type consumers are forced to use the service for shorter amounts of time and some even choose not to use it as a consequence. However, on aggregate, the consumer population as a whole benefits from these limits and the benefit grows as the potential for congestion increases.

More generally, these results highlight the difference in approaches for managing service systems with different objectives in mind. When the objective is to maximize revenue, our results show that in many cases a clever choice of a pricing scheme may be sufficiently successful even without time restrictions. By contrast, monetary policies do not work if the objective is to maximize surplus, but the implementation of time limits may improve it substantially to the point that consumers are not harmed, but can benefit from an increase in potential congestion.

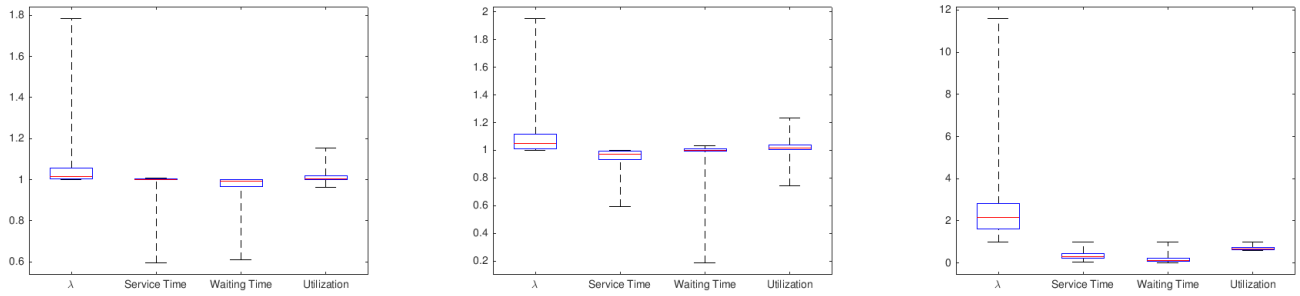
7.3 Operational Implications

Finally, we explore the operational implications of these mechanisms. We first investigate the effect of imposing a time limit on four operational measures: the actual arrival rate, λ, the expected service time conditional on joining, $\mathbb{E}[S]$ and the resulting expected waiting time, W, and utilization, U. To analyze the change in each of these performance metrics, we calculate the ratio of the measure after imposing the time limit over the measure without a time limit for all 2598 parameter combinations. Figure 13 presents the summary of these results for the optimal revenue with a price rate scheme (Figure 13(a)), the optimal revenue with per-use fees (Figure 13(b)) and the optimal consumer surplus for which the price is zero (Figure 13(c)).

The results in Figure 13(a) and Figure 13(b) tell a consistent story: when a firm imposes a time limit, more people join (λ increases), but expected time spent in service decreases because of two reasons: every customer that enters spends (weakly) less time in service because of the time limit and the additional customers that join are low type consumers who do not require much time in service. Overall, even though more customers enter, waiting time decreases, sometimes significantly (by 40% with a price rate scheme and 80% with a per-use fee). The effect on system utilization, however, is mixed, as the figure suggests. The figure highlights two interesting results. First, expected waiting time and utilization do not go hand in hand. It is possible that the system is more utilized, but consumers wait less time to be served. This phenomenon never occurs in service systems where customers do not value time in service and choose how long to stay. In those systems, only the rate with which customers join changes with the change of mechanism and this change affects both waiting time and utilization in the same way: if more customers join, waiting times will increase and so will system utilization. Here, what matters is not only how many people join, but also how long each customer stays in service and that may influence waiting times and utilization differently. Second, interestingly, the numerical study illustrates that while utilizations can be higher or lower with time limits, the firm sets a price and a time limit that keeps utilization levels similar to the setting without time limits, especially when the firm charges a price rate (the minimum ratio is $U_r^T/U_r = 96.3\%$ with price rates and $U_f^T/U_f = 74.1\%$ with per-use fees and the maximum ratio is $U_r^T/U_r = 115.5\%$ for price rate and $U_f^T/U_f = 123.5\%$ with per-use fees).

Figure 13(c) illustrates the analogous measures for consumer surplus maximization. The first set of results is equivalent, but more pronounced: time limits result in more consumers joining and in lower expected service and waiting times. The differences can be rather extreme when the potential for congestion is large (the maximum arrival rate ratio in the sample is $\lambda_{CS}^T/\lambda_{CS} = 115.8\%$ and would grow without bound as Λ increases, the minimum service time ratio is $\mathbb{E}\left[S_{CS}^T\right]/\mathbb{E}\left[S_{CS}\right] = 5.0\%$ and the minimum waiting time ratio is $W_{CS}^T/W_{CS} = 1.64\%$). However, compared with revenue maximization, the resulting utilizations in this case are always lower with time limits (with minimum utilization ratio of $U_{CS}^T/U_{CS} = 57.7\%$). The difference lies in the mechanism that a social planner uses to maximize consumer surplus. In the absence of prices as a lever, the social planner must set strict time limits to control congestion. Even though these result in greater demand, the shorter times in service inevitably drive utilizations down. Overall, the results in Figure 13 suggest that time limits lead to a socially desirable system with more consumers who enjoy service, but shorter waiting times, (generally) lower utilizations, higher revenues and higher surplus.

Next, we investigate the effect of charging price rates instead of per-use fees on operational performance. For a clean comparison, Figure 14 presents ratios of the four operational measures when the firm does not institute time limits. That is, we compare: percentage change of arrival rate, λ_r/λ_f, expected service time,



(a) Price rates **(b)** Per-use fees **(c)** Consumer surplus

Figure 13. Operational ratios for the effect of time limits. The bottom, middle and upper horizontal lines of the box correspond to the 1st quartile, median and 3rd quartile, respectively; the bottom and the top lines at the ends of the vertical lines are the minimum and maximum, respectively.

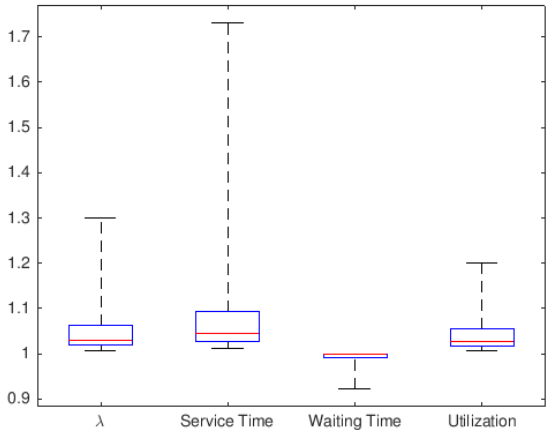


Figure 14. Operational ratios for the effect of the pricing policy. The bottom, middle and upper horizontal lines of the box correspond to the 1st quartile, median and 3rd quartile, respectively; the bottom and the top lines at the ends of the vertical lines are the minimum and maximum, respectively.

$\mathbb{E}[S_r]/\mathbb{E}[S_f]$, expected waiting time, W_r/W_f, and utilization, U_r/U_f. The directional results are intuitive: without a time limit, equilibrium consumer behavior only changes through the lower threshold, θ_l. A firm that sets a price rate is able to discriminate consumers through price and therefore attract more low type consumers. This results in lower average service times, but longer average waits as well as higher utilizations (both due to heavier load).

Finally, Figure 15 presents the effect on the system's overall performance when implementing both price rates and time limits compared to the base case with a per-use fee and no time limits. We have already seen that the impact on revenue can be significant. This is attributed not only from the ability to extract more rents from consumers, but also from increased demand through the ability to encourage more consumers to seek service. Time limits and the fact that more low type consumers join decrease the average service times

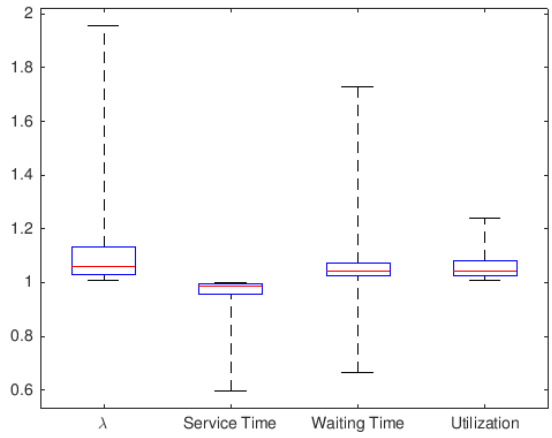


Figure 15. Operational ratios for the effect of the pricing policy. The bottom, middle and upper horizontal lines of the box correspond to the 1st quartile, median and 3rd quartile, respectively; the bottom and the top lines at the ends of the vertical lines are the minimum and maximum, respectively.

in the system. Therefore, on one hand, more consumers join, but on the other hand, they use the service for shorter time. This leads to increased utilizations, but the overall effect on average waiting times is mixed – in some instances the effect of increased demand dominates leading to higher waiting times, and in other, the effect of shorter service time dominates and leads to lower waiting times.

To summarize, we conclude that service systems where consumers value time in service and choose how long to stay are fundamentally different than settings in which the service provider controls service times and that different mechanisms may lead to sometimes counterintuitive and many times mixed implications on operational performance. Often, these implications will depend on market characteristics and the choice of mechanism. It is therefore important to first gather relevant market information that will help to choose the appropriate mechanism and understand the effects that will result.

