# Proactive customer service: operational benefits and economic frictions

Kraig Delana · Nicos Savva · Tolga Tezcan
London Business School, Regent's Park, London NW1 4SA, UK
kdelana@london.edu · nsavva@london.edu · ttezcan@london.edu

**Problem Definition:** We study a service setting where the provider may have advance information about customers' future service needs and may initiate service for such customers proactively if they are flexible with respect to the timing of service delivery.

**Academic / Practical Relevance:** Information about future customer service needs is becoming increasingly available due to better system integration coupled with advanced analytics and big-data methods. We contribute to the literature by presenting a systematic analysis of proactive service as a tool that can be used to better match service supply with demand.

**Methodology:** To study this setting, we combine (i) queueing theory, and in particular a diffusion approximation developed specifically for this problem, to quantify the impact of proactive service on customer delays with (ii) game theory to investigate economic frictions in a system with proactive service.

**Results:** We find that proactive service reduces average delays, which we quantify with a closed-form approximation. More specifically, we show that this reduction is increasing concave in the proportion of customers who can be served proactively. Nevertheless, customers might not benefit from proactive service due to economic frictions; in equilibrium more customers will join the system and fewer will be willing to be flexible compared to social optimum. This is due to a positive externality leading to free-riding behavior – customers who agree to be served proactively reduce waiting time for everyone else including those customers who do not agree to be served proactively.

**Managerial Implications:** Our results suggest that proactive service may have a large operational benefit, but caution that it may fail to fulfil its potential due to customer self-interested behavior.

*Key words*: Proactive Service, Queueing Theory, Game Theory, Healthcare Operations
*History*: Revised: June 9, 2018

## 1. Introduction

In many service settings (e.g., healthcare), demand is highly variable but capacity is relatively fixed over short periods of time, leading to delays for customers. To reduce such delays, service providers often implement mechanisms that aim to modulate customer demand. These include providing delay information to discourage customers from joining the system when congestion is high (Armony et al. 2009, Jouini et al. 2011, Ibrahim et al. 2016, Cui et al. 2014), offering customers the option to wait off-line or receive a call back (Kostami and Ward 2009, Armony and Maglaras 2004a,b), or offering customers priority if they arrive during pre-allotted times (De Lange

et al. 2013). In this paper, we investigate an alternative demand-modulation mechanism, proactive customer service, where the provider exploits information about customers' future service needs to proactively initiate service when there is idle server capacity.

Proactive service may find application in any a number of service systems. For example, this work was motivated by a healthcare setting, induction of labor – a procedure where the process of delivering a baby is started artificially (induced) through the use of pharmaceuticals in a hospital ward. In this setting, the schedule of pre-planned and elective patients provided a list of customers with future service needs that the hospital could serve proactively (see Appendix C for more details and a numerical illustration of this setting). Even in cases without a schedule, information on future customers may become available from other sources such as predictive models. For example, Jerath et al. (2015) develop a method to predict which customers are likely to contact a health insurance call center based on claims data. The authors go on to suggest that a potential application would be to make "advance outgoing calls to customers who have a high probability of calling," in other words, proactive customer service.

The first contribution of this paper is to formulate a model that captures the benefits associated with proactive service and to develop novel approximations that quantify the improvement in system performance. The model is based on a Markovian queueing system with two queues in tandem. Arrivals to the first queue, which we call the "orbit," represent virtual arrivals, or to be more exact, arrival of information about customers' future service needs that the provider may choose to fulfill proactively. Because these customers are willing to receive service at a moment chosen by the provider, we find it convenient to label them as "flexible." We assume the provider has minimal information about such flexible customers, having knowledge only of who requires service in the future but not when they would arrive. If a customer in orbit has not been served proactively, after a random amount of time they will transition to the second queue, which we label as the "service" queue. The service queue also experiences direct arrivals by customers who are not flexible, or equivalently, for whom the service provider does not have advance information about their service needs. Customers at the service queue are served in a first-come, first-served fashion (irrespective of whether they arrived directly or through orbit) by a single server. Naturally, proactive service only takes place if the server is idle (i.e., the service queue is empty) and there are customers in orbit.

Using this queueing model, we show that proactive service reduces service queue congestion in the first-order stochastic sense – proactive service exploits periods of idle capacity to bring forward arrivals who would have otherwise occurred at some point in the future. By doing so, proactive service smooths demand and, as a result, reduces delays for all customers, including those who are not flexible. Using a path-wise coupling argument, we show useful monotonicity results – the greater

the proportion of flexible customers and the earlier the provider knows about their service needs the lower the average congestion and delay in the service queue. Finally, we develop a diffusion approximation that allows us to derive closed-form expressions for the average steady-state waiting times. The approximation suggests that the reduction in delay associated with proactive service displays decreasing marginal returns in the proportion of flexible customers.

While the operational benefits associated with proactive service may be substantial, realizing them depends critically on customer behavior. On one hand, customers may refuse to be flexible, especially if there is an inconvenience cost associated with flexibility. On the other hand, the presence of proactive service, which reduces waiting times, may encourage customers to over-join the system compared to profit maximizing (or social optimal), thus eroding any of the associated benefits. Furthermore, the benefit of proactive service for each individual customer will depend on what other customers do, i.e., it is an equilibrium outcome. Therefore, to understand whether proactive service will indeed be beneficial requires a game-theoretic analysis.

The second contribution of this paper is to develop such a game-theoretic model to analyze customer behavior. To do so, we augment the standard "to queue or not to queue" dilemma (Hassin and Haviv 2003) with the additional option to join the queueing system but be flexible. The game theoretic analysis identifies two economic frictions. First, customers will under-adopt proactive service compared to the profit maximizing (or social) optimum. This result is driven by a positive externality which gives rise to free-riding behavior: a customer who agrees to be flexible will reduce the expected waiting time of everyone else but this is a benefit that she does not take into account when making her own decision. In fact, we find instances where this economic friction can be extreme in the sense that a profit-maximizing provider (or a central planner) would have wanted all of the customers to be flexible, but in equilibrium, no customer chooses to be so. Second, we find that given the option to be served proactively, customers will over-join the system compared to both the profit maximizing (or social optimal) joining rate, as well as compared to a system without proactive service. This is due to the well-known negative congestion-based externalities (e.g., Naor 1969) that proactive service exacerbates, i.e., for a given level of arrivals, proactive service reduces waiting times and, as a result, more customers would want to join compared to the case without proactive service. Interestingly, we find some surprising interactions between the positive and negative externalities. For example, an increase in the cost per unit of waiting time may lead to more customers joining the system. This is because the higher cost of waiting in the queue induces more customers to be flexible, which reduces waiting times, which in turn induces more customers to join the system.

We conclude the paper by presenting two extensions of the queueing model described above. The first covers the multiserver case and shows that the basic intuition and approximations developed for

the single-server case continue to hold in the multiserver case with minor modifications. The second extension examines the case where the information about future customer needs is imperfect. In this case, some of the customers served proactively did not require service. These "errors" could occur if customers may have their service need resolved through alternative channels (e.g., spontaneous labor), or because of errors in information systems or predictive models in determining customers with future service needs. We show that the diffusion-limit approximation we developed for the case without errors can be adapted to derive closed-form approximations of system performance in the presence of errors. Using this approximation, we derive conditions under which proactive service reduces waiting times despite errors, and show that for some model parameters, the system can handle more errors as system utilization increases. This seems counter-intuitive at first because in a more heavily utilized system one would expect errors to increase delays more than in a less utilized system. However, this can be explained by the fact that reduction in delays gained through proactive service grows as utilization increases.

Sketches of all proofs are presented in the Appendix of this paper. In the electronic companion (EC) we present detailed proofs, additional numerical/simulation analysis, and a numerical/simulation study with parameters calibrated to the induction of labor setting which motivated this work. This study suggests that proactive service can reduce average delays by up to 28% in this setting. Nevertheless, our analysis also suggests that the hospital should proceed carefully before implementing proactive service as economic frictions may lead to suboptimal voluntary adoption of the service.

## 2. Literature Review

The analysis of proactive service in this paper contributes to three streams of queueing literature which are connected by the objective of better matching service supply and demand. The first stream examines interventions that modulate service supply in response to an exogenous demand process. The second stream focuses on interventions that seek to actively manage endogenous demand by taking into account the economic incentives of strategic customers. The third stream builds on the first two by incorporating future demand information.

The first stream of literature considers supply-side interventions, e.g,. optimizing the number of servers, in response to exogenous changes in demand. The bulk of this literature is developed for call centers (see Gans et al. 2003, Aksin et al. 2007 for overviews) and has focused on topics ranging from long-term workforce-management planning (Gans and Zhou 2002), to medium-term shift staffing (Whitt 2006), down to short-term call routing policies (Gans and Zhou 2007), as well as combinations of short and medium-term solutions (Gurvich et al. 2010). Our work fits with the short-term strategies but, unlike the above-mentioned work, we assume that both system capacity

and the routing policy are fixed. One supply-side strategy that is closely related to proactive customer service is for idle servers to work on auxiliary tasks, such as emails in call centers (see, e.g., Gans and Zhou 2003 and Legros et al. 2015). In the case of proactive service, future customers can be thought of as the auxiliary tasks, however, this substantially changes the dynamics of the system by smoothing the demand process.

The second stream considers demand-side interventions that aim to influence strategic customers' (endogenous) decisions. See Hassin and Haviv (2003) and Hassin (2016) for a comprehensive review of the economics of queues and strategic customer decision-making. One important intervention is the use of pricing to control the overall level of demand. What makes pricing particularly important in service systems is a key observation, first made by Naor (1969), that utility-maximizing customers tend to over-utilize queueing systems compared to the socially optimal level. This is due to customers imposing a negative externality on each other in the form of delays, and as a consequence, the service provider can increase welfare by charging customers a toll for joining the system. This finding persists in multiple variants, e.g., when the queue is unobservable (Edelson and Hilderbrand 1975), and when customers are heterogenous or have multiple classes (Littlechild 1974, Mendelson and Whang 1990). Naturally, the negative externality and over-joining persists in the presence of proactive service. However, in this setting we also find a rare instance of a positive externality, where customers who agree to be flexible reduce the waiting time of everyone else.[1]

Beyond pricing, two other common demand-side interventions are delay announcements and multiple service priorities. Delay announcements encourage balking (Allon and Bassamboo 2011, Armony et al. 2009, Ibrahim et al. 2016, Jouini et al. 2011) or retrials (Cui et al. 2014), especially when the system is congested. Multiple service priorities encourage some customers to wait in low-priority queues (usually offline), thus reducing the waiting time of high priority customers (Engel and Hassin 2017, Armony and Maglaras 2004a,b, Kostami and Ward 2009). Our work is closer to the latter as one may think of customers who may be served proactively as arriving to a "low priority" queue. However, in contrast to the extant work, customers in this "low priority" queue may transition to the service system at any time, thus complicating the system dynamics.

The third stream of literature to which our work is related focuses on the setting where the provider has information about the future. The benefits of future (or advance) demand information on production and inventory systems (often modelled using queues) has been recognized by many

[1] We note that positive externalities are relatively rare in the literature of queueing games (Hassin 2016, §1.8). Three notable exceptions are: i) Engel and Hassin (2017), where customers that choose to join a low-priority queue reduce delays for customers that join the high-priority queue; ii) Nageswaran and Scheller-Wolf (2016), where allowing one class of customers to wait in multiple queues may, under some conditions, reduce waiting time for customers who are only able to wait in a single queue; iii) Hassin and Roet-Green (2011), where customers that pay to inspect the queue before making the decision to join or balk reduce waiting time for customers who do not inspect.

(e.g., Gallego and Özer 2001, Özer and Wei 2004, Papier and Thonemann 2010). More relevant is the work that considers customers who may accept product delivery early, i.e., are flexible to the timing of product delivery (Karaesmen et al. 2004, Wang and Toktay 2008). The main difference between this stream of work and ours is that production and inventory systems largely focus on different performance measures (e.g., cost of production, inventory cost, or stock-out costs as opposed to waiting times), and largely treat demand as exogenous. The study of future information in the context of service as opposed to inventory systems is more limited and has focused mainly on demand-side interventions in the form of admission control (e.g, Spencer et al. 2014, Xu 2015, Xu and Chan 2016). As far as we are aware, the only other work that studies proactive service is Zhang (2014). This work was motivated by computing applications (e.g., cache pre-loading or command pre-fetching) and differs from ours in a number of dimensions. We present a more detailed comparison once we introduce our model in §3.3.

## 3. Operational Analysis: Single-server Queueing Model

In this section, we present and analyze the proactive service system assuming there is a single server. The analysis has two goals: (i) to show that proactive service improves system performance, and (ii) to provide closed-form approximations that quantify the impact of proactive service on time-average measures of system performance.

### 3.1. Queueing Model

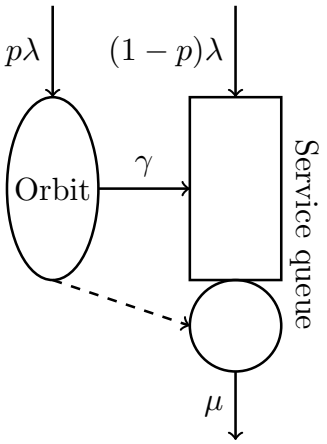
We assume that demand arrives to the system following a Poisson process with rate $\lambda$, and that there exists two types of customers who require service, "flexible" and "inflexible." The service times for both types of customers are assumed to be i.i.d. and exponential with parameter $\mu$; note we assume $\lambda < \mu$ throughout for stability. Inflexible customers make up a proportion $(1-p)$ of total demand and arrive to the service queue according to a Poisson process with rate $(1-p)\lambda$. Upon arrival they immediately begin service if the server is free or join the queue, which operates in a first-in, first-out manner. For flexible customers, we assume the service provider becomes aware of the customer's service need some time before they actually arrive to the service queue and the provider has the option of serving them proactively at any time after becoming aware of their service need. To capture this, we assume that flexible customers do not arrive directly to the service queue, but instead arrive to a virtual queue, which we refer to as "orbit." We assume arrivals to orbit follow a Poisson process with rate $p\lambda$. While in orbit customers may be served proactively if the server becomes idle, or, after a random period of time, which we assume to be i.i.d. and exponential with parameter $\gamma > 0$, they depart for the service queue on their own. Once at the service queue, flexible customers are served as any other customer who has arrived to the service queue directly. Together, these assumptions imply that the system may be modeled as two

Markovian queues in tandem linked by the proactive service mechanism, as depicted in Figure 1. We note that some of our results hold for more general time-in-orbit and service-time distributions. We indicate if this is the case when we state these results throughout the paper.

We will refer to the parameter $p$ as the proportion of flexible customers or, interchangeably, as the proportion of customers who have adopted proactive service. The average time flexible customers spend in orbit before transiting to the service queue on their own (i.e., $1/\gamma$) can be interpreted as the information lead time for flexible customers – this is the average time in advance the provider knows of a customer's service need before the customer arrives to the service queue.

We denote the occupancy of the orbit and the service queues at time $t > 0$ with $N_r(t)$ and $N_s(t)$, respectively. Similarly, we denote the steady-state average occupancy and steady-state distribution of the queue length processes (where they exist) with $\bar{N}_r$, $\bar{N}_s$, and $\pi = (\pi_r, \pi_s)$, respectively. Finally, we define the steady-state average time for each customer type $a \in \{r, s\}$, spent in each queue $b \in \{r, s\}$, with $\bar{T}_{ab}$, if this exists. For example, $\bar{T}_{rs}$ denotes the average time flexible customers spend in the service queue. We use the convention that a customer is assumed to be in the service queue while in service.

**Figure 1 Queueing model**




### 3.2. Impact of Proactive Service

In order to assess the impact of proactive service, we compare the system with proactive service to a benchmark system without this capability, all other things being equal. In the benchmark case, the whole system can be modeled as a Jackson network where orbit is an $M/M/\infty$ queue, the service queue is an $M/M/1$ queue, and all customers in orbit transition to the service queue. The steady-state distribution of queue lengths and waiting times for this system can easily be found in closed form (Kleinrock 1976, see §3.2 & §4.4). The steady-state distribution of total number of customers in the service queue follows the geometric distribution with parameter $1 - \rho$ where

$\rho := \lambda/\mu < 1$, and the steady-state distribution of the number of customers in orbit is Poisson with parameter $p\lambda/\gamma$. To denote the time average performance measures associated with the benchmark system, we append superscript $B$ to all the terms defined above; for example, $\bar{N}_s^B$ denotes the expected number of customers in the service queue in steady state for the benchmark case.

**Impact of proactive service on queue lengths.** We begin with the following result.

LEMMA 1. *In steady state, the total number of customers in the system with proactive service is equal in distribution to the number of customers in the service queue without proactive service, that is,* $\pi_r + \pi_s \stackrel{d}{=} \pi_s^B$.

Lemma 1 shows that the steady-state distribution of the total number of customers in the system with proactive service (that is the sum of customers in orbit and in the service queue) is equivalent to the steady-state distribution of number of customers in the service queue when proactive service is not possible. Interestingly, this implies that the distribution of the total number of customers in the system does not depend on the proportion of customers that is flexible (i.e., $p$) or the average information lead time (i.e., $1/\gamma$). This result immediately implies that the average total occupancy in the system with proactive service equals the average occupancy of the service queue in the benchmark case (i.e., $\bar{N}_r + \bar{N}_s = \bar{N}_s^B = \rho/(1-\rho)$). Furthermore, the non-negativity of the number of customers in orbit suggests that there is a stochastic ordering in the number of customers in the service queue, a result we present in Proposition 1. Throughout, we use $\preceq$ to denote stochastic ordering.

PROPOSITION 1.

*(i) The steady-state distribution of the occupancy of orbit in the system with proactive service is stochastically dominated by that of the system without proactive service:* $\pi_r \preceq \pi_r^B$.

*(ii) The steady-state distribution of the occupancy of the service queue in the system with proactive service is stochastically dominated by that of the system without proactive service:* $\pi_s \preceq \pi_s^B$.

The first part of the proposition establishes that the orbit is less occupied (in a stochastic sense) in the system with proactive service. This is not surprising. Since some customers are pulled from orbit to be served proactively, the time they spend in orbit is reduced and thus orbit becomes less congested compared to the system where proactive service is not possible. The second part of the proposition shows that the service queue is also less congested (in a stochastic sense) in the system with proactive service. Obviously, each part further implies that the time-average occupancy in both orbit and the service queue is reduced, that is, $\bar{N}_r \leq \bar{N}_r^B$ and $\bar{N}_s \leq \bar{N}_s^B$. We note here that Lemma 1 and Proposition 1 can be extended to the cases when time in orbit and/or service times are generally distributed.

**Impact of proactive service on wait times.** Because customers and service providers generally consider delay times and not system occupancy as the key metric of system performance, we now focus on the impact of proactive service on the expected time spent by each customer type in different parts of the system in steady state. The main result is provided in Proposition 2 and relies on Proposition 1 and the mean value approach (MVA) (Adan and Resing 2002, §7.6).

PROPOSITION 2. *Proactive service reduces delays for all customers in expectation:*

$$\textit{(i)}\ \bar{T}_{rr} \le \bar{T}^B_{rr}, \quad \textit{(ii)}\ \bar{T}_{ss} \le \bar{T}^B_{ss}, \quad \textit{(iii)}\ \bar{T}_{rs} \le \bar{T}^B_{rs},$$

*but more so for those customers who can be served proactively:*

$$\textit{(iv)}\ \bar{T}^B_{rs} - \bar{T}_{rs} \ge \bar{T}^B_{ss} - \bar{T}_{ss}.$$

*The difference in expected time spent by flexible vs. inflexible customers in the service queue is proportional to the expected time spent in orbit:*

$$\bar{T}_{ss} - \bar{T}_{rs} = \frac{\mu - \lambda}{\mu} \bar{T}_{rr} \ge 0. \tag{1}$$

Proposition 2 shows that proactive service benefits both flexible and inflexible customers. The fact that proactive service benefits flexible customers is not surprising – since some of them will be served proactively and receive service without having to wait in the service queue at all, proactive service will reduce the average waiting time for this class of customers. What is perhaps a little more surprising is that proactive service reduces waiting times for inflexible customers as well. This occurs because proactive service smooths demand by utilizing idle time to serve some customers early, thus, it reduces the likelihood that customers will arrive to a congested service queue. This reduction in congestion benefits all customers. However, Proposition 2 further implies that the benefit of proactive service is greater for flexible than inflexible customers.

**Impact of flexibility and information lead time.** So far we have shown proactive service decreases occupancy in both orbit and the service queue as well as average delays for all customers when compared to a benchmark system without proactive service. Next, we establish a partial answer to the question of how the performance of a system with proactive service changes as the proportion of flexible customers and the information lead time change in Proposition 3.

PROPOSITION 3.

*(i) The steady-state distribution of number of customers in orbit (i.e., $\pi_r$) is, in a stochastic ordering sense, increasing in $p$ and decreasing in $\gamma$.*

*(ii) The steady-state distribution of number of customers in the service queue (i.e., $\pi_s$) is, in a stochastic ordering sense, decreasing in $p$ and increasing in $\gamma$.*

**Table 1 Monotonic behavior of performance measures**

| | $\bar{N}_r$ | $\bar{N}_s$ | $\bar{T}_{rr}$ | $\bar{T}_{rs}$ | $\bar{T}_{ss}$ |
|---|---|---|---|---|---|
| $\gamma$ | $\downarrow$ | $\uparrow$ | $\downarrow$ | $\uparrow$ | $\uparrow$ |
| $p$ | $\uparrow$ | $\downarrow$ | ? | ? | $\downarrow$ |

The arrow $\uparrow$ ($\downarrow$) denotes that a given performance measure is increasing (decreasing) in $p$ or $\gamma$.

*(iii) The performance measures exhibit the monotonic behaviors summarized in Table 1.*

Proposition 3 relies on a pathwise coupling argument to show part $(i)$, specifically that there are more customers in orbit (in a stochastic sense) in steady state if a larger proportion of customers are flexible and fewer are in orbit if there is shorter information lead time. Combining this result with Lemma 1 immediately implies the opposite impact on the service queue, which is given in part $(ii)$. Together these results imply the monotonicity of performance measures presented in part $(iii)$: that more information lead time (i.e., smaller $\gamma$) reduces time in the service queue for both flexible and inflexible customers, and that a greater proportion of flexible customers (i.e., larger $p$) leads to greater occupancy of orbit and lower occupancy of the service queue. We note that it is not possible to use the MVA approach to derive monotonicity results for the waiting times of flexible customers with respect to the proportion of flexible customers ($p$). Therefore, we defer this to the next section where we develop diffusion limit approximations.[2]

### 3.3. Approximations Based on Diffusion Limits

In this section we present approximations based on diffusion limits for the performance measures we discussed in the previous section. To provide some intuition, in the diffusion limit, the primitive stochastic processes (e.g., arrivals and service completions) are replaced with appropriate limiting versions that make the occupancy processes more amenable to analysis. This enables the study of the macro-level behavior of the system over long periods of time and provides useful insights that are helpful in developing closed-form approximations of steady-state behavior (Chen and Yao 2013).

