Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Pooled vs. Dedicated Queues when Customers Are Delay-Sensitive

Nur Sunar, Yichen Tu, Serhan Ziya

To cite this article:
Nur Sunar, Yichen Tu, Serhan Ziya (2021) Pooled vs. Dedicated Queues when Customers Are Delay-Sensitive. Management Science 67(6):3785-3802. https://doi.org/10.1287/mnsc.2020.3663

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 67, No. 6, June 2021, pp. 3785–3802
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Pooled vs. Dedicated Queues when Customers Are Delay-Sensitive

Nur Sunar,[a] **Yichen Tu,**[b] **Serhan Ziya**[c]

[a] Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599; [b] Cox Automotive Inc., Atlanta, Georgia 30319; [c] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

Contact: Nur_Sunar@kenan-flagler.unc.edu, https://orcid.org/0000-0002-9306-1050 (NS); yichen1@live.unc.edu (YT); ziya@unc.edu, https://orcid.org/0000-0003-1558-6051 (SZ)

Received: July 20, 2017
Revised: July 17, 2018; February 10, 2019; November 29, 2019
Accepted: February 24, 2020
Published Online in Articles in Advance: January 20, 2021

https://doi.org/10.1287/mnsc.2020.3663

Abstract. It is generally accepted that operating with a combined (i.e., pooled) queue rather than separate (i.e., dedicated) queues is beneficial because pooling queues reduces long-run average sojourn time. In fact, this is a well-established result in the literature when jobs cannot make decisions and servers and jobs are identical. An important corollary of this finding is that pooling queues improves social welfare in the aforementioned setting. We consider an observable multiserver queueing system that can be operated with either dedicated queues or a pooled one. Customers are delay-sensitive, and they decide to join or balk based on queue length information upon arrival; they are not subject to an external admission control. In this setting, we prove that, contrary to the common understanding, pooling queues can increase the long-run average sojourn time so much that the pooled system results in strictly smaller social welfare (and strictly smaller consumer surplus) than the dedicated system under certain conditions. Specifically, pooling queues hurts performance when the arrival-rate-to-service-rate ratio is large (e.g., greater than one) and the normalized service benefit is also large. We prove that the performance loss due to pooling queues can be significant. Our numerical studies demonstrate that pooling queues can decrease the social welfare (and consumer surplus) by more than 95%. The benefit of pooling is commonly believed to increase with system size. In contrast, we show that when delay-sensitive customers make rational joining decisions, the magnitude of the performance loss due to pooling can strictly increase with the system size.

History: Accepted by Terry Taylor, operations management.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2020.3663.

Keywords: pooled queue • dedicated queues • customer balking • rational customers • delay-sensitive customers • observable queues • service management

1. Introduction

A fundamental question for services that are operated by multiple servers has been whether to run the system with separated queues or a combined one. These queueing configurations are called *dedicated* and *pooled*, respectively. It is not difficult to see why pooling separate queues might be appealing: a pooled system uses the available service capacity more efficiently because under pooling no server idles as long as there are customers waiting, a possibility that exists when individual queues are kept separated. The benefit of pooling is well established in the operations management literature: when servers are identical and customers are homogeneous in their service requirements, pooling queues is proven to improve efficiency by reducing idleness and the expected waiting time in the system.

When studying the age-old question of *to pool or not to pool*, the vast majority of the literature implicitly assumed that customers have no say in their joining decisions and they join a queue regardless of how long they wait. In fact, under this assumption, the well-established benefit of pooling has a key implication: pooling queues improves social welfare and consumer surplus. However, a common feature of many queueing systems in practice is that customers are delay-sensitive and decide whether to join a queue depending on how long they expect to wait. Thus, it is important to analyze systems with such customers and understand if pooling is still preferable when customers make their join or balk (i.e., not join) decisions. This is the primary objective of this paper.

The question of whether to operate a dedicated system or a pooled system is relevant in many service settings from shipping lines at the ports to voting lines in elections (Cattani and Schmidt 2005, *Financial Times* 2015, Hong et al. 2015, *New York Times* 2016, Karacostas 2018). Our paper studies this question by analyzing a model in which delay-sensitive customers have access to their delay information (e.g., through observing the queue length or by receiving real-time expected delay information) and make their joining decisions based on that information. Our model is motivated by various practical settings where the service is provided for free. Two of these settings are explained.

The first example is the design of call centers. Many organizations are grappling with the question of whether to consolidate their call centers or not (Southwest 2012, Xerox 2013, Rodriguez 2014). With consolidation, calls are processed in a single large call center, rather than separate, smaller, and typically region-specific call centers. In practice, the key benefit of consolidation is believed to be the efficient use of resources because of pooling, thereby improving customer satisfaction with the same or even smaller number of resources (Xerox 2013). However, these anticipated benefits do not take customer behavior into account. In many call centers, callers receive queue length or real-time expected delay information, and based on that information, they may choose not to join the system. (See Ibrahim 2018 for a literature review of such systems.)

The second example is the design of internal services in large organizations. For such organizations, there has long been a discussion on whether support services such as information technology, consulting, and purchasing should be shared across different units of the organization or administered in a decentralized manner where these services are provided within each individual unit (Schmidt 1997, Azziz 2014, Bondarouk 2014). Thus, in the management of internal services, the question of whether to operate a dedicated system or a pooled one is of paramount importance. Within many organizations, such as government agencies and universities, internal services are provided for free (see, e.g., Armbrüster 2006, p. 113 and UAFS 2018), and successful implementations of such services typically rely on information sharing, which enables members of the organization to observe and identify inefficiencies such as service congestion and delays (Campbell Public Affairs Institute 2017). Sharing support services is aimed to improve organizational efficiency by tapping into the operational benefit of pooling (Mader and Roth 2015, U.S. Department of the Treasury 2017). However, the design of such services also needs to account for the user behavior: if users within an organization face long delays in their service requests, they could give up solutions offered by the organization.

Motivated by these practical settings, our objective is to develop and analyze a formulation that centers on the following three questions. (i) How do delay-sensitive customers' rational joining decisions alter the basic calculus for the choice between pooled and dedicated systems in services with observable queue length (or delay information)? If pooling is not always preferable in such settings, what are the conditions under which the dedicated system is preferable? (ii) How large is the performance gain with a particular system design over the other? (iii) How does the system size impact such performance gain? We are not aware of any prior work that provides a theoretical analysis of the comparison between pooled and dedicated queues for an observable queueing system when customers make rational joining decisions.

This paper considers social welfare as the primary performance metric. In our setting, social welfare is equal to the consumer surplus, which is an important measure of customer satisfaction, and the long-run average sojourn time is one of the key determinants of social welfare. Thus, we will also consider the long-run average sojourn time as an auxiliary performance metric.

1.1. Summary of Main Results

Considering delay-sensitive customers' rational joining decisions in the comparison of pooled versus dedicated queues gives rise to the following three unexpected results for the observable systems.

First, Smith and Whitt (1981) establish that if every arrival joins the system (without making decisions), pooling queues is beneficial in the case of identical servers and jobs. In contrast, our paper proves (in Theorem 1(a)) that if arriving customers decide to join or balk, the dedicated system can outperform the pooled system depending on the following two factors: (i) *normalized benefit of service*, which is the ratio of service benefit to customer's expected cost of waiting in service, and (ii) *potential system load*, which is the ratio of arrival rate to service rate. Specifically, if both the normalized benefit of service and the potential system load are large, pooling queues strictly increases the average sojourn time, and this increase is so large that, compared with the dedicated system, the pooled system results in strictly smaller social welfare.

Second, in the case of nonstrategic and identical jobs and servers, the benefit of pooling queues is well known to increase with the number of servers (keeping the potential system load the same) (Calabrese 1992, Benjaafar 1995). In contrast, our paper proves (in Theorem 2) that when customers make their own joining decisions, the magnitude of the relative performance loss due to pooling can strictly increase with the number of servers (keeping the potential system load the same).

Third, our analysis and numerical studies show that the performance improvement due to separating queues can be drastic. Specifically, our paper proves (in Theorem 3) that the percentage increase in the social welfare with dedicated queues can be arbitrarily large, compared with the case with a pooled queue.

To provide a complete picture, our paper also identifies conditions under which the pooled system results in smaller average sojourn time and larger social welfare than the dedicated system. (Those conditions are provided in Theorem 1(b).)

Our paper also studies variants of the base model. Some of the key messages from this additional analysis (see, e.g., Section 4) are as follows. (i) The dedicated system may outperform the pooled system when customers are allowed to choose the shortest queue in the dedicated system, when customers are heterogeneous in the service benefit they receive, or when there is a fixed service fee. (ii) The following two criteria are necessary for our unexpected results to hold: the observability of queue length (or real-time expected delay information) and the lack of pricing control (or in general, the lack of admission control). Regarding the latter, when a social planner could charge a different service fee under each queue configuration to maximize social welfare, pooling queues improves the social welfare. Thus, the welfare advantage of pooling queues can be recovered if the social planner has the pricing lever.

1.2. Relevant Literature

Our paper belongs to the literature that studies pooled versus dedicated queues. To the best of our knowledge, there is no prior work that theoretically analyzes the comparison of pooled versus dedicated queues in an observable system when delay-sensitive customers make their own joining decisions. Our paper provides such an analysis.

The analysis of pooling queues has long been an interest in the queueing literature. To our knowledge, Smith and Whitt (1981) were the first to provide a mathematical investigation of pooling queues. Smith and Whitt (1981) showed that when jobs (e.g., customers) are homogeneous in their service requirements and servers are identical, pooling separate queues increases the system efficiency by reducing the expected steady-state waiting time. Since the publication of this seminal work, many articles studied the benefit of pooling queues in different contexts and under a variety of conditions. Calabrese (1992) provided an alternative proof to show the benefit of pooling for system efficiency. Benjaafar (1995) determined bounds on performance improvements through pooling. Gans et al. (2003) illustrated the benefits of pooling call centers (in different geographical locations) into one. Using approximation formulas for a two-server queueing system, van Dijk and van der Sluis (2008) made the observation that when customers are identical, a pooled system results in smaller long-run average waiting time than its dedicated counterpart. Andradóttir et al. (2017) showed that even if servers are subject to failures, pooling queues always results in smaller expected steady-state number of jobs in the system and hence, smaller long-run average waiting average time, compared to the system with dedicated queues.