8 Conclusion

Service systems in which customers value the time in service and can choose how long to spend there are special because service providers lose direct control of service times. Distinct from common service systems in which the provider can invest in capacity to reduce congestion, in these systems the ability to do so is limited. Fortunately, this settings put forth alternative mechanisms that are good at extracting revenue and decreasing waiting times. We introduce two such mechanisms: charge a rate per-unit of time and impose a limit on the maximum time a customer spends in service. We find that each mechanism provides a different advantage to the service provider. Time limits are effective at controlling congestion - they result in more

demand, but limit the time customers spend in service which decreases waiting times overall. By contrast, price rates are effective at extracting rents from heterogenous customers irrespective of the level of congestion in the system.

When the potential congestion is high, a firm that wants to maximize revenue generally benefits the most from using the two mechanisms combined. However, the two policies help the firm in different ways and neither dominates clearly. If implementation is costly, a firm that can only implement one of the two strategies should therefore take a close look at the market at which it operates: time limits are especially beneficial if the potential congestion is high and customers have flexible service needs. Otherwise, to maximize revenue, it is more important to focus on extracting rents which makes price rates an attractive mechanism.

The choice of how long to use service is not unique to service systems controlled by firms that wish to maximize revenue, but is common in common goods as well. The decision in these settings is more straightforward. A social planner should not charge customers for the use of the service, but time limits are an incredibly useful tool for controlling congestion and maximizing welfare–it keeps welfare high even when the potential congestion levels diminish consumer surplus in its absence. Of course, it may be challenging to force customers to adhere to time limits. Possible mechanisms to enforce these restrictions are fines or taxation. In the context of our model we assume that customers adhere to the restriction, but an analogous model is to couple time limits with very large fines.

Overall, our results highlight that it is important to consider appropriate mechanisms for controlling congestion in settings where consumers choose not only whether to join, but also how long to stay in service. We theoretically justify the implementation of price rates and time limits and argue that whether a service provider benefits from these strategies depends critically on consumers' flexibility in time needs, the potential congestion and the objective function in the absence of costs. Additional research could seek to determine if service providers that currently utilize these strategies indeed benefit from these strategies in practice and estimate the gains relative to the implementation costs in different markets.

References

Afèche, Philipp, Baron, Opher, Milner, Joseph, & Roet-Green, Ricky. 2018. *Pricing and Prioritizing Time-Sensitive Customers with Heterogeneous Demand Rates.* Tech. rept. Forthcoming in Operations Research.

Alizamir, Saed, de Vericourt, Francis, & Sun, Peng. 2013. Diagnostic Accuracy under Congestion. *Management Science*, **59**(1), 157–171.

Anand, Krishnan S., Paç, M Fazil, & Veeraraghavan, Senthil. 2011. Quality-speed conundrum: trade-offs in customer-intensive services. *Management Science*, **57**, 40–56.

Bowles, Samuel. 2009. *Microeconomics: behavior, institutions, and evolution.* Princeton University Press.

Cachon, Gérard P., & Feldman, Pnina. 2011. Pricing services subject to congestion: charge per-use fees or sell subscriptions? *Manufacturing & Service Operations Management*, **13**(2), 244–260.

Cachon, Gérard P., & Feldman, Pnina. 2017. Is advance selling desirable with competition? *Marketing Science*, **36**(2), 195–213.

Cachon, Gérard P., & Feldman, Pnina. 2018. *Pricing Capacity Over Time and Recourse Strategies: Facilitate Reselling, Offer Refunds/Options, or Overbook?* Tech. rept. Questrom School of Business, Boston University.

Cilantro. 2013. *Gym time limits.* Tech. rept. http://www.saltyrunning.com/gym-set-limit-treadmill-time/.

Cui, Shiliang, Su, Xuanming, & Veeraraghavan, Senthil K. 2016. A model of rational retrials in queues. *Working Paper, University of Pennsylvania, Philadelphia, PA.*

Cui, Shiliang, Wang, Zhongbin, & Yang, Luyi. 2018. The Economics of Line-Sitting. *Working paper.*

Dachis, Adam. 2012. *Circumvent Wi-Fi Time Limits at Coffee Shops by Spoofing Your MAC Address.* Tech. rept. https://lifehacker.com/5916339/circumvent-wi-fi-time-limits-at-coffee-shops-by-spoofing-your-mac-address.

Debo, Laurens, & Li, Cuihong. 2017. *Design and Pricing of Discretionary Service Lines.* Tech. rept. Tuck School of Business, Dartmouth. Working Paper.

DeGraba, Patrick. 1995. Buying frenzies and seller-induced excess demand. *The RAND Journal of Economics*, **26**(2), 331–342.

Ebben, Paula. 2013. *Coffee Shops Limit Wi-Fi To Discourage 'Laptop Hobos'.* Tech. rept. https://boston.cbslocal.com/2013/08/05/coffee-shops-limit-wi-fi-to-discourage-laptop-hobos/.

Feldman, Pnina, Li, Jun, & Tsai, Hsin-Tien. 2018. Welfare Implications of Congestion Pricing: Evidence from SFpark. *Working paper.*

Ha, Albert Y. 2001. Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science*, **47**(7), 915–930.

Hassin, Refael. 2016. *Rational Queueing*. Chapman and Hall/CRC.

Hassin, Refael, & Haviv, Moshe. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems*. Vol. 59. Springer Science & Business Media.

Haviv, Moshe. 2014. Regulating an M/G/1 queue when customers know their demand. *Performance Evaluation*, **77**, 57–71.

Hopp, Wally J., Irvani, S.M.R., & Yuen, G.Y. 2007. Operations systems with discretionary task completion. *Management science*, **53**(1), 61–77.

Kostami, Vasiliki, & Rajagopalan, Sampath. 2013. Speed–quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management*, **16**(1), 104–118.

Lowenhahl, B. 2005. *Strategic Management of Professional Service Firms*. Copenhagaen Business School Press, Denmark.

Masuda, Yasushi, & Whang, Seungin. 2006. On the Optimality of Fixed-up-to Tariff for Telecommunications Service. *Information Systems Research*, **17**(3), 247–253.

Naor, Pinhas. 1969. The regulation of queue size by levying tolls. *Econometrica: Journal of the Econometric Society*, 15–24.

Plambeck, Erica L., & Wang, Qiong. 2013. Implications of Hyperbolic Discounting for Optimal Pricing and Scheduling of Unpleasant Services That Generate Future Benefits. *Management Science*, **59**(8), 1927–1946.

Randhawa, Ramandeep S., & Kumar, Sunil. 2008. Usage Restriction and Subscription Services: Operational Benefits with Rational Users. *Manufacturing & Service Operations Management*, **10**(3), 429–447.

Stavins, Robert N. 2011. The Problem of the Commons: Still Unsettled after 100 Years. *American Economic Review*, **101**, 81–108.

Stoyan, Dietrich, & Daley, Deryl J. 1983. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley and Sons, Inc.

Tong, Chunyang, & Rajagopalan, Sampath. 2014. Pricing and Operational Performance in Discretionary Services. *Production and Opereations Management*, **23**, 689–703.

Van Mieghem, Jan A. 2000. Price and service discrimination in queuing systems: Incentive compatibility of scheduling. *Management Science*, **46**, 1249–1267.

Wang, Ruxian, Dada, Maqbool, & Sahin, Ozge. 2018. Pricing Ancillary Service Subscriptions. *Management Science, Forthcoming.*

Xie, J., & Shugan, Steve. 2001. Electronic tickets, smart cards, and online prepayments: When and how to advance sell. *Marketing Science*, **20**(3), 219–243.

Yang, Luyi, & Wu, C. 2018. Bundle Pricing of Congested Services. *Working paper.*

Online Appendix: Proofs of Theorems and Lemmas

Proof of Lemma 1. We prove that whenever $p \leq \bar{\theta}$ there exists a unique $\theta^0 \in \left[0, \bar{\theta}\right]$ that solves the equation

$$\theta^0 = p + cW\left(\theta^0\right), \tag{3}$$

where

$$W\left(\theta^0\right) = \frac{\lambda \mathbb{E}\left[S^2\right]}{2\left(1 - \lambda \mathbb{E}\left[S\right]\right)} = \frac{\Lambda \int_{\theta^0}^{\bar{\theta}} x^2 f\left(x\right) dx}{2\left(1 - \Lambda \int_{\theta^0}^{\bar{\theta}} x f\left(x\right) dx\right)}.$$

Since both $\Lambda \int_{\theta^0}^{\bar{\theta}} x^2 f\left(x\right) dx$ and $\Lambda \int_{\theta^0}^{\bar{\theta}} x f\left(x\right) dx$ are decreasing in θ^0, W is decreasing in θ^0. The waiting time function $W\left(\theta^0\right)$ is positive (and hence well defined) for all θ^0 that satisfy $\int_{\theta^0}^{\bar{\theta}} x f\left(x\right) dx \leq \frac{1}{\Lambda}$. If $\Lambda \leq \left(\int_0^{\bar{\theta}} x f\left(x\right) dx\right)^{-1}$ the waiting time is well defined for all $\theta^0 \in \left[0, \bar{\theta}\right]$. Otherwise, it is well defined for all θ^0such that $\theta^0 \geq \hat{\theta} > 0$, where $\hat{\theta}$ is implicitly defined by $\int_{\theta}^{\bar{\theta}} x f\left(x\right) dx = 1/\Lambda$. The l.h.s decreases with θ. At $\theta = 0$ it is greater than the r.h.s and at $\theta = \bar{\theta}$ it is smaller. Therefore there exists a unique $\hat{\theta}$ that solves it. Let

$$\tilde{\theta} = \begin{cases} 0 & \Lambda \leq \frac{1}{\int_0^{\bar{\theta}} x f(x) dx} \\ \hat{\theta} & otherwise \end{cases}.$$

Then $W\left(\theta^0\right)$ is positive for every $\theta^0 \in \left[\tilde{\theta}, \bar{\theta}\right]$. At $\theta^0 = \tilde{\theta}$ the l.h.s of equation (3) is strictly smaller than the r.h.s since if $\tilde{\theta} = 0$ then the l.h.s equals zero while the r.h.s is positive and if $\tilde{\theta} = \hat{\theta} > 0$ then $\lim_{\theta^0 \to \hat{\theta}} W\left(\theta^0\right) = \infty$. For $\theta^0 = \bar{\theta}$ the l.h.s is greater or equal to the r.h.s. Therefore there exists a unique θ^0 that solves the equation. Finally, when p increases the l.h.s increases and therefore θ^0 increases.

