Submitted to *Manufacturing & Service Operations Management*
manuscript MSOM-17-214

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 has information about some customers' future service needs and may initiate service for such customers proactively, if they agree to be flexible with respect to the timing of service delivery.

Academic / Practical Relevance: Information about future customer service needs is becoming increasingly available through remote monitoring systems and data analytics. However, the literature has not systematically examined proactive service as a tool that can be used to better match demand to service supply when customers are strategic.

Methodology: We combine i) queueing theory, and in particular a diffusion approximation developed specifically for this problem that allows us to derive analytic approximations for customer waiting times, with ii) game theory, which captures customer incentives to adopt proactive service.

Results: We show that proactive service can reduce customer waiting times, even if only a relatively small proportion of customers agree to be flexible, the information lead time is limited, and the system makes occasional errors in providing proactive service – in fact we show that the system's ability to tolerate errors increases with (nominal) utilization. Nevertheless, we show that these benefits may fail to materialize in equilibrium because of economic frictions: customers will under-adopt proactive service (due to free-riding) and over-join the system (due to negative congestion-based externalities). We also show that the service provider can incentivize optimal customer behavior through appropriate pricing.

Managerial Implications: Our results suggest that proactive service may offer substantial operational benefits, but caution that it may fail to fulfill its potential due to customer self-interested behavior.

Key words: Proactive Service, Flexible Customers, Queueing Theory, Game Theory, Service Operations.
History: Revised: April 16, 2019

1. Introduction

In many service settings, 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 scheduling customer arrivals (Cardoen et al. 2010, Cayirli and Veral 2003), providing delay information to discourage customers from joining when congestion is high (Armony et al. 2009, Jouini et al. 2011, Ibrahim et al. 2016), 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 models have found application in a number of settings. For example, BetterCloud, a tech company that was featured on the 2017 Forbes' Next Billion Dollar Startups list (Adams 2017), uses proactive service to address errors encountered by IT administrators as quickly as possible. In their words: *"instead of waiting for the customer to come to BetterCloud for help, the customer service team contacts customers"* (Hyken 2016, Stone 2015). Another application, in the context of equipment repair, is General Electric's (GE) OnWatch service (Roy et al. 2014). As GE explains, OnWatch provides fully automated and continuous monitoring of critical healthcare infrastructure systems such as computed tomography and magnetic resonance imaging scanners. Any deviations alert GE engineers who work with customers to repair the equipment with minimal disruption to operations. In addition to the examples above, which make use of technological monitoring and advanced analytics to predict future customer service needs, proactive service may be implemented by simpler methods. For example, a hospital the authors worked with considered implementing proactive service to reduce delays for patients undergoing induction of labor (IOL), a procedure where labor is pharmacologically induced in a hospital ward. (This procedure is indicated for women whose gestation period exceeds 40 weeks.) Following a system integration, where the hospital gained better visibility to women that would need the procedure in the future, the hospital considered bringing forward women to start the procedure up to 24 hours early, should there be empty induction beds.

In all of the examples described above the potential benefit of proactive service is clear: at least some customers will be served at an opportune time for the server, e.g., when the server would have otherwise been idle. Naturally, this eliminates waiting time for any proactively served customers. Furthermore, even customers not served proactively could benefit; by serving some customers proactively, the system reduces congestion and, through that, everyone else arriving to the system would experience, on average, shorter waiting times. Nevertheless, for the benefits of proactive service to materialize, two conditions need to be met. First, from the customers' perspective, enough of them need to be willing to accept being served proactively, that is, agree to be *flexible* as to the timing of the service. Customers may be reluctant to be flexible if there is an associated inconvenience cost (e.g., due to loss of autonomy in choosing when to receive service), as would be the case in at least some of the examples described above. This last point is further complicated by the fact that the benefit of proactive service depends on system congestion and on how many other customers decide to be flexible. In other words, it is an equilibrium outcome. Second, from the system's perspective, the prediction as to which customers have a service need must be sufficiently

accurate. Otherwise, the server will be reaching out to provide service to customers who do not need it – not only wasting resources but also increasing the system utilization and, through that, potentially increasing customer waiting times.