Unlike what has been established in this literature, our paper proves that when delay-sensitive customers make their joining decisions in an observable system, pooling queues can result in much worse performance than a dedicated system even with identical servers and homogeneous customers. We prove this result when there is no admission control (e.g., in the form of monopoly pricing).

There have been observations that pooling parallel queues is not always beneficial and may result in performance degradation; these observations are attributed to three main factors explained in (a)–(c) below. Our paper identifies a different factor not previously identified in the pooling literature: observable queue and customers' ability to make a joining/balking decision. We now explain the aforementioned three factors in (a)–(c) below and discuss the relevant literature.

a. If jobs are heterogeneous in their service requirements or servers are not identical, the pooled system may perform worse than the dedicated system. Smith and Whitt (1981) included a numerical example with heterogeneous servers to make this point. Rothkopf and Rech (1987) discussed that if jobs require different service times, combining separate queues into a single one can increase the average delay. Section 5.3 of Mandelbaum and Reiman (1998) briefly discussed this effect of heterogeneous servers in a parallel multiserver setting without providing proofs (as there are no exact formulas available in that setting). Using approximation formulas for queueing models, van Dijk and van der Sluis (2008, 2009) constructed numerical examples to illustrate the aforementioned effect of these factors.

b. Pooling queues may also result in worse performance (e.g., larger expected steady-state waiting time) because of server slowdown and other server-related issues. Rothkopf and Rech (1987) argued that when combining separate queues into a single one increases service times, the pooled system may result in larger average delay than the dedicated one. Gilbert and Weng (1998) studied a setting where there are two self-interested servers and a principle that compensates servers based on their performance. In this setting, authors established that pooling queues can be undesirable for the principle because of server incentives. Shunko et al. (2018) conducted controlled laboratory experiments to find evidence of server slowdown due to pooling queues. Jouini et al. (2008) numerically demonstrated that the dedicated system can outperform the pooled system if each agent works slower in the pooled system potentially because of decreased customer ownership. Song et al. (2015) empirically investigated the effects of pooling in an emergency department and found that the dedicated system is superior to the pooled system with respect to the average waiting time and the average length of stay. The paper attributes this to physicians' increased ownership of the patients under the dedicated system. Do et al. (2015) theoretically analyzed the implications of server slowdown due to pooling and showed that the pooled system can result in larger expected waiting time than the dedicated system. Using a data set from a supermarket, Wang and Zhou (2017) provided empirical evidence that pooling queues can increase the service time. The main driver of this finding was explained to be the social loafing effect with a pooled queue. Armony et al. (2017) considered a two-server queueing system where servers can choose their long-run average service rates and incur a cost for the expected workload or busyness. The authors showed that if servers are workload averse, pooling queues always achieves lower expected queue length but can result in larger expected work in process.

c. Apart from two factors explained in (a) and (b), Rothkopf and Rech (1987) conjectured that when jockeying (i.e., switching from one queue to another) among parallel queues is possible for customers, under very mild conditions, the average waiting time under the pooled system can be larger than that under the dedicated system.

It is worth emphasizing that none of the papers mentioned in (a)–(c) theoretically analyze customers who can make their own joining decisions. Unlike all of the papers mentioned above, our work provides a theoretical analysis of such self-optimizing customers in the context of pooling queues. In our problem formulation, to avoid any performance advantage to the dedicated system and to analyze the effect of customers' joining decisions in isolation, we will exclude the factors that were previously observed to cause pooling to potentially perform worse than the dedicated system.

Lu et al. (2013) empirically analyzed a data set from a supermarket's checkout line. Considering a specific queue setting, Lu et al. (2013) found evidence that the queue length can be an important driver of customers' purchasing behaviors. Based on this, Lu et al. (2013) argued that if there existed a practical setting in which customers' purchasing behaviors under pooled and dedicated systems were both the same as the one identified by the authors, pooling queues may decrease average waiting time because of balking. Lu et al. (2013) considered a specific queue setting in their study and hence, did not empirically investigate the trade-offs between pooled and dedicated systems. Unlike Lu et al. (2013), our paper theoretically compares pooled versus dedicated systems by considering rational customers' joining decisions. Moreover, the main performance metric in our paper is social welfare, which is the same as consumer surplus in our setting.

Our work is also relevant to the literature that studies delay-sensitive rational customers making their own decisions in observable queueing systems. The comparison of pooled versus dedicated queues has not been investigated in this literature. Our formulation of customers builds on the framework developed and analyzed in the seminal work by Naor (1969). Naor (1969) considered a single-server queue where customers can observe the queue length and decide whether to join the queue or balk depending on their expected net benefit of joining the queue. In his setting, a balking customer gains zero expected net benefit, whereas each joining customer incurs a constant waiting cost per unit of time spent in the system and receives a reward upon service completion. One of the main findings of Naor (1969) is that allowing customers to make their own decisions results in social welfare loss compared with the maximum achievable welfare. Many articles extend the model analyzed in Naor (1969) in various dimensions. The comprehensive review of these papers can be found in Hassin and Haviv (2003) and Hassin (2016).

1.3. Outline of the Paper

The remainder of our paper is organized as follows. Section 2 introduces the model and includes preliminary analysis. Section 3 includes main results and their interpretations. Section 4 studies several extensions of the base model. Section 5 provides additional discussions. Section 6 includes concluding remarks. Proofs of all formal results as well as supplementary materials are presented in the online appendix.

2. Model

Consider a first-come, first-served queueing system with $N \geq 2$ identical servers. The service time of each server is exponentially distributed with rate $\mu > 0$.[1] The system can be run with either dedicated queues or a pooled queue. These two alternatives will be called *dedicated* and *pooled systems* and indexed by $j = d$ and $j = p$, respectively.

The dedicated system contains N separate queues, each served by a separate server. In this setting, a server together with its queue is called a dedicated *subsystem*. In the dedicated system, customers arrive to each queue according to a Poisson process with rate $\Lambda_d = \lambda$, and a server provides service only to customers in his own queue.[2] In contrast, in the pooled system, separate queues are combined into a single one, and customers arrive to the queue according to a Poisson process with rate $\Lambda_p = N\lambda$. Whenever a server completes serving a customer, he serves the next customer waiting in the queue. Here, Λ_d and Λ_p can be interpreted as *potential arrival rate* for a queue in the associated system. In light of this, the ratio

$$\rho \doteq \lambda/\mu \tag{1}$$

is called the *potential system load*. As will be explained later, the actual arrival rate to a queue is different than the potential arrival rate because the former is determined by customers' joining decisions.

Customers make their own joining decisions. In both systems $j = d$ and $j = p$, upon arrival, each customer first observes the length of the queue she arrives and then decides whether to join the queue or balk.[3] If an arriving customer decides to join the queue, the customer incurs cost $c > 0$ per unit time she spends in the system. A customer gains a benefit R after service completion, and the service is free of charge. Considering a queueing system that provides free of charge service is common in the literature. Although their research questions are very different than ours, several studies analyze such systems (see, for instance, Hassin 1985, Armony et al. 2009, Gai et al. 2016, and Haviv and Oz 2016). All of our results and their proofs extend in a straightforward fashion if customers pay a fixed fee $f > 0$ upon service completion. In our formulation, all model parameters are common knowledge. Because μ and N are known, for customers, observing the queue length is the same as observing their real-time expected sojourn time.

As in Naor (1969), if an arriving customer decides to balk, she neither gets a benefit nor incurs a cost, and hence, she gains zero net benefit.[4] If a customer arrives to a particular queue, the customer receives the following expected net benefit by joining the queue:

$$\mathbb{E}\big[U(n;j)\big] = R - \bar{W}_j(n+1)c, \quad j \in \{d,p\}. \tag{2}$$

Here, $\bar{W}_j(n+1)$ represents the expected time spent by the arriving customer in the system; for the pooled system, n represents the number of customers who are already in the system, and for the dedicated system, n corresponds to the number of customers who are already in the arrived subsystem. A customer joins the queue if and only if her expected net benefit is nonnegative, which is equivalent to the following by (2):

$$\mathbb{E}\big[U(n;j)\big] = R - \bar{W}_j(n+1)c \geq 0;$$

otherwise, she balks. This suggests that an arriving customer optimally joins the queue if and only if the number of customers in the queue and its associated service is smaller than a threshold that depends on the system type; otherwise, the customer balks.

The aforementioned optimal threshold rule implies two key characteristics of the systems in our analysis. First, the rate at which customers join the queue, which is represented by $\lambda_{e,j}$, is always smaller than the potential arrival rate Λ_j for $j \in \{d,p\}$. Second, regardless of the value of the potential system load ρ, both pooled and dedicated systems are stable.[5]

Our primary goal is to analyze the implications of pooling for *social welfare*. In doing so, we will also study the implications of pooling for *average sojourn time*: that is, long-run average time spent in the system.

Denote by W_j the average sojourn time in system $j \in \{d,p\}$. In our setting, the social welfare equals the consumer surplus, which is the sum of long-run average net gains of all customers in a system. As a result, the social welfare in system $j \in \{d,p\}$ is equal to the multiplication of these two factors: (i) a single customer's long-run average net benefit $R - W_jc$ and (ii) the long-run average number of customers served; that is, *throughput*, θ_j:

$$\begin{aligned} SW_j &= (R - W_jc)\theta_j \\ &\doteq \begin{cases} (R - W_jc)\lambda_{e,j} = R\lambda_{e,j} - cL_j & \text{if } j = p, \\ (R - W_jc)\lambda_{e,j}N = R\lambda_{e,j}N - cL_jN & \text{if } j = d. \end{cases} \end{aligned} \tag{3}$$

Here, L_p is the (steady-state) average number of customers in the pooled system, L_d represents its counterpart in *one* of the N dedicated subsystems, and the throughput θ_j satisfies the following:

$$\theta_j = \begin{cases} \lambda_{e,j} & \text{if } j = p, \\ \lambda_{e,j}N & \text{if } j = d. \end{cases} \tag{4}$$

Note from (3) and (4) that social welfare can be expressed in terms of throughput and either average sojourn time or average number of customers in the system. In our paper, we will use both of these alternative representations.