□

Proof of Theorem 1. Differentiating $\pi\left(\theta^0\right) = \Lambda \bar{F}\left(\theta^0\right)\left(\theta^0 - cW\left(\theta^0\right)\right)$ and rearranging, we get

$$\left(1 - c\frac{W\left(\theta^0\right)}{\theta^0}\right)\frac{\theta^0 f\left(\theta^0\right)}{\bar{F}\left(\theta^0\right)} = 1 - cW'\left(\theta^0\right). \tag{4}$$

The l.h.s of increases in θ^0since F is IGFR and$W$ is decreasing in θ^0. Therefore $1 - cW\left(\theta^0\right)/\theta^0$ increases in θ^0. The r.h.s decreases with θ^0 if $W'\left(\theta^0\right)$ is increasing in θ^0. To see that write $W'\left(\theta^0\right)$ and rearrange:

$$W'\left(\theta^0\right) = -\frac{\Lambda\left(\theta^0\left(1 - \Lambda \int_{\theta^0}^{\bar{\theta}} x f\left(x\right) dx\right) + \Lambda \int_{\theta^0}^{\bar{\theta}} x^2 f\left(x\right) dx\right)}{2\frac{\left(1 - \Lambda \int_{\theta^0}^{\bar{\theta}} x f(x) dx\right)^2}{\bar{F}(\theta^0)}} \frac{\theta^0 f\left(\theta^0\right)}{\bar{F}\left(\theta^0\right)}$$

Since F is IGFR, $-\left(1 - \Lambda \int_{\theta^0}^{\bar{\theta}} x f\left(x\right) dx\right)^2 / \bar{F}\left(\theta^0\right)$ decreases with θ^0. Finally

$$\frac{d}{d\theta^0}\left(\theta^0\left(1 - \Lambda \int_{\theta^0}^{\bar{\theta}} x f\left(x\right) dx\right) + \Lambda \int_{\theta^0}^{\bar{\theta}} x^2 f\left(x\right) dx\right) = 1 - \Lambda \int_{\theta^0}^{\bar{\theta}} x f\left(x\right) dx + \Lambda\left(\theta^0\right)^2 f\left(\theta^0\right) - \Lambda\left(\theta^0\right)^2 f\left(\theta^0\right) > 0$$

Furthermore, when $\theta^0 = \tilde{\theta}$ the l.h.s of equation (4) is negative (if $\tilde{\theta} = 0$ then $\left(\theta^0 - cW\left(\theta^0\right)\right)\frac{f\left(\theta^0\right)}{\bar{F}(\theta^0)} = -cW\left(0\right)f\left(0\right) < 0$ and if $\tilde{\theta} = \hat{\theta} > 0$ then $\lim_{\theta^0 \to \hat{\theta}} W\left(\theta^0\right) = \infty$ and the l.h.s is negative) while the r.h.s is

positive (because $W'\left(\theta^0\right) < 0$). For $\theta^0 = \bar{\theta}$ we have $\lim_{\theta^0 \to \bar{\theta}} \left(\theta^0 - cW\left(\theta^0\right)\right) \frac{f\left(\theta^0\right)}{\bar{F}\left(\theta^0\right)} = \lim_{\theta^0 \to \bar{\theta}} \frac{\theta^0 f\left(\theta^0\right)}{\bar{F}\left(\theta^0\right)} = \infty$ while $\lim_{\theta^0 \to \bar{\theta}} \left(1 - cW'\left(\theta^0\right)\right) = 1 + c\Lambda\left(\bar{\theta}\right)^2 f\left(\bar{\theta}\right)/2 < \infty$. Therefore there exists a unique θ_f that solves equation (4). Since at $\theta^0 = 0$ the revenue function is increasing and θ_f is such that the corresponding price $p_f = \theta_f - cW\left(\theta_f\right)$ is positive.

□

Proof of Theorem 2. For consumer surplus we show that $CS_f\left(\theta^0\right)$ decreases in θ^0 and therefore the optimal $\hat{\theta}_f$ is the smallest, such that $p = \hat{\theta}_f - cW\left(\hat{\theta}_f\right) = 0$. Differentiating, we get $dCS_f/d\theta^0 = -\Lambda\theta^0 f\left(\theta^0\right) + \Lambda f\left(\theta^0\right)\theta^0 - \Lambda\bar{F}\left(\theta^0\right) < 0$. Since θ^0 increases in p there exists a unique, smallest, $\hat{\theta}_f$ for which $p = \hat{\theta}_f - cW\left(\hat{\theta}_f\right) = 0$ or $\hat{\theta}_f = cW\left(\hat{\theta}_f\right)$.

□

Proof of Lemma 2. We prove that whenever $p \leq 1$ there exists a unique $\theta^0 \in \left[0, \bar{\theta}\right]$ that solves $\theta^0(1-p) = cW\left(\theta^0\right)$. W decreases in θ^0 and therefore the l.h.s of (2) increases with θ^0while the r.h.s decreases. Moreover, when $\theta^0 = \tilde{\theta}$ the l.h.s is strictly smaller than the r.h.s (if $\tilde{\theta} = 0$ then the l.h.s is equal to zero while the r.h.s is positive, and if $\tilde{\theta} = \hat{\theta} > 0$ then $\lim_{\theta^0 \to \hat{\theta}} W\left(\theta^0\right) = \infty$ while the l.h.s is positive). When $\theta^0 = \bar{\theta}$ the l.h.s is larger or equal to the r.h.s (since $W\left(\bar{\theta}\right) = 0$). Therefore there exists a unique θ^0 that solves the equation. Finally, when p increases the l.h.s decreases and therefore θ^0 increases.

□

Proof of Theorem 3. Differentiating the revenue function $\pi_r\left(\theta^0\right)$ and rearranging we get:

$$\left(1 - c\frac{W\left(\theta^0\right)}{\theta^0}\right)\frac{\theta^0 f\left(\theta^0\right)}{\bar{F}\left(\theta^0\right)} = c\left(\frac{W\left(\theta^0\right)}{\theta^0} - W'\left(\theta^0\right)\right)\frac{\int_{\theta^0}^{\bar{\theta}} xf\left(x\right)dx}{\theta^0\bar{F}\left(\theta^0\right)} \tag{5}$$

The l.h.s of (5) increases with θ^0 since $W\left(\theta^0\right)$ decreases with θ^0 and F is IGFR. We show that the r.h.s decreases with θ^0. In the proof of Theorem (1) we showed that $W'\left(\theta^0\right)$ increases with θ^0. Moreover $W\left(\theta^0\right)/\theta^0$ decreases with θ^0. It remains to show that $\int_{\theta^0}^{\bar{\theta}} xf\left(x\right)dx/\theta^0\bar{F}\left(\theta^0\right) = \mathbb{E}\left[S|S \geq \theta^0\right]/\theta^0$ decreases with θ^0. If F is IGFR, $E\left[S - x|S > x\right]/x \geq E\left[S - (x+y)\,|S > x+y\right]/(x+y)$ (see e.g. Stoyan & Daley (1983)). Therefore $E\left[S|S > x\right]/x = 1 + E\left[S - x|s > x\right]/x \geq 1 + E\left[S - (x+y)\,|S > x+y\right]/(x+y) = E\left[S|S > x+y\right]/(x+y)$. Finally, at $\theta^0 = \tilde{\theta}$ we have that $\left(\theta^0 - cW\left(\theta^0\right)\right)f\left(\theta^0\right)/\bar{F}\left(\theta^0\right) < 0$. If $\tilde{\theta} = 0$ then

$$\lim_{\theta^0 \to 0}\left(c\left(\frac{W\left(\theta^0\right)}{\theta^0} - W'\left(\theta^0\right)\right)\frac{\int_{\theta^0}^{\bar{\theta}} xf\left(x\right)dx}{\theta^0\bar{F}\left(\theta^0\right)}\right) = c\int_0^{\bar{\theta}} xf\left(x\right)dx \cdot \left(W\left(0\right)\lim_{\theta^0 \to 0}\left(\frac{1}{\left(\theta^0\right)^2}\right) - \lim_{\theta^0 \to 0}\frac{W'\left(\theta^0\right)}{\theta^0}\right) > 0$$

since

$$\frac{W'\left(\theta^0\right)}{\theta^0} = -\frac{\Lambda f\left(\theta^0\right)\left(\theta^0\left(1 - \Lambda\int_{\theta^0}^{\bar{\theta}} xf\left(x\right)dx\right) + \Lambda\int_{\theta^0}^{\bar{\theta}} x^2 f\left(x\right)dx\right)}{2\left(1 - \Lambda\int_{\theta^0}^{\bar{\theta}} xf\left(x\right)dx\right)^2}$$

and

$$\lim_{\theta^0\to 0}\frac{W'\left(\theta^0\right)}{\theta^0}=-\frac{\Lambda^2 f\left(0\right)\int_0^{\bar{\theta}}x^2 f\left(x\right)dx}{2\left(1-\Lambda\int_0^{\bar{\theta}}xf\left(x\right)dx\right)^2}.$$