To proceed we need to define some additional notation. We focus on the system with proactive service (see Figure 1) and we define a sequence $\lambda^n = \mu - \frac{c}{\sqrt{n}}$ for some $c \geq 0$ and a sequence of systems indexed by $n$ with these arrival rates. We still assume that arrivals are flexible with probability $p$ and the departure rate from the service queue is $\mu$, but we let the departure rate of each customer from orbit to the service queue be $\gamma^n = \frac{\gamma}{\sqrt{n}}$. We further denote the number of customers in orbit and the service queue at time $t$ as $N_r^n(t)$ and $N_s^n(t)$, respectively.

[2] We note that the simpler case, where customers never transition from orbit to the service queue on their own (i.e., $\gamma = 0$), has been recently studied in Engel and Hassin (2017). In this case, the orbit becomes a low priority queue and the steady-state performance of the system can be obtained in closed form using exact analysis.

**Asymptotic analysis.** Observe that as $n$ increases, the total arrival rate ($\lambda^n$) approaches the service rate ($\mu$), which in turn implies that utilization goes to one. The part that is exploited by a diffusion limit is that utilization, and hence occupancy, grows at a specific rate. Knowing that the average number of customers in the $n^{th}$ system is ${\lambda^n}/{(\mu-\lambda^n)}$ is $\mathcal{O}(\sqrt{n})$ means that dividing through by $\sqrt{n}$ prevents the limit of the total occupancy process from going to infinity (and hence the limits of both the orbit and service queue occupancy processes as well). We further scale time by replacing $t$ by $nt$; this can be interpreted as the occupancy processes being observed over longer lengths of time as utilization approaches one to capture the macro-level behavior of the system. This leads to scaled occupancy processes $\hat{N}_r(t) = {N_r^n(nt)}/{\sqrt{n}}$ and $\hat{N}_s(t) = {N_s^n(nt)}/{\sqrt{n}}$. Defining $N_Q^n(t) = (N_r^n(t) + N_s^n(t) - 1)^+$ to be the total number of customers in the system but not in service at time $t$, then the asymptotic behavior of the scaled queue processes $\hat{N}_Q^n(t) = {N_Q^n(nt)}/{\sqrt{n}}$ is given by Theorem 1 below.

THEOREM 1. *Assume that* $\hat{N}_r^n(0) = \left(\hat{N}_Q^n(0) \wedge \frac{p\lambda^n}{\gamma}\right)$. *For any finite* $T > 0$,

$$\sup_{0\le t\le T}\left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t) \wedge \frac{p\lambda^n}{\gamma}\right)\right| \to 0 \text{ in probability as } n \to \infty.$$

Theorem 1, which relies on the Functional Strong Law of Large Numbers and the Functional Central Limit Theorem (Chen and Yao 2013), has a simple intuitive meaning. If the total scaled number of customers in the system excluding those receiving service, $\hat{N}_Q^n(t)$, is less than ${p\lambda^n}/{\gamma}$, then there are almost no customers waiting to receive service in the service queue (more precisely it is $o(\sqrt{n})$); alternatively if the total is greater than ${p\lambda^n}/{\gamma}$, then the scaled number in orbit is ${p\lambda^n}/{\gamma}$ and (almost) all others are in the service queue. More generally, Theorem 1 implies that, given the total number of customers in the limiting system, we now know how they are distributed between the orbit and the service queue. In other words, the state space collapses.

The state-space collapse result is similar to Proposition 3.1 in Armony and Maglaras (2004b), where the service provider offers customers call backs with a service guarantee. In their setting, customers who agree to receive a call back are also placed in a holding system akin to orbit in our setting. However, the driving mechanism and the proof techniques are significantly different. In our setting the orbit queue functions similarly to a low-priority queue in that if there are customers in the service queue, they are served exclusively; therefore, the service queue empties out faster than orbit. In contrast, in the setting of Armony and Maglaras (2004b), customers in orbit are sometimes given priority over the customers in the service queue (this happens when the number of customers in orbit exceeds the limit $\frac{p\lambda^n}{\gamma}$) otherwise the system would not meet the call-back guarantee. The diffusion limit presented above is also related to those developed in the queueing literature with abandonments; see Ward and Glynn (2003) and Borst et al. (2004). The main

difference in our model is that customers do not abandon but transition from orbit to the service queue. Therefore we need to use a different scaling to obtain meaningful limits approximations. For instance, if we used the scaling in Theorem 1.1. in Ward and Glynn (2003), the service queue would always be (asymptotically) empty, which does not lead to useful approximations. Hence we use an alternative scaling where the transition rate from orbit to the service queue scales at a faster rate; more specifically, it scales at the same rate as the utilization of the system. This scaling, however, introduces a technical difficulty because the orbit occupancy can change rapidly even in the limit. Nevertheless, we are able to prove that there is a state-space collapse in the limit, which leads to the diffusion limit presented above.

**Approximations.** In order to develop closed-form approximations for system performance, the next step is to apply the asymptotic result on the allocation of customers between the orbit and the service queue to a finite system. Since the exact results show that the total number of customers in the system is distributed geometrically, we apply the split of customers implied by Theorem 1 (assuming it holds for finite $n$). Computing the expected value of the occupancy of the service queue yields the following approximations,

$$\bar{N}_s \approx \rho + \rho^{\lfloor \frac{p\lambda}{\gamma}+1 \rfloor + 1} \left( \left\lfloor \frac{p\lambda}{\gamma} + 1 \right\rfloor - \frac{p\lambda}{\gamma} + \frac{\rho}{1-\rho} \right) \approx \frac{\rho}{1-\rho} \left( 1 - \rho \left( 1 - \rho^{\frac{p\lambda}{\gamma}} \right) \right), \tag{2}$$

where $\left\lfloor x \right\rfloor$ denotes the floor function. The second approximation follows from $\left\lfloor \frac{p\lambda}{\gamma} \right\rfloor \approx \frac{p\lambda}{\gamma}$. Utilizing MVA (see also Proposition 2), the approximation given by (2) can be used to estimate all other performance measures for queue lengths and wait times. By the PASTA property, the memoryless property of service times, and (2), the average time spent in the service queue for inflexible customers is

$$\bar{T}_{ss} = \frac{\bar{N}_s + 1}{\mu} \approx \frac{1}{\mu} \left( \frac{\rho}{1-\rho} \left( 1 - \rho \left( 1 - \rho^{\frac{p\lambda}{\gamma}} \right) \right) + 1 \right). \tag{3}$$

By (2) and the implication of Lemma 1 that $\bar{N}_s + \bar{N}_r = \frac{\rho}{1-\rho}$, we have that,

$$\bar{N}_r = \frac{\rho}{1-\rho} - \bar{N}_s \approx \frac{\rho}{1-\rho} \rho \left( 1 - \rho^{\frac{p\lambda}{\gamma}} \right). \tag{4}$$

By Little's Law and (4), we have that the average time spent in orbit is

$$\bar{T}_{rr} = \frac{\bar{N}_r}{p\lambda} \approx \frac{1}{p(\mu-\lambda)} \rho \left( 1 - \rho^{\frac{p\lambda}{\gamma}} \right), \tag{5}$$

and finally by equation (1) and the approximations (5) and (3), we can find an approximation of the average time spent in the service queue for flexible customers $\bar{T}_{rs}$.

The approximation given by equation (2) for the average number of customers in the service queue has an intuitive appeal. It is equal to the average number of customers at the service queue

in the absence of proactive service ($\rho/(1-\rho)$), multiplied by a constant, $\left(1-\rho(1-\rho^{p\lambda/\gamma})\right) \leq 1$, that represents the benefit of proactive service. As expected, this benefit disappears (i.e., the constant goes to one) if there are no flexible customers (i.e., $p=0$) or the average information lead time goes to zero (i.e., $1/\gamma \to 0$).

Furthermore, the approximations above allow us to derive additional properties of performance measures that could not be derived using exact analysis (see Table 1). For instance, using equation (2), we can show that the service queue occupancy decreases exponentially with $p/\gamma$, which implies there are decreasing marginal benefits in the proportion of customers that are flexible and the average information lead time. In addition, using equation (5) we can show that $\bar{T}_{rr}$ is monotonic decreasing in $p$. Also $\bar{T}_{rs}$ is monotonic decreasing in $p$ provided $\gamma \geq \mu - \lambda$ and $\rho > .2$.

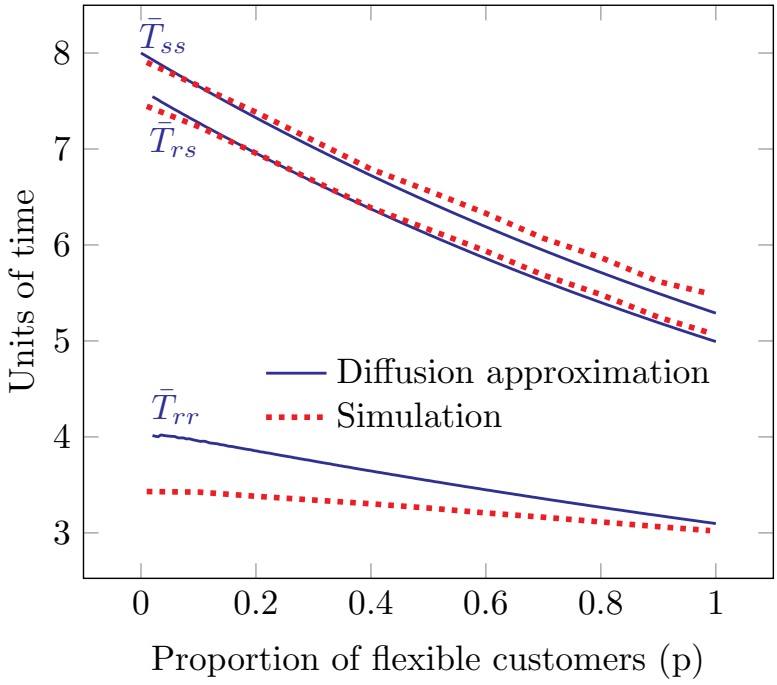
**Verifying the accuracy of the approximations:** Because the approximations presented above are based on an asymptotic result, in this section we examine their accuracy in finite systems where utilization is less than one. For instance, Figure 2 depicts the comparison of the diffusion approximations and simulated average delays for the case when $\lambda = 0.875$, $\gamma = .2$, and $\mu = 1$. A full sensitivity analysis of the accuracy of the approximations is given in Appendix B.1.1. In general, the approximations perform remarkably well for all values of $p$ when utilization is high (i.e., $\rho \in (.75, 1)$), and information lead time is not too large (i.e., $(\mu-\lambda)/\gamma \leq 1$). This is not surprising given the asymptotic regime deployed to develop the approximations assumed that $\mu - \lambda^n \to 0$ at the same rate as $\gamma^n \to 0$.[3]

Figure 2 also serves to illustrate the substantial reduction in delays derived from proactive service. For instance, if all customers are flexible (i.e., $p=1$) the total average delay in the service queue is reduced by 38.7% (from 8.0 to 4.90 time units). Even if only half of customers are flexible (i.e., $p=0.5$), the average time in the service queue is reduced by 22.2% (from 8.0 to 6.23 time units). This reduction in delays is achieved even though the average information lead time is relatively short (only 62.5% of the expected delay in the benchmark case). In other words, relatively little information lead time goes a long way when customers can be served proactively.

Remark 1 (When time in orbit is a constant). Although not essential for the rest of our analysis, we compare the reduction in wait times achieved with proactive service when time in orbit is exponentially distributed, to the case when time in orbit is deterministic. The latter case is studied in Zhang (2014) and can be interpreted as a case when the provider possesses more information about the customers' future service needs compared to the former. The setting in Zhang (2014) has two additional differences. First, it assumes that a customer does not have to be present

[3] As the diffusion limit presented above fails when the information lead time increases (i.e., $1/\gamma \to \infty$), not surprisingly, the approximation does not collapse to the exact analysis presented in Engel and Hassin (2017) where the authors assume that $\gamma = 0$. Therefore, the two results can be seen as applicable to different parameter regions.

**Figure 2** **Customer delays in the proportion of flexible customers** $(p)$**, where** $\lambda = .875$**,** $\gamma = .2$**,** $\mu = 1$**.**




for service. Hence, the waiting time measure they consider does not include service time, only time in queue. It is straightforward to modify their approach to include service time as well. This is the approximation we present here. Second, Zhang (2014) assumes that inflexible customers have preemptive priority over flexible customers. This assumption is essential for his analysis technique, however, it is not realistic for service systems. Hence, we only compare our results to his when $p = 1$, in which case preemptive priority does not matter because there are no inflexible customers. Let $w$ denote the time customers spent in orbit before they transition to the service queue. Under this assumption, Zhang (2014) shows that the average amount of time customers spend in the service queue (excluding time spend in service) when $p = 1$ is $\bar{\bar{T}}^q_{ss} = \frac{\rho}{\mu-\lambda} e^{-\mu(1-\rho)w}$. Based on (2), with $p = 1$ and $\gamma = 1/w$, we arrive at the following approximation, $\bar{T}^q_{ss} = \frac{\rho}{\mu-\lambda}\rho^{\lambda w}$. Let $\Delta(\rho) = \bar{\bar{T}}^q_{ss}/\bar{T}^q_{ss}$, then we have $\Delta(\rho) = e^{-w}\left(\frac{e}{\rho}\right)^{\rho w}$. It can be shown that $\lim_{\rho\to 0}\Delta(\rho) = 0$, $\lim_{\rho\to 1}\Delta(\rho) = 1$, and that $\Delta$ is (strictly) increasing in $\rho$. Therefore, knowing exactly when customers would transition from the orbit to the service queue helps further reduce average time spent in service queue. However, this additional reduction in waiting time decreases as the system reaches heavy traffic.

## 4. Economic Analysis: Endogenous Decision-Making and Welfare

The queueing analysis thus far has shown the significant potential of proactive service to improve operational performance. However, it assumes exogenous customer arrival rates to both orbit and service queue. This is unlikely to be realistic in many service settings because it is customers who choose to join the queue and/or to be flexible based on the costs and benefits of each option.

Therefore, to understand the benefits of proactive service, we need to consider the decisions of strategic customers in equilibrium.

To do so we build off a standard queueing game (e.g., Hassin and Haviv 2003) where, in addition to the option of joining or not, customers need to also choose if they accept to be flexible. In such a system, customer self-interested behavior generates two distinct economic frictions. The first has to do with the customer joining decision – any customer who joins the system increases the waiting time for everyone else. This is a negative externality customers do not take into account when deciding to join the system, leading to customers over-utilizing the system compared to the social (or profit maximizing) optimal (Naor 1969). The option to be flexible introduces a second friction. Any customer that chooses to be flexible reduces the waiting time for everyone else. This is a positive externality that customers do not take into account when deciding whether to be flexible or not, leading to customers under-adopting proactive services. Moreover, as we show in the next section, these two opposing externalities interact in non-trivial ways.

### 4.1. Customer Utility and Equilibrium Demand

To facilitate a game theoretic analysis, we assume that there exists a large population of potential customers who are homogeneous, rational, and risk-neutral economic agents. We also assume that customer waiting times are accurately approximated by the (smooth version of the) closed-form diffusion approximations of §3.

Each customer has some small exogenous probability of requiring service such that, in aggregate, customer service needs can be modelled by a Poisson process with rate $\Lambda$. Receiving service is valued at $v$ and each customer also has access to an outside service option whose value we normalize to zero. Customers decide whether to join, and if they join whether to be flexible, by examining the *expected* cost of these choices which we assume is common knowledge. More specifically, we assume that real-time waiting time information is not available but customers have an accurate belief about average waiting times; see Chapter 3 of Hassin and Haviv (2003) for an extensive review of the theory and applications of unobservable queues. The expected costs have three sources. First, all customers are averse to waiting at the service queue and incur a waiting-time cost $w_s \geq 0$ per unit of time spent there (waiting or receiving service). Second, flexible customers need to be ready to "answer the call" from the idle service provider at any time and therefore incur i) an *opportunity cost* $0 \leq w_r \leq w_s$ per unit time spent in orbit that reflects any inconvenience associated with "waiting" to commence service early; ii) a fixed *inconvenience cost* $h \geq 0$, which can be interpreted as the cost of giving up autonomy/spontaneity in the timing of joining the queue. Third, customers may need to pay prices $c_r \geq 0$ and $c_s \geq 0$ set by the provider for flexible and inflexible customers, respectively. Given the assumptions, the expected utility of customers who choose to join but

not to be flexible is $v - c_s - w_s\bar{T}_{ss}$, the expected utility of customers who join and are flexible is $v - c_r - h - w_r\bar{T}_{rr} - w_s\bar{T}_{rs}$, and the utility of customers who do not join is zero.

Customers choose to (1) not join, (2) join and be flexible, or (3) join and be inflexible, based on the option with the greatest expected utility. For notational convenience, we let $\lambda \leq \Lambda$ represent the effective demand (i.e., arrival) rate to the system such that $J = \lambda/\Lambda \in [0,1]$ gives the proportion of customers who join the system, and $p \in [0,1]$ represents the proportion of customers who choose to be flexible conditional on joining. Because customers are homogeneous, we are interested in symmetric Nash Equilibria where, given that all other customers play a mixed strategy represented by $(J,p)$ (i.e., join with probability $J$ and are flexible with probability $p$), each customer's best response is to also play strategy $(J,p)$. For the rest of the analysis we restrict our attention to $\lambda$ rather than $J$ as there is a one-to-one correspondence between the two.

### 4.2. Unregulated Customer Equilibrium

To study the incentives introduced by proactive service, we examine the case where customers make their own utility-maximizing decisions in an unregulated system, i.e., where $c_s = c_r = 0$. Under mild assumptions, Proposition 4 establishes the existence and uniqueness of equilibrium as well as comparative statics.

PROPOSITION 4. *If* $\frac{\Lambda}{\mu} \geq .75$, $v \geq 4\frac{w_s}{\mu}$ *and* $\gamma \geq \frac{w_s}{v}$, *then:*

*i. There exists a unique symmetric Nash Equilibrium* $(p_e, \lambda_e)$ *for customer flexibility and joining behavior.*

*ii. The equilibrium strategy is such that:*

(a) *The proportion of flexible customers* $p$ *and the arrival rate* $\lambda_e$ *are non-increasing in the costs of flexibility* $h$ *and* $w_r$.

(b) *The proportion of flexible customers* $p_e$ *is non-increasing in customer valuation* $v$, *and the arrival rate* $\lambda_e$ *is non-decreasing in customer valuation* $v$.

(c) *The proportion of flexible customers* $p_e$ *is non-decreasing in the waiting-time cost* $w_s$, *but the arrival rate* $\lambda_e$ *can be decreasing or increasing in the waiting-time cost* $w_s$. *Specifically, if all strategies are played with positive probability so that* $\lambda_e < \Lambda$ *and* $p_e \in (0,1)$, *then the arrival rate* $\lambda_e$ *is increasing in the waiting-time cost* $w_s$, *otherwise* $\lambda_e$ *is decreasing in the waiting-time cost* $w_s$.

The conditions under which this proposition holds also ensure that utilization is relatively high and information lead time is relatively low, therefore ensuring that the diffusion approximations are a good representation of the system performance. We prove Part i by construction, considering all possible cases and proving uniqueness and existence though enumeration. The comparative statics results in Part ii rely on the monotonicity of delays in both the arrival rate and the proportion

of flexible customers, and are largely in line with intuition. Part iia shows that, as the costs of flexibility $(h, w_r)$ increase, fewer customers agree to be flexible and fewer customers join, just as one might expect. Part iib shows that, as customer valuation for service $(v)$ increases, more customers join, which is also as expected. More interestingly, Part iib also shows that, as customer valuation $(v)$ increases, a smaller proportion of those who join choose to be flexible, which suggests a non-obvious interaction of externalities. Specifically, this happens because as congestion increases (i.e., more customers join due to their valuations $(v)$ being higher) the value of free riding (i.e., the value a customer gets when other customers choose to be flexible) also increases. As a result, the proportion of customers who agree to be flexible becomes smaller. Perhaps even more surprising is Part iic, which shows that, as the cost of time spent in the service queue $(w_s)$ increases, the arrival rate may actually increase. The reason is that the increase in waiting-time cost $(w_s)$ induces more customers to be flexible, which generates a positive externality (i.e., reduces average waiting time), and in turn induces more customers to join. In this way one can clearly see the positive externality associated with flexibility interacts with the negative externality associated with congestion.

### 4.3. Customer suboptimal behavior: Over-utilization and Free-Riding

Next, we seek to understand how customer decisions in an unregulated equilibrium compare to those that a profit-maximizing service provider would want. The provider seeks to maximize the revenue rate from prices paid by customers subject to customers' equilibrium behavior,[4] i.e.,

$$\max_{c_r, c_s \geq 0} \lambda(pc_r + (1-p)c_s) \quad \text{subject to: } (p, \lambda) \text{ is an equilibrium.} \tag{6}$$

Because customers are homogeneous, a profit maximizing provider will not find it optimal to set prices such that customers are left with a positive surplus in equilibrium; if this was the case the provider would be able to increase prices without impacting customer decisions (Hassin and Haviv 2003, §1.3). Therefore, the profit maximizer will set prices such that $c_s = v - w_s\bar{T}_{ss}(p_e, \lambda_e)$ and $c_r = v - h - w_r\bar{T}_{rr}(p_e, \lambda_e) - w_s\bar{T}_{rs}(p_e, \lambda_e)$. Given this, the provider's revenue can be rewritten as,

$$W(p, \lambda) = \lambda \left[ p(v - h - w_r\bar{T}_{rr}(p, \lambda) - w_s\bar{T}_{rs}(p, \lambda)) + (1-p)(v - w_s\bar{T}_{ss}(p, \lambda)) \right]. \tag{7}$$

An interesting observation is that the profit of the provider in this case is equal to the average welfare rate of customers, an observation also made by (Hassin and Haviv 2003, §1.3). In other words,

[4] We note that, since proactive service may require a one-off cost to implement and perhaps a variable cost to monitor customer service needs, the provider will need to compare any increase in revenue to the implementation and running-costs to determine whether or not to implement proactive service. However, since this is a straightforward comparison, we do not model this explicitly and assume proactive service can be implemented at zero cost. We will therefore use revenue and profit interchangeably.

the profit maximizer would want customers to behave in exactly the same way as a benevolent social planner (whose aim is to maximize welfare). The only difference is that the profit maximizer would set the prices so as to extract all of the customer surplus, while the social planner, for whom prices are an internal transfer, would be indifferent between any price. Therefore, to understand whether customers' autonomous joining decisions of §4.2 are suboptimal for the profit maximizer, it would suffice to compare them to those of a benevolent social planner.

The existence of the optimal solution to the social planner's problem is guaranteed by the fact that the action space is compact and the objective function is continuous.[5] We compare the socially optimal customer actions with the equilibrium customer decisions of Proposition of §4.2 in the next result.