2.1. Preliminary Analysis

To avoid trivialities, this paper focuses on a setting where

$$k \doteq \left\lfloor \frac{R\mu}{c} \right\rfloor \geq 1. \tag{5}$$

Here, $\lfloor\cdot\rfloor$ is the standard floor function. Condition (5) implies that an arriving customer always joins an empty system. Define the normalized benefit of service as

$$\nu \doteq R\mu/c. \tag{6}$$

Lemma 1. *In the pooled system, an arriving customer joins the queue if and only if the number of customers already in the system is* $n \le K-1$, *where*

$$K \doteq \left\lfloor \frac{RN\mu}{c} \right\rfloor. \tag{7}$$

Furthermore, in the pooled system, for $K > N$, *the average sojourn time and social welfare are, respectively, given by*

$$W_p = \frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}i\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} i\rho^i}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)N\lambda}, \tag{8}$$

$$SW_p = \left(1 - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i}\right)RN\lambda - \frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}i\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} i\rho^i}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i}c, \tag{9}$$

where ρ *is as defined in* (1).

Lemma 2. *In the dedicated system, an arriving customer joins the queue if and only if the number of customers already in that subsystem is* $n \le k-1$, *where* k *is as defined in* (5). *Furthermore, in the dedicated system, the average sojourn time and social welfare are, respectively, given by*

$$W_d = \frac{\sum_{i=0}^{k} i\rho^i}{\left(\sum_{i=0}^{k-1}\rho^i\right)\lambda} \quad \text{and}$$

$$SW_d = \left(1 - \frac{\rho^k}{\sum_{i=0}^{k}\rho^i}\right)RN\lambda - \frac{\sum_{i=0}^{k} i\rho^i}{\sum_{i=0}^{k}\rho^i}Nc, \tag{10}$$

where ρ *is as defined in* (1).

Based on Lemmas 1 and 2, hereafter, we refer to k as the *balking threshold* in the dedicated system and K as the *balking threshold* in the pooled system. Note that the balking thresholds satisfy

$$K \ge Nk. \tag{11}$$

3. Analysis

We will first establish our key result. That is, when customers make their own joining decisions, pooling queues can strictly decrease the social welfare even with identical servers and customers. Specifically, we will prove in Theorem 1(a) that such a result arises when $\rho > 1$ and $\nu > \eta$ for some finite η. To do so, we will provide a step-by-step analysis for Theorem 1(a) in Section 3.1. Recall that the social welfare (3) is determined by the throughput and the average sojourn time (or the average number of customers in the system). In light of this, the key findings in Section 3.1 are as follows. The pooled system results in strictly larger throughput than the dedicated system. But, when $\rho > 1$ and $\nu > \eta$, pooling queues increases average sojourn time (average number of customers in the system) so much that the pooled system results in strictly smaller social welfare than the dedicated system.

3.1. Step-by-Step Analysis to Establish Theorem 1(a)

Because the throughput is an important determinant of social welfare by (3), we first present the following result.

Proposition 1. *The dedicated system results in strictly smaller throughput than the pooled system* (*i.e.*, $\theta_d < \theta_p$).

The rationale behind Proposition 1 is as follows. The dedicated system has N subsystems, and each of them is a single-server queueing system with a balking threshold k. Thus, there can be a situation where a customer arrives to a dedicated subsystem and finds out that there are already k customers in the subsystem, whereas other dedicated subsystems have not reached their balking thresholds. Such a situation does not arise in the pooled system because the pooled system has a single queue with balking threshold K, which is more than N times the balking threshold k of the dedicated system by (11). Moreover, the pooled system also reduces idleness; this makes it less likely for the pooled system to operate at the balking threshold than the dedicated system. Because of all these reasons, the pooled system results in strictly smaller balking probability and hence, strictly larger throughput than the dedicated system, as proved in Proposition 1.

Proposition 1 and the form of the social welfare in (3) suggest that the dedicated system can outperform the pooled system in terms of social welfare only when the former has a sufficiently lower average sojourn time (or sufficiently lower average number of customers in the system) that offsets the lower throughput.

We now introduce a "*scaled queueing system*" (i.e., *SQ system*) as a bridge for the comparison between the dedicated and pooled systems. We consider the SQ system because comparing the dedicated and SQ systems or comparing the pooled and SQ systems is analytically more tractable than directly comparing the dedicated and pooled systems.

Definition 1. An *SQ system* is a single-server queueing system indexed by $j = s$ with the following properties. (a) Customers arrive to the system according to a Poisson process with rate λN. (b) The service time has an exponential distribution with rate μN. (c) Each arriving customer balks if and only if the number of customers already in the system is larger than K (as defined in (7)); otherwise, she joins the system.

In the remainder of this section, Proposition 2 will prove that the SQ system always dominates the pooled system (in terms of $\theta_\cdot$, $W_\cdot$, and $SW_\cdot$). Proposition 3 will identify the conditions under which the dedicated system results in strictly larger social welfare than the SQ system. Combining these two results will give us the conditions under which the dedicated system dominates the pooled system in terms of social welfare.

In the following results, θ_s, W_s, and SW_s represent the throughput, average sojourn time, and social welfare in the SQ system, respectively.

Proposition 2 (SQ vs. Pooled). *Compared with the pooled system, the SQ system results in* (a) *strictly larger throughput (i.e., $\theta_s > \theta_p$),* (b) *strictly smaller average sojourn time (i.e., $W_s < W_p$), and* (c) *strictly larger social welfare (i.e., $SW_s > SW_p$).*

There are two key observations related to the SQ system. (i) The SQ and pooled systems have the same balking threshold K. (ii) The service rate in the SQ system is (weakly) larger than the one in the pooled system for any given number of customers in the system, and the former is strictly larger than the latter when the number of customers in the system is small. Then, by (i) and (ii), the SQ system results in strictly smaller average sojourn time than the pooled system, as proved in Proposition 2(b). This and Proposition 2(a) immediately imply Proposition 2(c).

To state Proposition 3, denote by L_s the steady-state average number of customers in the SQ system and recall that L_d is the steady-state average number of customers in each of the N dedicated subsystems.

Proposition 3 (SQ vs. Dedicated). (a) *The dedicated system results in strictly smaller throughput than the SQ system (i.e., $\theta_s > \theta_d$). Moreover, compared with the SQ system, the dedicated system results in* (b) *significantly smaller average number of customers (i.e., $L_s - NL_d > (N-1)/(2(\rho-1)) > 0$) and* (c) *strictly larger social welfare (i.e., $SW_d > SW_s$) if*

$$\rho > 1 \quad \textit{and} \quad \nu > \eta, \tag{12}$$

where ν is as in (6), and η is finite when $\rho > 1$ and does not depend on either R or c.

Let us explain the rationale behind Proposition 3. Part (a) follows from Propositions 1 and 2(a). The condition (12) guarantees that L_s is significantly larger than NL_d, as proved in Proposition 3(b), and $\theta_s - \theta_d > 0$ is small. From this, Proposition 3(c) follows because $SW_d = R\theta_d - cNL_d$ and $SW_s = R\theta_s - cL_s$.

We now explain how (12) guarantees the aforementioned two properties. The condition $\nu > \eta$ implies that $\theta_s - \theta_d$ is small. The reason is when $\nu > \eta$, the balking thresholds in each dedicated subsystem and the SQ system are both very large. Hence, the throughputs θ_s and θ_d are very close to each other. To explain the rationale for Proposition 3(b), let $\pi_d(i)$ ($\pi_s(i)$) denote the steady-state probability of having i customers in a dedicated subsystem (in the SQ system). For $\rho > 1$, steady-state probabilities $\pi_d(i)$ and $\pi_s(i)$ are convex increasing in i, the number of customers. Then, when $\nu > \eta$ in addition to $\rho > 1$, $\pi_s(\cdot)$ puts relatively more weight to larger values of i, compared with $\pi_d(\cdot)$, because the support of $\pi_s(\cdot)$ extends to much larger values of i than $\pi_d(\cdot)$ when $\nu > \eta$. This and (11) immediately imply Proposition 3(b).

Remark 1. (a) Under (12), $W_s > W_d$, which is proved at the end of Online Appendix D. Proposition 3(a) and the form of social welfare imply that under (12), $W_s - W_d$ is so large that we have Proposition 3(c). (b) There exists a constant $\widetilde{\eta}$ such that if $\rho < 1$ and $\nu > \widetilde{\eta}$, $W_d > W_s$ and $SW_d < SW_s$. Proposition EC.1 in Online Appendix E formalizes this result.

3.2. The Formal Statement and Discussion of Theorem 1

In light of Propositions 1–3, we now state our key result.

Theorem 1. *There exist constants η and $\bar{\eta}$ such that the following results hold.*

a. *The dedicated system results in* (i) *strictly smaller average sojourn time and* (ii) *strictly larger social welfare than the pooled system: that is, $W_d < W_p$ and $SW_d > SW_p$, respectively, if*

$$\rho > 1 \quad \textit{and} \quad \nu > \eta, \tag{13}$$

where η is as in Proposition 3, and ρ and ν are defined in (1) and (6), respectively.

b. *The pooled system results in (i) smaller average sojourn time and (ii) strictly larger social welfare than the dedicated system: that is, $W_p \le W_d$ and $SW_p > SW_d$, respectively, if either*

$$\nu < (N+1)/N \tag{14}$$

or

$$\rho < 1 \qquad \textit{and} \qquad \nu > \bar{\eta}, \tag{15}$$

where $\bar{\eta}$ is finite when $\rho < 1$, and does not depend on either R or c.