Moreover, if $\tilde{\theta}=\bar{\theta}>0$ then $\lim_{\theta^0\to\hat{\theta}}W\left(\theta^0\right)=\infty$ and the r.h.s is positive. When $\theta^0=\bar{\theta}$ the l.h.s of (5) goes to infinity while for the r.h.s we have

$$\lim_{\theta^0\to\bar{\theta}}\left(c\left(\frac{W\left(\theta^0\right)}{\theta^0}-W'\left(\theta^0\right)\right)\frac{\int_{\theta^0}^{\bar{\theta}}xf\left(x\right)dx}{\theta^0\bar{F}\left(\theta^0\right)}\right)=c\Lambda\frac{f\left(\bar{\theta}\right)\left(\bar{\theta}\right)^2}{2}>0$$

Therefore there exists a unique θ_r that solves equation (5). Since at $\theta^0=0$ the revenue function is increasing θ_r is such that the corresponding price $p_r=1-cW\left(\theta_r\right)/\theta_r$ is positive. The function $CS_r\left(\theta^0\right)=\Lambda cW\left(\theta^0\right)\left(\frac{1}{\theta^0}\int_{\theta^0}^{\bar{\theta}}xf\left(x\right)dx-\bar{F}\left(\theta^0\right)\right)$ decreases in θ^0 so the optimal $\hat{\theta}_r$ is the smallest, such that $p_r=1-cW\left(\theta_r\right)/\theta_r=0$. Since $dW\left(\theta^0\right)/d\theta_0<0$, we have

$$\frac{dCS_r}{d\theta^0}=\Lambda c\left(\frac{dW\left(\theta^0\right)}{d\theta_0}\left(\frac{1}{\theta^0}\int_{\theta^0}^{\bar{\theta}}xf\left(x\right)dx-\bar{F}\left(\theta^0\right)\right)+W\left(\theta^0\right)\left(-\frac{1}{\left(\theta^0\right)^2}\int_{\theta^0}^{\bar{\theta}}xf\left(x\right)dx\right)\right)<0.$$

Moreover, since θ^0 increases in p there exists a unique, smallest, $\hat{\theta}_r$ for which $p_r=1-cW\left(\theta_r\right)/\theta_r=0$ or $\hat{\theta}_r=cW\left(\hat{\theta}_r\right)$. Therefore $\hat{\theta}_r=\hat{\theta}_f$.

□

Proof of Lemma 3. Assuming there exist two unique thresholds $0\leq\theta_l\leq T$ and $T\leq\theta_h\leq\bar{\theta}$ such that customers with $\theta\in[\theta_l,T]$ stay for θ units of time, and customers with $\theta\in[T,\theta_h]$ use the service for T units of time, the expected waiting time is:

$$W_T\left(\theta_l,\theta_h\right)=\frac{\lambda\mathbb{E}\left[S^2\right]}{2\left(1-\lambda\mathbb{E}\left[S\right]\right)}=\frac{\Lambda\left(\int_{\theta_l}^{T}x^2f\left(x\right)dx+T^2\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)}{2\left(1-\Lambda\left(\int_{\theta_l}^{T}xf\left(x\right)dx+T\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)\right)}.$$

When $p>T$ no consumer has a positive utility from service (since in this case for all $0\leq\theta_l\leq T$ we have $\theta_l-p<0$ and for all $T\leq\theta_h\leq\bar{\theta}$ we have $\left(T-\beta\theta_h\right)/\left(1-\beta\right)-p<0$). Therefore no consumer joins and $\theta_l=\theta_h=T$. Assume therefore that $p\leq T$. Suppose first that $0\leq\beta<1$. The types $0\leq\theta_l\leq T$ and $T\leq\theta_h\leq\bar{\theta}$ are defined by the equations $\left(T-\beta\theta_h\right)/\left(1-\beta\right)-p-cW_T\left(\theta_l,\theta_h\right)=0$ and $\theta_l-p-cW_T\left(\theta_l,\theta_h\right)=0$. Types with $\theta\geq T/\beta$ will not join and therefore $T\leq\theta_h\leq\min\left\{T/\beta,\bar{\theta}\right\}$. Let $\theta_l\left(\theta_h\right)=\left(T-\beta\theta_h\right)/\left(1-\beta\right)$ then for $T\leq\theta_h\leq\min\left\{T/\beta,\bar{\theta}\right\}$ we get that $\max\left\{0,\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right)\right\}\leq\theta_l\left(\theta_h\right)\leq T$. We can therefore express the two equations as a function of θ_h:

$$\frac{T-\beta\theta_h}{1-\beta}-p=cW_T\left(\frac{T-\beta\theta_h}{1-\beta},\theta_h\right). \tag{6}$$

The l.h.s of (6) decreases with θ_h. Moreover, $\frac{dW}{d\theta_h}=\frac{\partial W}{\partial\theta_l}\frac{d\theta_l}{d\theta_h}+\frac{\partial W}{\partial\theta_h}$. Since $d\theta_l/d\theta_h=-\beta/\left(1-\beta\right)<0$,

$$\frac{\partial W}{\partial\theta_l}=-\frac{\Lambda\left(\theta_l\left(1-\Lambda\left(\int_{\theta_l}^{T}xf\left(x\right)dx+T\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)\right)+\Lambda\left(\int_{\theta_l}^{T}x^2f\left(x\right)dx+T^2\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)\theta_lf\left(\theta_l\right)}{2\left(1-\Lambda\left(\int_{\theta_l}^{T}xf\left(x\right)dx+T\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)\right)^2}<0,$$

and

$$\frac{\partial W}{\partial \theta_h}=\frac{\Lambda\left(T\left(1-\Lambda\left(\int_{\theta_l}^{T}xf\left(x\right)dx+T\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)\right)+\Lambda\left(\int_{\theta_l}^{T}x^2f\left(x\right)dx+T^2\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)Tf\left(\theta_h\right)}{2\left(1-\Lambda\left(\int_{\theta_l}^{T}xf\left(x\right)dx+T\left(F\left(\theta_h\right)-F\left(T\right)\right)\right)\right)^2}>0$$

we have that the r.h.s of (6) increases with θ_h. At $\theta_h=T$ the l.h.s is positive and the r.h.s is zero. For $T\leq\beta\bar{\theta}$ we have that at $\theta_h=T/\beta$ the l.h.s is negative while the r.h.s is positive and therefore there exists a unique solution for equation (6), $T\leq\theta_h\leq T/\beta$ and $\theta_l=\theta_l\left(\theta_h\right)$. For $T\geq\beta\bar{\theta}$ a sufficient condition for a unique solution is that at $\theta_h=\bar{\theta}$ we have $\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right)-p\leq cW_T\left(\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right),\bar{\theta}\right)$ or

$$p\geq\frac{T-\beta\bar{\theta}}{1-\beta}-cW_T\left(\frac{T-\beta\bar{\theta}}{1-\beta},\bar{\theta}\right). \tag{7}$$

When (7) holds we have a unique solution for equation (6), $T\leq\theta_h\leq\bar{\theta}$ and $\theta_l\left(\theta_h\right)$. The r.h.s of (7) is negative for $T=\beta\bar{\theta}$ and equal to $\bar{\theta}$ for $T=\bar{\theta}$. Moreover $\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right)-cW_T\left(\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right),\bar{\theta}\right)\leq T$. When (7) does not hold there exists no $T\leq\theta_h\leq\bar{\theta}$ that solves (6) and all consumers with $T\leq\theta\leq\bar{\theta}$ would like to use the service for T units of time. In this case we have $\theta_h=\bar{\theta}$ and $0\leq\theta_l\leq T$ is defined by

$$\theta_l-p=cW_T\left(\theta_l,\bar{\theta}\right) \tag{8}$$

Note that the l.h.s of (8) increases with θ_l and the r.h.s decreases with θ_l. At $\theta_l=0$ the l.h.s is negative and the r.h.s is positive. Thus, a sufficient condition for the existence of a unique solution $0\leq\theta_l\leq T$ for equation (8) is that at $\theta_l=T$ we have $p\leq T-cW_T\left(T,\bar{\theta}\right)$. Note that for all $\beta\bar{\theta}\leq T\leq\bar{\theta}$, $T-cW_T\left(T,\bar{\theta}\right)\geq\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right)-cW_T\left(\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right),\bar{\theta}\right)$ since $T\geq\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right)$ and W decreases in θ_l. Therefore when $p\leq\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right)-cW_T\left(\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right),\bar{\theta}\right)$ we have $\theta_h=\bar{\theta}$ and θ_l is uniquely defined by (8). For $\beta=0$ condition (7) becomes $p\geq T-cW_T\left(T,\bar{\theta}\right)$. Therefore, for each such p we have $\theta_l=T$ and θ_h is defined by $T-p=cW_T\left(T,\theta_h\right)$. Note that in this case, although all consumers with $\theta\geq T$ enjoy the same utility if they join, an equilibrium exists when only those with $\theta\leq\theta_h$ enter. All others do not join. The utility of all consumers in this equilibrium is zero. When $p\leq T-cW_T\left(T,\bar{\theta}\right)$ we have $\theta_h=\bar{\theta}$ and θ_l is defined by equation (8). For $\beta=1$ we have $\theta_h=T$. No consumer with $\theta>T$ joins since her utility is zero. In this case θ_l is defined implicitly by $\theta_l-p=cW_T\left(\theta_l,T\right)$. The l.h.s of this equation increases with θ_l and the r.h.s decreases. Moreover, at $\theta_l=0$ the l.h.s is negative while the r.h.s is positive. Finally at $\theta_l=T$ the l.h.s is positive while the r.h.s is equal to zero and uniquness is guaranteed.