To quantify the benefit of proactive service, and further investigate the two impediments associated with its implementation described above, we begin by formulating a stylized Markovian queueing model of proactive service. We first assume that customers' arrival rate to the system and whether they are flexible or not are exogenous, and that all customers are served by a single server that, when reaching out to serve customers proactively, never makes errors (i.e., perfect information.) We use this model to show that proactive service reduces congestion in the first-order stochastic sense. Using a pathwise coupling argument, we establish useful monotonicity results – the greater the proportion of flexible customers and the earlier the provider knows about the customers' service needs, the lower the average congestion and waiting time. Subsequently, to quantify the benefit of proactive service we develop a novel diffusion approximation that allows us to estimate average steady-state waiting times in closed form. The approximation suggests that the reduction in delay associated with proactive service displays decreasing marginal returns in the proportion of flexible customers. This is important from a managerial perspective as it suggests that even a little customer flexibility may lead to substantial waiting-time reduction.

Having established the benefit of proactive service, we then relax the assumption that customer behavior is exogenous. 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 to 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 welfare-maximizing 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 socially 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, that is, 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 show that the positive and negative externalities interact, giving rise to surprising comparative statics. 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. Our research shows that, as a result of these economic frictions, even in cases where proactive service can significantly reduce delays, it may fail to be adopted because of customers' self-interested behavior. Furthermore, our work suggests that, with the right financial incentives in place, such economic frictions could be overcome.

We then focus on the problem where the system's ability to predict customers' future service needs is imperfect. In this case, some of the customers served proactively will not have required the service. These "errors" could occur if customers may have their service need resolved through alternative channels (e.g., by visiting the company's website or using internal resources), or because of prediction errors. We show that the diffusion-limit approximation developed for the case without errors can be adapted to derive closed-form approximations of system performance in the presence of errors. Not surprisingly, the approximation suggests that there exists a threshold such that, if the proportion of errors is no greater than this threshold, then proactive service will continue to reduce waiting times. What is perhaps more surprising is that, in most cases, this threshold increases in system utilization. This seems counterintuitive 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. From a practical perspective, this result suggests that predictive errors are not a serious impediment to implementing proactive service, especially if the system is highly utilized.

We conclude with one piece of anecdotal evidence that supports the theoretical findings described above. As reported in Roy et al. (2014), after GE implemented the OnWatch proactive service system described earlier for 136 CT scanners, they observed an overall increase in service events by 19% despite a decrease in *user-initiated* customer requests. This is likely due to a combination of errors in identifying instances of proactive service and an increase in customer utilization due to proactive service. However, despite this increase in utilization, the report finds that OnWatch also reduced average time to service completion for user-initiated calls by 21%. This suggests that OnWatch had a positive impact on GE's customers despite the marked increase in system utilization.

The rest of the paper is structured as follows. Section 2 presents a review of the related literature. Section 3 presents the analysis of proactive customer service for the case of a single server with exogenous demand and perfect information about future customer service needs. Section 4 presents the economic analysis of endogenous customer demand and Section 5 relaxes the assumption of perfect information. Section 6 concludes with a short discussion and directions for future work. Sketches of all proofs are presented in the Appendix of this paper. Detailed proofs are presented in

the electronic companion (EC). A detailed numerical investigation of the diffusion approximation developed in Section 3, an extension of the model to multiple servers, and a simulation study using parameters calibrated to the example of IOL are available from the authors.

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 complicates the dynamic evolution of the system.

The second stream considers demand-side interventions. Perhaps the simplest intervention is to schedule customers' arrivals. This is certainly possible in some service settings, for example, operating rooms and outpatient doctor visits (Cardoen et al. 2010, Cayirli and Veral 2003). When scheduling is not possible, providers may try to influence customers' (endogenous) decisions on whether or when to join the queue; 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 there is also a positive externality; customers who agree to be flexible reduce the waiting time for everyone else. (We note that positive externalities are relatively rare in the literature of queueing games (Hassin 2016, §1.8). For notable exceptions see Engel and Hassin (2017), Nageswaran and Scheller-Wolf (2016), Hassin and Roet-Green (2011), Cui et al. (2014).)