In our setting, which allows for customer balking in a stable system, if (13) holds, $W_d < W_s < W_p$ and $SW_d > SW_s > SW_p$ (by Propositions 2 and 3 and Remark 1(a)). Thus, we have Theorem 1(a). It is worth noting that the classical understanding, which assumes no customer balking in a stable system, suggests that $W_s < W_p < W_d$ and $SW_s > SW_p > SW_d$, where $j = s$ here is the modified scaled system that assumes no balking and satisfies properties (a) and (b) in Definition 1.[6]

We now explain the conditions in Theorem 1(b). If (14) holds, a joining customer immediately enters the service in both dedicated and pooled systems because $k = 1$ and $K = N$ under that condition. Thus, dedicated and pooled systems have the same average sojourn time W_j and provide the same long-run average net benefit to each joining customer. This and strictly larger throughput in the pooled system (by Proposition 1) imply strictly larger social welfare for the pooled system if (14) holds.

The conditions in (15) can be explained as follows. When the benefit is large (i.e., $\nu > \bar{\eta}$), balking thresholds are large in both systems. With a relatively small potential load (i.e., $\rho < 1$), dedicated and pooled systems barely achieve their balking thresholds, implying very small expected number of balking customers for both systems. Thus, the pooled and dedicated systems have very close throughputs under (15). Moreover, under these conditions, there is significant idleness in the dedicated system. As a result, pooling results in smaller average sojourn time by reducing idleness in the system. This and Proposition 1 imply larger social welfare for the pooled system.

3.3. Numerical Comparison of the Pooled and Dedicated Systems

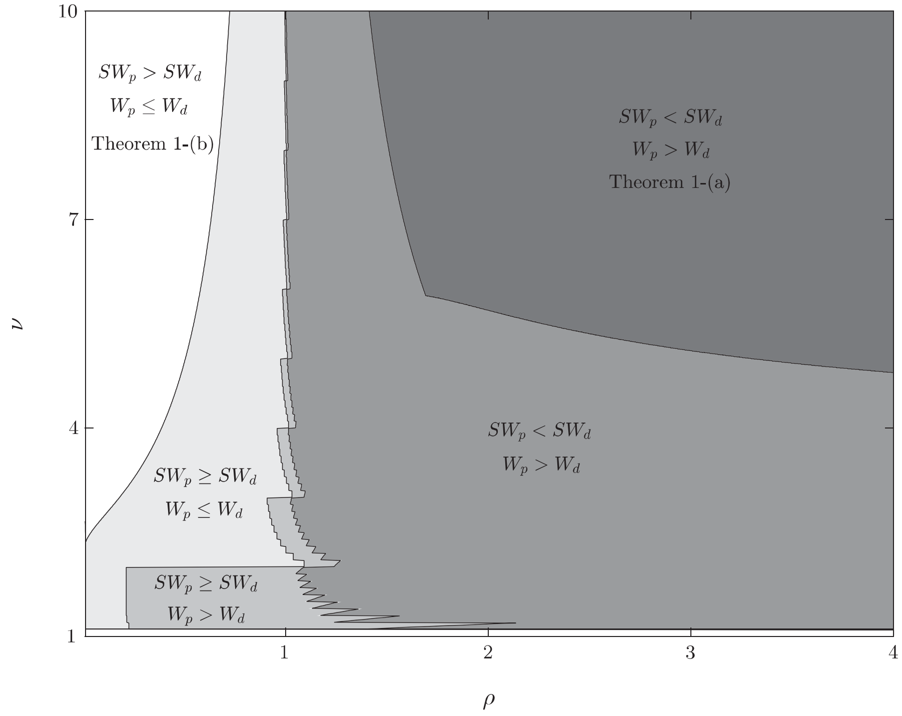
Figure 1 pictures the conditions under which the dedicated system outperforms the pooled system for a numerical example. From this example, we can observe that for a wide range of parameters, the dedicated system results in strictly larger social welfare than the pooled system. As indicated in the figure, the darkest-shaded region corresponds to the parameter space identified by (13) in Theorem 1(a). We can see that the sufficient condition (13) constitutes a large portion of the parameter set in which $SW_d > SW_p$. (Similar figures can also be obtained for other N values; see, e.g., Figure EC.1 in Online Appendix N for a comparison of pooled and dedicated systems when $N = 2$.)

Figure 1 suggests that, for a given service rate, $SW_d > SW_p$ if and only if ν is not too small and ρ is larger than a threshold. Moreover, the parameter region in which $SW_d > SW_p$ is a subset of the parameter region in which $W_d < W_p$. This is because if the dedicated system results in larger social welfare than the pooled system at a given ρ, then, by (3) and Proposition 1, it must also result in strictly smaller average sojourn time than the pooled system at the same ρ.

Observe from Figure 1 that it is possible to have $W_p > W_d$ when $\rho < 1$. For other numerical examples, we also observed that it is possible to have $SW_d > SW_p$ for $\rho < 1$ when ν is not too large. This means that $\rho > 1$ is not a necessary condition for the superior performance of the dedicated system. However, we would like to note that the parameter region in which $SW_d > SW_p$ with $\rho < 1$ is much smaller compared with the one with $\rho > 1$. This implies that when $\rho < 1$, the traditional superiority of the pooled system over the dedicated one is mainly recovered.

Theorem 1 and Figure 1 demonstrate that λ plays an important role in the comparison between the

Figure 1. Comparison of Pooled and Dedicated Systems when $c = 1$, $\mu = 1$, and $N = 10$



Note. Displayed boundaries between regions, except the one for the darkest-shaded region, have jumps because of the floor function in k and K.

dedicated system and the pooled system through ρ. Figure 2 sheds more light on the effect of ρ on the comparison between the pooled and dedicated systems.

Figure 2 demonstrates W_p/W_d and SW_d/SW_p for a numerical example with a given service rate μ. A key message from this figure is that the dedicated system can result in significantly larger social welfare than the pooled system for large ρ. In fact, for this example, as $\rho \to \infty$, SW_d/SW_p converges to approximately 65. The value $\lim_{\rho\to\infty} SW_d/SW_p$ can be analytically verified as follows. For a given service rate, as $\rho \to \infty$, both dedicated and pooled systems mostly operate at their balking thresholds. This means that in the limit (i.e., $\rho \to \infty$), every joined customer joins as the last customer before the system reaches its balking threshold and hence, experiences the longest (feasible) expected sojourn time in the system with probability 1. Furthermore, as $\rho \to \infty$, because servers are busy with probability 1, throughputs of the dedicated and pooled systems are the same and equal to the total service rate ($N\mu$). Combining these and (3), we conclude that $\lim_{\rho\to\infty} SW_d = N\mu(R - ck/\mu)$ and $\lim_{\rho\to\infty} SW_p = N\mu(R - cK/(N\mu))$.[7] Among other properties, the steep increase in W_p/W_d around $\rho = 1$ in Figure 2 shows that when ρ is close to 1, the average sojourn time can increase in the potential load significantly faster under the pooled system, than under the dedicated system. This increase eventually leads to welfare loss under pooling. Note from Figure 2 that the explained sojourn time phenomenon cannot be observed as $\rho \to 0$ or $\rho \to \infty$. (The aforementioned sojourn time observations are analytically verified by (EC.73), (EC.75), and (EC.76) in Lemma EC.7, which is in Online Appendix G.) Overall, Figure 2 underscores the importance of judiciously evaluating the pooled and dedicated systems for services, as the relative performance of a system can be very sensitive to a change in ρ.

Remark 2. Any result or numerical observation about SW_d/SW_p and W_p/W_d can easily be expressed in terms of the percentages $\beta_{SW} \doteq ((SW_d - SW_p)/SW_p) \times 100\%$ and $\beta_W \doteq ((W_p - W_d)/W_d) \times 100\%$, respectively, because $\beta_{SW} = SW_d/SW_p - 1$ and $\beta_W = W_p/W_d - 1$.

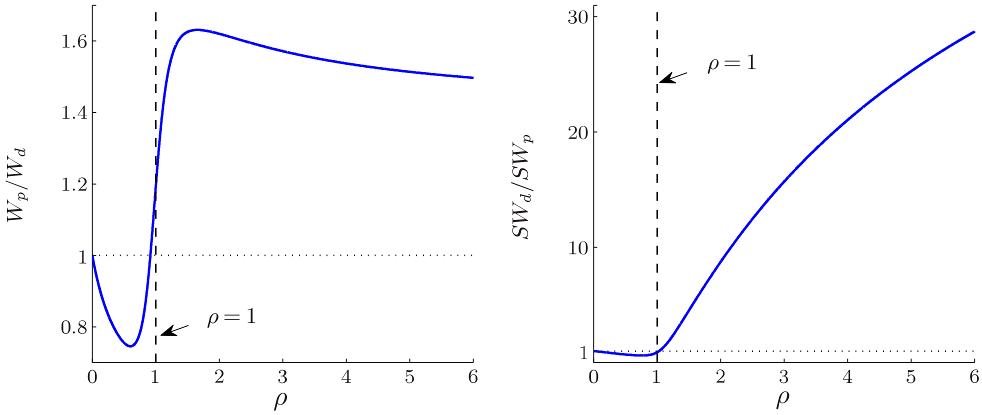
3.4. The Impact of Number of Servers *N* and the Magnitude of Performance Gain

We now analyze the impact of N on the comparison of pooled and dedicated systems. Consider a sequence of systems indexed by $n = \{2, 3, \ldots\}$ such that in the nth system, there are $N = n$ servers, and the total potential arrival rate in the system is $n\lambda$. In this context, n can be seen as a proxy for the *system size*. In light of this and Remark 2, Theorem 2 proves that when customers make their own joining decisions, pooling a larger system results in *larger* percentage *loss* in social welfare under certain conditions. In Theorem 2 and related discussions, we will include N as an argument of the performance metric under consideration to emphasize its dependence on N. Among the considered metrics, $W_d(N)$ is the only one that does not change with N; we include N as an argument of W_d just for notational consistency.