PROPOSITION 5. *If* $\frac{\Lambda}{\mu} \geq .75$, $v \geq 16\frac{w_s}{\mu}$, $\gamma \geq \sqrt{\mu w_s/v}$, *then for any socially optimal/profit-maximizing solution* $(p_{so}, \lambda_{so})$,

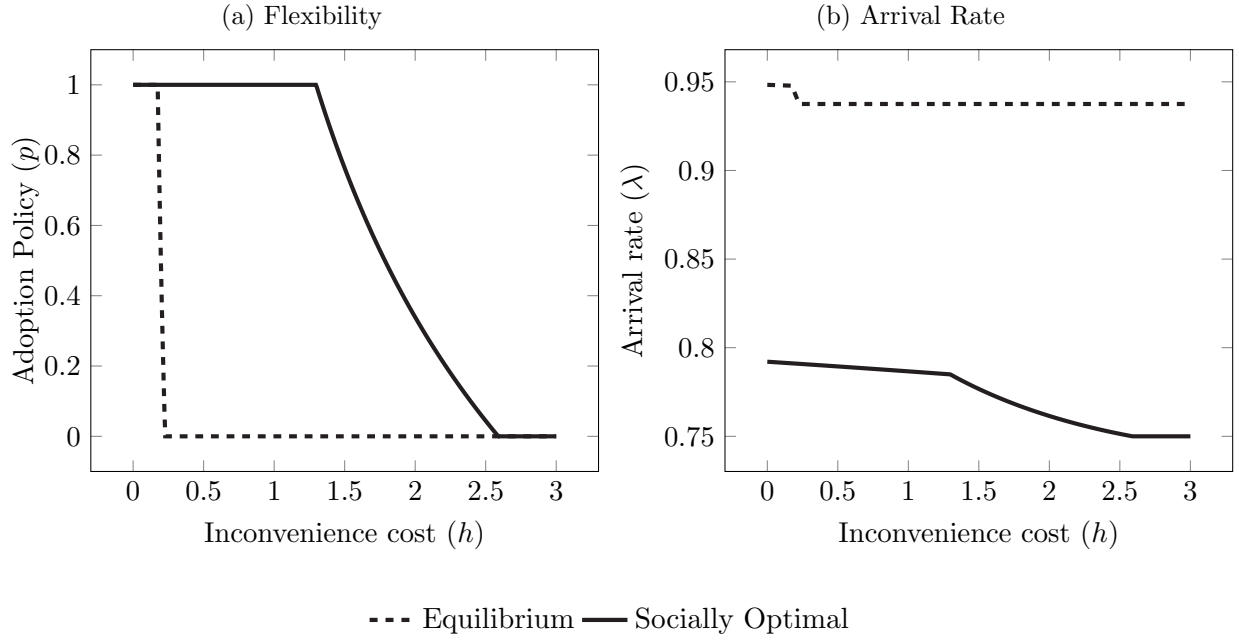
*i. Customers over-utilize the system compared to the socially optimal/profit-maximizing solution,* $(\lambda_{so} \leq \lambda_e)$.

*ii. Customers under-adopt proactive service compared to the socially optimal/profit-maximizing solution,* $(p_{so} \geq p_e)$. *In particular, there exist thresholds of the flexibility cost* $h$ *denoted by* $\underline{h}$ *and* $\bar{h}$, *where* $0 < \underline{h} < \bar{h}$, *such that if* $h \geq \underline{h}$ *then* $p_e = 0$ *and if* $h \leq \bar{h}$ *then* $p_{so} = 1$. *This implies that if* $\underline{h} \leq h \leq \bar{h}$ *then* $p_e = 0$ *and* $p_{so} = 1$, *i.e., no customer would choose to be flexible in equilibrium but the social planner (or profit maximizer) would designate all customers who join to be flexible.*

The conditions under which this proposition holds are a subset of the conditions of Proposition 4 and, as was the case there, they also ensure that the diffusion approximations are a good representation of the system performance. Proposition 5 shows that customers will over-utilize a service system with proactive service and under-adopt proactive service (the option to be flexible) compared to the socially optimal or the profit-maximizing solution. Figure 3 illustrates this point by showing the equilibrium strategy and the socially optimal/profit-maximizing strategy as a function of the fixed cost of flexibility ($h$) for a specific example. As can be seen in Figure 3a, the under-adoption of proactive service can be substantial in the sense that there exists a region ($0.3 < h < 1.4$ in Figure 3a), where the central planner would have dictated that all customers who join be flexible but in equilibrium flexibility is a strictly dominated strategy. Proposition 5 Part ii shows that such a region is not specific to this example but always exists. In this region, customers would be better off if they collectively chose to be flexible, but because customers are individually better off by

[5] We note that we are unable to prove that the social planner's objective $W(p, \lambda)$ is concave. Nevertheless, in numerical experiments we find the first order conditions are both necessary and sufficient.

**Figure 3** **Comparison of Equilibrium Strategy and Socially Optimal:** $\Lambda = .95, \gamma = .25, \mu = 1, w_s = 1, w_r = 0, v = 16$.



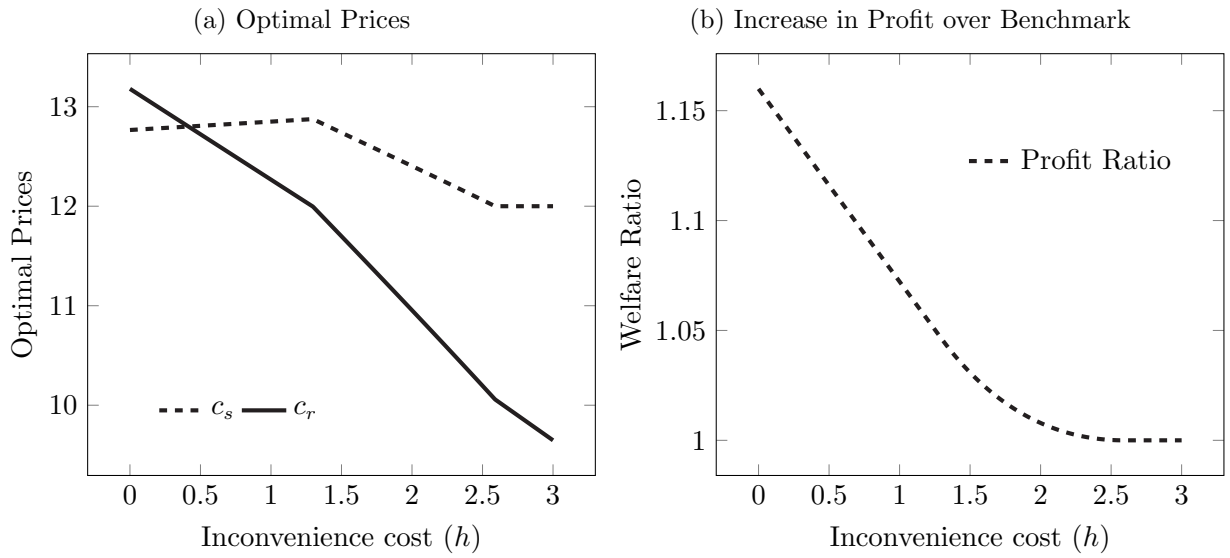

free-riding, no-one chooses to be flexible. Similarly, Figure 3b shows that customers over-utilize the system in general, but when being flexible is optimal for at least some customers (i.e., $h < 0.3$), it also exacerbates customer over-utilization through an increase in the equilibrium arrival rate.

The result of Proposition 5 suggests that, although the operational benefit from proactive service could be substantial, realizing it will not necessarily take place automatically because of customer self-interested behavior. As in the case where proactive service is not possible (e.g., Naor 1969), the profit maximizer can overcome this problem by setting prices/tolls $c_r$ and $c_s$, for flexible and inflexible customers, respectively, that incentivize optimal joining behavior. Figure 4a depicts these prices against the fixed cost of flexibility ($h$) for the same example as Figure 3.[6] Similarly, Figure 4b shows the improvement in provider revenue against the fixed cost of flexibility $h$ by showing the ratio of the optimal revenue in the case with proactive service over the benchmark case without proactive service.

What is important to note from Figure 4a is that, in different regions of the fixed cost of flexibility ($h$), the price for flexible customers ($c_r$) and inflexible customers ($c_s$) play different roles depending on the combination of economic frictions faced. For example, $c_r$ is lower than $c_s$ for all cases where $h$ is high enough such that not all customers would autonomously choose to be flexible ($0.3 < h < 2.6$). This must be the case to incentivize customers to be flexible despite their free-riding

[6] We note that for some values of $h$ there exist multiple prices that are optimal and revenue equivalent. For these cases we show the lowest price. More specifically, when $h < 1.4$ ($h > 2.6$) any price $c_s$ ($c_r$) greater than the one depicted in the figure would also be optimal. Since no customer would choose to be inflexible in the case when $h < 1.4$ (flexible in the case when $h > 2.6$), this would not make a difference to the revenue.

**Figure 4** **Optimal Pricing and Welfare:** $\Lambda = .95, \gamma = .25, \mu = 1, w_s = 1, w_r = 0, v = 16$




incentives. Since the incentive to free-ride grows as $h$ increases, so must the gap between $c_s$ and $c_r$. Furthermore, in the regions where at least some customers choose to be flexible (i.e., $0 < h < 2.6$) $c_r$ is decreasing in $h$. This happens because, as the fixed cost of flexibility $h$ increases, the toll that the profit maximizer needs to impose to prevent customers from over-joining (while at the same time extracting all rents) is lower. Similarly, in the regions where some (but not all) customers choose to be inflexible (i.e., $1.3 < h < 2.6$) $c_s$ is also decreasing in $h$. This happens because, as $h$ increases, the provider will find it optimal to incentivize fewer customers to be flexible. Since fewer customers are flexible, the waiting time of inflexible customers will increase and therefore the price that the profit maximizer will need to impose to prevent over-joining (and extract all rents) will have to decrease. Finally, what is important to take away from Figure 4b is first, that proactive service can substantially increase the revenue (or welfare) in a system that offers proactive customer service compared to one that does not, and second, that the lower the costs of flexibility for customers, the more valuable proactive service.

## 5. Generalizations and Extensions of the Queueing Model

In this section we explore two extensions to the queueing analysis presented in the previous sections. First, we examine a setting with more than one server. Second, we examine the case when future information is imperfect, specifically the case that by serving customers proactively the provider can make "errors" and serve customers who would not have been served in the absence of proactive service. We find that key results, pertaining to the benefit of proactive service and the economic incentives to adopt it, continue to hold under these extensions.

### 5.1. Multiserver Setting

Due to the fact that queueing systems exhibit economies of scale, the results of the single-server case cannot be taken for granted in the multiserver setting. To extend the analysis, we assume that there are $m > 1$ identical servers who may serve customers proactively should the service queue be empty. As we show below, the results of Lemma 1 and Proposition 1 extend directly, as do parts $(i)$ and $(ii)$ of Proposition 3. Changes in the the mean value relations between performance measures, however, complicate the analysis. Nevertheless, we are still able to show that proactive service creates a benefit for both flexible and inflexible customers. Finally, the scalings used in Theorem 1 are no longer helpful. However, alternative scalings enable us to establish a similar state-space collapse from which we are able to suggest approximations for the performance measures of interest. For the rest of this section we add the superscript $m$ to the performance measure notation to indicate the multiserver setting.

**Direct extension of Lemma 1, Proposition 1, and Proposition 3 parts $(i)$ and $(ii)$:** The system without proactive service is still a Jackson network where orbit is an $M/M/\infty$ and the service queue is an $M/M/m$ queue. As such, the stability condition is that $\lambda < m\mu$. Lemma 1 applies in the multiserver setting as well; the proof is essentially identical. This leads to the extension of Proposition 1. Lastly, the pathwise coupling argument used to prove Proposition 3 parts $(i)$ and $(ii)$ applies independently of the number of servers and hence extends directly as well.

**Modified results from the single-server case:** The first change in the analysis from the single-server case is that the MVA changes slightly. Specifically, by Lemma 1 we have that $\bar{N}_r^m + \bar{N}_s^m = \bar{N}_s^{mB}$ and, although Little's Law and PASTA continue to apply, the expected time in system for inflexible customers is more complicated than the single-server setting and equation (1) no longer applies. However, because Proposition 1 is still valid in the case with multiple servers, Proposition 2(i)–(iii) continues to hold; that is, the average delay in each queue is shorter in a system with proactive service and that proactive service reduces service queue delays for inflexible customers. While this result is not as detailed as the analysis in the single-server case, it does demonstrate that proactive service benefits both flexible and inflexible customers.

The next step is to develop a diffusion approximation for the multiserver case. To do so, we start from the notation used in §3.3, but in place of the scalings of Theorem 1, we use the standard Halfin–Whitt multiserver scalings (Halfin and Whitt 1981). That is, we consider a sequence of systems indexed by $n$ and the arrival rate and the number of servers $\lambda^n = n\mu(1 - \frac{\beta}{\sqrt{n}})$, $m^n = n$, respectively, and additionally we scale the orbit departure rate as $\gamma^n = \sqrt{n}\gamma$ in the $n$th system. Rescaling the orbit occupancy, service queue occupancy (excluding customers in service), and the

total number of customers in the system (again excluding customers in service) as $\hat{N}_r^n(t) = \frac{N_r^n(t)}{\sqrt{n}}$, $\hat{N}_s^n(t) = \frac{N_s^n(t)-n}{\sqrt{n}}$, and $\hat{N}_q^n(t) = \frac{N_r^{(m)n}(t)+N_s^n(t)-n}{\sqrt{n}}$, respectively, then we have the following result.

THEOREM 2. *If* $\hat{N}_r^n(0) = \left(\hat{N}_q^n(0) \wedge \frac{p\lambda^n}{\gamma^n}\right)$ *then*

$$\sup_{0\le t\le T} \left| \hat{N}_r^n(t) - \left(\hat{N}_q^n(t) \wedge \frac{p\lambda^n}{\gamma^n}\right)\right| \to 0 \text{ in probability as } n \to \infty.$$

Theorem 2 shows that under the updated scalings, the state-space collapse argument holds under the Halfin–Whitt scaling as it did under the scaling of Theorem 1. This allows us to generalize the intuition of the single-server setting regarding how customers are distributed between orbit and the service queue. Using this intuition, and the fact that the total number of customers in the (finite) system follows the same distribution as the occupancy of the standard $M/M/m$ queue, we can then apply the split of customers implied by Theorem 2 to compute the expected number of customers in the orbit and the service queue. To compute the expected waiting times in different parts of the system, we use this approximation and Little's Law as follows. By a direct application of Little's Law we find the expected time in orbit for flexible customers $\bar{T}_{rr}^m$. The expected time in the system, denoted by $\bar{T}^m$, can be found using Little's Law and the expected number of customers in system. To find the average waiting time of inflexible customers, $\bar{T}_{ss}^m$, we use an MVA approach similar to that in §5.2 of Adan and Resing (2002). To use this approach we need the probability that there are more than $m$ customers in the system, which can be calculated using the classic Erlang-C formulas. Finally, the expected time flexible customers spend in the service queue, $\bar{T}_{rs}$, is then determined from $\bar{T}^m$, $\bar{T}_{ss}^m$ and $\bar{T}_{rr}^m$. The accuracy of these approximations is similar to the single-server case and is examined using simulation in the online Appendix B.1.2 and applied in the example of induction of labor in Appendix C.

### 5.2. Imperfect Information about Future Service Needs

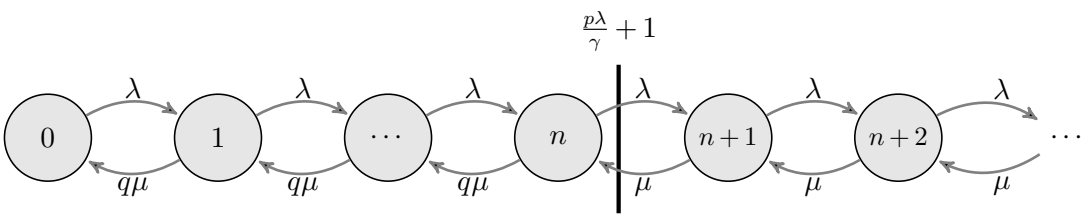
Up to this point we have assumed that the provider has perfect information about future customer service needs. This assumption may not always be true. For example, in a healthcare setting, the needs of patients who have a planned procedure may change, e.g., patients scheduled for induction of labor may go into labor spontaneously and hence no longer require the procedure. In other settings where the future information comes from predictive models, the models could make errors in predicting customers with future service needs or, alternatively, customers who are predicted to have a service need may manage to have it resolved though alternative channels, for example, the firm's website. By serving such customers proactively, the provider serves customers who, in the absence of proactive service, would not have entered the service system, thus increasing utilization and hence congestion and delays. The goal of this section is to examine the impact of imperfect information on the performance of proactive service.

To model imperfect information, we return to the single-server model depicted in Figure 1. Incorporating the potential "error" customers into the model requires additional assumptions, however, the main idea we wish to investigate is under what conditions proactive service is still beneficial despite errors. To do so with an analytical model, we assume that every time the server pulls a customer from orbit it makes an error with probability $1-q$. That is, a proportion $1-q$ of customers served proactively would not have transitioned to the service system had they not been pulled and, therefore, would not have been served at all. As a consequence, errors increase the effective arrival rate to the system.

By construction, under these assumptions the analysis of the benchmark case (where there is no proactive service) does not change. The steady-state occupancies of the orbit and the service queue follow the Poisson and Geometric distributions with parameters $\frac{p\lambda}{\gamma}$ and $\rho$, respectively. However, the exact analysis of the system with proactive service is substantially more challenging. Lemma 1 no longer holds because, unlike in the case with no errors, the total number of customers in the system in steady state now depends on how frequently the server pulls from the orbit due to errors. Therefore, there is no longer a guarantee that proactive service will lead to shorter waiting times.

The asymptotic analysis, however, can be used to develop approximations of system performance. Taking the case of a single server, it can be shown that Theorem 1 holds in this case as well because the service rate for customers from orbit does not play a role in the proof. (Note: Theorem 2 would apply in the multiserver case by the same reasoning.) Now, assuming that Theorem 1 also holds for finite systems as well, allows us to model the system using a birth-death process as follows. Let $N_Q$ denote the total number of customers in the queue. Since this is a Markovian system, the birth rate (i.e., the rate of transition from $N_Q$ to $N_Q+1$) is given by $\lambda$. If $N_Q \geq p\lambda/\gamma$, then orbit occupancy is $p\lambda/\gamma$ customers and the rest of the customers are waiting in the service queue. In this case, the departure rate from $N_Q$ to $N_Q-1$ is given by the service rate $\mu$ – since the server picks customers from the service queue it never makes mistakes. If $0 < N_Q < p\lambda/\gamma$ then all customers are in orbit. Hence, the departure rate is $\mu q$, since there is a probability $1-q$ that the server will pull a customer in error. Therefore, the total number of customers in the system (including those in service) can be modeled as the birth-death process pictured in Figure 5 and the occupancy distribution can be estimated using simple recursive equations.

**Figure 5 Birth-Death Transition**



From this, we can estimate delays in the system as a whole. To estimate delays at the service system and the orbit we need to use Theorem 1 again. More specifically, for every state $N_Q$ of the system described by Figure 5, Theorem 1 implies that the number of customers in orbit is $N_Q - 1$ if $0 < N_Q < p\lambda/\gamma$ and $p\lambda/\gamma$ if $N_Q > p\lambda/\gamma$ and all other customers are in the service queue. With this, we can then estimate the average occupancy of the service queue. If $q \neq \rho$, then

$$\bar{N}_s \approx \frac{\rho}{1-\rho}\frac{P_0(n)}{q-\rho}\left(1-\rho+(q-\rho)\left(\frac{\rho}{q}\right)^n\left(n-\frac{p\lambda}{\gamma}+\frac{q-1}{q-\rho}+\frac{\rho}{1-\rho}\right)\right), \tag{8}$$

$$\approx \frac{\rho}{1-\rho}\frac{P_0(\frac{p\lambda}{\gamma}+1)}{q-\rho}\left(1-\rho+(q-\rho)\left(\frac{\rho}{q}\right)^{\left(\frac{p\lambda}{\gamma}+1\right)}\left(1+\frac{q-1}{q-\rho}+\frac{\rho}{1-\rho}\right)\right), \tag{9}$$

where $n := \lfloor \frac{p\lambda}{\gamma}+1 \rfloor$, $P_0(x) = \left(\frac{q}{q-\rho}\left(1-\left(\frac{\rho}{q}\right)^{x+1}\right)+\frac{\rho}{1-\rho}\left(\frac{\rho}{q}\right)^x\right)^{-1}$, and if $q=\rho$, then

$$\bar{N}_s \approx \frac{\rho}{1-\rho}\left(\frac{1}{n+\frac{1}{1-\rho}}\right)\left(\frac{n(1-\rho)}{\rho}+\left(n+1-\frac{p\lambda}{\gamma}+\frac{\rho}{1-\rho}\right)\right), \tag{10}$$

$$\approx \frac{\rho}{1-\rho}\left(\frac{1}{(\frac{p\lambda}{\gamma}+1)+\frac{1}{1-\rho}}\right)\left(\left(\frac{p\lambda}{\gamma}+1\right)\frac{1-\rho}{\rho}+\left(2+\frac{\rho}{1-\rho}\right)\right). \tag{11}$$

The approximations given by (9) and (11), which are smooth in the proportion of flexible customers $p$, follow from those of (8) and (10), respectively, by letting $\lfloor \frac{p\lambda}{\gamma} \rfloor = \frac{p\lambda}{\gamma}$.

Estimates for delays of inflexible customers in the service queue can then be estimated as $\bar{T}_{ss} = (\bar{N}_s + 1)/\mu$, which follows from the PASTA property and the memorylessness of service times. To get an approximation for the delays of flexible customers in the service queue $\bar{T}_{rs}$, we make the assumption that the arrival process of flexible customers to the service queue can be approximated by a Poisson process. Given this approximation, the delays for flexible customers in the service queue are equal to those of inflexible customers. Naturally, this approximation becomes more accurate as utilization increases and fewer customers are served proactively. We illustrate the performance of these approximations with a specific example in Figure 6. This example, and a more extensive numerical comparison, suggests that the approximations work well when $\rho \geq 0.75$ and $(\mu-\lambda)/\gamma \leq 1$.[7] Using these approximations, one can make several interesting observations, which we summarize with the following proposition.

[7] We compare this approximation to simulations for the cases when $\lambda/\mu \in \{.75, .8, .825, .85, .875, .9, .925, .95, .975, .99\}$, $\mu = 1$, $\gamma \in \{.25, .2, .15, .1, , .05, .025\}$, $p \in \{.001, .1, .2, ..., .9, .999\}$, and $q \in \{.1, .2, ..., .9, 1\}$. Each combination is simulated for 30 replications, each replication is simulated for 100,000 units of time, of which the first 20% is considered a warm-up period and is subsequently excluded from estimation of performance measures. For the cases when $\gamma > \mu - \lambda$, the simulated $\bar{T}_{ss}$ and $\bar{T}_{rs}$ are respectively within 10% absolute percent deviation from the approximation 95.7% and 93.4% of the time, where the approximation follows from (8) and (10) for $\bar{N}_s$ and $\bar{T}_{rs} = \bar{T}_{ss} = (\bar{N}_s + 1)/\mu$. The average absolute percentage deviations for $\bar{T}_{ss}$ and $\bar{T}_{rs}$ are 3.36% and 4.18% (respectively).