□

Proof of Theorem 4. If the firm sets a price $p\geq T$ then its revenue is zero. We therefore assume that $p<T$. Whenever $p\leq\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right)-cW_T\left(\left(T-\beta\bar{\theta}\right)/\left(1-\beta\right),\bar{\theta}\right)$, we have that $\theta_h=\bar{\theta}$ and θ_l is uniquely defined by $\theta_l-p=cW_T\left(\theta_l,\bar{\theta}\right)$. For $T=\bar{\theta}$ the inequality holds since $p<\bar{\theta}$. Therefore there exists a $\tilde{T}\left(p\right)<\bar{\theta}$ such that the inequality holds for all $\tilde{T}\leq T\leq\bar{\theta}$. For all such T we have $\pi_f\left(p,T\right)=\Lambda p\left(1-F\left(\theta_l\right)\right)$. Then $\partial\pi_f/\partial T=-\Lambda pf\left(\theta_l\right)d\theta_l/dT$ and $d\theta_l/dT=c\left(\partial W/\partial T\right)/\left(1-c\partial W/\partial\theta_l\right)$. Finally

$$\frac{\partial W}{\partial T}=\frac{\Lambda\left(2T\left(1-\Lambda\left(\int_{\theta_l}^{T}xf\left(x\right)dx+T\left(1-F\left(T\right)\right)\right)\right)+\Lambda\left(\int_{\theta_l}^{T}x^2f\left(x\right)dx+T^2\left(1-F\left(T\right)\right)\right)\left(1-F\left(T\right)\right)}{2\left(1-\Lambda\left(\int_{\theta_l}^{T}xf\left(x\right)dx+T\left(1-F\left(T\right)\right)\right)\right)^2}>0$$

while $\partial W/\partial\theta_l<0$. Therefore $d\theta_l/dT>0$. We conclude that for every p we have $\partial\pi_f/\partial T<0$ for all $\tilde{T}\left(p\right)\leq T\leq\bar{\theta}$ and therefore the optimal time limit $T_f\leq\tilde{T}\left(p\right)<\bar{\theta}$.

□

Proof of Theorem 5. For $\beta = 1$ and uniform distribution we have that $\theta_h = T$ and

$$W_T(\theta_l, T) = \frac{\Lambda \frac{1}{3}\left(T^3 - \theta_l^3\right)}{2\left(\bar{\theta} - \Lambda \frac{1}{2}\left(T^2 - \theta_l^2\right)\right)}$$

The firm's revenue is $\pi_f(p,T) = \Lambda p(T - \theta_l)/\bar{\theta}$, where θ_l is defined by $\theta_l - p = cW_T(\theta_l, T)$. Then $\partial \pi_f / \partial T = \Lambda p(1 - d\theta_l/dT)/\bar{\theta}$ and $d\theta_l/dT = c\partial W_T(\theta_l, T)/\partial T/(1 - c\partial W_T(\theta_l, T)/\partial \theta_l) > 0$. Moreover $\partial \pi_f / \partial T \geq 0$ for all T if and only if

$$\frac{c\frac{\partial W_T(\theta_l, T)}{\partial T}}{1 - c\frac{\partial W_T(\theta_l, T)}{\partial \theta_l}} \leq 1 \Longleftrightarrow \frac{\partial W_T(\theta_l, T)}{\partial T} + \frac{\partial W_T(\theta_l, T)}{\partial \theta_l} \leq \frac{1}{c}.$$

In equilibrium $\theta_l - p = cW_T(\theta_l, T)$ and $1/c = p/c\theta_l + W/\theta_l$. Therefore at the optimal price it holds that

$$\frac{\partial W_T(\theta_l, T)}{\partial T} + \frac{\partial W_T(\theta_l, T)}{\partial \theta_l} \leq \frac{p}{c\theta_l} + \frac{W}{\theta_l}$$

Now, $\partial \pi_f / \partial p = \Lambda(T - \theta_l - p \cdot d\theta_l/dp)/\bar{\theta}$. Since at $p = 0$ and $p = T$ the revenue is zero, the optimal price is defined by the f.o.c $T - \theta_l - p \cdot d\theta_l/dp = 0$ or $p = (T - \theta_l)(1 - c\partial W_T(\theta_l, T)/\partial \theta_l)$. Therefore after rearranging, we show that

$$\theta_l \frac{\partial W_T(\theta_l, T)}{\partial T} + T\frac{\partial W_T(\theta_l, T)}{\partial \theta_l} - W \leq \frac{(T - \theta_l)}{c}. \tag{9}$$

Finally, inequality (9) holds because

$$\theta_l \frac{\partial W_T(\theta_l, T)}{\partial T} + T\frac{\partial W_T(\theta_l, T)}{\partial \theta_l} - W = -\frac{\frac{1}{3}\Lambda(T - \theta_l)^3}{2\left(\bar{\theta} - \Lambda \frac{1}{2}\left(T^2 - \theta_l^2\right)\right)} < 0$$

For $\beta = 0$ if $p \geq T - cW_T\left(T, \bar{\theta}\right)$ then in equilibrium $\theta_l = T$ and θ_h is uniquely defined by $T - p = cW_T(T, \theta_h)$. In this case, the firm's revenue is $\pi_f(p,T) = 2p(1 - p/T)/(2(T - p) + cT)$. The optimal price given that $p \geq T - cW_T\left(T, \bar{\theta}\right)$) is $\max\left\{T - cW_T\left(T, \bar{\theta}\right), T\left(c + 2 + \sqrt{c(c+2)}\right)/2\right\}$, where $T\left(c + 2 + \sqrt{c(c+2)}\right)/2 \geq T - cW_T\left(T, \bar{\theta}\right)$ for every T in the interval $\left(\bar{\theta} \pm \sqrt{\bar{\theta}^2 - 4\bar{\theta}\left(1 - \sqrt{c/(c+2)}\right)}/\Lambda\right)/2$. The interval exists whenever $\Lambda \geq 4\left(1 - \sqrt{c/(c+2)}\right)/\bar{\theta}$ and for all such T the revenue is $\pi_f(p,T) = c + 1 - \sqrt{c(c+2)}$. If

$$p \leq T - cW_T\left(T, \bar{\theta}\right) = T - c\frac{\Lambda T^2\left(\bar{\theta} - T\right)}{2\left(\bar{\theta} - \Lambda T\left(\bar{\theta} - T\right)\right)}. \tag{10}$$

and $\Lambda \leq 8/\bar{\theta}(2 + c)$ the r.h.s is non negative for all $0 \leq T \leq \bar{\theta}$. For $\Lambda > 8/\bar{\theta}(2 + c)$ the r.h.s is positive only when $0 \leq T \leq \left(\bar{\theta} - \sqrt{\bar{\theta}(1 - 8/\Lambda(2 + c))}\right)/2$ or $\left(\bar{\theta} - \sqrt{\bar{\theta}(1 - 8/\Lambda(2 + c))}\right)/2 \leq T \leq \bar{\theta}$. Moreover $\pi_f = \Lambda p\left(1 - \theta_l/\bar{\theta}\right)$ and $\partial \pi_f / \partial T = -\Lambda p(d\theta_l/dT)/\bar{\theta}$. Therefore, given $p \leq T - cW_T\left(T, \bar{\theta}\right)$ the optimal T is the smallest such that condition (10) holds. The equality $T - cW_T\left(T, \bar{\theta}\right) = p$ can be written as a cubic polynomial in T: $\Lambda(c + 2)T^3 - \Lambda\left(2p + (2 + c)\bar{\theta}\right)T^2 + 2\bar{\theta}(1 + p\Lambda)T - 2\bar{\theta}p = 0$. At $T = 0$ this polynomial is negative while at $T = \bar{\theta}$ it is equal to $2\bar{\theta}\left(\bar{\theta} - p\right) > 0$. Therefore there either exists a unique real root of the polynomial, T_1, such that for all $T_1 \leq T \leq \bar{\theta}$ condition (10) holds or there exists three real roots T_1, T_2, T_3 such that condition (10) holds for every $T_1 \leq T \leq T_2$ and $T_3 \leq T \leq \bar{\theta}$. In either case there exists a unique smallest $T(p)$ for which condition (10) holds and this is therefore the optimal time limit, under condition (10). At this optimal time limit it is therefore true that $p = T - cW_T\left(T, \bar{\theta}\right)$ which implies that $\theta_l = T$. All consumers with $T \leq \theta \leq \bar{\theta}$ join and use it for T units of time and no other consumer joins. In this case we can express the revenue as a function of only the time limit T: $\pi_f = \Lambda\left(T - cW_T\left(T, \bar{\theta}\right)\right)\left(\bar{\theta} - T\right)/\bar{\theta}$. This

revenue function is symmetric around $\bar{\theta}/2$ since $\pi_f(T) = \pi_f\left(\bar{\theta} - T\right)$. Moreover

$$\pi_f'(T) = \frac{\left(\bar{\theta} - 2T\right)\left(-T\Lambda\left(\bar{\theta} - T\right)(c+2)\left(2\bar{\theta} - T\Lambda\left(\bar{\theta} - T\right)\right) + 2\left(\bar{\theta}\right)^2\right)}{2\left(\bar{\theta} - T\Lambda\left(\bar{\theta} - T\right)\right)^2}.$$