Theorem 2. *Suppose that ν is an integer.* (a) *There exists a constant η_1 such that $W_p(N)/W_d(N)$ is larger than 1 and strictly increases in the system size if $\rho > 1$ and $\nu > \eta_1$. The constant η_1 is finite when $\rho > 1$ and does not depend on either R or c.* (b) *There exists a constant η_2 such that $SW_d(N)/SW_p(N)$ is larger than 1 and strictly increases in the system size if $\rho > 1$ and $\nu > \eta_2$. The constant η_2 is finite when $\rho > 1$ and does not depend on either R or c.*

Remark 3. For expositional brevity, Theorem 2 states the result for integer ν. Online Appendix H provides a proof for the generalized version of Theorem 2 that is also valid for noninteger ν.

Figure 2. (Color online) Sensitivity of the Performance Ratios with Respect to ρ



Note. The following parameters are used: $R = 75$, $c = 4$, $N = 10$, and $\mu = 0.15$.

There are two main drivers of Theorem 2. (i) In both the dedicated subsystem and pooled system, the stationary probability of having l customers in the system is convex and increasing in l for $\rho > 1$, and (ii) the balking threshold in the pooled system increases in N. Based on (1) and (ii), when there is a performance loss because of pooling, the loss is exacerbated even more with an increase in the system size.

It is well established in the literature that pooling queues in a larger system provides larger performance benefits. For instance, Benjaafar (1995) demonstrates that when there is no balking, the average delay decreases with the system size when multiple M/M/1 systems are combined and run as a pooled system (see Kulkarni 2010 and Sztrik 2012 for the fundamentals of the M/M/1 queueing system). An important implication of this observation in their setting is that the social welfare benefit of pooling also increases with the system size. In contrast, by Remark 2, Theorem 2 proves that percentage *gain* in social welfare due to separating queues can strictly increase with the system size when customers make their own joining decisions. The reason for the contrast between Theorem 2 and the classical finding in Benjaafar (1995) about the effect of system size on the benefit of pooling is the following. In Benjaafar (1995), customers are not delay-sensitive and join the system regardless. Therefore, that study assumes an infinite buffer size and $\rho < 1$ for stability. In contrast, our paper considers rational joining decisions of delay-sensitive customers (which imply a finite balking threshold) and allows for $\rho > 1$.

Figure 3 displays for a numerical example the impact of the system size on the performance ratios when ν is not an integer. Figure 3 shows that when ν is not an integer, the ratios demonstrate a generally increasing trend in N. Online Appendix H provides a proof for this observation. Note also that both of these ratios display a certain nonmonotone pattern in Figure 3. In particular, the ratios increase with N for five data points, switch to a different level, and then increase with N for five more data points. This pattern repeats itself. Such a pattern is observed when ν is not an integer (because of the floor function in k and K); if ν is an integer, both ratios strictly increase with N under the conditions identified in Theorem 2.

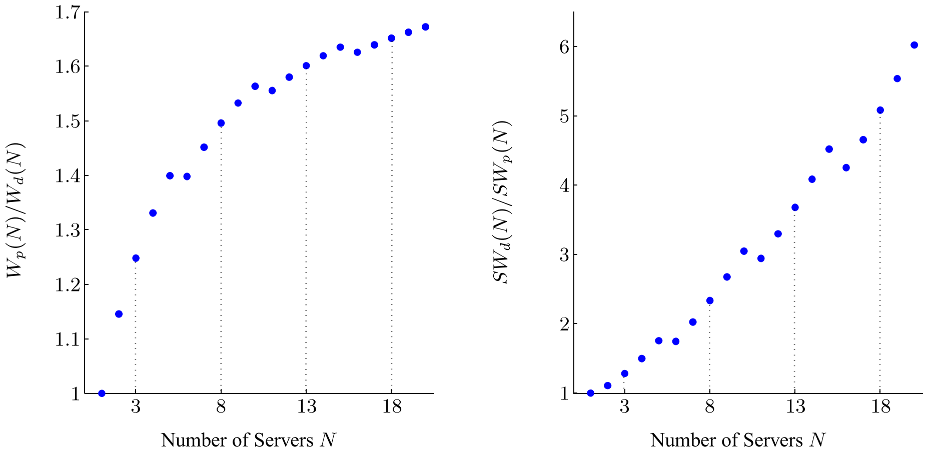
Theorem 2 and Remark 2 proved that for a given R, social welfare improvement under the dedicated system strictly increases with the system size under certain conditions. This naturally brings forth the following question: compared with the pooled system, how much can the dedicated system improve the social welfare SW? Theorem 3 answers this question by identifying a lower bound on the achievable SW_d/SW_p. In the statement of Theorem 3, we include R as an argument of $SW(\cdot)$ to emphasize its dependence on R.

Theorem 3. *The social welfare ratio satisfies the following for $\rho > 1$:*

$$\max_R \{SW_d(R)/SW_p(R)\} > (N-1). \tag{16}$$

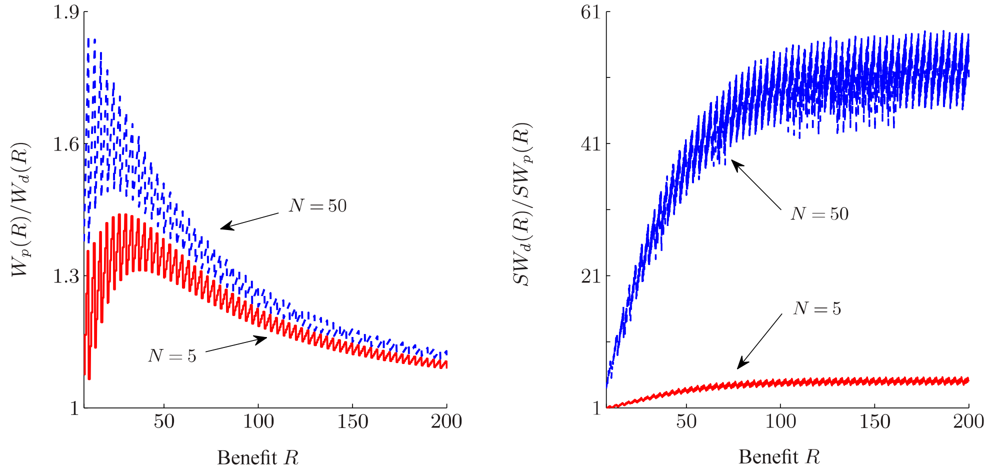
Theorem 3 and Remark 2 imply that there exist systems in which separating queues improves the social welfare by more than $(N-2)100\%$ compared with pooling them, and R is a key determinant of the existence of such systems. Figure 4 displays plots for two systems that satisfy (16). Based on Theorem 3 and Figure 4, we can see that operating a system with dedicated queues rather than a pooled one can significantly improve the social welfare even in small-scale systems. For example, in this example, the dedicated system can improve the social welfare as large as 466% and 5,710.2% when $N = 5$ and $N = 50$, respectively. Figure 4 suggests that the dedicated system can achieve the large performance gain on the

Figure 3. (Color online) The Effect of the Number of Servers N on the W and SW Ratios



Note. The parameters are as follows: $\lambda = 0.35$, $c = 1$, $\mu = 0.3$, and $R = 16$.

Figure 4. (Color online) The Effect of the Benefit R on the W and SW Ratios



Notes. The parameters are as follows: $\lambda = 0.35$, $c = 1$, and $\mu = 0.3$. In this example, the displayed functions are nonsmooth because of the floor function in k and K. Total potential arrival rate in the system is λN. Thus, N can be seen as the scale of the system.

right-hand side of (16) when R is large. However, welfare benefit of the dedicated system can still be significant when R is not very large. For example, when $N = 50$, the percentage improvement in social welfare under the dedicated system is larger than 100% for $R \geq 7.5$. The reason is that pooling queues can drastically increase the average sojourn time W even at moderate benefit R; in fact, the maximum percentage increase in W is typically observed at moderate R, as suggested by Figure 4 and Remark 2.

An important corollary of Theorem 3 is the following. Compared with the pooled system, the dedicated system may improve social welfare in a way that the percentage improvement in social welfare eventually takes values larger than any fixed value as the system size goes to infinity. Thus, the achievable percentage improvement in the social welfare under the dedicated system (compared with the pooled system) can indeed be drastically large for very large systems.

4. Extensions

In this section, we aim to check the robustness of our key result in Theorem 1(a). This analysis will also help us further investigate what causes the dedicated system to outperform the pooled system in terms of social welfare.

4.1. Optimal Pricing

Suppose that each customer pays a fee upon service completion in the system $j \in \{d, p\}$, and this service fee is set to either maximize the service provider's revenue or the social welfare. All other modeling elements are the same as in Section 2.

Based on this, we will compare the pooled and dedicated systems under the following two formulations.

i. *Welfare maximization.* For each system $j \in \{d, p\}$, a service fee f_j is set to maximize the social welfare:

$$\max_{f_j \geq 0} \; SW_j \doteq \left(R - cW_j(f_j)\right)\theta_j(f_j), \quad j \in \{d, p\}, \tag{17}$$

where $\theta_j(\cdot)$ is the throughput in the system $j \in \{d, p\}$. The social welfare does not include the term $f_j\theta_j(f_j)$ because the total collected fee is just a transfer between customers and the fee collector.

ii. *Revenue maximization.* For each system $j \in \{d,p\}$, a service fee f_j is set to maximize the service provider's revenue:

$$\max_{f_j \geq 0} \; RV_j \doteq f_j\theta_j(f_j), \quad j \in \{d, p\}. \tag{18}$$

Proposition 4. (a) *Under the welfare maximization formulation* (17), *the maximum social welfare in the pooled system is greater than or equal to that in the dedicated system.* (b) *Under the revenue maximization formulation* (18), *the maximum revenue in the pooled system is greater than or equal to that in the dedicated system.*

Let us explain the rationale behind Proposition 4(a). By setting the fee, the social planner prevents the system from becoming too congested, and hence, customers cannot overutilize the system.[8] In that case, the pooled system improves the system efficiency by reducing idleness in the system. Thus, under formulation (17), in the pooled system, by setting the fee, the social planner not only changes customers' joining behaviors in a socially optimal way but also achieves the system efficiency. As a result, as proved in Proposition 4(a), the pooled system outperforms the dedicated system when formulation (17) is considered.