PROPOSITION 6. *There exists a threshold* $0 < \bar{q} < 1$ *such that the system with proactive service generates lower waiting times compared to the system without proactive service if, and only if, the proportion of errors is less than* $1 - \bar{q}$*. Furthermore,*

*(i) If* $\rho > \frac{1}{2} + \frac{\gamma}{2p\lambda}$*, then the maximum proportion of errors the system can tolerate* $(1 - \bar{q})$ *is greater than the system's idle time* $(1 - \rho)$*.*

*(ii) There exist combinations of parameters* $(p, \lambda, \gamma, \mu)$ *such that the threshold* $\bar{q}$ *is decreasing in utilization (i.e., as the service rate* $\mu$ *approaches the arrival rate* $\lambda$*).*

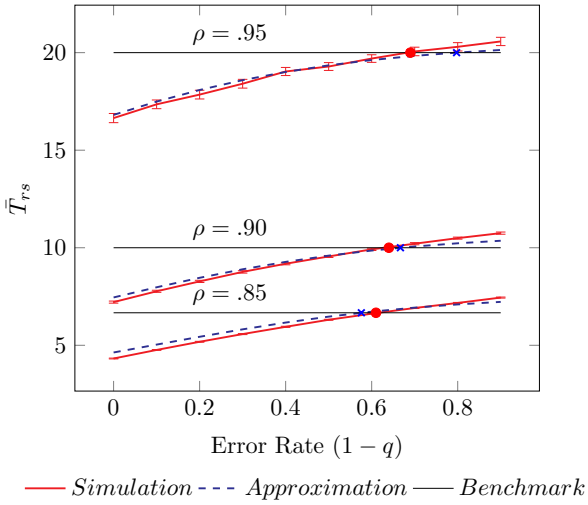
The proposition establishes the intuitive result that proactive service reduces waiting times only if the proportion of errors is below a critical threshold. What is more interesting is that, provided the system is relatively highly utilized (i.e., $\rho > \frac{1}{2} + \frac{\gamma}{2p\lambda}$), then proactive service may reduce waiting times even if the proportion of errors is greater than the system's idle capacity. Furthermore, the proposition shows that there exist cases such that at higher system utilization (i.e., as $\mu$ approaches $\lambda$) the system is able to handle more errors. In fact, we empirically observe that the system's tolerance for error increase with utilization in 99.49% of numerical experiments described in footnote 7. The finding that a more heavily utilized system is able to handle more errors and still benefit customers compared to the benchmark case is illustrated in Figure 6. Figure 6 depicts the delays for flexible customers in the service queue when $p = 1$ as a function of the error rate $(1 - q)$ and shows that at a utilization of $\rho = 85\%$ the system can tolerate an error rate as high as 60% before the benefits of proactive service are eroded by errors, and when utilization increases to $\rho = 95\%$ the system can handle an error rate of almost 70% before delays are greater than the benchmark case. Furthermore, as Figure 6 also makes clear, this result is not an artifact of the approximation as it is confirmed by the simulation study. This finding seems counterintuitive at first because, in a more heavily utilized system, one would expect errors to increase delays more than in a system at lower utilization. However, this can be explained by the fact that reduction in delays gained through proactive service grows as utilization increases.

## 6. Discussion

This paper set out to explore two high-level questions: (i) What is the operational impact of proactive service, and (ii) are there any economic frictions associated with proactive service?

From an operational perspective there are two contributions. The first is to show that proactive service can substantially reduce delays for both flexible and inflexible customers. This is the case in both single- and multiserver settings and remains true even if the proportion of flexible customers is limited or proactive service may result in sometimes serving customers who would not have otherwise required service. The second contribution is the diffusion approximation, based on novel asymptotic limits, that allow us to quantify these benefits with closed-form expressions.

**Figure 6** $\bar{T}_{rs}$ **vs Proportion of Errors** $(1-q)$**, where** $p=1$**,** $\gamma=.25$**,** $\mu=1$**,** $\lambda \in \{.85, .90, .95\}$**.**




From an economic perspective, the most important contribution is to show that proactive service is likely to be under-adopted due to a free-riding problem and may also exacerbate customers' incentives to over-join the system. Furthermore, the equilibrium behaviour of the system with proactive service generates counterintuitive findings. For example, as the waiting time cost increases the arrival rate may increase – due to the higher cost of waiting in the queue more customers decide to adopt proactive service, which in turn reduces average waiting time and, thus, induces more customers to join the system. These results underline the importance of formally modelling proactive service and suggest that, in order to realize the operational benefits of proactive service, providers will need to offer incentives (e.g., tolls) that alleviate the economic frictions of free-riding and over-utilization.

# A. Appendix

## A.1. Proof of Lemma 1:

The result follows from the fact that the system without proactive service and the service queue when proactive service is used can be modeled by Markov chains with identical transition rates. □

## A.2. Proof of Proposition 1:

We prove part $(i)$ of the proposition using a path-wise coupling of stochastic processes on a common probability space. We provide the sketch of the proof since this approach is standard in queueing literature (See Levin et al. 2009, Chapters 4,5 for an introduction). Fix a sample-path (i.e., the sequence of customer inter-arrival times) to both locations, the sequence of information lead-times (times in orbit), and the sequence of service times. Now, given this sample path, we compare the resultant orbit occupancy processes in a system with proactive service to that without on the same probability space (i.e., $N_r(t)$ and $N_r^B(t)$). By examining all possible events (e.g., arrivals, customer departures from orbit, and customer departures from the service queue), one can show that $N_r(t) \leq N_r^B(t)$ in each such sample path. The result follows then by Shaked and Shanthikumar 2007, Theorem 1.A.1.

Part $(ii)$ of the proposition is an immediate consequence of Lemma 1 and the non-negativity of $N_r(t)$. $\square$

### A.3. Proof of Proposition 2:

We first establish four equalities using MVA and then prove the desired results using these equalities. Because external arrivals follow a Poisson process, by the PASTA property Tijms (2003) and the fact that service times are exponential, we have $(a)$, $\bar{T}_{ss} = \frac{\bar{N}_s+1}{\mu}$. By Little's Law (Kleinrock 1976, eq.2.25 on pg.17) we have the following identities, $(b)$, $\bar{N}_r = p\lambda\bar{T}_{rr}$ and $(c)$ $\bar{N}_s = \lambda\left((1-p)\bar{T}_{ss} + p\bar{T}_{rs}\right)$. Finally Lemma 1 yields $(d)$, $\bar{N}_r + \bar{N}_s = \frac{\lambda}{\mu-\lambda}$.

By Proposition 1(i) and (b), we have (i). Similarly (ii) follows from Proposition 1(ii) and $(a)$. Observing that, for the benchmark system with no proactive service that $\bar{T}^B_{ss} = \bar{T}^B_{rs}$, then equation (1) implies (iv), and combining this with (ii) yields (iii). Next we prove (1) to conclude the proof. By $(c)$ and $(a)$ we have $(\mu-\lambda)T_{ss} - 1 = -\lambda p(T_{ss} - T_{rs})$, and plugging in $(b)$ and $(c)$ for $N_r$ and $N_s$ in $(d)$, respectively, we obtain $(\mu-\lambda)T_{ss} - 1 = (\mu-\lambda)p(T_{ss} - T_{rs}) - p(\mu-\lambda)T_{rr}$, combining this with $(\mu-\lambda)T_{ss} - 1 = -\lambda p(T_{ss} - T_{rs})$ yields (1). $\square$

### A.4. Sketch for Proof of Proposition 3:

We provide a sketch of the couplings used in the proof of part $(i)$ which, when combined with Lemma 1 implies part $(ii)$, and from both parts $(i)$ and $(ii)$ the monotone results follow. Full details are provided in the online appendix. We note that the exponential assumptions of inter-event times are necessary for the coupling in this proof.

To show that $\pi_r$ is increasing in $p$ (in a stochastic ordering sense), we couple the arrival and service queue departure events (service completions) so that arrivals and departures are synchronized across two versions of the Markov Chain, representing the state of the processes (note: this means the number of customers in each version is identical also). We further couple the customer types such that a flexible arrival in the first version implies a flexible arrival in the second, but an inflexible customer arrival in the first may result in a flexible customer arrival in the second (this captures the difference in $p$ across versions). Lastly the epochs when a flexible customer in orbit self-transitions to the service queue are also coupled such that when there are more people in orbit in the second version (because more flexible customers have arrived there), then customers may depart in the second version but not the first. However, if the number in orbit across versions is identical then these events are synchronized across versions.

To show that $\pi_r$ is decreasing in $\gamma$ (in a stochastic ordering sense) we couple arrivals, customer types, and service queue departure events so that arrivals and departures are synchronized across two versions of the Markov Chain representing the state of the processes. We then vary the rate at which customers depart from orbit to the service queue on their own across versions so that customers depart orbit faster in the second version. By coupling the self-transitions from orbit to the service queue such that the common (minimum) self-transition rate across versions (at a given point in time) is captured by one exponential variable, and the difference in transition rates across versions is captured by another exponential variable, we couple the self-transition events such that when the number in orbit is identical across systems it cannot be that the system with a slower transition rate (smaller $\gamma$) has a departure when the faster version does not.

### A.5. Sketch for proof of Theorem 1:

We provide a sketch of the proof; full details are provided in the online appendix. We prove the result in two steps, and in each step we use the approach in Reiman (1984). First we prove that for any $\epsilon > 0$

$$P\left\{\sup_{0\le t<1} \hat{N}_r^n(t) > \frac{p\lambda^n}{\gamma} + \epsilon\right\} \to 0, \text{ as } n \to \infty. \tag{12}$$

This result implies that the number of customers in the orbit is almost always less than $\frac{p\lambda^n}{\gamma}$, therefore bounded. We prove this result by showing that whenever the number of customers in the orbit is more than $\frac{p\lambda^n}{\gamma}$, then the rate customers leave the orbit is much higher that the rate that they arrive to the orbit, regardless of the number of customers in the service queue.

In the second step we focus on the service queue, assuming $\hat{N}_r^n(t) \le \frac{p\lambda^n}{\gamma} + \epsilon/4$ for all $t$ and arbitrary $\epsilon > 0$. We know from (12) that the probability that this assumption holds goes to 1 as $n \to \infty$. Next we prove that, under this assumption, if

$$\left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t) \wedge \frac{p\lambda^n}{\gamma}\right)\right| > \epsilon. \tag{13}$$

that is, the claimed state-space collapse result does not hold, then

$$\hat{N}_r^n(t) < {}^{p\lambda^n}\!/_{\gamma} - {}^{\epsilon}\!/_{2}, \text{ and } \hat{N}_s^n(t) > \frac{\epsilon}{2}. \tag{14}$$

In other words, (33) implies that the number of customers in orbit is strictly less than the upper bound $\frac{p\lambda^n}{\gamma^n}$ and the number of customers in the service queue is non-negative. Because the service queue has priority this implies that whenever (13) holds, the server will only serve the service queue. We then show that the service queue must therefore reach zero quickly. But if the service queue is empty, then $\left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t) \wedge \frac{p\lambda^n}{\gamma}\right)\right| < \epsilon/4$ and so (13) cannot hold. Since $\epsilon > 0$ is arbitrary, this proves the desired result.

### A.6. Sketch for Proof of Proposition 4:

There are six possible types of equilibrium strategies which are the combinations of $\lambda_e < \Lambda$ or $\lambda_e = \Lambda$ with $p_e = 0$ or $0 < p_e < 1$ or $p_e = 1$. We show each type of equilibrium corresponds to a given region of the parameter space in $v$ and $h$ which can be expressed in terms of the other model primitives $\Lambda$, $\gamma$, $\mu$, $w_s$, and $w_r$. To prove part (i.) on the uniqueness and existence of equilibrium, we show that the regions are mutually exclusive and collectively exhaustive. The cases (unique equilibrium solution and region) are:

**Case 1:** $p_e = 0$ **and** $\lambda_e = \Lambda$, if $\Lambda < \mu$, $v \ge \hat{v}_0 := \frac{w_s}{\mu - \Lambda}$ and $h \ge \hat{h}_\Lambda := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \Lambda}\right)\frac{\Lambda^2}{\gamma\mu}(-\ln\frac{\Lambda}{\mu})$.

**Case 2:** $p_e = 0$ **and** $\lambda_e = \lambda_0 := \mu - \frac{w_s}{v} < \Lambda$, if either $\Lambda \ge \mu$ or $v < \hat{v}_0$ and $h \ge \hat{h}_{\lambda_0} := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \lambda_0}\right)\frac{\lambda_0^2}{\gamma\mu}(-\ln\frac{\lambda_0}{\mu})$.

**Case 3:** $p_e = 1$ **and** $\lambda_e = \Lambda$, if $\Lambda < \mu$, $v \ge \hat{v}_1 := \frac{w_s}{\mu - \Lambda} + h - \frac{w_s - w_r}{\mu - \Lambda}\frac{\Lambda}{\mu}\left(1 - ({}^{\Lambda}\!/_{\mu})^{\Lambda/\gamma}\right)$ and $h \le \check{h}_\Lambda := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \Lambda}\right)\frac{\Lambda}{\mu}\left(1 - ({}^{\Lambda}\!/_{\mu})^{\Lambda/\gamma}\right)$.

**Case 4:** $p_e = 1$ **and** $\lambda_e = \lambda_1 < \Lambda$, if either $\Lambda \ge \mu$ or $v < \hat{v}_1$ and $h \le \check{h}_{\lambda_1} := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \lambda_1}\right)\frac{\lambda_1}{\mu}\left(1 - ({}^{\lambda_1}\!/_{\mu})^{\lambda_1/\gamma}\right)$, where $\lambda_1$ is implicitly defined by $v = \frac{w_s}{\mu - \lambda_1} + h - \frac{w_s - w_r}{\mu - \lambda_1}\frac{\lambda_1}{\mu}\left(1 - ({}^{\lambda_1}\!/_{\mu})^{\lambda_1/\gamma}\right)$.

**Case 5:** $0 < p_e = \tilde{p} < 1$ **and** $\lambda_e = \Lambda$, if $\Lambda < \mu$, $v \geq \hat{v}_p := \frac{w_s}{\mu - \Lambda}\left(1 - (\Lambda/\mu)^2\left(1 - (\Lambda/\mu)^{\tilde{p}\Lambda/\gamma}\right)\right)$, and $\check{h}_\Lambda < h < \hat{h}_\Lambda$, where $\tilde{p}$ is implicitly defined by $h = \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \Lambda}\right)\frac{\Lambda}{\tilde{p}\mu}\left(1 - (\Lambda/\mu)^{\tilde{p}\Lambda/\gamma}\right)$.

**Case 6:** $0 < p_e = \tilde{p} < 1$ **and** $\lambda_e = \tilde{\lambda} < \Lambda$, if either $\Lambda > \mu$ or $v < \hat{v}_p$, and $\check{h}_{\lambda_1} < h < \hat{h}_{\lambda_0}$, where $(\tilde{p}, \tilde{\lambda})$ solve,

$$v = \frac{w_s}{\mu - \tilde{\lambda}}\left(1 - \left(\frac{\tilde{\lambda}}{\mu}\right)^2\left(1 - \left(\frac{\tilde{\lambda}}{\mu}\right)^{\frac{\tilde{p}\tilde{\lambda}}{\gamma}}\right)\right), \tag{15}$$

$$h = \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \tilde{\lambda}}\right)\frac{\tilde{\lambda}}{\tilde{p}\mu}\left(1 - \left(\frac{\tilde{\lambda}}{\mu}\right)^{\frac{p\tilde{\lambda}}{\gamma}}\right). \tag{16}$$

To prove part (ii.) we use the equilibrium condition equations defined in each case to derive the comparative statics results by taking the full derivatives of the equations.

### A.7. Sketch for Proof of Proposition 5:

To prove part (i.) we show that the partial derivative of the welfare function with respect to $\lambda$ is negative for all equilibrium arrival rates. To prove part (ii.) we show that, for a fixed exogenous arrival rate, the result as stated holds, in particular for the case when $\lambda = \lambda_{so}$. Then we use the fact that $\lambda_e > \lambda_{so}$ from part (i.) to show that this extends to the case when the arrival rates are different because, as more customers join in equilibrium, a smaller proportion agree to be flexible.

### A.8. Proof of Theorem 2:

The proof is almost identical to that of Theorem 1 and is omitted in the interest of brevity.

### A.9. Proof of Proposition 6:

For the first part of the proof we start from the fact that, when proactive service makes no errors (i.e., $q = 1$), the system with proactive service generates lower waiting times compared to the system without proactive service (i.e., $N_s(1) < \frac{\rho}{1-\rho}$, where $N_s(q)$ is given by the approximation of (2)). If $q = 0$, one can use the birth-death process of Figure 5 to show that waiting times will be approximately $N_s(0) = \frac{\rho}{1-\rho} + (n - p\lambda/\gamma) > \frac{\rho}{1-\rho}$. Therefore, to complete the first part of the proof, we will need to show that $\frac{dN_s(q)}{dq} < 0$ for all $0 < q < 1$. From the birth-death process depicted in Figure 5, we have that

$$P_0\left(\sum_{i=0}^{n}\left(\frac{\rho}{q}\right)^i + \left(\frac{\rho}{q}\right)^n\sum_{i=n+1}^{\infty}\rho^{i-n}\right) = 1,$$

and

$$N_s(q) = 1 - P_0 + P_0\left(\frac{\rho}{q}\right)^n\sum_{i=n+1}^{\infty}\rho^{i-n}(i - \frac{p\lambda}{\gamma} - 1).$$

Therefore, $\frac{q}{P_0}\frac{dP_0}{dq} = P_0\left(\sum_{i=0}^{n} i\left(\frac{\rho}{q}\right)^i + n\left(\frac{\rho}{q}\right)^n\sum_{i=n+1}^{\infty}\rho^{i-n}\right) > 0$. Furthermore, with some algebraic manipulation, $\frac{q}{P_0}\frac{dP_0}{dq} = n - P_0\sum_{i=0}^{n}(n-i)\left(\frac{\rho}{q}\right)^i$.

$$\frac{dN_s}{dq} = -\frac{dP_0}{dq} + \left(\frac{q}{P_0}\frac{dP_0}{dq} - n\right)\frac{P_0}{q}\left(\frac{\rho}{q}\right)^n\sum_{i=n+1}^{\infty}\rho^{i-n}(i - \frac{p\lambda}{\gamma} - 1)$$

$$= -\frac{dP_0}{dq} - \left( P_0 \sum_{i=0}^{n} (n-i) \left(\frac{\rho}{q}\right)^i \right) \frac{P_0}{q} \sum_{i=n+1}^{\infty} \rho^{i-n} (i - \frac{p\lambda}{\gamma} - 1) < 0.$$

Since $N_s$ is monotonic decreasing in $q$, this implies that there exists a threshold $0 < \bar{q} < 1$ such that $N_s(q) < \frac{\rho}{1-\rho}$ if and only if $q > \bar{q}$.

To show that if $\rho > \frac{1}{2} + \frac{\gamma}{2p\lambda}$ then $\bar{q} < \rho$, we start from the approximation for $N_s$ when $q = \rho$, given by (10). Using this approximation, the system with proactive service reduces waiting time despite errors if $\rho > \frac{n}{p\lambda/\gamma + n}$. Since $n := \lfloor \frac{p\lambda}{\gamma} + 1 \rfloor$, any $\rho > \frac{1}{2} + \frac{\gamma}{2p\lambda}$ would satisfy this.

For the last part of the proposition, we use the approximation of $N_s$ given by (8), which implies that $\bar{q}$ is the unique solution in the interval $(0,1)$ of the following polynomial equation: $q^{n+1} - q^n(1-\rho) - q\rho^n(n - p\lambda/\gamma + 1) + \rho^n(\rho(n - p\lambda/\gamma - 1) + 1) = 0$. We know the solution $\bar{q}$ exists and is unique from the first part of the proposition. As $\mu$ approaches $\lambda$ the system utilization $\rho$ increases but $n$ and $\lambda$ remain unchanged. Differentiating the polynomial equation with respect to $\rho$ gives $\frac{d\bar{q}}{d\rho} = -\frac{b(\bar{q},\rho)}{a(\bar{q},\rho)}$, where $a(q,\rho) = q^n \left(1 + n - n\frac{1-\rho}{q} - (2-\delta)(\frac{\rho}{q})^n\right)$, $b(q,\rho) = q^n \left(1 - \left(\frac{\rho}{q}\right)^n \left(n\frac{q}{\rho}(2 - \delta - 1/q) + \delta(n+1)\right)\right)$, where $\delta := p\lambda/\gamma + 1 - n$. Assume that no combination of parameters $(p, \lambda, \gamma, \mu)$ exist such that $\frac{d\bar{q}}{d\rho} < 0$, by counter example we arrive at a contradiction – let $p = 1$, $\lambda = .85$, $\gamma = .25$, and $\mu = 1$, then $\bar{q} = .422959$ and $\frac{d\bar{q}}{d\rho} = -0.3$. Thus there exist combinations of parameters $(p, \lambda, \gamma\mu)$ such that the threshold $\bar{q}$ is decreasing in utilization.

## References


Adan, I., J. Resing. 2002. *Queueing theory*. Eindhoven University of Technology, Eindhoven NL.

Aksin, Z., M. Armony, V. Mehrotra. 2007. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Allon, G., A. Bassamboo. 2011. The impact of delaying the delay announcements. *Operations Research* **59**(5) 1198–1210.

Armony, M., C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545.

Armony, M., C. Maglaras. 2004b. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *Operations Research* **52**(2) 271–292.

Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.

Borst, S., A. Mandelbaum, M. Reiman. 2004. Dimensioning large call centers. *Operations Research* **52**(1) 17–34.

Chen, H., D. Yao. 2013. *Fundamentals of queueing networks: performance, asymptotics, and optimization*, vol. 46. Springer-Verlag, New York NY, USA.

Cui, S., X. Su, S.K. Veeraraghavan. 2014. A model of rational retrials in queues. *Working Paper* .

De Lange, Robert, Ilya Samoilovich, Bo van der Rhee. 2013. Virtual queueing at airport security lanes. *European Journal of Operational Research* **225**(1) 153–165.

Edelson, N., D. Hilderbrand. 1975. Congestion tolls for poisson queueing processes. *Econometrica: Journal of the Econometric Society* 81–92.

Engel, R., R. Hassin. 2017. Customer equilibrium in a single-server system with virtual and system queues. *Queueing Systems* **87**(1-2) 161–180.

Gallego, G., Ö. Özer. 2001. Integrating replenishment decisions with advance demand information. *Management science* **47**(10) 1344–1360.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

Gans, N., Y. Zhou. 2002. Managing learning and turnover in employee staffing. *Operations Research* **50**(6) 991–1006.

Gans, N., Y. Zhou. 2003. A call-routing problem with service-level constraints. *Operations Research* **51**(2) 255–271.

Gans, N., Y. Zhou. 2007. Call-routing schemes for call-center outsourcing. *Manufacturing & Service Operations Management* **9**(1) 33–50.

Gurvich, I., J. Luedtke, T. Tezcan. 2010. Staffing call centers with uncertain demand forecasts: a chance-constrained optimization approach. *Management Science* **56**(7) 1093–1115.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3) 567–588.

Hassin, R. 2016. *Rational queueing*. CRC press, Boca Raton FL, USA.

Hassin, R., M. Haviv. 2003. *To queue or not to queue: equilibrium behavior in queueing systems*, vol. 59. Kluwer Academic Publishers, Norwell MA, USA.

Hassin, R., R. Roet-Green. 2011. Equilibrium in a two dimensional queueing game: When inspecting the queue is costly. Tech. rep., Citeseer.

Ibrahim, R., M. Armony, A. Bassamboo. 2016. Does the past predict the future? The case of delay announcements in service systems. *Management Science, Forthcoming* .