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), especially when the system is congested. Multiple service priorities encourage some customers to wait in low-priority queues (usually off-line), 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 modeled using queues) has been recognized by many (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, that is, are flexible to the timing of product delivery (Karaesmen et al. 2004, Wang and Toktay 2008). 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

This section presents a stylized model of proactive service assuming there is a single server, demand is exogenous, and the predictions of future customer service needs are perfect. 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

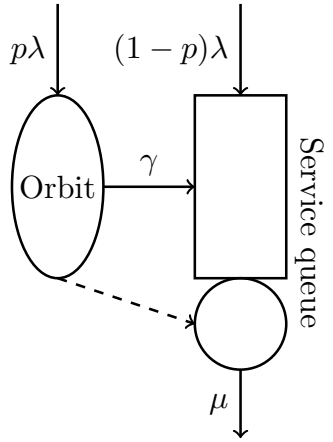
We assume that demand arrives to the system following a Poisson process with rate λ, and that there exists two types of customers who require service – "flexible" and "inflexible." The service times for both types of customers are assumed to be independent and identically distributed (i.i.d.) and exponential with parameter μ; 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 that 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 that 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.

We note that the special 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). This assumption

Figure 1 Queueing model



simplifies the dynamic evolution of the system (effectively, the orbit becomes a low-priority queue) and the steady-state performance of the system can be obtained in closed form using exact analysis. This is not the case in general (i.e., when $\gamma > 0$).

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 (i.e., 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 the orbit and the service queue is reduced, that is, $\bar{N}_r \le \bar{N}_r^B$ and $\bar{N}_s \le \bar{N}_s^B$. We note 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. Turning to the impact of proactive service on the expected time spent by each customer type in different parts of the system in steady state, we use Proposition 1 and the mean value approach (MVA) (Adan and Resing 2002, §7.6) to derive the following results.

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

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

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

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

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

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

Table 1 Monotonic behavior of performance measures

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

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

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 that 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., π_r) is, in a stochastic ordering sense, increasing in p and decreasing in γ.

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

(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 γ) 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.

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 μ, 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.

Asymptotic analysis. Observe that as n increases, the total arrival rate (λ^n) approaches the service rate (μ), 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 with 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* $T > 0$,

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

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 diffusion 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 Poisson Arrivals See Time Averages

(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 (3) and (5), 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/γ, which implies there 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 $\rho > .2$.

Although not essential for the rest of our analysis, we note two observations. First, with a relatively small tweak to the scaling regime, the diffusion approximation developed in this section for the single-server system can be readily extended to the multiserver case. The specific details are available from the authors. Second, the case where time in orbit is deterministic (as opposed to an exponentially distributed random variable) was studied in Zhang (2014). It is therefore interesting to compare the waiting-time reduction associated with these two different assumptions. The setting in Zhang (2014) has two additional differences. First, it assumes that a customer does not have to be present 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 Zhang's when $p=1$, in which case preemptive priority does not matter. 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 Δ is (strictly) increasing in ρ. Therefore, knowing exactly when customers would transition from the orbit to the service queue (i.e., deterministic service time) helps further reduce average time spent in service queue. However, this additional reduction in waiting time decreases as the system reaches heavy traffic.