Proposition 4(b) follows from two facts. (i) For any fixed service fee, the pooled system results in strictly larger throughput than the dedicated system. The rationale behind this fact is the same as the one explained for Proposition 1. This fact implies that when fees in both systems are set to any given fee (e.g., the optimal fee for the dedicated system), the revenue under the pooled system is strictly larger than that under the dedicated system. (ii) Under formulation (18), the optimal fee for the dedicated system is feasible but not necessarily optimal for the pooled system.

In light of the setting in Hassin (1986), it might also be of interest to analyze the social welfare under formulation (18). One can show that under formulation (18), the social welfare under the pooled system is strictly smaller than the one under the dedicated system when ν is moderate.

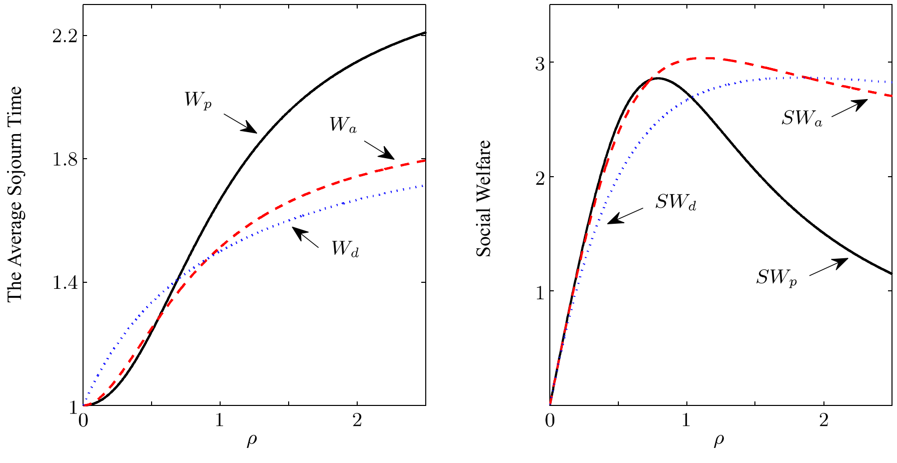
4.2. Join-the-Shortest-Queue Policy

In the base model described in Section 2, there is a separate arrival stream for each queue in the dedicated system. Consider an alternative dedicated queueing system where an arriving customer observes the number of customers in each of N queues and then decides whether to join a queue or balk. This system is called the *alternative system* and denoted by the index $j = a$. In the alternative system, if an arriving customer decides to join, she optimally chooses the shortest queue.[9] In this setting, an arriving customer optimally balks if and only if each of the N dedicated subsystems already has k customers.

It is well known that the exact analysis of the *join-the-shortest-queue* (*JSQ*) *policy* in a first-come, first-served queueing system is typically intractable (Gupta et al. 2007).[10] Thus, the vast majority of the literature focuses on approximations or numerical analysis to evaluate the performance of the JSQ policy (see, e.g., Grassmann 1980, Rao and Posner 1987, and Nelson and Philips 1993). In light of this, we will present our numerical insights in this section.

Figure 5 depicts social welfare performances of three systems (i.e., pooled, dedicated, and alternative systems) for a numerical example.[11] Note from this figure that under certain conditions, the alternative system outperforms the pooled system, and thus, the key result in Theorem 1(a) extends to this setting. Figure 5 provides a key observation. Among the three systems, the one that achieves the largest social welfare is (a) the pooled system when the potential system load ρ is small (i.e., $\rho \in (0, 0.74]$), (b) the alternative system when ρ is moderate (i.e., $\rho \in (0.74, 1.88]$), and (c) the dedicated system when ρ is large (i.e., $\rho > 1.88$). We observed this structure for many additional numerical examples as well. We now explain the rationale behind this observed structure. In this example, the throughput under the system $j \in \{d, p, a\}$ satisfies $\theta_p \geq \theta_a > \theta_d$ for any ρ, and the positive throughput difference between any two systems increases with ρ for small ρ. Based on this, the reason for (a) is as follows. When ρ is small, the average sojourn time under the pooled system is either the smallest among the three systems or very close to the ones under the other two systems (see left panel of Figure 5). When it is the former, (a) follows from the aforementioned throughput relation. When it is the latter, (a) holds because the throughput under pooled system is considerably larger than the one under any of the other two systems. There are two reasons for (b). (i) When ρ is moderate, the pooled system yields much larger average sojourn time W than the alternative system. The resulting increase in W under the pooled system is so large that the alternative system results in strictly larger social welfare than the pooled system. (ii) When ρ is moderate, compared with the dedicated system, the alternative system significantly improves the throughput by giving customers more discretion in their joining decisions. The significantly larger throughput in the alternative system translates into the larger social welfare for the alternative system.

Figure 5. (Color online) The Effect of ρ on the Average Sojourn Time and Social Welfare



Note. The parameters are as follows: $\lambda = 0.35$, $c = 1$, and $\mu = 0.3$.

Finally, (c) holds because relatively, the average sojourn time under the dedicated system is much lower than the one under the pooled or the alternative system.

4.3. Partial Pooling

This section allows for partial pooling, which refers to combining only some of the separate queues (instead of all queues) to form a single line. To focus on reasonable number of partial pooling scenarios, this section considers symmetric partial pooling, meaning that each pooled subsystem (within the partially pooled system) contains the same number of servers. Thus, under partial pooling, every M (which can be any divisor of N except one or N) queues of the N queues are combined into a separate single queue.

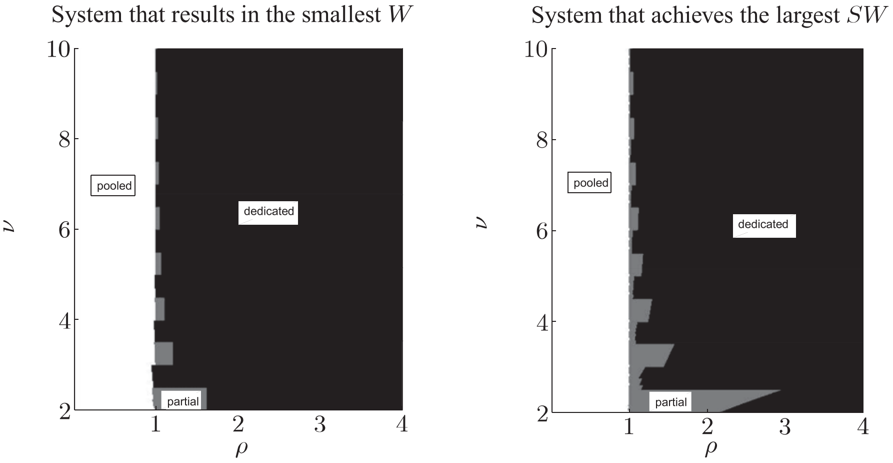
Figure 6 pictures a numerical example that allows for partial pooling. This figure demonstrates that the dedicated system generates the maximum social welfare among all systems when ν is not too small and ρ is larger than a threshold. Theorem 2(b) has an important implication for this setting: the dedicated system outperforms partial pooling in terms of social welfare when the conditions in Theorem 2(b) hold. In fact, under these conditions, the dedicated system performs the best among all system designs. Thus, the dedicated system outperforms all other system designs under conditions similar to the ones in Theorem 1(a). The reason why our key result—the superiority of dedicated system under some conditions—extends to the partial pooling setting is as follows. When a system is partially pooled, there is no interaction between distinct subsystems of pooled queues, and thus, each subsystem of pooled queues can be viewed as an independent system of pooled queues. Consequently, the comparison between a partially pooled system and its dedicated counterpart is equivalent to the comparison between each pooled subsystem and the dedicated counterpart of that subsystem.[12]

Let us explain when partial pooling performs the best among all system designs. From Figure 6, we observe that partial pooling outperforms both the dedicated and pooled systems when ν is moderate and ρ is strictly larger than one but not very large. For example, in Figure 6, at $\nu = 2.2$ and $\rho = 1.8$, partial pooling with $M = 2$ maximizes the social welfare among all systems. The reason behind these numerical observations is as follows. Under the aforementioned conditions, the system throughput concavely increases with M. More specifically, throughput significantly increases with M when M grows from $M = 1$ to small values of $M > 1$ (e.g., $M = 2$), whereas it does not change too much with M for moderate or large values of M. Thus, compared with the dedicated system, any partially pooled system or the pooled system results in significantly larger throughput. Moreover, under the stated conditions, compared with the dedicated system, complete pooling (i.e., $M = N$) significantly increases the average sojourn time, whereas a partially pooled system with a small or moderate M results in the average sojourn time being very close to the one under the dedicated system. Thus, under the stated conditions, when separate queues are partially pooled, the system throughput significantly increases without a considerable increase in the average sojourn time. As a result, partial pooling maximizes social welfare when ν is moderate, and ρ is strictly larger than one but not very large.

4.4. Unobservable System

This section studies an unobservable system setting where the queue length information or real-time expected delay information is not available to customers. (See Hassin and Haviv 2003 and Hassin 2016 for literature reviews on unobservable service systems. Also, see, e.g., Afèche 2013, Yang et al. 2017, and Ravner and Shamir 2020 for some of the novel problems studied in that context.) All other modeling

Figure 6. The Following Parameters Are Used: $c = 1$, $\mu = 1$, $N = 20$



elements are the same as the ones in Section 2. There are two key insights in this section: if queue length information is not available to the customers, (i) the social welfare under the pooled system is greater than or equal to the social welfare under the dedicated system when the service is free of charge, and (ii) the pooled system still dominates the dedicated system in terms of social welfare (revenue) when a fee is set to maximize social welfare (the provider's revenue).

Our formulation is as follows. Arriving customers decide whether to join or balk based on the potential arrival rate Λ_j and average sojourn time in the system $j \in \{d, p\}$. A customer's joining/balking strategy is determined by a joining probability q_j for $j \in \{d, p\}$; a customer joins the queue with probability q_j and balks with probability $(1 - q_j)$. The unique equilibrium in this setting is characterized and explained in Online Appendix K.1; that section also includes additional details about the setting.