$\pi_f'(T) = 0$ for $T = \bar{\theta}/2$ and for T that solves the equation $-T\Lambda\left(\bar{\theta} - T\right)(c+2)\left(2\bar{\theta} - T\Lambda\left(\bar{\theta} - T\right)\right) + 2\left(\bar{\theta}\right)^2 = 0$. The s.o.c is $\pi_f''(T)|_{T=\frac{\bar{\theta}}{2}} = -\left(\left(\bar{\theta}\right)^2(2+c)\Lambda^2 - 8\bar{\theta}(2+c)\Lambda + 32\right)/\left(4 - \bar{\theta}\Lambda\right)^2$. $\pi_f''(T)|_{T=\bar{\theta}/2} < 0$ iff $\Lambda \leq 4\left(1 - \sqrt{c/(2+c)}\right)/\bar{\theta}$ or $\Lambda \geq 4\left(1 + \sqrt{c/(2+c)}\right)/\bar{\theta}$. Thus, the optimal time limit is $T = \bar{\theta}/2$ whenever $\Lambda \leq 4\left(1 - \sqrt{c/(2+c)}\right)/\bar{\theta}$ or $\Lambda \geq 4\left(1 + \sqrt{c/(2+c)}\right)/\bar{\theta}$. Note that $4\left(1 - \sqrt{c/(2+c)}\right)/\bar{\theta} \leq 8/\left(\bar{\theta}(2+c)\right) \leq 4\left(1 + \sqrt{c/(2+c)}\right)/\bar{\theta}$. We show that in this case revenue is not higher for $p \geq T - cW_T\left(T, \bar{\theta}\right)$. When $\Lambda \leq 4\left(1 - \sqrt{c/(2+c)}\right)/\bar{\theta}$ we have $T\left(c + 2 + \sqrt{c(c+2)}\right)/2 \leq T - cW_T\left(T, \bar{\theta}\right)$ for every T and therefore the optimal price is $p(T) = T - cW_T\left(T, \bar{\theta}\right)\ \forall T$ and the same revenue is obtained as in $p \leq T - cW_T\left(T, \bar{\theta}\right)$. For $\Lambda \geq 4\left(1 - \sqrt{c/(2+c)}\right)/\bar{\theta}$ while the optimal price when $p \leq T - cW_T\left(T, \bar{\theta}\right)$ is $p = T - cW_T\left(T, \bar{\theta}\right)\ \forall T$, we showed that when $p \geq T - cW_T\left(T, \bar{\theta}\right)$, revenue is always higher as long as $T\left(c + 2 + \sqrt{c(c+2)}\right)/2 \geq T - cW_T\left(T, \bar{\theta}\right)$. Thus, any time limit T such that $\left(\bar{\theta} - \sqrt{\bar{\theta}\left(\bar{\theta} - 4\left(1 - \sqrt{c/(2+c)}\right)/\Lambda\right)}\right)/2 \leq T \leq \left(\bar{\theta} + \sqrt{\bar{\theta}\left(\bar{\theta} - 4\left(1 - \sqrt{c/(2+c)}\right)/\Lambda\right)}\right)/2$ and the corresponding price $p(T) = T\left(c + 2 + \sqrt{c(c+2)}\right)/2$ is optimal. The optimal thresholds are $\theta_l = T$ and $\theta_h = T + \bar{\theta}\left(1 - \sqrt{c/(c+2)}\right)/T\Lambda$. The optimal revenue is $\pi_f^T = c + 1 - \sqrt{c(c+2)}$. In all these optimal combinations the firm extracts all surplus from consumers (all consumers receive 0 utility).

□

Proof of Theorem 6. We start by showing that the optimal price is $\hat{p}_f = 0$ for $0 \leq \beta < 1$. Consumer surplus is

$$CS(p,T) = \Lambda\left(\int_{\theta_l}^{T} \theta f(\theta)\, d\theta + \int_{T}^{\theta_h} \frac{T - \beta\theta}{1-\beta} f(\theta)\, d\theta - \left(p + cW_T(\theta_l, \theta_h)\right)\left(F(\theta_h) - F(\theta_l)\right)\right).$$

When T is small enough such that $\theta_h < \bar{\theta}$, the following equations hold in equilibrium: $\left(T - \beta\theta_h\right)/(1-\beta) - p = cW_T(\theta_l, \theta_h)$ and $\theta_l = \left(T - \beta\theta_h\right)/(1-\beta)$. We can rewrite consumer surplus as

$$CS(p,T) = \Lambda\left(\int_{\theta_l}^{T} \theta f(\theta)\, d\theta - \frac{1}{1-\beta} T\left(F(T) - F(\theta_l)\right) + \frac{\beta}{1-\beta}\left(\theta_h\left(F(\theta_h) - F(\theta_l)\right) - \int_{T}^{\theta_h} \theta f(\theta)\, d\theta\right)\right).$$

Differentiating consumer surplus as a function of p and rearranging:

$$\frac{\partial CS}{\partial p} = \Lambda \frac{\beta}{1-\beta}\left(F(\theta_h) - F(\theta_l)\right)\frac{d\theta_h}{dp}.$$

Since $d\theta_h/dp = -\left(\beta/(1-\beta) + c\partial W/\partial\theta_h\right)^{-1}$ and $dW/d\theta_h = -\partial W/\partial\theta_l \cdot \beta/(1-\beta) + \partial W/\partial\theta_h > 0$, $\partial CS/\partial p < 0$ and the optimal price, given the time limit is $\hat{p}_f = 0$. If T is such that $\theta_h = \bar{\theta}$ then

$$CS(p,T) = \Lambda\left(\int_{\theta_l}^{T} \theta f(\theta)\, d\theta + \int_{T}^{\bar{\theta}} \frac{T - \beta\theta}{1-\beta} f(\theta)\, d\theta - \left(p + cW_T\left(\theta_l, \bar{\theta}\right)\right)\left(1 - F(\theta_l)\right)\right),$$

where $\theta_l - p = cW_T\left(\theta_l, \bar{\theta}\right)$ and therefore $CS\left(p,T\right) = \Lambda\left(\int_{\theta_l}^{T} \theta f\left(\theta\right) d\theta + \int_{T}^{\bar{\theta}} \left(T - \beta\theta\right)/\left(1-\beta\right) f\left(\theta\right) d\theta - \theta_l\left(1 - F\left(\theta_l\right)\right)\right)$. Now $\partial CS/\partial p = \Lambda\left(-\left(1-F\left(\theta_l\right)\right) d\theta_l/dp\right)$ and $d\theta_l/dp = \left(1 - c\partial W/\partial\theta_l\right)^{-1} > 0$. Therefore for such T, $\partial CS/\partial p < 0$ and $\hat{p}_f = 0$.

When $\beta = 1$, $\theta_h = T$ and $CS\left(p,T\right) = \Lambda\left(\int_{\theta_l}^{T} \theta f\left(\theta\right) d\theta - \left(p + cW_T\left(\theta_l, T\right)\right)\left(F\left(T\right) - F\left(\theta_l\right)\right)\right)$, where $\theta_l - p = cW_T\left(\theta_l, T\right)$. Then $CS\left(p,T\right) = \Lambda\left(\int_{\theta_l}^{T} \theta f\left(\theta\right) d\theta - \theta_l\left(F\left(T\right) - F\left(\theta_l\right)\right)\right)$ and $\partial CS/\partial p = -\Lambda\left(F\left(T\right) - F\left(\theta_l\right) d\theta_l/dp\right) < 0$. Therefore $\hat{p}_f = 0$ for $\beta = 1$ as well. We therefore set $p = 0$ for the rest of the proof. Next we prove that for a uniform $\left[0, \bar{\theta}\right]$ distribution when $\beta = 1$, there exists $\Lambda_1 > 0$ such that if $\Lambda \leq \Lambda_1$, consumer surplus is maximized when $\hat{T}_f = \bar{\theta}$ and otherwise the social planner sets a time limit, i.e., $\hat{T}_f < \bar{\theta}$. We have

$$\frac{dCS}{dT} = \Lambda\left(\left(T - \theta_l\right) f\left(T\right) - \frac{d\theta_l}{dT}\left(F\left(T\right) - F\left(\theta_l\right)\right)\right) = \Lambda\frac{1}{\bar{\theta}}\left(T - \theta_l\right)\left(1 - \frac{d\theta_l}{dT}\right)$$

where

$$\frac{d\theta_l}{dT} = \frac{c\frac{\partial W}{\partial T}}{1 - c\frac{\partial W}{\partial \theta_l}} = \frac{c\frac{\Lambda\left(T^2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right) + \frac{1}{3}\left(T^3 - \theta_l^3\right)T\Lambda\right)}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)^2}}{1 + c\frac{\Lambda\left(\theta_l^2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right) + \frac{1}{3}\left(T^3 - \theta_l^3\right)\theta_l\Lambda\right)}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)^2}} \leq 1 \Leftrightarrow$$

$$\Lambda\left(\frac{\left(T^2 - \theta_l^2\right)\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right) + \frac{1}{3}\left(T^3 - \theta_l^3\right)\left(T - \theta_l\right)\Lambda}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)^2}\right) \leq \frac{1}{c}.$$