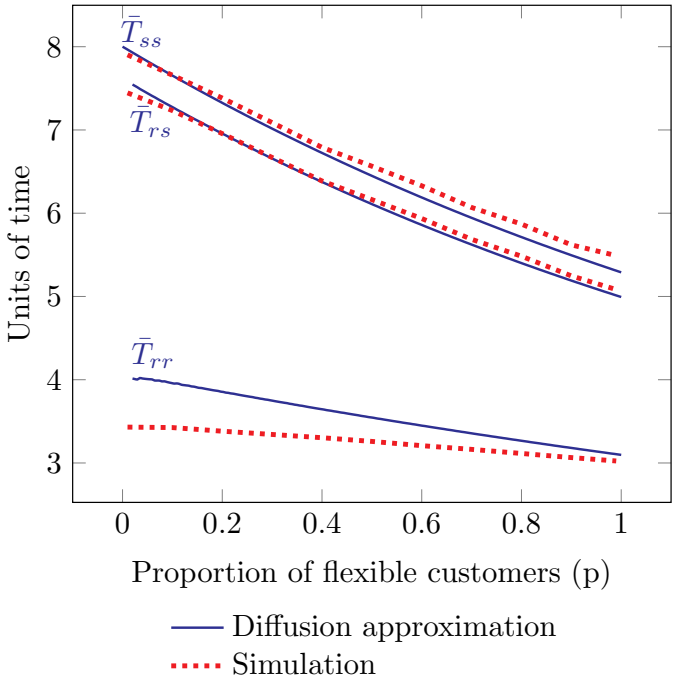
Accuracy of the approximations and managerial implications: Because the approximations presented above are based on an asymptotic result, in this section we examine their accuracy in finite systems where utilization is less than one. For instance, Figure 2 depicts the comparison of the diffusion approximations 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 available from the authors. 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 $\gamma^n \to 0$. Because the diffusion limit presented above fails when the information lead time increases (i.e., $1/\gamma \to \infty$), it is not surprising that 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.

From a managerial perspective, 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 for a relatively small proportion of flexible customers goes a long way when the system can serve customers proactively.

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

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



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. For example, in the case of OnWatch, at least some customers agreed to be served proactively (Roy et al. 2014). This was not so easy in the case of IOL, where expectant mothers interviewed by the hospital expressed reservations as to having the service brought forward. Therefore, to understand the benefits of proactive service, we need to consider the decisions of 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.

4.1. Customer Demand and Utility

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 modeled by a Poisson process with rate Λ. 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), each customer's best response is to also play strategy (J,p). For the rest of the analysis we restrict our attention to λ 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, that is, 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, λ_e) *for customer flexibility and joining behavior.*

ii. The equilibrium strategy is such that:

(a) *The proportion of flexible customers* p_e *and the arrival rate* λ_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* λ_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 λ_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 λ_e is increasing in the waiting-time cost w_s, otherwise λ_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 the diffusion approximations offer an accurate representation of the system performance. 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 because customers are willing to wait for longer, 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. 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. Clearly, in this case, 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 a profit-maximizing service provider may seek to influence customer behavior through the use of prices/tolls. More specifically, the provider seeks to maximize the revenue rate from prices paid by customers subject to customers' equilibrium behavior as shown in (6) below. (We note that for this analysis we use revenues and profits interchangeably. We can do this because we have assumed that there are no costs associated with implementing proactive customer service. In reality, there may be fixed and variable costs associated with implementing proactive service and monitoring customer service needs. Since it would be straightforward to include these costs in the analysis, we do not model them explicitly.)

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

Because customers are homogeneous, a profit-maximizing provider will not find it optimal to set prices that leave customers with 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, §3.1.3). 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 objective can be rewritten as,

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

Since customers are left with zero surplus, the provider's objective, as given by (7), is also equal to total welfare. (This is the case because the payments c_s and c_r are an internal transfer between customers and the service provider.) Therefore, in this case, the objective of the provider happens to coincide with that of a benevolent social planner who can dictate customers' joining decisions. This result, which follows from the assumption that customers are homogeneous, is well known in the literature of queueing games (see (Hassin and Haviv 2003, §1.3)), and it implies that, in order to understand whether customers' autonomous joining decisions presented in §4.2 are suboptimal for the profit maximizer, it would suffice to compare them to the actions chosen by a benevolent social planner.

The existence of a solution to the optimization problem above is guaranteed by the fact that the action space is compact and the objective function is continuous. However, we note that we are unable to prove that $W(p, \lambda)$ is concave. Nevertheless, in numerical experiments we find the first order conditions are both necessary and sufficient. We compare the socially optimal customer (or profit-maximizing) actions with the equilibrium customer decisions of Proposition of §4.2 in the next result.