Proposition 5. (a) *When service is free of charge, in equilibrium, the social welfare under the unobservable pooled system is greater than or equal to the social welfare under the unobservable dedicated system.* (b) *When the fee is set to maximize the social welfare (the provider's revenue) for each system $j \in \{d, p\}$, in equilibrium, the social welfare at the welfare-maximizing fee (at the revenue-maximizing fee) under the unobservable pooled system is greater than or equal to the social welfare at the welfare-maximizing fee (at the revenue-maximizing fee) under the unobservable dedicated system.*

The proof of Proposition 5 is presented for a more general case with any given fixed fee $f \geq 0$. Because R is constant, setting a fee to maximize social welfare is equivalent to setting a fee to maximize the provider's revenue; in the latter case, the provider can extract the entire consumer surplus. An immediate corollary of Proposition 5(b) is that when the fee is set to maximize the provider's revenue, the maximum revenue under the unobservable pooled system is larger than that under the unobservable dedicated system in equilibrium.[13] This and Proposition 5 suggest that the observability of queue (or customers having access to their real-time expected delay information) is a necessary condition for the dedicated system to outperform the pooled system in terms of social welfare.

Proposition 5 is in contrast with Theorem 1(a) that studies the observable queue setting. The reason is as follows. When queues are unobservable, (i) the throughput under the dedicated system is (weakly) smaller than the one under the pooled system (see the proof of Proposition 5(a)) and (ii) the average sojourn time under the pooled system is smaller than the one under the dedicated system in equilibrium. Different from Theorem 1(a), in this setting, we have (ii) because in equilibrium, every customer's joining probability is such that the effective system load is always strictly smaller than one. Because of this and the fact that the buffer size is unlimited in the unobservable system, pooled and dedicated systems in the unobservable case behave similar to the ones studied by Smith and Whitt (1981). Hence, the classical benefit of pooling is recovered in the unobservable setting.

4.5. Observability as a System Feature

There could be practical scenarios in which running systems with observable or unobservable queues are both feasible options. In such cases, a system can be run in one of the following four alternative ways: pooled observable, pooled unobservable, dedicated observable, and dedicated unobservable. Considering these four alternatives, we proved Proposition 6. This result and the discussion following it complement the literature that studies an M/M/1 setting to understand if revealing queue length information improves the social welfare (see, for instance, Hassin 1986, Hassin and Roet-Green 2017, and Hu et al. 2018).

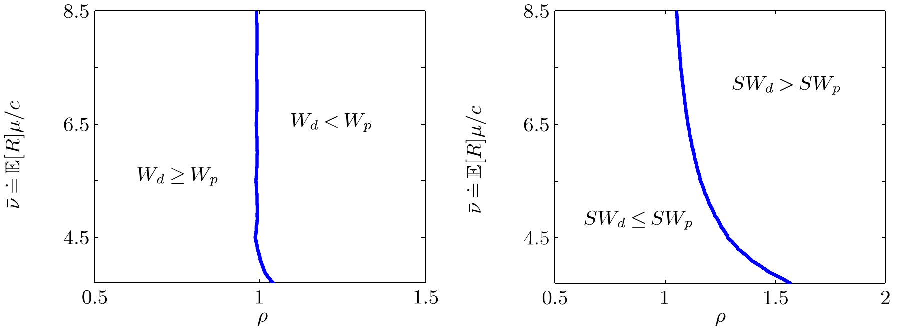
Proposition 6. (a) *When there is no service fee, the social welfare under the observable pooled system is greater than or equal to the social welfare under both the unobservable pooled system and th unobservable dedicated system in equilibrium.* (b) *When a fee is set to maximize the social welfare in each system, the observable pooled system results in the maximum social welfare among the four systems.*

A key implication of Proposition 6(a) is that when there is no service fee, hiding queue length or the real-time expected delay information never improves the social welfare (or consumer surplus). Furthermore, Theorem 1(a) and Proposition 6(a) imply that when there is no service fee, the observable dedicated system results in maximum social welfare among the four systems if the conditions in (13) hold. By Proposition 6(b), when all of the four systems are feasible options and a service fee is set to maximize the social welfare, hiding queue length cannot be welfare-maximizing. This insight differs from the one in Hassin (1986). The reason is that Hassin (1986) considers a different setting than ours; Hassin (1986) compares the welfare under observable and unobservable M/M/1 systems when the service provider sets a fee to maximize its revenue.

4.6. Heterogeneous Benefits

This section extends the results in Theorem 1(a) and Proposition 5(a) to a setting where any two customers can receive different benefits from the same service. An important insight is that the dedicated system can outperform the pooled one in the observable setting with heterogeneous service benefits. Figure 7

Figure 7. (Color online) The Following Parameters Are Used: $c = 1$, $\mu = 1$, $N = 5$, $R \sim U[r, r + 1]$ for $r > 0$



demonstrates this via a numerical example where customer benefit R is uniformly distributed on $[r, r + 1]$ for $r > 0$, and each customer's service benefit is independent of others' benefits. When the variance of R is small, the customer's joining behavior in the system $j \in \{d, p\}$ is similar to the one with the homogeneous service benefit (in Section 3). Thus, in Figure 7, similar to the observations in Section 3, the dedicated system outperforms the pooled system when ρ and the expected normalized benefit of service ($\mathbb{E}[R]\mu/c$) are large. Our further numerical studies suggest that the parameter region in which the dedicated system outperforms the pooled system gets smaller when the variance of R increases. The reason is as follows. When the variance of R is larger, the variance in balking thresholds is also larger and thus, compared with the setting in Section 3, fewer customers join a system when it is very congested. Hence, when the variance of R increases, the dedicated system's benefit, that is, preventing customers from experiencing long real-time expected delay in an already-congested system, gets smaller.

Apart from these, Proposition EC.2 in Online Appendix M extends Proposition 5(a) by proving that the unobservable pooled system performs better than the unobservable dedicated system when service benefits are heterogeneous. The rationale behind this result is as follows. In the unobservable setting, there exists a threshold benefit for each system $j \in \{d, p\}$ such that in equilibrium, only customers with a benefit larger than the threshold join that system. The threshold benefit determines the effective arrival rate in each system. When unobservable pooled and dedicated systems are run with the equilibrium effective arrival rate of the latter system, the classical benefit of pooling is recovered as pooling reduces the average sojourn time $\widehat{W}$. (This is because in an unobservable system $j \in \{d, p\}$, there is no limit on the buffer size, and the effective arrival rate is smaller than the service rate, as in the classical case.) Based on this, if the unobservable pooled and dedicated systems are run with the dedicated system's equilibrium threshold benefit (which implies the same effective arrival rate under both systems), the unobservable pooled system results in larger expected net benefit than the unobservable dedicated system. As a result, customers in the unobservable pooled system are more likely to join than the ones in the unobservable dedicated system, resulting in a smaller equilibrium threshold benefit under pooling. Then, because the threshold benefit equals customer's average waiting cost $c\widehat{W}$ under each system in equilibrium, pooling also results in smaller $c\widehat{W}$ in equilibrium. Combining the effect of joining threshold on the effective arrival rate and the aforementioned ordering of $c\widehat{W}$ under the two systems in equilibrium, Proposition EC.2 immediately follows.

5. Discussions on Other Operational Levers for Performance Improvement

Theorem 1(a) establishes the superiority of the dedicated system over the pooled one under (13). When pooling queues is inevitable and pricing control is not feasible for an observable queueing system, under (13), there could be other operational levers to improve social welfare and consumer surplus in the pooled system. (Social welfare and consumer surplus are equal in our setting.)

In some practical settings, the number of servers could be a feasible operational lever. By changing the number of servers in the pooled system, one might improve the social welfare (and customer surplus) under the pooled system. In fact, based on our numerical study, when the dedicated system outperforms the pooled system, by sufficiently increasing the number of servers in the pooled system, the performance of the pooled system achieves or exceeds the performance of the dedicated system. We also numerically observe that the minimum number of additional servers required in the pooled system to achieve or exceed the dedicated system's performance increases with the potential system load. This means that, when there is an increase in the potential system load, more server addition is necessary for the pooled system to perform as well as the dedicated system.

Adding servers is typically a costly strategy. If the cost of adding servers exceeds the overall improvement in the consumer surplus, which is an important measure of customer satisfaction, adding servers might not be economically justified. Thus, in such cases, it might be optimal to operate the system with a large potential load (as in (13)). For example, in the context of customer service, not being able to receive service from a particular channel, say, a call center, is not always equivalent to giving up service entirely. In various practical settings (e.g., e-commerce), a customer who does not join the call center queue can still receive service via a less desirable alternative channel such as a web form or email, which might provide less value to customers. The strategy of trying to meet some of the customer service demand through these less desirable channels instead of adding servers to the call center could be a reason why some call centers might operate with a very large potential load. In fact, such a strategy could be justifiable based on the capacity cost: if the cost of adding servers to the call center is larger than the potential improvement in the consumer surplus due to that addition, having less desirable customer service channels might be more favorable than adding servers to the call center. On the other hand, if the cost of having additional servers is very small, the system might benefit from adding servers to the call center and eliminating the less desirable customer service channels.

Another operational lever for performance improvement in the pooled system could be limiting the queue length under pooling. For settings in which pricing is not feasible, this can be implemented by setting a buffer size and rejecting arrivals after the buffer is full. Our numerical studies suggest that when the potential system load is large, the social welfare (and consumer surplus) in the pooled system can be improved by choosing an appropriate buffer size for the pooled system. Such a queue length control improves the system performance by mitigating overutilization in the system, especially when the potential system load is very large. Our numerical studies also show that the appropriate buffer size for the pooled system is a moderate one, which is typically smaller than the maximum number of customers in the dedicated system (see, e.g., Figure EC.2 in Online Appendix N). We note that choosing a buffer size to maximize the social welfare is equivalent to setting a service fee to maximize the social welfare. Thus, Proposition 4 implies that applying an admission control in the pooled system, either in the form of limiting the buffer size or imposing the socially optimal price, restores the classical performance superiority of pooling. If neither of these operational levers is feasible and (13) holds, running the system as a dedicated one (rather than a pooled one) strictly improves the social welfare.