Jerath, K., A. Kumar, S. Netessine. 2015. An information stock model of customer behavior in multichannel customer support services. *Manufacturing & Service Operations Management* **17**(3) 368–383.

Jouini, O., Z. Akşin, Y. Dallery. 2011. Call centers with delay information: models and insights. *Manufacturing & Service Operations Management* **13**(4) 534–548.

Karaesmen, F., G. Liberopoulos, Y. Dallery. 2004. The value of advance demand information in production/inventory systems. *Annals of Operations Research* **126**(1-4) 135–157.

Kleinrock, L. 1976. *Queueing Systems: Theory*. No. v. 1 in A Wiley-Interscience publication, John Wiley and Sons, New York NY, USA.

Kostami, V., A. Ward. 2009. Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management* **11**(4) 644–656.

Legros, B., O. Jouini, G. Koole. 2015. Adaptive threshold policies for multi-channel call centers. *IIE Transactions* **47**(4) 414–430.

Levin, D., Y. Peres, E. Wilmer. 2009. *Markov chains and mixing times*. American Mathematical Society, Providence RI, USA.

Littlechild, SC. 1974. Optimal arrival rate in a simple queueing system. *International Journal of Production Research* **12**(3) 391–397.

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

Nageswaran, L., A. Scheller-Wolf. 2016. Queues with redundancy: is waiting in multiple lines fair? *Working Paper* .

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica: Journal of the Econometric Society* **37** 15–24.

Özer, Ö., W. Wei. 2004. Inventory control with limited capacity and advance demand information. *Operations Research* **52**(6) 988–1000.

Papier, F., U. Thonemann. 2010. Capacity rationing in stochastic rental systems with advance demand information. *Operations research* **58**(2) 274–288.

Reiman, M. 1984. Some diffusion approximations with state space collapse. *Modelling and performance evaluation methodology* 207–240.

Royal College of Obstetricians and Gynaecologists. 2008. Induction of labour - clinical guideline. Tech. rep., The National Institute for Health and Care Excellence. URL `https://www.nice.org.uk/guidance/cg70/evidence/full-guideline-241871149`. Last Accessed: 16/15/2017.

Shaked, M., G. Shanthikumar. 2007. *Stochastic orders*. Springer Science & Business Media, New York NY, USA.

Spencer, J., M. Sudan, K. Xu. 2014. Queueing with future information. *ACM SIGMETRICS Performance Evaluation Review* **41**(3) 40–42.

Tijms, H. 2003. *A first course in stochastic models*. John Wiley and Sons, Chichester, England.

Wang, T., B. Toktay. 2008. Inventory management with advance demand information and flexible delivery. *Management Science* **54**(4) 716–732.

Ward, A., P. Glynn. 2003. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems* **43**(1) 103–128.

Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* **15**(1) 88–102.

Xu, K., C. Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management* **18**(3) 314–331.

Xu, Kuang. 2015. Necessity of future information in admission control. *Operations Research* **63**(5) 1213–1226.

Zhang, S. 2014. Proactive serving decreases user delay exponentially. Ph.D. thesis, The Chinese University of Hong Kong.

# B. Online Appendix

## B.1. Performance of delay approximations compared to simulation

**B.1.1. Single-server Approximation** We simulate the single-server system with proactive service at 792 different parameter settings, specifically all combinations of $p \in \{.01, .1, .2, ..., .8, .9, .99\}, \lambda \in \{.45, .55, ..., .85, .95\}, \gamma \in \{.001, .01, .2, .4, ..., 1.8, 2\}$, and $\mu = 1$. Each combination is simulated for 100,000 units of time of which the first 20% is considered a warm-up period and is not included in the estimation of performance measures. Each combination is further run for 30 replications from which we compute simulation errors and 95% confidence intervals.

Figure 7 shows a comparison of the simulated and the approximated customer delays for a subset of the simulation parameters. The first row (Figures 7a–7c) shows that the accuracy of the approximation improves as utilization ($\lambda$) increases, for example, for $p = .5$ and $\gamma = 0.2$ when $\lambda = .45$, the % error of delays in the service queue for flexible customers (i.e., $\bar{T}_{rs}$) is -15.12%; when utilization increases and $\lambda = .95$, that decreases to 1.6%. The second row (Figures 7d–7f) shows the approximations for time spent in the service queue (i.e., $T_{ss}$ and $T_{rs}$) work well for any value of $p$, however the approximation overestimates time spent in orbit (i.e. $\bar{T}_{rr}$) if $p$ and $\gamma$ are relatively small. The third row (Figures 7g–7i) shows that the accuracy of the approximation deteriorates as $\gamma$ decreases.

To further understand the degradation of the accuracy of the approximation, especially at small values of $\gamma$, we run additional simulations. Specifically, for any value of $\lambda \in \{.45, .55, ..., .85, .95\}$, we run simulations where $\gamma$ is set such that $\frac{\mu-\lambda}{\gamma} \in \{.01, .2, .4, ..., 2, 2.2\}$. Figures 8a–8c show that the accuracy of the approximation (measured as relative error $\left(\frac{Approx.-Simulated}{Simulated}\right)$) deteriorates as $\frac{\mu-\lambda}{\gamma}$ increases. Furthermore, the figures demonstrate that if $\rho \geq .75$ and $\frac{\mu-\lambda}{\gamma} \leq 1$, the approximations for time spent in the service queue (i.e., $T_{ss}$ and $T_{rs}$) are quite accurate – the error is less than 5.6% – and the error for the time spent in orbit (i.e., $T_{rr}$) is no more than 13.8% (but we note that time spent in orbit is small by comparison to time spend in the service queue, making the relative error large).

**B.1.2. Multiserver Approximation** The comparison of approximations to simulated performance in the multiserver case is carried out to the same technical specifications as the single-server case with minor changes in the parameter space. In the multiserver case, utilization is now expressed as $\rho = \frac{\lambda}{m\mu}$. We fix $\mu = 1$ and vary $p \in \{.01, .1, .2, ..., .8, .9, .99\}, \rho \in \{.45, .55, ..., .85, .95\}$, $\gamma \in \{0.1, .03, .45, .6, .85, 1, 2, 3, 4, 5\}$, and $m \in \{1, 2..., 40\}$. Figure 9 shows that the performance of the approximation improves as the number of servers increase and, as in the single-server case, it also improves as information lead time decreases. Further, we again find that the accuracy of the approximation improves as utilization increases and deteriorates as information lead time grows large (i.e., $\gamma$ small). In all simulation instances where $\frac{\lambda}{\gamma \bar{N}_s^{mB}} \leq 1$ and $\rho \geq .75$, the approximations for $\bar{T}_{ss}$ and $\bar{T}_{rs}$ are within 3.86% and 4.31% of the respective simulated values. The approximation for $\bar{T}_{rr}$ can have a relative error as large as 206.85% even if $\frac{\lambda}{\gamma \bar{N}_s^{mB}} \leq 1$ and $\rho \geq .75$; however, this is due to the fact that $\bar{T}_{rr}$ can be small itself and we note that, if $\frac{\lambda}{\gamma \bar{N}_s^{mB}} \leq 1$ and $\rho \geq .75$,

**Figure 7** **Single-server Approximation Performance – Simulated values are shown as solid (red) lines with 95% confidence intervals, and approximations are shown as dashed (blue) lines.**

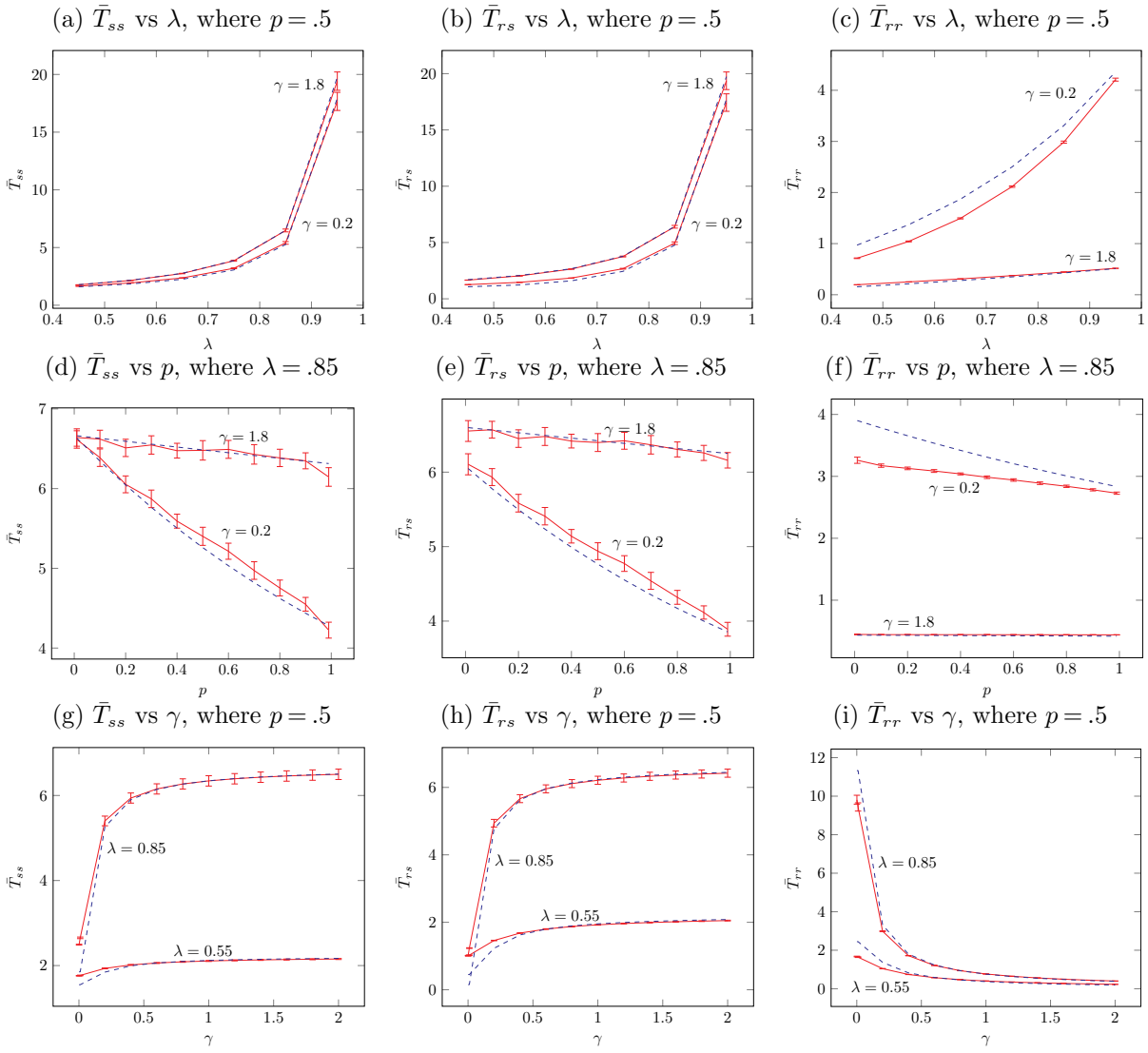



**Figure 8** **Single-server approximation: Simulation errors with 95% confidence intervals for $p = .5$.**

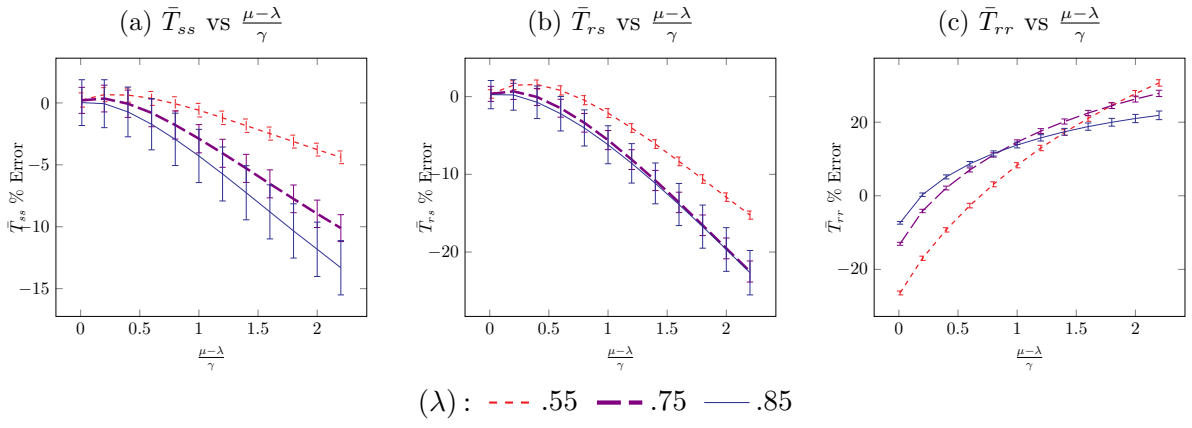
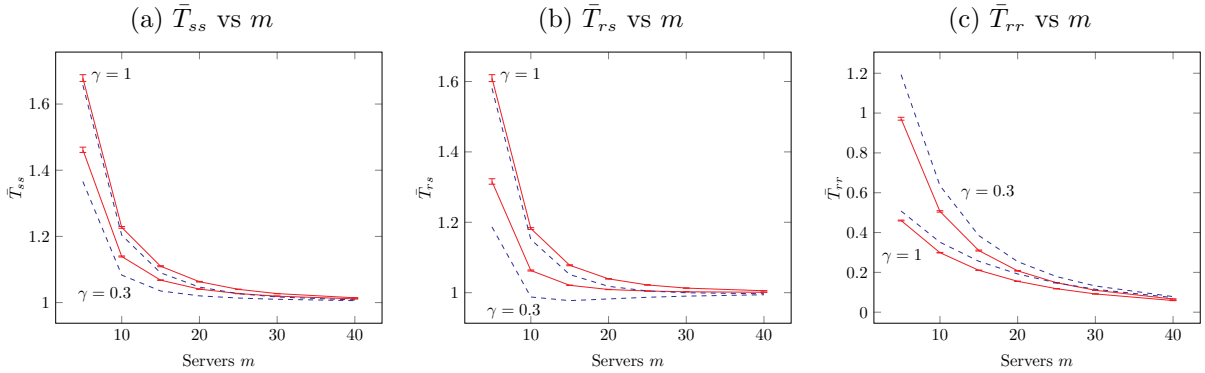


**Figure 9** **Multiserver Approximation Performance – Simulated values are shown as solid red lines with 95% confidence intervals, and approximations are shown as dashed (blue) lines for $p = .5$, $\rho = .85$.**




the average absolute error is 0.024 units of time (i.e., 2.4% of the expected service time).

## B.2. Proofs

### B.2.1. Detailed Proof of Proposition 3

We use two different coupling arguments for ordering in $p$ and $\gamma$; however, the proof ideas are similar, therefore we mainly focus on ordering in $p$ and explain the differences for $\gamma$ at the end. We prove the result in two steps: a) we first build the underlying coupled queue length processes in a common probability space such that the marginal distribution of each individual process is the same as the original processes, and b) we show that the orbit queue length processes are ordered as desired w.p.1 using an induction argument at the transition times in the coupled processes.

Before we present the details of coupling, note that given parameters $p, \lambda, \gamma$, and $\mu$, the state of a proactive service system at time $t$, denoted by $\mathbf{N}(t) := (N_r(t), N_s(t))$, is a continuous time Markov chain. Specifically if $N_r(t) = n_r$ and $N_s(t) = n_s$ the transition rates of this Markov chain with the associated event are given as follows;

$$\text{Arrival to orbit: To } (n_r + 1, n_s) \text{ with } p\lambda \tag{17}$$

$$\text{Arrival rate to service queue: To } (n_r, n_s + 1) \text{ with } (1-p)\lambda \tag{18}$$

$$\text{Orbit jump rate to service: To } (n_r - 1, n_s + 1) \text{ with } \gamma n_r \tag{19}$$

$$\text{Service completion: To } (n_r, n_s - 1) \text{ if } n_s - 1 \geq 1 \text{ or to } (n_r - 1, n_s - 1) \text{ if } n_s - 1 = 1, \text{ with } \mu \mathbb{1}\{N_s(t) > 0\}. \tag{20}$$

**Proof of ordering in $p$:** *Step I: (Coupling)* Consider two systems with proactive service as depicted in Figure 1 with orbit arrival probabilities $0 \leq p^{(1)} < p^{(2)} \leq 1$ respectively, with all other parameters (i.e., $\lambda, \gamma$, and $\mu$) equal. Our first objective is to construct two stochastic processes $\mathbf{N}^{(1)}$ and $\mathbf{N}^{(2)}$ in the same probability space using a coupling argument such that for $j = 1, 2$, $\mathbf{N}^{(j)}$ has the same transition rates as in (17)-(20) (given their parameters). This then implies that each of the individual occupancy processes has the same marginal distribution as that without a coupling.

Let $\ell = \{\ell_i, i = 1, 2, ..\}$ and $u = \{u_i, i = 1, 2, ..\}$ be independent sequences of i.i.d exponentially distributed random variables with rates $\lambda$ and $\mu$ respectively. Let $g^{(j)} = \{g_i^{(j)}, i = 1, 2, ..\}$, $j = 1, 2$ denote two independent i.i.d. sequences of exponentially distributed random variables both with rates $\gamma$. Let $b^{(1)} = \{b_i^{(1)}, i = 1, 2, ..\}$ denote a sequence of i.i.d. Bernoulli random variables with parameter $p^{(1)}$. Lastly, construct a second sequence (independent from $\ell, u$ and $g^{(j)}$) $b^{(2)} = \{b_i^{(2)}, i = 1, 2, ..\}$ such that if $b_i^{(1)} = 1$ then $b_i^{(2)} = 1$, but if $b_i^{(1)} = 0$ then $b_i^{(2)} = 1$ with probability $\frac{p^{(2)} - p^{(1)}}{1 - p^{(1)}}$, and $b_i^{(2)} = 0$ with probability $\frac{1 - p^{(2)}}{1 - p^{(1)}}$. Hence $b_i^{(2)} = 1$ with probability $p^{(2)}$ and 0 with probability $1 - p^{(2)}$.

We define the coupled processes at the transition times, $\tau_0, \tau_1, \ldots$ (how these times are chosen explained below) as follows. Consider the sequence of event times $\tau_k$ when the state changes (e.g., an arrival) and set $\tau_0 = 0$. The system state stays at the same level until the next event. Let $\delta_1 = (N_r^{(1)}(\tau_k) \wedge N_r^{(2)}(\tau_k))$ and $\delta_2 = (N_r^{(1)}(\tau_k) \vee N_r^{(2)}(\tau_k)) - \delta_1$ and

$$G_k^{(1)} = \frac{g_k^{(1)}}{\delta_1} \text{ and } G_k^{(2)} = \frac{g_k^{(2)}}{\delta_2}, \tag{21}$$

where by convention we take $./0 = \infty$. Also let

$$\tilde{\mu}_k = \frac{\mu_k}{\mathbb{1}\left\{N_s^{(1)}(\tau_k) + N_s^{(2)}(\tau_k) \geq 1\right\}}. \tag{22}$$

Now define

$$\Delta_k = \min\left\{\ell_k, G_k^{(1)}, G_k^{(2)}, \tilde{\mu}_k\right\}. \tag{23}$$

Based on which term is the minimizer we update the system state for $j = 1, 2$ as follows. Set $\tau_{k+1} = \tau_k + \Delta_k$.

- If $\Delta_k = \ell_k$ then for $j = 1, 2$:
  — if $N_s^{(j)}(\tau_k) \geq 1$ and $b_k^{(j)} = 1$, then set $N_r^{(j)}(\tau_{k+1}) = N_r^{(j)}(\tau_k) + 1$,
  — if $b_k^{(j)} = 0$ or if $N_s^{(j)}(\tau_k) = 0$ and $b_k^{(j)} = 1$, then set $N_s^{(j)}(\tau_{k+1}) = N_s^{(j)}(\tau_k) + 1$.
- If $\Delta_k = \tilde{\mu}_k$ then for $j = 1, 2$:
  — if $N_s^{(j)}(\tau_k) > 1$, then set $N_s^{(j)}(\tau_{k+1}) = N_s^{(j)}(\tau_k) - 1$
  — if $N_s^{(j)}(\tau_k) = 1$ and $N_r^{(j)}(\tau_k) > 0$, then set $N_r^{(j)}(\tau_{k+1}) = N_r^{(j)}(\tau_k) - 1$
  — if $N_s^{(j)}(\tau_k) = 0$, then no change in the state .
- If $\Delta_k = G_k^{(1)}$ then for $j = 1, 2$:
  — set $N_s^{(j)}(\tau_{k+1}) = N_s^{(j)}(\tau_k) + 1$ and $N_r^{(j)}(\tau_{k+1}) = N_r^{(j)}(\tau_k) - 1$.
- If $\Delta_k = G_k^{(2)}$,
  — if $N_r^{(1)}(\tau_k) > N_r^{(2)}(\tau_k)$, then set $N_s^{(1)}(\tau_{k+1}) = N_s^{(1)}(\tau_k) + 1$ and $N_r^{(1)}(\tau_{k+1}) = N_r^{(1)}(\tau_k) - 1$ and $N_s^{(2)}(\tau_{k+1}) = N_s^{(2)}(\tau_k)$ and $N_r^{(2)}(\tau_{k+1}) = N_r^{(2)}(\tau_k)$.
  — if $N_r^{(1)}(\tau_k) < N_r^{(2)}(\tau_k)$, then set $N_s^{(2)}(\tau_{k+1}) = N_s^{(2)}(\tau_k) + 1$ and $N_r^{(2)}(\tau_{k+1}) = N_r^{(2)}(\tau_k) - 1$ and $N_s^{(1)}(\tau_{k+1}) = N_s^{(1)}(\tau_k)$ and $N_r^{(1)}(\tau_{k+1}) = N_r^{(1)}(\tau_k)$.

We claim that $N^{(j)}$ is a Markov chain with the transition rates given in (17)-(20) with $p = p^{(j)}$. This follows from the standard uniformization argument for Markov processes

(see Section 6.7 in Ross probability models) and the fact that $\{N^{(j)}(\tau_k)\}$ is a discrete time Markov chain for $j=1,2$. In addition, because both chains have the interarrival and service times sequences:

$$N_r^{(1)}(t)+N_s^{(1)}(t)=N_r^{(2)}(t)+N_s^{(2)}(t),\ \forall t\geq 0. \tag{24}$$

*Step II: (Induction)* Now we use an induction argument and prove that $N_r^{(1)}(\tau_k)\leq N_r^{(2)}(\tau_k)$ for $k=1,2,\ldots$ when $p^{(1)}\leq p^{(2)}$. Assume without loss of generality, because both chains are positive recurrent, that $N^{(1)}(0)=N^{(2)}(0)$ .

*Induction base case.* Given $N_r(0)=N_r(0)$, by the coupling above we have $G_k^{(2)}=\infty$, so jumps from orbit are synchronized which implies $N_r^{(1)}(\tau_1)=N_r^{(2)}(\tau_1)$ if $\Delta_k=G_k^{(1)}$. Further service departures are also synchronized, thus any incidents of proactive service departures from orbit are synchronized, which implies $N_r^{(1)}(\tau_1)=N_r^{(2)}(\tau_1)$ if $\Delta_k=\tilde{\mu}_k$. Finally, arrival events are synchronized; however, they can differ on arrival location. Since $b_k^{(1)}\leq b_k^{(2)}$ it must be that $N_r^{(1)}(\tau_1)\leq N_r^{(2)}(\tau_1)$.