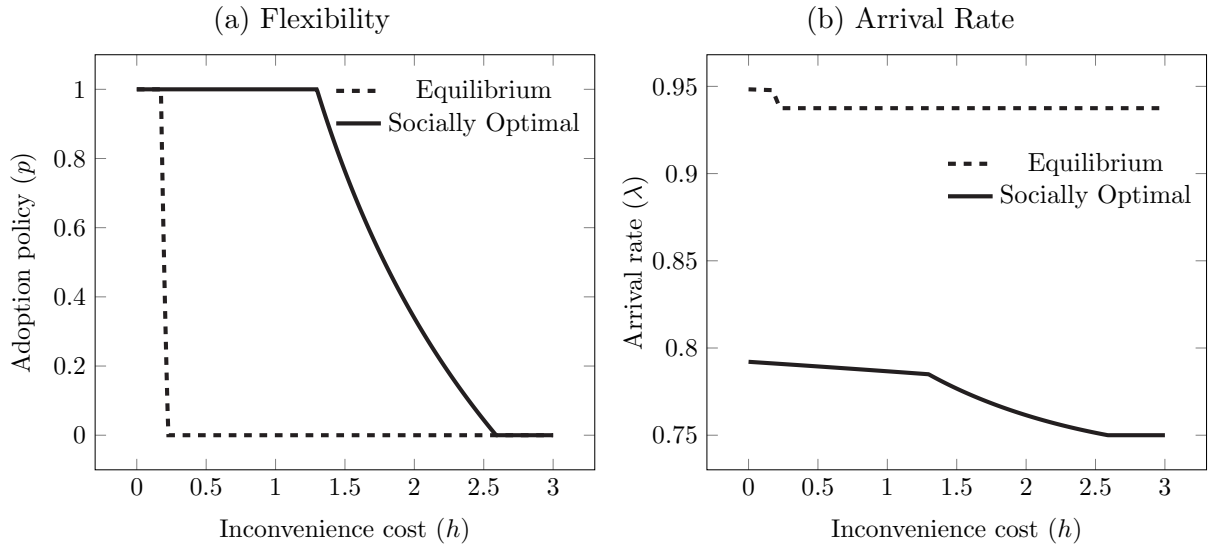
PROPOSITION 5. *If* $\frac{\Lambda}{\mu} \geq .75$, $v \geq 16\frac{w_s}{\mu}$, $\gamma \geq \sqrt{\mu w_s / v}$, *then for any socially optimal/profit-maximizing solution* (p_{so}, λ_{so}),

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* $p_{so} = 1$, *that is, 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 system performance. Proposition 5 shows that customers will over-utilize the service system

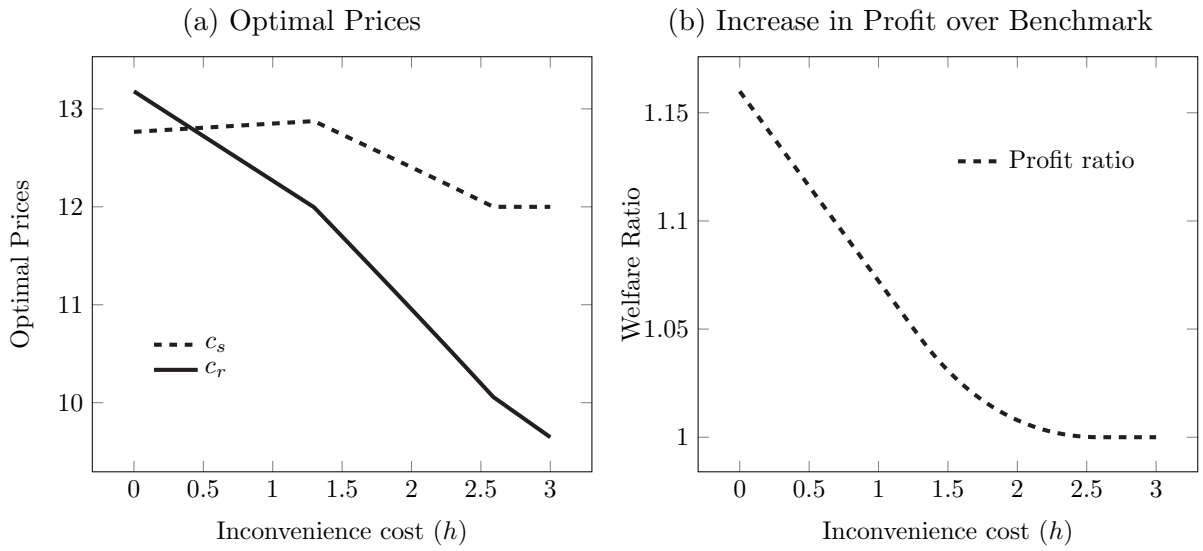
Figure 3 Comparison of Equilibrium Strategy and Socially Optimal: $\Lambda = .95, \gamma = .25, \mu = 1, w_s = 1, w_r = 0, v = 16$.