In equilibrium we have $1/c = W_T\left(\theta_l, T\right)/\theta_l$. Therefore

$$\frac{d\theta_l}{dT} \leq 1 \Leftrightarrow \Lambda\left(\frac{\left(T^2 - \theta_l^2\right)\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right) + \frac{1}{3}\left(T^3 - \theta_l^3\right)\left(T - \theta_l\right)\Lambda}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)^2}\right) \leq \frac{\Lambda\frac{1}{3}\left(T^3 - \theta_l^3\right)}{2\theta_l\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)}$$

$$\Leftrightarrow \frac{\Lambda\frac{1}{3}\left(T^3 - \theta_l^3\right)}{2\theta_l\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)} \leq \frac{\left(\frac{1}{3}\left(T^3 - \theta_l^3\right) - \left(T^2 - \theta_l^2\right)\theta_l\right)}{2\left(T - \theta_l\right)\theta_l^2}.$$

The l.h.s is equal to $1/c$, so the condition becomes

$$\frac{1}{c} \leq \frac{\left(\frac{1}{3}\left(T^3 - \theta_l^3\right) - \left(T^2 - \theta_l^2\right)\theta_l\right)}{2\left(T - \theta_l\right)\theta_l^2} \Leftrightarrow 0 \leq \frac{\theta_l}{T} \leq \frac{\sqrt{3c\left(c+2\right)}}{2\left(c+3\right)} < 1.$$

Now $d\left(\theta_l/T\right)/dT = \left(T \cdot d\theta_l/dT - \theta_l\right)/T^2$ where

$$\frac{d\theta_l}{dT}T - \theta_l = \frac{c\Lambda\left(\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right) + \frac{1}{3}\Lambda\left(T^2 - \theta_l^2\right)\right)\left(T^3 - \theta_l^3\right) - \theta_l\left(2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)^2\right)}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right)^2 + c\Lambda\left(\theta_l^2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)\right) + \frac{1}{3}\left(T^3 - \theta_l^3\right)\theta_l\Lambda\right)} \geq 0$$

since by rearranging we get

$$\frac{d\theta_l}{dT}T - \theta_l \geq 0 \Leftrightarrow \frac{3\left(\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right) + \frac{1}{3}\Lambda\left(T^2 - \theta_l^2\right)\right)}{\bar{\theta} - \Lambda\frac{1}{2}\left(T^2 - \theta_l^2\right)} \geq 1$$

which holds. Therefore θ_l/T increases in T. If $\theta_l\left(\bar{\theta}\right)/\bar{\theta} \leq \sqrt{3c\left(c+2\right)}/\left(2\left(c+3\right)\right)$, we conclude that $dCS/dT \geq 0\ \forall T$ and the optimal T is $\hat{T}_f = \bar{\theta}$. Otherwise, there exists an optimal T such that $\hat{T}_f < \bar{\theta}$. Finally note that $\theta_l\left(\bar{\theta}\right)/\bar{\theta}$ increases in Λ since $d\theta_l\left(\bar{\theta}\right)/d\Lambda = c\partial W_{\bar{\theta}}\left(\theta_l, \bar{\theta}\right)/\partial\Lambda/\left(1 - c\partial W_{\bar{\theta}}\left(\theta_l, \bar{\theta}\right)/\partial\theta_l\right) > 0$. As $\lim_{\Lambda\to 0}\theta_l\left(\bar{\theta}\right)/\bar{\theta} = 0$ and $\lim_{\Lambda\to\infty}\theta_l\left(\bar{\theta}\right)/\bar{\theta} = 1$, there exists $\Lambda_1 > 0$ such that if $\Lambda \leq \Lambda_1$, consumer surplus is maximized when $\hat{T}_f = \bar{\theta}$ and otherwise the social planner sets $\hat{T}_f < \bar{\theta}$. Finally we prove that for a uniform distribution on $[0, \bar{\theta}]$ and for $\beta = 0$ there exist Λ_2, Λ_3 and Λ_4 with $0 < \Lambda_2 < \Lambda_3 < \Lambda_4$ such that if $\Lambda \leq \Lambda_2$, consumer surplus is maximized when $\hat{T}_f = \bar{\theta}$ and if $\Lambda \in (\Lambda_3, \Lambda_4]$, then$\hat{T}_f < \bar{\theta}$. When $\beta = 0$ there exists an equilibrium with $\theta_h\left(T\right) = \bar{\theta}$ and θ_l defined by $\theta_l = cW_T\left(\theta_l, \bar{\theta}\right)$. Then

$CS(T) = \Lambda\left(\frac{1}{2}\left(T^2 - (\theta_l)^2\right) + T\left(\bar{\theta} - T\right) - \theta_l\left(\bar{\theta} - \theta_l\right)\right)/\bar{\theta}$ and $dCS/dT = \Lambda\left(\left(\bar{\theta} - T\right) - d\theta_l/dT\left(\bar{\theta} - \theta_l\right)\right)/\bar{\theta}$, where

$$\frac{d\theta_l}{dT} = \frac{c\frac{\partial W}{\partial T}}{1 - c\frac{\partial W}{\partial \theta_l}} = \frac{c\frac{\Lambda\left(2T\left(\bar{\theta} - \Lambda\left(\frac{1}{2}\left(T^2 - \theta_l^2\right) + T\left(\bar{\theta} - T\right)\right)\right) + \Lambda\left(\frac{1}{3}\left(T^3 - \theta_l^3\right) + T^2\left(\bar{\theta} - T\right)\right)\right)\left(\bar{\theta} - T\right)}{2\left(\bar{\theta} - \Lambda\left(\frac{1}{2}\left(T^2 - \theta_l^2\right) + T\left(\bar{\theta} - T\right)\right)\right)^2}}{1 + c\frac{\Lambda\left(\theta_l^2\left(\bar{\theta} - \Lambda\left(\frac{1}{2}\left(T^2 - \theta_l^2\right) + T\left(\bar{\theta} - T\right)\right)\right) + \Lambda\theta_l\left(\frac{1}{3}\left(T^3 - \theta_l^3\right) + T^2\left(\bar{\theta} - T\right)\right)\right)}{2\left(\bar{\theta} - \Lambda\left(\frac{1}{2}\left(T^2 - \theta_l^2\right) + T\left(\bar{\theta} - T\right)\right)\right)^2}}$$

Since $\lim_{\Lambda \to 0} d\theta_l/dT\left(\bar{\theta} - \theta_l\right) = 0$, there exists $0 < \Lambda_2$ such that if $\Lambda \leq \Lambda_2$, consumer surplus is maximized when $\hat{T}_f = \bar{\theta}$. Moreover $d^2CS/dT^2 = \Lambda\left(-1 - d^2\theta_l dT^2\left(\bar{\theta} - \theta_l\right) + \left(d\theta_l/dT\right)^2\right)/\bar{\theta}$, $\lim_{T \to \bar{\theta}} d\theta_l/dT = 0$ and

$$\lim_{T \to \bar{\theta}} \frac{d^2\theta_l}{dT^2} = -\frac{c\Lambda\frac{\left(2\bar{\theta}\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right) + \frac{1}{3}\Lambda\left(\bar{\theta}^3 - \theta_l^3\right)\right)}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right)^2}}{1 + c\Lambda\frac{\left(\theta_l\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right) + \frac{1}{3}\Lambda\left(\bar{\theta}^3 - \theta_l^3\right)\right)\theta_l}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right)^2}},$$

where $\theta_l = \theta_l\left(\bar{\theta}\right)$. Then

$$\lim_{T \to \bar{\theta}} \frac{d^2CS}{dT^2} = \Lambda\frac{1}{\bar{\theta}}\left(-1 + \frac{c\Lambda\frac{\left(2\bar{\theta}\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right) + \frac{1}{3}\Lambda\left(\bar{\theta}^3 - \theta_l^3\right)\right)}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right)^2}}{1 + c\Lambda\frac{\left(\theta_l\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right) + \frac{1}{3}\Lambda\left(\bar{\theta}^3 - \theta_l^3\right)\right)\theta_l}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right)^2}}\left(\bar{\theta} - \theta_l\right)\right).$$

Now

$$\lim_{T \to \bar{\theta}} \frac{d^2CS}{dT^2} > 0 \Leftrightarrow \frac{\Lambda\left(\left(2\bar{\theta}\left(\bar{\theta} - \theta_l\right) - \theta_l^2\right)\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right) + \frac{1}{3}\Lambda\left(\bar{\theta}^3 - \theta_l^3\right)\left(\bar{\theta} - 2\theta_l\right)\right)}{2\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right)^2} > \frac{1}{c}.$$

Since in equilibrium, $1/c = W_{\bar{\theta}}\left(\theta_l\left(\bar{\theta}\right), \bar{\theta}\right)/\theta_l\left(\bar{\theta}\right)$, it follows that

$$\lim_{T \to \bar{\theta}} \frac{d^2CS}{dT^2} > 0 \Leftrightarrow \frac{\frac{1}{3}\Lambda\left(\bar{\theta}^3 - \theta_l^3\right)}{\left(\bar{\theta} - \Lambda\frac{1}{2}\left(\bar{\theta}^2 - \theta_l^2\right)\right)} > \frac{\left(\frac{1}{3}\left(\bar{\theta}^3 - \theta_l^3\right) - \left(2\bar{\theta}\left(\bar{\theta} - \theta_l\right) - \theta_l^2\right)\theta_l\right)}{\theta_l\left(\bar{\theta} - 2\theta_l\right)}.$$