*Assume inductively* that $N_r^{(1)}(\tau_k)\leq N_r^{(2)}(\tau_k)$ is true for $k=1,2,..,n$ and we now show this inequality holds for $k=n+1$. If $N_r^{(1)}(\tau_n)=N_r^{(2)}(\tau_n)$ then by (24), $N_s^{(1)}(\tau_n)=N_s^{(2)}(\tau_n)$. Therefore $N_r^{(1)}(\tau_{n+1})\leq N_r^{(2)}(\tau_{n+1})$ by the same reasoning as in the base case. If $N_r^{(1)}(\tau_n)<N_r^{(2)}(\tau_n)$ then, because each process $N_r^{(j)}$ can change by at most 1 when any event occurs (note a departure from orbit and arrival cannot co-occur), then $N_r^{(1)}(\tau_{n+1})\leq N_r^{(2)}(\tau_{n+1})$. This completes the proof of ordering in $p$.

**Proof of ordering in $\gamma$:** To prove the ordering result in $\gamma$ we need to modify the coupling argument. Once the coupling is modified, the proof follows using the same argument in the proof of ordering in $p$. Specifically we need to alter the transition rates associated with jumps from the orbit as follows. Set $\gamma^{(1)}<\gamma^{(2)}$, all other parameters $p$, $\lambda$ and $\mu$, being equal. Let $\ell$ and $u$ be defined as above and $\{b_i, i=1,2,..\}$ denote a sequence of i.i.d Bernoulli random variables with parameter $p$. Finally we let $g^{(j)}=\{g_i^{(j)}, i=1,2,...\}$ denote a sequence of i.i.d. exponential random variables with parameter 1 for $j=1,2$.

Consider the sequence of event times $\tau_k$ when the state changes as above. Let $\delta_1=(\gamma^{(1)}N_r^{(1)}(\tau_k)\wedge\gamma^{(2)}N_r^{(2)}(\tau_k))$ and $\delta_2=(\gamma^{(1)}N_r^{(1)}(\tau_k)\vee\gamma^{(2)}N_r^{(2)}(\tau_k))-\delta_1$ and

$$G_k^{(1)}=\frac{g_k^{(1)}}{\delta_1}\text{ and }G_k^{(2)}=\frac{g_k^{(2)}}{\delta_2}. \tag{25}$$

With $\ell_k$, $\tilde{\mu}_k$ and $\Delta_k$ defined as in the proof of *part (i).* Note that both systems have the same probability that an arrival is an arrival to the orbit in this case, hence we do not need two different sequences $b^{(1)}$ and $b^{(2)}$. The transition rates are altered as follows.

- If $\Delta_k=G_k^{(1)}$ then for $j=1,2$:
  — set $N_s^{(j)}(\tau_{k+1})=N_s^{(j)}(\tau_k)+1$ and $N_r^{(j)}(\tau_{k+1})=N_r^{(j)}(\tau_k)-1$.
- If $\Delta_k=G_k^{(2)}$,
  — if $\gamma^{(1)}N_r^{(1)}(\tau_k)>\gamma^{(2)}N_r^{(2)}(\tau_k)$, then set $N_s^{(1)}(\tau_{k+1})=N_s^{(1)}(\tau_k)+1$ and $N_r^{(1)}(\tau_{k+1})=N_r^{(1)}(\tau_k)-1$ and $N_s^{(2)}(\tau_{k+1})=N_s^{(2)}(\tau_k)$ and $N_r^{(2)}(\tau_{k+1})=N_r^{(2)}(\tau_k)$.

—if $\gamma^{(1)}N_r^{(1)}(\tau_k) < \gamma^{(2)}N_r^{(2)}(\tau_k)$, then set $N_s^{(2)}(\tau_{k+1}) = N_s^{(2)}(\tau_k) + 1$ and $N_r^{(2)}(\tau_{k+1}) = N_r^{(2)}(\tau_k) - 1$ and $N_s^{(1)}(\tau_{k+1}) = N_s^{(1)}(\tau_k)$ and $N_r^{(1)}(\tau_{k+1}) = N_r^{(1)}(\tau_k)$.
Observing this coupling means that, when $N_r^{(1)}(\tau_k) = N_r^{(2)}(\tau_k)$ then $\gamma^{(1)}N_r^{(1)}(\tau_k) < \gamma^{(2)}N_r^{(2)}(\tau_k)$, thus $N_r^{(1)}$ cannot decrease below $N_r^{(2)}$ via a spontaneous jump from orbit. Furthermore, because (24) still holds, if $N_r^{(1)}(\tau_k) = N_r^{(2)}(\tau_k)$, then $N_s^{(1)}(\tau_k) = N_s^{(2)}(\tau_k)$, hence there cannot be a service completion and proactive service event such that $N_r^{(1)}$ decreases when $N_r^{(2)}$ does not. Proof of part (ii) follows from part (i) and (24).

**Proof of part (iii):** The monotonicity of $\bar{N}_r$ and $\bar{N}_s$ in $p$ and $\gamma$ given in Table 1 are immediately implied by parts $(i)$ and $(ii)$. That $\bar{T}_{rr}$ is decreasing in $\gamma$ then follows from the application of Little's Law to orbit (i.e., $\bar{N}_r = p\lambda\bar{T}_{rr}$) and that $\bar{N}_r$ is decreasing in $\gamma$. Similarly, that $\bar{T}_{ss}$ is increasing in $\gamma$ then follows from the relation between service queue occupancy and delays for inflexible customers (i.e., $\bar{T}_{ss} = (1/\mu)(\bar{N}_s + 1)$) and that $\bar{N}_s$ is increasing in $\gamma$. That $\bar{T}_{rs}$ is increasing in $\gamma$ then follows from $\bar{T}_{rs} = \bar{T}_{ss} - \frac{\mu-\lambda}{\mu}\bar{T}_{rr}$ and that $\bar{T}_{rr}$ and $\bar{T}_{ss}$ are respectively decreasing and increasing in $\gamma$. Lastly that $\bar{T}_{ss}$ is decreasing in $p$ then follows from the relation between service queue occupancy and delays for inflexible customers (i.e., $\bar{T}_{ss} = (1/\mu)(\bar{N}_s + 1)$) and that $\bar{N}_s$ is decreasing in $p$. □

**B.2.2. Proof of Theorem 1:** We begin by defining the necessary notation. Let $A_r^n(t)$ and $A_s^n(t)$ to be Poisson Processes with rates $p\lambda^n$ and $(1-p)\lambda^n$, respectively. For a given $n$, $A_r^n(t)$ and $A_s^n(t)$ represent the number of arrivals to the orbit and to the service queue, respectively, up to time $t$. Further let $C(t)$ and $S(t)$ be Poisson processes with rate one and let $D^n(t)$ denote the total number of customers who are served proactively (i.e., pulled from orbit) up to time $t$. Then:

$$N_r^n(t) = N_r^n(0) + A_r^n(t) - C\left(\gamma^n \int_0^t N_r^n(s)ds\right) - D^n(t), \tag{26}$$

$$N_s^n(t) = N_s^n(0) + A_s^n(t) + C\left(\gamma^n \int_0^t N_r^n(s)ds\right) + D^n(t) - S\left(\mu \int_0^t \mathbb{1}_{[N_s^n(s)>0]}ds\right) \tag{27}$$

The scaled occupancy processes can now be expressed as:

$$\hat{N}_r^n(t) := \frac{N_r^n(nt)}{\sqrt{n}} = \frac{N_r^n(0)}{\sqrt{n}} + \frac{A_r^n(nt)}{\sqrt{n}} - \frac{C\left(n\gamma\int_0^t \hat{N}_r^n(u)du\right)}{\sqrt{n}} - \frac{D^n(nt)}{\sqrt{n}}, \tag{28}$$

$$\hat{N}_s^n(t) := \frac{N_s^n(nt)}{\sqrt{n}} = \frac{N_s^n(0)}{\sqrt{n}} + \frac{A_s^n(nt)}{\sqrt{n}} + \frac{C\left(n\gamma\int_0^t \hat{N}_r^n(u)du\right)}{\sqrt{n}} + \frac{D^n(nt)}{\sqrt{n}} - \frac{S\left(n\mu\int_0^t \mathbb{1}_{[\hat{N}_s^n(u)>0]}du\right)}{\sqrt{n}}. \tag{29}$$

Assume that $\hat{N}_r^n(0) = \left(\hat{N}_Q^n(0) \wedge \frac{p\lambda^n}{\gamma}\right)$. We prove the result in two steps. We first show that (12) holds for any $\epsilon > 0$. Using an idea similar to Reiman (1984), assume that $\hat{N}_r^n(t) > p\lambda^n/\gamma + \epsilon$ for some $t \in [0,1]$ and let $T^n = \inf_{0<t\leq 1}\left\{t : \hat{N}_r^n(t) > p\lambda^n/\gamma + \epsilon\right\}$. Define $\tau^n = \sup_{0<t<T}\left\{t : \hat{N}_r^n(t) \leq p\lambda^n/\gamma + \epsilon/2\right\}$. Hence i) between times $\tau_n$ and $T_n$ that arrivals to orbit exceed the departures from orbit by $\epsilon/2$ as well as

ii) $\hat{N}_r^n(t) \geq {}^{p\lambda^n}/_{\gamma} + {}^{\epsilon}/_{2}$, $\forall\, t \in [\tau^n, T^n]$. Thus, defining $f(t_1;t_2) = f(t_2) - f(t_1)$ for a process $f$ with $t_1 < t_2$, we have that $P\left\{\sup_{0\leq t<1} \hat{N}_r^n(t) > \frac{p\lambda}{\gamma} + \epsilon\right\} \leq P\left\{\sup_{0\leq t_1\leq t_2<1} \frac{A_r^n(nt_1;nt_2)}{\sqrt{n}} - \frac{C\left(n\gamma\int_0^{t_2}\left(\frac{p\lambda^n}{\gamma}+\frac{\epsilon}{2}\right)du\right) - C\left(n\gamma\int_0^{t_1}\left(\frac{p\lambda^n}{\gamma}+\frac{\epsilon}{2}\right)du\right)}{\sqrt{n}} > \frac{\epsilon}{2}\right\}$.

Let

$$\tilde{A}_r^n(t_1;t_2) := \sqrt{n}\left(\frac{A_r^n(nt_1;nt_2)}{n} - p\lambda^n(t_2-t_1)\right), \text{ and}$$

$$\tilde{C}^n(t_1;t_2) := \sqrt{n}\left(\frac{C\left(n\gamma\int_0^{t_2}\left(\frac{p\lambda^n}{\gamma}+\frac{\epsilon}{2}\right)du\right) - C\left(n\gamma\int_0^{t_1}\left(\frac{p\lambda^n}{\gamma}+\frac{\epsilon}{2}\right)du\right)}{n} - \gamma(t_2-t_1)\left(\frac{p\lambda^n}{\gamma}+\frac{\epsilon}{2}\right)\right).$$

We have:

$$P\left\{\sup_{0\leq t_1\leq t_2<1} \frac{A_r^n(nt_1;nt_2)}{\sqrt{n}} - \frac{C\left(n\gamma\int_0^{t_2}\left(\frac{p\lambda^n}{\gamma}+\frac{\epsilon}{2}\right)du\right) - C\left(n\gamma\int_0^{t_1}\left(\frac{p\lambda^n}{\gamma}+\frac{\epsilon}{2}\right)du\right)}{\sqrt{n}} > \frac{\epsilon}{2}\right\} \leq$$

$$P\left\{\sup_{\substack{0\leq t_1\leq t_2<1\\ t_2-t_1\leq\delta}} \tilde{A}_r^n(t_1;t_2) - \tilde{C}^n(t_1;t_2) > \frac{\epsilon}{2}\right\} + P\left\{\sup_{\substack{0\leq t_1\leq t_2<1\\ t_2-t_1>\delta}} \tilde{A}_r^n(t_1;t_2) - \tilde{C}^n(t_1;t_2) - \sqrt{n}\frac{\gamma\epsilon(t_2-t_1)}{2} > \frac{\epsilon}{2}\right\}.$$

The process $\sqrt{n}(\frac{A_r^n(nt)}{n} - p\lambda^n t)$ converges weakly to a Brownian Motion with drift 0 and variance $p\lambda$ as $n$ goes to infinity by the FCLT. Similarly $\sqrt{n}(\frac{C^n(nt)}{n} - t)$ converges weakly to a standard Brownian Motion. Therefore by the continuous mapping theorem, the first term on the right-hand side is arbitrarily small for small $\delta$, and the second term goes to zero as $n$ goes to infinity because the term $\sqrt{n}\frac{\gamma\epsilon(t_2-t_1)}{2}$ goes to negative infinity as $n$ goes to infinity. This yields (12).

The next step is to show that,

$$P\left\{\sup_{0\leq t<1}\left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t)\wedge\frac{p\lambda^n}{\gamma}\right)\right| > \epsilon\right\} \to 0 \text{ as } n\to\infty. \tag{30}$$

Assume that there exists $t \in [0,1]$ such that $\left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t)\wedge\frac{p\lambda^n}{\gamma}\right)\right| > \epsilon$. Define $T_n = \inf_{0\leq t}\{t : \left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t)\wedge\frac{p\lambda^n}{\gamma}\right)\right| > \epsilon\}$ and $\tau^n = \sup_{0<t<T}\left\{t : \left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t)\wedge\frac{p\lambda^n}{\gamma}\right)\right| > {}^{\epsilon}/_{2}\right\}$. Hence $\left|\hat{N}_r^n(t) - \left(\hat{N}_Q^n(t)\wedge\frac{p\lambda^n}{\gamma}\right)\right| > {}^{\epsilon}/_{2}$, $\forall\, t \in [\tau_n, T_n]$. Also assume that:

$$\sup_{0\leq t\leq 1} \hat{N}_r^n(t) < \frac{p\lambda^n}{\gamma} + \epsilon/4. \tag{31}$$

We claim that (31) implies:

$$\hat{N}_r^n(t) < {}^{p\lambda^n}/_{\gamma} - {}^{\epsilon}/_{2}, \text{ and } \hat{N}_s^n(t) > \frac{\epsilon}{2}, \ \forall t \in [\tau_n, T_n]. \tag{32}$$

To prove (32) first suppose that for some $t \in [\tau_n, T_n]$

$$\hat{N}_r^n(t) - \left( \hat{N}_Q^n(t) \wedge \frac{p\lambda^n}{\gamma} \right) > \epsilon/2. \tag{33}$$

If $\hat{N}_Q^n(t) < \frac{p\lambda^n}{\gamma}$, then (33) would imply $\hat{N}_r^n(t) > \hat{N}_Q^n(t)$ which by construction cannot occur. On the other hand, if $\frac{p\lambda^n}{\gamma} < \hat{N}_Q^n(t)$ then (33) contradicts (31). Hence (33) cannot hold if (31) holds. Therefore for all $t \in [\tau^n, T^n]$

$$\left( \hat{N}_Q^n(t) \wedge \frac{p\lambda^n}{\gamma} \right) - \hat{N}_r^n(t) > \epsilon/2.$$

This clearly implies (32).

Now (32) implies that during $[\tau_n, T_n]$ the server queue (excluding those in service) is never empty and so no customers are served proactively, i.e., $D(nT_n) - D(n\tau_n) = 0$. Then by (28), (29), and (32),

$$P\left\{ \sup_{0\le t<1} \left| \hat{N}_r^n(t) - \left( \hat{N}_Q^n(t) \wedge \frac{p\lambda^n}{\gamma} \right) \right| > \epsilon \right\} \le$$
$$P\left\{ \sup_{0\le t_1\le t_2<1} \frac{A_s^n(nt_1; nt_2)}{\sqrt{n}} + \frac{C\left(n\gamma \int_0^{t_2} \left( \frac{p\lambda^n}{\gamma} - \frac{\epsilon}{2} \right) ds \right) - C\left(n\gamma \int_0^{t_1} \left( \frac{p\lambda^n}{\gamma} - \frac{\epsilon}{2} \right) ds \right)}{\sqrt{n}} \right.$$
$$\left. - \frac{S(n\mu t_2) - S(n\mu t_1)}{\sqrt{n}} > \epsilon \right\} + P\left\{ \sup_{0\le t<1} \hat{N}_r^n(t) > \frac{p\lambda}{\gamma} + \epsilon/4 \right\}.$$

By similar arguments we used in proving (12), and by (12) the right-hand side goes to zero as $n \to \infty$ proving the result. □

### B.3. Proof of Proposition 4

**B.3.1. Part i.** We begin by establishing the preliminary result that if an equilibrium $(p_e, \lambda_e)$ exists then $\frac{\Lambda}{\mu} \ge .75$, $v \ge 4\frac{w_s}{\mu}$ $\gamma \ge \frac{w_s}{v}$ imply that $\frac{\lambda_e}{\mu} \ge .75$ and $\gamma \ge \mu - \lambda_e$. Consider the case where customers only choose between balking or joining and being inflexible, then $\underline{\lambda} := \min\{\mu - \frac{w_s}{v}, \Lambda\}$ is the equilibrium arrival rate, see (Hassin and Haviv 2003, Chapter 3, Section 1.1) for the details. Given this, a little algebra shows that, in this case, if $\frac{\Lambda}{\mu} \ge .75$, then $v \ge 4\frac{w_s}{\mu}$ implies $\underline{\lambda} \ge .75\mu$ and $\gamma \ge \frac{w_s}{v}$ implies $\underline{\lambda} \ge \mu - \gamma$. Therefore, to establish this preliminary result, we must show that no fewer customers will join the system when customers can also join and be flexible in addition to the options to balk or join and be inflexible.

In the case when joining and being flexible is dominated, i.e. $p_e = 0$, then $\lambda_e = \underline{\lambda}$ and $v - ws\bar{T}_{ss}(0, \underline{\lambda}) \ge 0$ where the inequality holds strictly only in cases where $\underline{\lambda} = \Lambda$. In the case when joining and being flexible is not dominated, i.e. $p_e > 0$, we show $\lambda_e > \underline{\lambda}$ by contradiction. Assume by contradiction that for some $\lambda_e < \underline{\lambda}$ and $p_e \in (0, 1]$ is an equilibrium. Then by $\lambda < \underline{\lambda}$ we have that $v - w_s\bar{T}_{ss}(0, \lambda) > v - w_s\bar{T}_{ss}(0, \underline{\lambda}) \ge 0$ and by Proposition 3, specifically that $\bar{T}_{ss}$ is decreasing in $p$, we have that $v - w_s\bar{T}_{ss}(p, \lambda) > v - w_s\bar{T}_{ss}(0, \lambda) \ge 0$.

Therefore, relative to the option to balk, the option to join and be inflexible has strictly higher utility than in the case when no-one is flexible and customers join at rate $\lambda$. This cannot be an equilibrium as customers have incentive to deviate, thus we have the contradiction we seek and it cannot be that strictly fewer customers join in equilibrium when customers are given the option to join and be flexible. Going forward we have that, for any potential equilibrium arrival rate $\lambda_e$, it is such that $\lambda_e \geq .75\mu$ and $\lambda_e \geq \mu - \gamma$.

There are six possible types of equilibrium strategies which are the combinations of $\lambda_e < \Lambda$ or $\lambda_e = \Lambda$ with $p_e = 0$ or $0 < p_e < 1$ or $p_e = 1$. We show that each type of equilibrium corresponds to a given region of the parameter space in $v$ and $h$ which can be expressed in terms the other model primitives $\Lambda$, $\gamma$, $\mu$, $w_s$, and $w_r$. To prove the uniqueness and existence of equilibrium, we show that the regions are mutually exclusive and collectively exhaustive. The cases (unique equilibrium solution and region) are:

**Case 1:$p_e = 0$ and $\lambda_e = \Lambda$.** For this to be an equilibrium, it must be that $\Lambda < \mu$ so that the system is stable if all customers join, $v \geq w_s \bar{T}_{ss}(0, \Lambda)$ so that customers have no incentive to balk, and $h + w_r \bar{T}_{rr}(0, \Lambda) + w_s \bar{T}_{rs}(0, \Lambda) \geq w_s \bar{T}_{ss}(0, \Lambda)$ so that customers have no incentive to be flexible. These conditions are respectively equivalent to $\Lambda < \mu$, $v \geq \hat{v}_0 := \frac{w_s}{\mu - \Lambda}$ and $h \geq \hat{h}_\Lambda := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \Lambda}\right) \frac{\Lambda^2}{\gamma\mu}(-\ln \frac{\Lambda}{\mu})$, which provides the region where $p_e = 0$ and $\lambda_e = \Lambda$ is an equilibrium strategy.

**Case 2: $p_e = 0$ and $\lambda_e = \lambda_0 := \mu - \frac{w_s}{v} < \Lambda$.** Where $\lambda_0$ is such that $0 = v - w_s \bar{T}_{ss}(0, \lambda_0)$ so that customers are indifferent between balking, and joining and being inflexible. In terms of model primitives, this indifference can be expressed as $v = \frac{w_s}{\mu - \lambda_0}$. For this to be an equilibrium it must be either that $\Lambda \geq \mu$ or $v < w_s \bar{T}_{ss}(0, \Lambda)$ so that if all customers joined there would be incentive for some to balk, and $h + w_r \bar{T}_{rr}(0, \lambda_0) + w_s \bar{T}_{rs}(0, \lambda_0) \geq w_s \bar{T}_{ss}(0, \lambda_0)$ so that customers have no incentive to be flexible. These conditions are respectively equivalent to either $\Lambda \geq \mu$ or $v < \hat{v}_0$ and $h \geq \hat{h}_{\lambda_0} := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \lambda_0}\right) \frac{\lambda_0^2}{\gamma\mu}(-\ln \frac{\lambda_0}{\mu})$, which provides the region where $p_e = 0$ and $\lambda_e = \lambda_0$ is an equilibrium strategy.

**Case 3: $p_e = 1$ and $\lambda_e = \Lambda$.** For this to be an equilibrium it must be that $\Lambda < \mu$ so that the system is stable if all customers join, $h + w_r \bar{T}_{rr}(1, \Lambda) + w_s \bar{T}_{rs}(1, \Lambda) \leq v$ so that customers have no incentive to balk when all customers join and are flexible, and $w_s \bar{T}_{ss}(1, \Lambda) \geq h + w_r \bar{T}_{rr}(1, \Lambda) + w_s \bar{T}_{rs}(1, \Lambda)$ so that customers have no incentive to be inflexible. These conditions are respectively equivalent to $\Lambda < \mu$, $v \geq \hat{v}_1 := \frac{w_s}{\mu - \Lambda} + h - \frac{w_s - w_r}{\mu - \Lambda} \frac{\Lambda}{\mu} \left(1 - (\Lambda/\mu)^{\Lambda/\gamma}\right)$ and $h \leq \check{h}_\Lambda := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \Lambda}\right) \frac{\Lambda}{\mu} \left(1 - (\Lambda/\mu)^{\Lambda/\gamma}\right)$, which provides the region where $p_e = 1$ and $\lambda_e = \Lambda$ is an equilibrium strategy.