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 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.

As in the case where proactive service is not possible (e.g., Naor 1969), it would be possible to incentivize optimal customer behavior by setting different prices/tolls c_r and c_s, for flexible and inflexible customers, respectively. Figure 4a depicts the optimal prices c_r and c_s against the fixed cost of flexibility (h) for the same example as Figure 3. (We note that, for some values of h there exist multiple prices that are optimal and revenue equivalent. 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 (flexible) in this case, choosing any such price would not make a difference to the revenue. For these cases, the figure depicts the lowest price.) Similarly, Figure 4b shows the improvement in provider revenue against the fixed cost of flexibility h by

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



showing the ratio of the optimal revenue in the case with proactive service over the benchmark case without proactive service.

As illustrated in Figure 4a, for 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 in order to incentivize customers to be flexible despite their free-riding 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.

Managerial Implications: The analysis presented in this section has identified two inefficiencies associated with customer self-interested behavior: customers will under-adopt proactive service

and will over-utilize the system compared to the social (or profit maximizing) optimal. Therefore, the real-world benefits of proactive service are likely to be lower than those suggested by the operational analysis of §3, which assumed exogenous arrival rates. Furthermore, this analysis shows that there are instances where proactive service will not be adopted at all even though all customers would be better off by doing so. Clearly, such negative results are not easy to verify empirically, but may offer an explanation as to why proactive service is not more widely adopted in practice. Nevertheless, our research suggests that in settings where the provider has the ability to set differential prices for flexible/inflexible customers, such frictions may be overcome.

5. Imperfect Information about Customer 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 the case of OnWatch, the service team may contact customers who did not actually have a service need (i.e., the prediction that they needed service was erroneous), or serve customers who could have solved the problem on their own. In the case of IOL, patients induced proactively may have otherwise gone into labor spontaneously. 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. This could erode the operational benefits of proactive service.

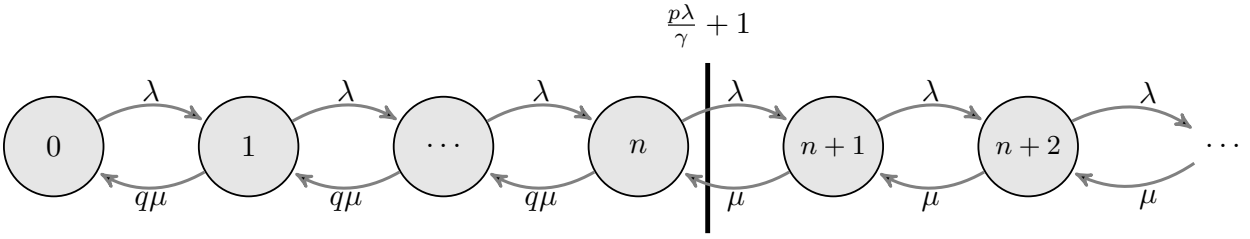
To investigate whether proactive service is still beneficial despite errors with an analytical model, we assume that every time the server pulls a customer from orbit it makes an error with probability $1-q$. That is, a proportion $1-q$ of customers served proactively would not have transitioned to the service system had they not been pulled and, therefore, would not have been served at all.