Rearranging we get that for $\lim_{T \to \bar{\theta}} d^2CS/dT^2 > 0$ holds iff $-\frac{2}{3}(c+6)\left(\theta_l/\bar{\theta}\right)^3 + 2(1-c)\left(\theta_l/\bar{\theta}\right)^2 + 2c\left(\theta_l/\bar{\theta}\right) - c/3 > 0$. This cubic polynomial is negative at $\theta_l = 0$ and at $\theta_l/\bar{\theta} = 1$. It has a unique local maximum between 0 and 1 and it can be easily shown that at this maximum it is indeed positive. Therefore there exist two roots of this polynomial: $0 < z_1 < z_2 < 1$ such that it is positive for every $z_1 < \theta_l/\bar{\theta} < z_2$. Finally since $\theta_l/\bar{\theta}$ increases with Λ and $\lim_{\Lambda \to 0} \theta_l\left(\bar{\theta}\right)/\bar{\theta} = 0$ and $\lim_{\Lambda \to \infty} \theta_l\left(\bar{\theta}\right)/\bar{\theta} = 1$, there exist Λ_3 and Λ_4 with $0 < \Lambda_3 < \Lambda_4$ such that if $\Lambda \in (\Lambda_3, \Lambda_4]$, then $\lim_{T \to \bar{\theta}} d^2CS/dT^2 > 0$ and therefore $\hat{T}_f < \bar{\theta}$ (since $\lim_{T \to \bar{\theta}} d\theta_l/dT = 0$ and therefore $d\theta_l/dT$ must be negative for some $T < \bar{\theta}$ and since it is positive for very small T it must be equal to zero at some $\hat{T}_f < \bar{\theta}$).

□

Proof of Theorem 7. When $\theta \sim U\left[0, \bar{\theta}\right]$ and $\beta = 1$, $\theta_h = T$ and $\pi_r(p, T) = \Lambda p\left(T^2 - \theta_l^2\right)/2\bar{\theta}$ where $\theta_l(p, T)$ is defined by $\theta_l(1-p) = cW_T(\theta_l, T)$. Then $\partial\pi_r/\partial T = \Lambda p\left(T - \theta_l\partial\theta_l/\partial T\right)/\bar{\theta}$ and $\partial\theta_l/\partial T = c\partial W/\partial T/\left(1 - p - c\partial W/\partial\theta_l\right)$. We wish to show that $T - \theta_l\partial\theta_l/\partial T \geq 0 \Leftrightarrow \partial\theta_l/\partial T \leq T/\theta_l \ \forall T$ and therefore the optimal T is $\bar{\theta}$. Now $\partial\theta_l/\partial T \leq T/\theta_l \Leftrightarrow \theta_l\partial W/\partial T + T\partial W/\partial\theta_l \leq T(1-p)/c$. In equilibrium, $(1-p)/c = W_T(\theta_l, T)/\theta_l$. Therefore we have to show that $\theta_l\partial W/\partial T + T\partial W/\partial\theta_l \leq TW/\theta_l$. In the proof of Theorem 5 we showed that $\theta_l\partial W/\partial T + T\partial W/\partial\theta_l < W$ and since $\theta_l \leq T$ the inequality holds. When $\theta \sim U\left[0, \bar{\theta}\right]$ and $\beta = 0$ we have: If $p \leq 1 - cW_T\left(T, \bar{\theta}\right)/T$ then $\theta_h = \bar{\theta}$ and the revenue is $\pi_r(p, T) = \Lambda p\left(\frac{1}{2}\left(T^2 - \theta_l^2\right) + T\left(\bar{\theta} - T\right)\right)/\bar{\theta}$, where θ_l is defined by $\theta_l(1-p) = cW_T\left(\theta_l, \bar{\theta}\right)$. Then $\partial\pi_r/\partial T = \Lambda p\left(\bar{\theta} - T - \theta_l\partial\theta_l/\partial T\right)/\bar{\theta}$. In equilibrium

$1-p=cW/\theta_l$. Therefore $d\theta_l/dT=\partial W/\partial T/\left(W/\theta_l-\partial W/\partial\theta_l\right)$. We have,

$$\begin{aligned}\frac{\partial\pi_r}{\partial T}&=\frac{1}{\bar{\theta}}\Lambda p\left(\bar{\theta}-T-\theta_l\frac{d\theta_l}{dT}\right)\\&=\frac{\Lambda p\left(\bar{\theta}-\Lambda\left(\frac{1}{2}\left(T^2-\theta_l^2\right)+T\left(\bar{\theta}-T\right)\right)\right)\left(\frac{1}{3}\left(T^3-\theta_l^3\right)+T^2\left(\bar{\theta}-T\right)+\theta_l^3-2T\theta_l^2\right)}{\bar{\theta}\left(\frac{1}{3}\left(T^3-\theta_l^3\right)+T^2\left(\bar{\theta}-T\right)\right)\left(\bar{\theta}-\Lambda\left(\frac{1}{2}\left(T^2-\theta_l^2\right)+T\left(\bar{\theta}-T\right)\right)+\Lambda\theta_l^2\right)+\theta_l^3\left(\bar{\theta}-\Lambda\left(\frac{1}{2}\left(T^2-\theta_l^2\right)+T\left(\bar{\theta}-T\right)\right)\right)}.\end{aligned}$$

Therefore $\partial\pi_r/\partial T\geq 0$ which implies $\left(T^3-\theta_l^3\right)/3+T^2\left(\bar{\theta}-T\right)+\theta_l^3-2T\theta_l^2\geq 0$. The polynomial $2\theta_l^3/3-2T\theta_l^2+T^2\left(\bar{\theta}-2T/3\right)$ decreases in θ_l and equal $T^2\left(\bar{\theta}-2T\right)$ at $\theta_l=T$. When $T\leq\bar{\theta}/2$ then $\partial\pi_r/\partial T\geq 0$. When $T\geq\bar{\theta}/2$ there exists a unique $\tilde{\theta}_l$ that solves $2\theta_l^3/3-2T\theta_l^2+T^2\left(\bar{\theta}-2T/3\right)=0$ and $\partial\pi_r/\partial T\geq 0$ if $\theta_l\leq\tilde{\theta}_l$. Now $d\theta_l/d\Lambda>0$ and $\lim_{\Lambda\to 0}\theta_l=0$. Therefore there exists Λ_1 such that for all $\Lambda\leq\Lambda_1$ it holds that $\theta_l\leq\tilde{\theta}_l$ for $T\geq\bar{\theta}/2$. Then $\partial\pi_r/\partial T\geq 0\ \forall T$ and the optimal T is $T=\bar{\theta}$. When $\lim_{T\to\bar{\theta}}\partial\pi_r/\partial T<0$ there exists an optimal $T_r<\bar{\theta}$ (for every p it holds that $p\leq 1-cW_T\left(T,\bar{\theta}\right)/T$ at $T=\bar{\theta}$). Now $\lim_{T\to\bar{\theta}}\partial\pi_r/\partial T<0$ and $\lim_{T\to\bar{\theta}}\partial\pi_r/\partial T<0\Leftrightarrow 2\bar{\theta}_l^3/3-2\bar{\theta}\theta_l^2+\bar{\theta}^3/3<0$. There exists a unique $\tilde{\theta}_l\left(\bar{\theta}\right)$ such that the equation holds and therefore when $\lim_{T\to\bar{\theta}}\theta_l>\tilde{\theta}_l\left(\bar{\theta}\right)$ we have $T_r<\bar{\theta}$. Finally, since $\theta_l\left(\bar{\theta}\right)$ increases with Λ there exists a Λ_2 such that when $\Lambda>\Lambda_2$ we have $T_r<\bar{\theta}$. We denote $\Lambda_r=\min\left\{\Lambda_1,\Lambda_2\right\}$.

□

Theorem 8. *Assume θ is uniformly distributed on $\left[0,\bar{\theta}\right]$. The social planner does not charge a price rate (i.e., $p_r=0$) if $\beta=0$ or $\beta=1$.*

Proof. $\beta=1$: $\theta_h=T$ and $\theta_l\left(1-p\right)=cW_T\left(\theta_l,T\right)$. Therefore $CS\left(p,T\right)=\Lambda\left(1-p\right)\left(T-\theta_l\right)^2/2\bar{\theta}$. The f.o.c $\partial CS/\partial p=\Lambda\left(-\left(T-\theta_l\right)^2-2\left(1-p\right)\left(T-\theta_l\right)d\theta_l/dp\right)/2\bar{\theta}<0$ and therefore $p_r=0$. $\beta=0$: for $p\leq 1-cW_T\left(T,\bar{\theta}\right)/T$, $\theta_h=\bar{\theta}$ and $\theta_l\left(1-p\right)=cW_T\left(\theta_l,\bar{\theta}\right)$. Therefore $CS\left(p,T\right)=\Lambda\frac{1}{2\theta}\left(1-p\right)\left(2\bar{\theta}-\left(T+\theta_l\right)\right)\left(T-\theta_l\right)$. The f.o.c $\partial CS/\partial p=\Lambda\left(-\left(2\bar{\theta}-\left(T+\theta_l\right)\right)\left(T-\theta_l\right)-2\left(1-p\right)d\theta_l/dp\left(\bar{\theta}-\theta_l\right)\right)/2\bar{\theta}<0$ and therefore $p_r=0$. Finally, when $p>1-cW_T\left(T,\bar{\theta}\right)/T$, $\theta_l=T$ and $T\left(1-p\right)=cW_T\left(T,\theta_h\right)$. Therefore $CS\left(p,T\right)=0$ and the optimal $p_r=0$.

□