**Case 4: $p_e = 1$ and $\lambda_e = \lambda_1 < \Lambda$.** Where $\lambda_1$ (see below for proof of existence and uniqueness) is such that $0 = v - h + w_r \bar{T}_{rr}(1, \lambda_1) + w_s \bar{T}_{rs}(1, \lambda_1)$ so that customers are indifferent between balking, and joining and being flexible. In terms of model primitives, this indifference can be expressed as $v = \frac{w_s}{\mu - \lambda_1} + h - \frac{w_s - w_r}{\mu - \lambda_1} \frac{\lambda_1}{\mu} \left(1 - (\lambda_1/\mu)^{\lambda_1/\gamma}\right)$. For this to be an equilibrium it must be either that $\Lambda \geq \mu$ or $v < h + w_r \bar{T}_{rr}(1, \Lambda) + w_s \bar{T}_{rs}(1, \Lambda)$, so that if all customers join and are flexible there would be incentive for some to balk, and $h + w_r \bar{T}_{rr}(1, \lambda_1) + w_s \bar{T}_{rs}(1, \lambda_1) \leq w_s \bar{T}_{ss}(1, \lambda_1)$ so that customers have no incentive to be inflexi-

ble. These conditions are respectively equivalent to either $\Lambda \geq \mu$ or $v < \hat{v}_1$ and $h \leq \check{h}_{\lambda_1} := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu-\lambda_1}\right)\frac{\lambda_1}{\mu}\left(1 - \left({}^{\lambda_1}\!/_{\mu}\right)^{\lambda_1/\gamma}\right)$, which provides the region where $p_e = 1$ and $\lambda_e = \lambda_1$ is an equilibrium strategy.

**Existence and uniqueness of $\lambda_1$.** To see that $\lambda_1$ exists and is unique, note that for any $\lambda < \mu$,

$$\frac{d}{d\lambda}\left[\frac{w_s}{\mu-\lambda} + h - \frac{w_s - w_r}{\mu-\lambda}\frac{\lambda}{\mu}\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{\lambda}{\gamma}}\right)\right] =$$
$$\frac{w_s}{(\mu-\lambda)^2} - (w_s - w_r)\left(\frac{1}{(\mu-\lambda)^2}\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{\lambda}{\gamma}}\right) - \frac{\lambda}{\mu(\mu-\lambda)}\left(\left(\frac{\lambda}{\mu}\right)^{\frac{\lambda}{\gamma}}\frac{1}{\gamma}\left(1 + \ln\frac{\lambda}{\mu}\right)\right)\right),$$

and that ${}^{\lambda}\!/_{\mu} > e^{-1} \approx .368$ is sufficient for this to be positive, which we have by assumption. Therefore, if $v \leq \check{v}_1$, then $\lambda_1$ must exist, and the monotonicity ensures $\lambda_1$ is unique when it exists.

**Case 5: $0 < p_e = \tilde{p} < 1$ and $\lambda_e = \Lambda$.** Where $\tilde{p}$ (see below for proof of existence and uniqueness) is such that $v - w_s\bar{T}_{ss}(p,\Lambda) = v - h + w_r\bar{T}_{rr}(p,\Lambda) + w_s\bar{T}_{rs}(p,\Lambda)$ so that customers are indifferent between joining and choosing to be inflexible vs being flexible when all customers join. In terms of model primitives, this indifference can be expressed as $h = \left(\frac{w_s}{\mu} - \frac{w_r}{\mu-\Lambda}\right)\frac{\Lambda}{\tilde{p}\mu}\left(1 - \left({}^{\Lambda}\!/_{\mu}\right)^{\tilde{p}\Lambda/\gamma}\right)$. For this to be an equilibrium it must be that $\Lambda < \mu$ and $v \geq w_s\bar{T}_{ss}(\tilde{p},\Lambda)$ so that customers have no incentive to balk when everyone joins and a proportion $\tilde{p} \in (0,1)$ are flexible, and that $w_s\bar{T}_{ss}(p,\Lambda) = h + w_r\bar{T}_{rr}(p,\Lambda) + w_s\bar{T}_{rs}(p,\Lambda)$ so customers are indifferent between choosing to be flexible and inflexible. The region where $p_e = \tilde{p}$ and $\lambda_e = \Lambda$ is an equilibrium strategy, is when $\Lambda < \mu$, $v \geq \hat{v}_p := \frac{w_s}{\mu-\Lambda}\left(1 - \left({}^{\Lambda}\!/_{\mu}\right)^2\left(1 - \left({}^{\Lambda}\!/_{\mu}\right)^{\tilde{p}\Lambda/\gamma}\right)\right)$, and $\check{h}_\Lambda < h < \hat{h}_\Lambda$, where this last pair of inequalities is derived in the proof of existence and uniqueness below.

**Existence and uniqueness of $\tilde{p}$.** Such a $\tilde{p}$ will exist (and be unique) if $h \in [\hat{h}_\Lambda, \check{h}_\Lambda]$ because $\left(\frac{w_s}{\mu} - \frac{w_r}{\mu-\Lambda}\right)\frac{\Lambda}{p\mu}\left(1 - \left({}^{\Lambda}\!/_{\mu}\right)^{p\Lambda/\gamma}\right)$ is decreasing in $p$ (see B.3.5), from $\hat{h}_\Lambda$ (when $p = 0$) to $\check{h}_\Lambda$ (when $p = 1$).

**Case 6: $0 < p_e = \tilde{p} < 1$ and $\lambda_e = \tilde{\lambda} < \Lambda$.** Where $(\tilde{p}, \tilde{\lambda})$ (see below for proof of existence and uniqueness) are the $(p,\lambda)$ such that $0 = v - w_s\bar{T}_{ss}(p,\lambda)$ and $0 = v - (h + w_r\bar{T}_{rr}(p,\lambda) + w_s\bar{T}_{rs}(p,\lambda))$ so that customers are indifferent between balking, joining and being flexible, and joining and and being inflexible. In terms of model primitives, this can be expressed as:

$$v = \frac{w_s}{\mu - \tilde{\lambda}}\left(1 - \left(\frac{\tilde{\lambda}}{\mu}\right)^2\left(1 - \left(\frac{\tilde{\lambda}}{\mu}\right)^{\frac{\tilde{p}\tilde{\lambda}}{\gamma}}\right)\right), \tag{34}$$

$$h = \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \tilde{\lambda}}\right)\frac{\tilde{\lambda}}{\tilde{p}\mu}\left(1 - \left(\frac{\tilde{\lambda}}{\mu}\right)^{\frac{p\tilde{\lambda}}{\gamma}}\right). \tag{35}$$

The region where $p_e = \tilde{p}$ and $\lambda_e = \tilde{\lambda}$ is an equilibrium strategy, is when either $\Lambda > \mu$ or $v < \hat{v}_p$, and $\check{h}_{\lambda_1} < h < \hat{h}_{\lambda_0}$, where these inequalities are derived in the proof of existence and uniqueness below.

**Existence and uniqueness of $(\tilde{p}, \tilde{\lambda})$.** To see that this system of equations has, at most, one solution when $\lambda/\mu \geq .75$ and $\gamma \geq (\mu - \lambda)$, assume by contradiction that there exist two such solutions $(p, \lambda)$ and $(p', \lambda')$ such that $p \neq \tilde{p}$ and $\lambda \neq \tilde{\lambda}$; then,

$$\frac{w_s}{\mu - \lambda}\left(1 - \left(\frac{\lambda}{\mu}\right)^2\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right) = \frac{w_s}{\mu - \lambda'}\left(1 - \left(\frac{\lambda'}{\mu}\right)^2\left(1 - \left(\frac{\lambda'}{\mu}\right)^{\frac{p'\lambda'}{\gamma}}\right)\right), \tag{36}$$

$$\left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \lambda}\right)\frac{\lambda}{p\mu}\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right) = \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \lambda'}\right)\frac{\lambda'}{p'\mu}\left(1 - \left(\frac{\lambda'}{\mu}\right)^{\frac{p'\lambda'}{\gamma}}\right). \tag{37}$$

Without loss of generality assume $\lambda < \lambda'$. Now we will come to a contradiction that (36) implies that $p' > p$ while (37) implies that $p' < p$. To see that (36) implies that $p' > p$, observe that the right-hand side of (34) is increasing in $\lambda$ and decreasing in $p$ (see B.3.2 and B.3.3). Therefore, since $\lambda < \lambda'$, for (36) to hold it must be that $p' > p$. To see that (37) implies that $p' < p$, observe that when $\frac{\lambda}{\mu} > .75$ and $\gamma > \mu - \lambda$ the right-hand side of equation (35) is decreasing in $\lambda$ and decreasing in $p$ (see B.3.4 and B.3.5). Therefore, since $\lambda < \lambda'$, for (37) to hold it must be that $p' < p$. Hence we have contradiction, and therefore, at most one solution exists.

To see the inequalities $v < \hat{v}_p$ and $\check{h}_{\lambda_1} < h < \hat{h}_{\lambda_0}$ imply that a solution exists, observe that $v < \hat{v}_p$ implies that, if everyone joins and customers are indifferent between choosing to be flexible and inflexible, then customers have incentive to balk. Since the right-hand side of (34) is increasing in $\lambda$, for any $p \in (0, 1)$ there exists a unique $\lambda_p < \Lambda$ such that $v = \frac{w_s}{\mu - \lambda_p}\left(1 - (\lambda_p/\mu)^2\left(1 - (\lambda_p/\mu)^{p\lambda_p/\gamma}\right)\right)$. Note, this implies $\lambda_p$ is increasing in $p$ because the right-hand side is decreasing in $p$. Lastly, observing that $\left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \lambda_p}\right)\frac{\lambda_p}{p\mu}\left(1 - (\lambda_p/\mu)^{p\lambda_p/\gamma}\right)$ is decreasing in both $p$ and $\lambda_p$ from $\hat{h}_{\lambda_0}$ (when $p = 0$ and $\lambda = \lambda_0$) to $\check{h}_{\lambda_1}$ (when $p = 1$ and $\lambda = \lambda_1$), the fact that $\check{h}_{\lambda_1} < h < \hat{h}_{\lambda_0}$ ensures the existence of a unique solution for the double $(\tilde{p}, \tilde{\lambda})$.

**Cases 1-6 are mutually exclusive and collectively exhaustive.** By the fact that the regions given in cases 1-6 are collectively exhaustive in the space of $v$ and $h$, it must be that at least one case applies. By the fact that the regions given in cases 1-6 are also mutually exclusive, it must be that one, and only one, case applies. Therefore, by the fact that each case has a unique potential equilibrium, and that one, and only one case applies, we have that the equilibrium exists and is unique.

**The supporting monotonicity results for Parts i and ii.**

### B.3.2. Monotonic non-decreasing behavior of the RHS of (34) in $\lambda$

$$\frac{d}{d\lambda}\left[\frac{w_s}{\mu-\lambda}\left(1-\left(\frac{\lambda}{\mu}\right)^2\left(1-\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right)\right] = w_s\frac{d}{d\lambda}\left[\frac{\mu+\lambda}{\mu^2}+\frac{1}{\mu-\lambda}\left(\frac{\lambda}{\mu}\right)^2\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right] \tag{38}$$
$$= w_s\left[\frac{1}{\mu^2}+\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\left(\frac{\lambda(2\mu-\lambda)}{\mu^2(\mu-\lambda)^2}+\frac{\lambda^2}{\mu^2(\mu-\lambda)}\left[\frac{p}{\gamma}\left(1+\ln\frac{\lambda}{\mu}\right)\right]\right)\right]$$

A sufficient condition for this to be positive is $1+\ln\frac{\lambda}{\mu}>0$ or $\frac{\lambda}{\mu}>e^{-1}\approx.368$, which we have by assumption.

### B.3.3. Monotonic non-increasing behavior of the RHS of (34) in $p$

$$\frac{d}{dp}\left[\frac{w_s}{\mu-\lambda}\left(1-\left(\frac{\lambda}{\mu}\right)^2\left(1-\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right)\right] = \frac{w_s}{\mu-\lambda}\left(\frac{\lambda}{\mu}\right)^2\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\frac{\lambda}{\gamma}\left(\ln\frac{\lambda}{\mu}\right) \tag{39}$$

Which is negative because $\left(\ln\frac{\lambda}{\mu}\right)<0$.

### B.3.4. Monotonic non-increasing behavior of the RHS of (35) in $\lambda$

$$\frac{d}{d\lambda}\left[\left(\frac{w_s}{\mu}-\frac{w_r}{\mu-\lambda}\right)\frac{\lambda}{p\mu}\left(1-\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right] =$$
$$-\frac{w_r}{(\mu-\lambda)^2}\frac{\lambda}{p\mu}\left(1-\rho^{\frac{p\lambda}{\gamma}}\right)+\left(\frac{w_s}{\mu}-\frac{w_r}{\mu-\lambda}\right)\frac{1}{p\mu}\left[1-\rho^{\frac{p\lambda}{\gamma}}-\frac{p\lambda}{\gamma}\rho^{\frac{p\lambda}{\gamma}}\left(\ln\rho+1\right)\right] \tag{40}$$

Note, we are only interested in the case where (35) holds, therefore we consider the case where $\left(\frac{w_s}{\mu}-\frac{w_r}{\mu-\lambda}\right)>0$. The first term of the derivative is non-positive and a sufficient (but not necessary) condition for the second term to also be non-positive is that the sub-term in square brackets is non-positive. The term in square brackets is non-positive if $\frac{\gamma}{p\lambda}\left(\rho^{-\frac{p\lambda}{\gamma}}-1\right)\leq\ln\rho+1$. Note $\frac{\gamma}{p\lambda}\left(\rho^{-\frac{p\lambda}{\gamma}}-1\right)$ is decreasing in $\gamma$ and increasing in $p$, the relevant derivatives are:

$$\frac{d}{d\gamma}\left[\frac{\gamma}{p\lambda}\left(\rho^{-\frac{p\lambda}{\gamma}}-1\right)\right]=\frac{1}{p\lambda}\left[\rho^{-\frac{p\lambda}{\gamma}}\left(1+\frac{p\lambda}{\gamma}\ln\rho\right)-1\right],$$

and

$$\frac{d}{dp}\left[\frac{\gamma}{p\lambda}\left(\rho^{-\frac{p\lambda}{\gamma}}-1\right)\right]=-\frac{\gamma}{p^2\mu}\left[\rho^{-\frac{p\lambda}{\gamma}}\left(1+\frac{p\lambda}{\gamma}\ln\rho\right)-1\right].$$

To establish the sign of these derivatives we need to establish the sign of the term in brackets. The term in brackets is non-positive if $\ln\rho^{\frac{p\lambda}{\gamma}}\leq\rho^{\frac{p\lambda}{\gamma}}-1$, or equivalently $\ln x\leq x-1$, which is true for all $x\in[0,1]$. Therefore, in the domain $p\leq1$ and ${}^{(\mu-\lambda)}/_{\gamma}\leq1$, the left-hand side is maximized at $p=1$ and $\gamma=\mu-\lambda$.

Given this, we establish that $\left[1-\rho^{\frac{p\lambda}{\gamma}}-\frac{p\lambda}{\gamma}\rho^{\frac{p\lambda}{\gamma}}\left(\ln\rho+1\right)\right]$ is non-positive if $\frac{1-\rho}{\rho}\left(\rho^{-\frac{\rho}{1-\rho}}-1\right)\leq\ln\rho+1$, which is true for all $\rho\in(.676,1)$, yielding the result.

**B.3.5. Monotonic non-increasing behavior of the RHS of (35) in $p$**

$$\frac{d}{dp}\left[\left(\frac{w_s}{\mu}-\frac{w_r}{\mu-\lambda}\right)\frac{\lambda}{p\mu}\left(1-\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right]=-\left(\frac{w_s}{\mu}-\frac{w_r}{\mu-\lambda}\right)\frac{1}{p^2}\frac{\lambda}{\mu}\left[1-\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}+\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\ln\left(\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right] \tag{41}$$

Note, we are only interested in the case where (35) holds, therefore we consider the case where $\left(\frac{w_s}{\mu}-\frac{w_r}{\mu-\lambda}\right)>0$. Substituting $x=(\lambda/\mu)^{p\lambda/\gamma}$ into the term in brackets, this term is positive because $1-x+x\ln x\geq 0$ for all $x\in[0,1]$. Therefore, (41) is negative.

**B.3.6. Part ii.** Note that in cases 1 and 3, $p_e$ and $\lambda_e$ are fixed, thus it suffices to check cases 2, 4, 5, and 6. For notational convenience let the right-hand side of (34) be denoted as $\alpha$, the right-hand side of (35) be denoted as $\beta$, and the right-hand side of $v=\left[\frac{w_s}{\mu-\lambda_e}+h-\frac{w_s-w_r}{\mu-\lambda_e}\frac{\lambda_e}{\mu}\left(1-\left(\frac{\lambda_e}{\mu}\right)^{\frac{\lambda_e}{\gamma}}\right)\right]$ be denoted as $\delta$. Note that $\frac{\partial\alpha}{\partial\lambda},\frac{\partial\delta}{\partial\lambda}$ are positive, by B.3.2 and the analysis in case 4 used to show $\lambda_1$ is unique, respectively. Also, $\frac{\partial\alpha}{\partial p},\frac{\partial\beta}{\partial\lambda},\frac{\partial\beta}{\partial p}$ are all negative by B.3.3,B.3.4 and B.3.5, respectively.

**B.3.7. Part ii.a <u>Comparative statics with respect to $h$.</u>**

**Case 2:** $p_e=0$ and $\lambda_e<\Lambda$. Then, $v=\alpha$ which is independent of $h$ when $p=0$, thus $\frac{d\lambda}{dh}$ equals zero.

**Case 4:** $p_e=1$ and $\lambda_e<\Lambda$. Then, $v=\delta$, and taking the derivative with respect to $h$ yields, $0=1+\frac{\partial\delta}{\partial\lambda}\frac{d\lambda}{dh}$. This implies that $\frac{d\lambda}{dh}$ is negative because $\frac{\partial\delta}{\partial\lambda}$ is positive.

**Case 5:** $p_e=\in(0,1)$ and $\lambda_e=\Lambda$. Then $h=\beta$, and taking the derivative with respect to $h$ yields, $1=\frac{\partial\beta}{\partial h}+\frac{\partial\beta}{\partial p}\frac{dp}{dh}$. This implies that $\frac{dp}{dh}$ is negative because $\frac{\partial\beta}{\partial h}$ equals zero, and $\frac{\partial\beta}{\partial p}$ is negative.

**Case 6:** $p_e=\in(0,1)$ and $\lambda_e<\Lambda$. Then $v=\alpha$ and $h=\beta$, and taking the derivative with respect to $h$ yields,

$$0=\frac{\partial\alpha}{\partial h}+\frac{\partial\alpha}{\partial\lambda}\frac{d\lambda}{dh}+\frac{\partial\alpha}{\partial p}\frac{dp}{dh}, \tag{42}$$

$$1=\frac{\partial\beta}{\partial h}+\frac{\partial\beta}{\partial\lambda}\frac{d\lambda}{dh}+\frac{\partial\beta}{\partial p}\frac{dp}{dh}. \tag{43}$$

Equation (42) implies that $\frac{d\lambda}{dh}$ and $\frac{dp}{dh}$ are of the same sign (both are positive or both are negative) because $\frac{\partial\alpha}{\partial h}$ equals zero, so $\frac{\partial\alpha}{\partial\lambda}\frac{d\lambda}{dh}=-\frac{\partial\alpha}{\partial p}\frac{dp}{dh}$, and $\frac{\partial\alpha}{\partial p}$ is negative, and $\frac{\partial\alpha}{\partial\lambda}$ is positive. Equation (43) implies both $\frac{dp}{dh}$ and $\frac{d\lambda}{dh}$ must be negative because $\frac{\partial\beta}{\partial h}=0$ and both $\frac{\partial\beta}{\partial\lambda}$ and $\frac{\partial\beta}{\partial p}$ are negative.

**<u>Comparative statics with respect to $w_r$.</u>**

**Case 2:** $p_e=0$ and $\lambda_e<\Lambda$. Then, $v=\alpha$ which is independent of $w_r$ when $p=0$, thus $\frac{d\lambda}{dw_r}$ equals zero.

**Case 4:** $p_e=1$ and $\lambda_e<\Lambda$. Then, $v=\delta$, and taking the derivative with respect to $w_r$ yields, $0=\frac{\partial\delta}{\partial w_r}+\frac{\partial\delta}{\partial\lambda}\frac{d\lambda}{dw_r}$. This implies that $\frac{d\lambda}{dw_r}$ is negative because $\frac{\partial\delta}{\partial w_r}$ and $\frac{\partial\delta}{\partial\lambda}$ are positive.

**Case 5:** $p_e\in(0,1)$ and $\lambda_e=\Lambda$. Then $h=\beta$, and taking the derivative with respect to $w_r$ yields, $0=\frac{\partial\beta}{\partial w_r}+\frac{\partial\beta}{\partial p}\frac{dp}{dw_r}$. This implies that $\frac{dp}{dw_r}$ is negative because $\frac{\partial\beta}{\partial wr}$ and $\frac{\partial\beta}{\partial p}$ are negative.

**Case 6:** $p_e \in (0,1)$ and $\lambda_e < \Lambda$. Then $v = \alpha$ and $h = \beta$, and taking the derivative with respect to $w_r$ yields,

$$0 = \frac{\partial \alpha}{\partial w_r} + \frac{\partial \alpha}{\partial \lambda}\frac{d\lambda}{dw_r} + \frac{\partial \alpha}{\partial p}\frac{dp}{dw_r}, \tag{44}$$

$$0 = \frac{\partial \beta}{\partial w_r} + \frac{\partial \beta}{\partial \lambda}\frac{d\lambda}{dw_r} + \frac{\partial \beta}{\partial p}\frac{dp}{dw_r}. \tag{45}$$

Equation (44) implies that $\frac{d\lambda}{dw_r}$ and $\frac{dp}{dw_r}$ are of the same sign (both are positive or both are negative) because $\frac{\partial \alpha}{\partial w_r}$ equals zero, so $\frac{\partial \alpha}{\partial \lambda}\frac{d\lambda}{dw_r} = -\frac{\partial \alpha}{\partial p}\frac{dp}{dw_r}$, and $\frac{\partial \alpha}{\partial \lambda}$ is positive and $\frac{\partial \alpha}{\partial p}$ is negative. Equation (45) implies both $\frac{dp}{dw_r}$ and $\frac{d\lambda}{dw_r}$ must be negative because $\frac{\partial \beta}{\partial w_r}$, $\frac{\partial \beta}{\partial \lambda}$, and $\frac{\partial \beta}{\partial p}$ are all negative.

### B.3.8. Part ii.b Comparative statics with respect to $v$.

**Case 2:** $p_e = 0$ and $\lambda_e < \Lambda$. Then, $v = \alpha$, and taking the derivative with respect to $v$ yields $1 = \frac{\partial \alpha}{\partial v} + \frac{\partial \alpha}{\partial \lambda}\frac{d\lambda}{dv}$. This implies $\frac{d\lambda}{dv}$ is positive because $\frac{\partial \alpha}{\partial \lambda}$ is positive and $\frac{\partial \alpha}{\partial v}$ equals zero.