By construction, under this additional assumption, 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 ρ, 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. Assuming that the asymptotic result of Theorem 1 also holds for finite systems as well, we can 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 λ. 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 (by Theorem 1). In this case, the departure rate from N_Q to N_Q-1 is given by the service rate μ – 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 μq, 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 available from the authors, suggest that the approximations work well when $\rho \geq 0.75$ and $(\mu-\lambda)/\gamma \leq 1$. Using these approximations, one can prove the following result.

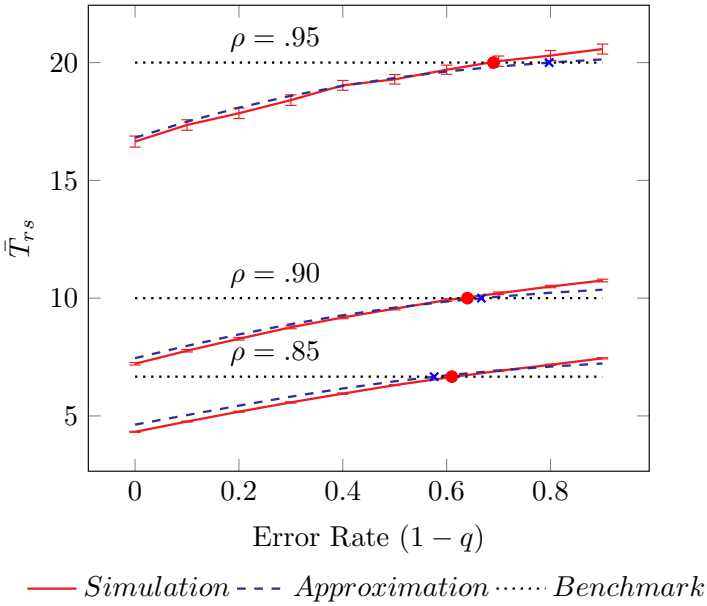
PROPOSITION 6. *There exists a threshold* $0 < \bar{q} < 1$ *such that the system with proactive service generates lower waiting times compared to the system without proactive service if, and only if, the proportion of errors is less than* $1 - \bar{q}$*. Furthermore,*

(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* μ *approaches the arrival rate* λ*).*

The proposition establishes the intuitive result that proactive service reduces waiting times only if the proportion of errors is below a critical threshold, $1 - \bar{q}$. 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, as the system utilization increases (i.e., as μ approaches λ), the system is able to handle even more errors. (In fact, we numerically observe that this is almost always the case.) 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 $p = 1$ (i.e., when all customers are flexible) 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 completely eliminated 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 and holds true in almost all numerical experiments for both flexible and inflexible customers. This finding may seem counterintuitive at first because in

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



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.

From a practical perspective, the results of this section suggest that, unless the error rate is extreme (i.e., greater than the critical threshold $(1-\bar{q})$), proactive service continues to offer operational benefits, especially for systems that are highly utilized. Furthermore, provided the error rate is below this critical threshold, the economic frictions associated with the adoption of proactive service described in the previous section (i.e., over-joining the system and under-adopting proactive service) will remain qualitatively unchanged in the presence of errors.

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 practical impediments (e.g., economic frictions and prediction errors) that limit its potential to make an impact in practice?

From an operational perspective, we find that proactive service can substantially reduce delays for both flexible and inflexible customers. This is the case even if the proportion of flexible customers is limited and the information lead time relatively short, and even if proactive service ends up serving a moderate proportion of customers that would otherwise not have needed the service. To derive these results, we develop a novel diffusion approximation that other researchers may find useful. From an economic perspective, we show that proactive service is likely to be under-adopted (due to a free-riding problem) and may also exacerbate customers' tendency to over-join

the system (due to congestion-based negative externalities). These economic frictions may provide an explanation as to why proactive service has not been more widely adopted in practice and suggest that, in order to realize the operational benefits of proactive service, providers will need to offer incentives (e.g., in the form of differentiated prices) to alleviate the economic frictions of free-riding and over-utilization.