**Case 4:** $p_e = 1$ and $\lambda_e < \Lambda$. Then, $v = \delta$, and taking the derivative with respect to $v$ yields $1 = \frac{\partial \delta}{\partial v} + \frac{\partial \delta}{\partial \lambda}\frac{d\lambda}{dv}$. This implies that $\frac{d\lambda}{dv}$ is positive because $\frac{\partial \delta}{\partial \lambda}$ is positive (as shown in analysis of case 4) and $\frac{\partial \delta}{\partial v}$ equals zero.

**Case 5:** $p_e = \in (0,1)$ and $\lambda_e = \Lambda$. Then $h = \beta$, and taking the derivative with respect to $v$ yields $0 = \frac{\partial \beta}{\partial v} + \frac{\partial \beta}{\partial p}\frac{dp}{dv}$. This implies that $\frac{dp}{dv} = 0$ because $\frac{\partial \beta}{\partial v} = 0$ and $\frac{\partial \beta}{\partial p} < 0$.

**Case 6:** $p_e = \in (0,1)$ and $\lambda_e < \Lambda$. Then $v = \alpha$ and $h = \beta$, and taking the derivative with respect to $v$ yields,

$$1 = \frac{\partial \alpha}{\partial v} + \frac{\partial \alpha}{\partial \lambda}\frac{d\lambda}{dv} + \frac{\partial \alpha}{\partial p}\frac{dp}{dv}, \tag{46}$$

$$0 = \frac{\partial \beta}{\partial v} + \frac{\partial \beta}{\partial \lambda}\frac{d\lambda}{dv} + \frac{\partial \beta}{\partial p}\frac{dp}{dv}. \tag{47}$$

Noting that $\frac{\partial \alpha}{\partial v} = \frac{\partial \beta}{\partial v} = 0$, solving (47) for $\frac{d\lambda}{dv}$ and substituting into the first equation yields:

$$1 = \left( -\frac{\partial \alpha}{\partial \lambda}\frac{\frac{\partial \beta}{\partial p}}{\frac{\partial \beta}{\partial \lambda}} + \frac{\partial \alpha}{\partial p} \right)\frac{dp}{dv}. \tag{48}$$

Since $\frac{\partial \alpha}{\partial \lambda}$ is positive, $\frac{\partial \beta}{\partial p}/\frac{\partial \beta}{\partial \lambda}$ is positive ( negative divided by a negative), and $\frac{\partial \alpha}{\partial p}$ is negative, it must be that $\frac{dp}{dv}$ is negative. Given that $\frac{dp}{dv}$ is negative, (46) then implies that $\frac{d\lambda}{dv}$ is positive because $\frac{d\beta}{dv}$ equals zero, $\frac{\partial \beta}{\partial p}\frac{dp}{dv}$ is positive, and $\frac{\partial \beta}{\partial \lambda}$ is negative.

### B.3.9. Part ii.c Comparative statics with respect to $w_s$.

**Case 2:** $p_e = 0$ and $\lambda_e < \Lambda$. Then, $v = \alpha$, and taking the derivative with respect to $w_s$ yields $0 = \frac{\partial \alpha}{\partial w_s} + \frac{\partial \alpha}{\partial \lambda}\frac{d\lambda}{dw_s}$. This implies that $\frac{d\lambda}{dw_s}$ is negative because, $\frac{\partial \alpha}{\partial w_s}$ and $\frac{\partial \alpha}{\partial \lambda}$ are positive.

**Case 4:** $p_e = 1$ and $\lambda_e < \Lambda$. Then, $v = \delta$, and taking the derivative with respect to $w_s$ yields, $0 = \frac{\partial \delta}{\partial w_s} + \frac{\partial \delta}{\partial \lambda}\frac{d\lambda}{dw_s}$. This implies that $\frac{d\lambda}{dw_s}$ is negative because $\frac{\partial \delta}{\partial w_s}$ and $\frac{\partial \delta}{\partial \lambda}$ are positive.

**Case 5:** $p_e \in (0,1)$ and $\lambda_e = \Lambda$. Then $h = \beta$, and taking the derivative with respect to $w_s$ yields, $0 = \frac{\partial \beta}{\partial w_s} + \frac{\partial \beta}{\partial p}\frac{dp}{dw_s}$. This implies that $\frac{dp}{dw_s}$ is positive because $\frac{\partial \beta}{\partial w_s}$ is positive and $\frac{\partial \beta}{\partial p}$ are negative.

**Case 6:** $p_e \in (0,1)$ and $\lambda_e < \Lambda$. Then $v = \alpha$ and $h = \beta$, and taking the derivative with respect to $w_s$ yields,

$$0 = \frac{\partial \alpha}{\partial w_s} + \frac{\partial \alpha}{\partial \lambda}\frac{d\lambda}{dw_s} + \frac{\partial \alpha}{\partial p}\frac{dp}{dw_s}, \tag{49}$$

$$0 = \frac{\partial \beta}{\partial w_s} + \frac{\partial \beta}{\partial \lambda}\frac{d\lambda}{dw_s} + \frac{\partial \beta}{\partial p}\frac{dp}{dw_s}. \tag{50}$$

Solving (49) for $\frac{dp}{dw_s}$ and substituting into (50) yields,

$$\frac{\partial \beta}{\partial p}\frac{\frac{\partial \alpha}{\partial w_s}}{\frac{\partial \alpha}{\partial p}} - \frac{\partial \beta}{\partial w_s} = \left(\frac{\partial \beta}{\partial \lambda} - \frac{\partial \beta}{\partial p}\frac{\frac{\partial \alpha}{\partial \lambda}}{\frac{\partial \alpha}{\partial p}}\right)\frac{d\lambda}{dw_s}. \tag{51}$$

The term in parenthesis on the right-hand side is negative (negative minus a positive) because $\frac{\partial \beta}{\partial \lambda}$, $\frac{\partial \beta}{\partial p}$, and $\frac{\partial \alpha}{\partial p}$ are negative and $\frac{\partial \alpha}{\partial \lambda}$ is positive. Hence, given the sign of the left-hand side, then $\frac{d\lambda}{dw_s}$ has the oppositive sign.

The left-hand side (positive minus a positive) is a negative if, $\frac{\partial \beta}{\partial w_s} > \frac{\partial \beta}{\partial p}\frac{\frac{\partial \alpha}{\partial w_s}}{\frac{\partial \alpha}{\partial p}}$, which expressed in terms of model primitives is equivalent to:

$$\frac{1}{\mu}\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right) \geq \left(\frac{1}{\mu} - \frac{w_r}{w_s(\mu - \lambda)}\right)\left[1 - \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}} + \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}} \ln\left(\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right] \frac{\left(1 - \left(\frac{\lambda}{\mu}\right)^2\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right)}{\left(\frac{\lambda}{\mu}\right)^2\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\left(-\ln\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)}. \tag{52}$$

Observe that, if this inequality holds for the case when $w_r = 0$, it is true for all $w_r > 0$ because the right-hand side is decreasing in $w_r$. To see this note that all the terms multiplied by $w_r$ are positive, to see the term in brackets is positive, let $x = (\lambda/\mu)^{p\lambda/\gamma}$ and note that $1 - x + x\ln x > 0,\ \forall x \in [0,1]$. Hence, letting $w_r = 0$ and substituting $x$ in for notational convenience the inequality reduces to,

$$(\lambda/\mu)^2 \geq \frac{1 - x(1 - \ln x)}{(1-x)(1 - x\ln x)} \tag{53}$$

The right-hand side is decreasing in $x$, therefore if this inequality holds when $p = 1$ and $\gamma = \mu - \lambda$ it holds for all $p \in [0,1]$ and all $\gamma > \mu - \lambda$. Hence, letting $x = (\lambda/\mu)^{\frac{p\lambda}{\gamma}} \to (\lambda/\mu)^{\frac{\lambda}{\mu-\lambda}} = \rho^{\frac{\rho}{1-\rho}}$ we have the result when

$$\rho^2 \geq \frac{1 - \rho^{\frac{\rho}{1-\rho}}(1 - \ln \rho^{\frac{\rho}{1-\rho}})}{(1 - \rho^{\frac{\rho}{1-\rho}})(1 - \rho^{\frac{\rho}{1-\rho}}\ln \rho^{\frac{\rho}{1-\rho}})}. \tag{54}$$

A sufficient condition for this inequality to hold is $\rho > .5$ which we have by assumption. Hence $\frac{d\lambda}{dw_s} > 0$.

Given $\frac{d\lambda}{dw_s} > 0$, (49) implies that $\frac{dp}{dw_s}$ is positive. To see this observe $\left(\frac{\partial \alpha}{\partial w_s}\right)$ is positive, $\left(\frac{\partial \alpha}{\partial \lambda}\frac{d\lambda}{dw_s}\right)$ is the product of two positives, therefore the third term must be negative, and since $\frac{\partial \alpha}{\partial p}$ is negative, it must be that $\frac{dp}{dw_s}$ is positive.

### B.4. Proof of Proposition 5

We first show that $\frac{\Lambda}{\mu} \in [.75, 1)$, $v > 16^{w_s}/\mu$, $\gamma \geq \sqrt{\mu w_s/v}$ imply that $\lambda_{so} > .75\mu$, and $\lambda_{so} \geq \mu - \gamma$, which we use in the proof below. First, if no customers can be flexible ($p = 0$), the socially optimal arrival rate is $\lambda^0_{so} := \min\{\Lambda, \mu - \sqrt{\frac{w_s\mu}{v}}\}$ (see (Hassin and Haviv 2003, Chapter 3, Section 1.2) for the details). Given this, a little algebra shows that, in this case, if $\Lambda\mu \geq .75$ then, $v > 16^{w_s}/\mu$ implies $\lambda^0_{so} \geq 0.75\mu$ and $\gamma \geq \sqrt{\mu w_s/v}$ implies $\lambda^0_{so} \geq \mu - \gamma$. Therefore, as long as a central planner, given the option to dictate customer flexibility $p \in [0, 1]$ in addition to the arrival rate $\lambda$, would not dictate fewer customers join than in the case when $p$ is restricted to zero, we have the preliminary result. This is obvious as proactive service enables a provider to reduce delays while serving the same amount of customers (because delays are decreasing in $p$), thus, given the option to dictate flexibility, it can never be optimal to serve fewer customers as it would be dominated by the case where the same number of customers are served as the benchmark case and some positive proportion of customers are flexible. Therefore, for any socially optimal proportion of flexible customers and socially optimal arrival rate $(p_{so}, \lambda_{so})$ we have that $\lambda_{so} > .75\mu$, and $\lambda_{so} \geq \mu - \gamma$.

**B.4.1. Part 1:** We next prove that $\lambda_{so} \leq \lambda_e$. We do it by showing that $\lambda_{so} \leq \lambda_0 \leq \lambda_e$, where $\lambda_0 = \min\{\Lambda, \mu - \frac{w_s}{v}\}$ is the equilibrium arrival rate when the proportion of flexible customers is fixed at zero, and the second inequality follows from the fact that the equilibrium arrival rate ($\lambda_e$) is no smaller than $\lambda_0$ (see the preliminary result in the proof of Proposition 4). Note in the case where $\lambda_e = \lambda_0 = \Lambda$, the inequality is trivially satisfied, therefore we will focus on the case where $\lambda_0 = \mu - \frac{w_s}{v} < \Lambda$. To prove the first inequality we show that, for any fixed $p$, the partial derivative of the welfare function (7) with respect to $\lambda$ (given below), is negative for all $\lambda \geq \lambda_0$.

$$\frac{\partial}{\partial \lambda} W(p,\lambda) = \overbrace{v - \frac{w_s}{\mu - \lambda}}^{A} - \overbrace{\left( ph + \frac{w_r}{(\mu-\lambda)^2}\frac{\lambda(2\mu-\lambda)}{\mu}\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right) + \frac{w_s - w_r}{\mu - \lambda}\frac{\lambda}{\mu}\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\frac{p\lambda}{\gamma}\left(1 + \ln\frac{\lambda}{\mu}\right)\right)}^{B} \tag{55}$$

$$\underbrace{- \frac{w_s\lambda}{(\mu-\lambda)^2}\left[1 - \frac{(2\mu-\lambda)}{\mu}\left(1 - \left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\right)\right]}_{C} \tag{56}$$

To show that this derivative is negative if $\lambda \geq \lambda_0$ we show that A is non-positive, and B is non-negative (-B is non-positive) and C is positive (-C is negative). To see that A is non-positive observe that $v - \frac{w_s}{\mu - \lambda_0} = 0$, therefore for $\lambda \geq \lambda_0$ this term is non-positive. To

see that B is non-negative, note the first two terms within B are trivially non-negative and that $\lambda/\mu > e^{-1}$ is a sufficient condition for the third term to be non-negative which we have by assumption, therefore B is non-negative. To see that $-C$ is negative, observe that this is true when the term within the square brackets is positive. Noting that the term in brackets is decreasing in $p$, letting $p=1$ minimizes this term and therefore, as long as it is positive when $p=1$, it is positive for all $p \in [0,1]$. Further noting that term in brackets is increasing in $\gamma$ then letting $\gamma = \mu - \lambda$ minimizes this term (note from the assumption that $\gamma \geq \sqrt{\mu w_s/v}$, and $v > w_s/\mu$ we have that $\gamma \geq \sqrt{\mu w_s/v} \geq \sqrt{w_s^2/v^2} = w_s/v = \mu - \lambda_0$, therefore for all $\lambda \geq \lambda_0$ we have that $\gamma \geq \mu - \lambda$). With these substitutions for $p$ and $\gamma$, the term in brackets is positive if $1 - (2-\rho)\left(1 - \rho^{\frac{\rho}{1-\rho}}\right)$ is positive. Observing this is true for all $\rho \in [0,1]$ we have the result that $\lambda_{so} \leq \lambda_e$.

**B.4.2. Part 2:** Now we show that $p_{so} \geq p_e$. Fix $\lambda > 0$. The partial derivative of the welfare function (given by equation 7) with respect to $p$ is

$$\frac{\partial}{\partial p} W(p,\lambda) = -\lambda(h - \xi(p,\lambda)) \tag{57}$$

where

$$\xi_s(p,\lambda) = \frac{w_s - w_r}{\mu - \lambda}\left(\frac{\lambda^2}{\gamma\mu}\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\left(-\ln\frac{\lambda}{\mu}\right)\right), \tag{58}$$

The second partial derivative of $W(p,\lambda)$ with respect to $p$ is $-\frac{w_s - w_r}{\mu-\lambda}\left(\frac{\lambda^4}{\gamma^2\mu}\left(\frac{\lambda}{\mu}\right)^{\frac{p\lambda}{\gamma}}\left(\ln\frac{\lambda}{\mu}\right)^2\right)$, which is non-positive by the assumption that $w_r \leq w_s$. Therefore, for fixed $\lambda \in (0,\Lambda]$, the welfare function is concave with respect to $p$. Let $p_{so}(\lambda)$ denote the proportion of flexible customers that maximizes the welfare function when the arrival rate is $\lambda$.

By concavity of the welfare function, for fixed $\lambda$ if (57) is positive when $p=1$, i.e.,

$$h \leq \xi_s(1,\lambda) \tag{59}$$

then $p_{so}(\lambda) = 1$, i.e., it is social optimal for everyone to be flexible.

Now consider unregulated customer equilibrium where $c_r = c_s = 0$ and $\lambda$ is fixed. Taking the difference in the utility of inflexible and flexible customers (given by $v - c_s - w_s\bar{T}_{ss}$ and $v - c_r - h - w_r\bar{T}_{rr} - w_s\bar{T}_{rs}$ respectively), we have (after some algebra) that if

$$h > \xi_e(p,\lambda) := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu-\lambda}\right)\frac{1}{\mu-\lambda}\bar{T}_{rr}(p,\lambda), \tag{60}$$

then the utility of inflexible customers is greater than that of flexible customers. Since $\bar{T}_{rr}$ is decreasing in the proportion of flexible customers $p$ (which follows from the approximation given in (5)), if $h \geq \xi_e(0,\lambda)$, then choosing to be flexible is a dominated strategy. Substituting the approximation of $\bar{T}_{rr}$ given by (5) into 60 and applying L'Hôpital's rule, flexibility is dominated in equilibrium (for fixed $\lambda$) if,

$$h \geq \xi_e(0,\lambda) \tag{61}$$

where

$$\xi_e(0,\lambda) = \left(\frac{w_s}{\mu} - \frac{w_r}{\mu-\lambda}\right)\frac{\lambda^2}{\gamma\mu}\left(-\ln\frac{\lambda}{\mu}\right). \tag{62}$$

We prove below that if $\gamma \geq \mu - \lambda$ and $\frac{\lambda}{\mu} \geq .75$,

$$\xi_s(1,\lambda) \geq \xi_e(0,\lambda). \tag{63}$$

This together with the fact that $\xi_e$ is decreasing in $\lambda$ (see B.3.4) and the fact that $\lambda_{so} < \lambda_e$ (from Part 1 above), implies that

$$\xi_s(1,\lambda_{so}) \geq \xi_e(0,\lambda_{so}) \geq \xi_e(0,\lambda_e). \tag{64}$$

Then the desired result follows by setting $\underline{h} := \xi_e(0,\lambda_e)$ and $\bar{h} := \xi_s(1,\lambda_{so})$.

To complete the proof, we now prove (63). We note that (63) is (after some algebra) equivalent to

$$\rho^{\frac{\lambda}{\gamma}} > \frac{w_s(1-\rho) - w_r}{w_s - w_r}. \tag{65}$$

Because LHS of this inequality is increasing in $\gamma$,

$$\rho^{\frac{\lambda}{\gamma}} \geq \rho^{\frac{\lambda}{\mu-\lambda}} \tag{66}$$

under the assumption $\gamma \geq \mu - \lambda$. Additionally, the RHS is decreasing in $w_r$, hence

$$\frac{w_s(1-\rho) - w_r}{w_s - w_r} \leq \frac{w_s(1-\rho)}{w_s} \tag{67}$$

Therefore (65) holds if $\rho + \rho^{\frac{\rho}{1-\rho}} > 1$ and the latter holds for all $\rho > .5$. Since we assume that $\lambda/\mu \geq 0.75$, we have (63).

## C. Application to Induction of Labor

This Appendix presents a numerical illustration of the model of proactive service for the case of induction of labor, a medical procedure performed at large UK-based maternity hospital, that motivated this work. The management of the hospital wanted to tackle delays in induction of labor (IOL), where childbirth is pharmacologically initiated in specialized beds on the antenatal (pre-birth) ward. This procedure takes 12–36 hours to complete and is medically indicated for overdue or higher-risk pregnancies. Although some patients in need of emergency induction arrive to the ward with little advance warning, for many patients, IOL is booked by community midwives anytime between one to seven days in advance. Demand for the procedure is highly variable and so are service times, leading to significant delays in starting the process – in some cases patients needed to wait up to 3–4 days. Such delays are not only unpleasant for the patients but also increase the risk of medical complications. One way to tackle these unpleasant waiting times would be to

increase capacity (e.g., beds and staffing), at least at times when demand is high. Nevertheless, financial, human resource, and space constraints made this approach infeasible. Instead, the maternity hospital contemplated an alternative approach: call patients in to undergo IOL proactively when there were available beds. Bringing the procedure forward by 1-2 days is considered medically safe for patients who are overdue (Royal College of Obstetricians and Gynaecologists 2008, pg.8). Since a large proportion of the patients live within a short distance of their planned birth hospital, it was also considered practical, especially if the expecting mothers were told to be prepared for the event.

We calibrate model parameters from hospital data to estimate the potential benefits of proactive service in this setting. The hospital has $m = 5$ beds specially equipped and reserved for IOL patients. The arrival rate for induction patients is $\lambda = 3.97$ per day. Precise data (e.g., time stamps) is not available for waiting times and service times, but on average, patients spend 2 days in the antenatal ward. Imputation using $M/M/m$ queueing formulas suggests that the average service time would be approximately $\mu^{-1} = 25.66$ hours[8]. The average bed utilization is 85%, which is consistent with a busy hospital unit. We take the average information lead time to be $\gamma^{-1} = 1$ days as the hospital does not want to bring forward patients' procedures by too much. From data we estimate that the probability that the delivery takes place naturally (i.e., without induction) on the next day conditional on the expecting mother being in or beyond the 39th week of pregnancy to be 8.3%. Therefore, we estimate the performance of the system, including the extension where customers who are served proactively may not actually have needed the service, i.e., the case of imperfect information with $q = .917$. Note that this means the approximation for $\bar{T}_{rs}$ is the same as that for $\bar{T}_{ss}$ (see §5.2).

**Figure 10** **Delays for Induction of Labor where** $\lambda = 3.97, \mu = .935, \gamma = 1, m = 5, q = 0.917$

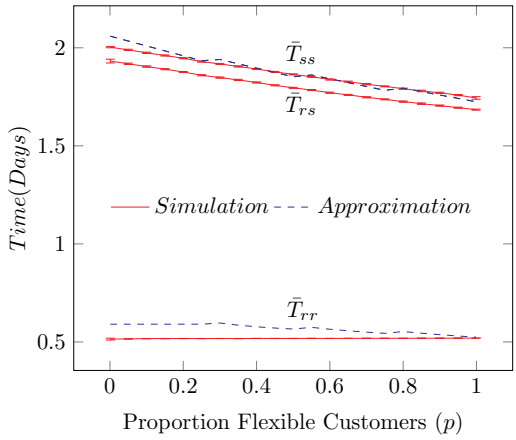



Figure 10 illustrates the operational benefits in terms of reduction in time spent on the

[8] In practice, IOL starts at specific points in time (usually at 8am and 3pm). We abstracted from this starting-time batching as it would further complicate the analysis. Due to this, our results may be viewed as an upper bound on the value of proactive service.

antenatal ward for varying levels of adoption of proactive service ($p$). The average time patients spend on the service queue in the benchmark case without proactive service is 2 days (of which approximately .93 days is waiting and 1.07 days is service). If all customers were flexible, the length of stay at the antenatal unit is reduced to 1.74 days and the reduction to delays ahead of service is 28%. Furthermore, we note that the results based on the multiserver approximation modified to include proactive service closely match those of a simulation model. Finally, we note that, although the model is highly stylized and these numbers are likely to overestimate the benefit of proactive service (e.g., the model assumes that the hospital would could call expecting mothers to be induced even late at night), these results suggest that it may be an effective way to reduce delays without requiring additional capacity[9].

We next investigate whether patients would be willing to participate in proactive service. For the purposes of this application we assume the demand rate is exogenous to delays, a realistic assumption in this setting. Normalizing the cost of time spent on the antenatal ward to one unit per day (i.e., $w_s = 1$) and assuming that waiting in orbit, which in this case is equivalent to waiting at home, is costless (i.e., $w_r = 0$), then the equilibrium strategy is for all patients to adopt if the fixed cost of flexibility $h < .0596$ (i.e., less than the cost of about 85 minutes waiting on the unit), and no patient will adopt if $h > .0701$ (i.e., more than the cost of 100 minutes waiting on the unit). In contrast, the central planner would have dictated that all patients adopt if $h < .25$. These findings suggest that it may well be the case that the maternity hospital is in the inefficient region where all patients would benefit if they agreed to be flexible, but fewer patients than optimal actually do so. These results suggest that the management of the hospital should be cautious before implementing proactive service. Hence, they should engage in a discussion with patients to understand how willing they would be to participate in a flexible scheduling policy for inductions.

[9] We note that a more detailed simulation-based analysis of the unit that captures a number of features that go beyond this stylized study, such as batching, estimates that proactive service can reduce waiting times by approximately 25%.