We conclude by noting that there are several practically relevant directions where the study of proactive service, using the tools developed in this paper, could be extended. One obvious direction is the study of multiple customer priorities. Such a system will require us to appropriately modify the heavy-traffic approximations developed by this paper. Furthermore, in such a system the service provider may be in a position to offer higher priority as an additional incentive to induce customers to adopt proactive service. This incentive may be more straightforward to implement in settings where prices are exogenous (e.g., healthcare). Another promising direction is to extend the economic analysis from homogeneous customers to the more realistic case of heterogeneous customers. We leave these extensions to future work.

References

Adams, S. 2017. The next billion-dollar startups 2017. URL `https://www.forbes.com/sites/susanadams/2017/09/26/the-next-billion-dollar-startups-2017/#57259bd04447`.

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.

Cardoen, Brecht, Erik Demeulemeester, Jeroen Beliën. 2010. Operating room planning and scheduling: A literature review. *European journal of operational research* **201**(3) 921–932.

Cayirli, Tugba, Emre Veral. 2003. Outpatient scheduling in health care: a review of literature. *Production and Operations Management* **12**(4) 519–549.

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.

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. *Working Paper* .

Hyken, S. 2016. Five ways to deliver proactive customer service. URL `https://www.forbes.com/sites/shephyken/2016/12/10/five-ways-to-deliver-proactive-customer-service/#2a0d31e79686`.

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

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.

Roy, A, Payronnet J., Sievevpiper C. 2014. Determining the benefits of proactive digital service for computed tomography scanners. URL `http://www3.gehealthcare.co.uk/~/media/downloads/uk/services/services%20_%20onwatch_white%20paper-v3.pdf`.

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.

Stone, M. 2015. Beyond satisfactory: Why bettercloud introduced proactive support. URL `https://www.bettercloud.com/monitor/proactive-support/`.

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.

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 pathwise 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 samplepath (i.e., the sequence of customer interarrival 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 then follows 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 $N_r(t)$. □

A.3. Proof of Proposition 2:

We first establish four equalities using Mean Value Approach (MVA) and then prove the desired results using these equalities. Because external arrivals follow a Poisson process, by the Poisson Arrivals See Time Averages (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 $\bar{T}_{ss}^B = \bar{T}_{rs}^B$, 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). □

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 electronic companion. We note that the exponential assumptions of inter-event times are necessary for the coupling in this proof.

To show that π_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 π_r is decreasing in γ (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 γ) 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 electronic companion. 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 than 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 $\hat{N}_r^n(t) \leq \frac{p\lambda^n}{\gamma} + \epsilon/4$ 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, (14) 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 Λ, γ, μ, 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 $\Lambda < \mu$, $v \geq \hat{v}_0 := \frac{w_s}{\mu - \Lambda}$ and $h \geq \hat{h}_\Lambda := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \Lambda} \right) \frac{\Lambda^2}{\gamma\mu} (-\ln \frac{\Lambda}{\mu})$.

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

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

Case 4: $p_e = 1$ **and** $\lambda_e = \lambda_1 < \Lambda$, if either $\Lambda \geq \mu$ or $v < \hat{v}_1$ and $h \leq \check{h}_{\lambda_1} := \left(\frac{w_s}{\mu} - \frac{w_r}{\mu - \lambda_1} \right) \frac{\lambda_1}{\mu} \left(1 - ({}^{\lambda_1}/_{\mu})^{\lambda_1/\gamma} \right)$, where λ_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 λ 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 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 μ approaches λ the system utilization ρ increases but n and λ remain unchanged. Differentiating the polynomial equation with respect to ρ 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, combinations of parameters $(p, \lambda, \gamma, \mu)$ exist such that the threshold $\bar{q}$ is decreasing in utilization.