6. Concluding Remarks

Our paper provides key insights for service management. Our analysis (in Theorems 1–3) suggests that the pooling option should be evaluated very carefully in queueing systems. This is because our results show that when customers' joining decisions are considered, pooling queues may significantly hurt the social welfare (and consumer surplus) by considerably increasing the average sojourn time. Services with large benefit and large potential load are particularly prone to the potential harm of pooling when the queue length information is available to the customers and admission control (e.g., monopoly pricing) is not feasible. For these types of services, the magnitude of the performance loss due to pooling can be even larger for larger systems.

Acknowledgments

The authors thank the Department Editor Terry Taylor, the associate editor, and three referees for their valuable feedback.

Endnotes

[1] Our model considers identical servers to tease out the effect of customers' joining decisions; heterogeneous servers were already observed to cause pooling to potentially perform worse than the dedicated system.

[2] Considering a dedicated arrival stream for each server is common in the formulation of dedicated queueing systems. See, for instance, Smith and Whitt (1981) and Yu et al. (2015). Section 4.2 demonstrates that if customers are allowed to choose the shortest queue in the dedicated system, the key phenomenon proved in Theorem 1(a) extends under certain conditions.

[3] Recall footnote 2.

[4] See, for example, Debo and Veeraraghavan (2014) and Cui and Veeraraghavan (2016) for some of the recent papers with this modeling feature.

[5] Similarly, the M/M/1 system studied by Naor (1969) is stable regardless of ρ. This property was further explained by Gilboa-Freedman et al. (2014).

[6] When customers cannot balk, throughputs are the same for $j \in \{s, d, p\}$ because every arrival joins. As the service rate is larger in the modified scaled system than in the pooled system for any given number of customers in the system, $W_s < W_p$ in the absence of customer balking. (The proof of this statement is similar to the proof of Proposition 2, and hence omitted.) We already know from Smith and Whitt (1981) that $W_p < W_d$ when customers cannot balk. Combining these, we have $W_s < W_p < W_d$. Then, $SW_s > SW_p > SW_d$ in the absence of customer balking because $SW_j = N\lambda(R - cW_j)$ for $j \in \{s, d, p\}$ in that setting.

[7] For this numerical example, SW_d/SW_p is bounded as $\rho \to \infty$. One can show that for a given service rate μ, the ratio SW_d/SW_p can be unbounded as $\rho \to \infty$ (e.g., when $R\mu/c - k = 1/N$).

[8] Under formulation (17), the service fee affects the social welfare only through the resulting balking threshold.

[9] When multiple queues are in a tie, we assume that the customer chooses the queue with the smallest index. Another way to break the

tie is to pick a queue randomly with equal probability. This alternative tie-breaking rule would not alter any of our insights.

[10] Even the analysis of the JSQ policy with two servers was found to be difficult in the literature (Selen et al. 2016).

[11] This figure uses balance equations to identify the steady-state queue length distribution in the alternative system.

[12] To be more precise, the dedicated system with N servers outperforms a partially pooled system with N servers and N/M symmetric subsystems of pooled queues if and only if the dedicated system with M servers outperforms a completely pooled system with M servers.

[13] This is consistent with numerical observations by Ros and Tuffin (2004) that consider a queueing system with two "divisible" servers.

References

Afèche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing Service Oper. Management* 15(3):423–443.

Andradóttir S, Ayhan H, Down DG (2017) Resource pooling in the presence of failures: Efficiency vs. risk. *Eur. J. Oper. Res.* 256(1):230–241.

Armbrüster T (2006) *The Economics and Sociology of Management Consulting* (Cambridge University Press, Cambridge, UK).

Armony M, Roels G, Song H (2017) Pooling queues with discretionary service capacity. Working paper, NYU Stern School of Business.

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

Azziz R (2014) Implementing shared services in higher education. Accessed July 8, 2018, https://www.universitybusiness.com/article/shared-services.

Benjaafar S (1995) Performance bounds for the effectiveness of pooling in multi-processing systems. *Eur. J. Oper. Res.* 87(2):375–388.

Bondarouk T (2014) *Shared Services as a New Organizational Form* (Emerald Group Publishing, Bingley, UK).

Calabrese JM (1992) Optimal workload allocation in open networks of multiserver queues. *Management Sci.* 38(12):1792–1802.

Campbell Public Affairs Institute (2017) Considering Shared Government Services in New York State: A Guide for Citizens and Public Officials. Accessed July 8, 2018, https://www.maxwell.syr.edu/uploadedFiles/campbell/NYS-Shared-Services-Guide.pdf.

Cattani K, Schmidt GM (2005) The pooling principle. *INFORMS Trans. Ed.* 5(2):17–24.

Cui S, Veeraraghavan S (2016) Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Sci.* 62(12):3656–3672.

Debo L, Veeraraghavan S (2014) Equilibrium in queues under unknown service times and service value. *Oper. Res.* 62(1):38–57.

Do H, Shunko M, Lucas MT, Novak D (2015) On the pooling of queues: How server behavior affects performance. Working paper, Foster School of Business, University of Washington.

Financial Times (2015) Big ships leave ports awash with problems. www.ft.com/content/ce9a61e0-8705-11e4-982e-00144feabdc0.

Gai Y, Liu H, Krishnamachari B (2016) A packet dropping mechanism for efficient operation of M/M/1 queues with selfish users. *Comput. Networks* 98(April 7):1–13.

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

Gilbert SM, Weng ZK (1998) Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Sci.* 44(12 part 1):1662–1669.

Gilboa-Freedman G, Hassin R, Kerner Y (2014) The price of anarchy in the Markovian single server queue. *IEEE Trans. Automatic Control* 59(2):455–459.

Grassmann WK (1980) Transient and steady state results for two parallel queues. *Omega* 8(1):105–112.

Gupta V, Harchol Balter M, Sigman K, Whitt W (2007) Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation* 64(9-12):1062–1081.

Hassin R (1985) On the optimality of first come last served queues. *Econometrica* 53(1):201–202.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.

Hassin R (2016) *Rational Queueing*, 1st ed., CRC Series in Operations Research (Chapman & Hall, Boca Raton, FL).

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, vol. 59 (Springer Science & Business Media, New York).

Hassin R, Roet-Green R (2017) The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Oper. Res.* 65(3):804–820.

Haviv M, Oz B (2016) Regulating an observable M/M/1 queue. *Oper. Res. Lett.* 44(2):196–198.

Hong LJ, Xu X, Zhang SH (2015) Capacity reservation for time-sensitive service providers: An application in seaport management. *Eur. J. Oper. Res.* 245(2):470–479.

Hu M, Li Y, Wang J (2018) Efficient ignorance: Information heterogeneity in a queue. *Management Sci.* 64(6):2650–2671.

Ibrahim R (2018) Sharing delay information in service systems: A literature survey. *Queueing Systems* 89(1-2):49–79.

Jouini O, Dallery Y, Nait-Abdallah R (2008) Analysis of the impact of team-based organizations in call center management. *Management Sci.* 54(2):400–414.

Karacostas C (2018) Voting wait time at FAC for primaries 14 times longer than Travis County as a whole. Accessed July 1, 2018, http://www.dailytexanonline.com/2018/03/22/voting-wait-time-at-fac-for-primaries-14-times-longer-than-travis-county-as-a-whole.

Kulkarni VG (2010) *Modeling and Analysis of Stochastic Systems* (CRC Press, New York).

Lu Y, Musalem A, Olivares M, Schilkrut A (2013) Measuring the effect of queues on customer purchases. *Management Sci.* 59(8):1743–1763.

Mader D, Roth DT (2015) Scaling implementation of shared services. Accessed July 8, 2018, https://obamawhitehouse.archives.gov/blog/2015/10/22/scaling-implementation-shared-services.

Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. *Management Sci.* 44(7):971–981.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Nelson RD, Philips TK (1993) An approximation for the mean response time for shortest queue routing with general interarrival and service times. *Performance Evaluation* 17(2):123–139.

New York Times (2016) Why long voting lines could have long-term consequences. (November 8), https://www.nytimes.com/2016/11/09/upshot/why-long-voting-lines-today-could-have-long-term-consequences.html.

Rao B, Posner M (1987) Algorithmic and approximation analyses of the shorter queue model. *Naval Res. Logist.* 34(3):381–398.

Ravner L, Shamir N (2020) Pricing strategy, capacity level and collusion in a market with delay sensitivity. *Naval Res. Logist.*, ePub ahead of print March 1, https://doi.org/10.1002/nav.21894.

Rodriguez S (2014) Verizon wireless to close five call centers, consolidate seven others. *Los Angeles Times* (February 12), http://articles.latimes.com/2014/feb/12/business/la-fi-tn-verizon-call-centers-moving-employees-20140212.

Ros D, Tuffin B (2004) A mathematical model of the Paris metro pricing scheme for charging packet networks. *Comput. Networks* 46(1):73–85.

Rothkopf MH, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. *Oper. Res.* 35(6):906–909.

Schmidt J (1997) Breaking down fiefdoms. *Management Rev.* 86(1):45–49.

Selen J, Adan I, Kapodistria S, van Leeuwaarden J (2016) Steady-state analysis of shortest expected delay routing. *Queueing Systems* 84(3-4):309–354.

Shunko M, Niederhoff J, Rosokha Y (2018) Humans are not machines: The behavioral impact of queueing design on service time. *Management Sci.* 64(1):453–473.

Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. *Bell System Tech. J.* 60(1):39–55.

Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Sci.* 61(12):3032–3053.

Southwest (2012) Investor Relations: Southwest Airlines Announces Reservations Consolidation into New Atlanta Call Center. Accessed July 1, 2018, http://investors.southwest.com/news-and-events/news-releases/2012/28-11-2012.

Sztrik J (2012) *Basic Queueing Theory* (University of Debrecen, Debrecen, Hungary).

UAFS (2018) Marketing & Communication Toolbox - Design Services. Accessed July 8, 2018, https://uafs.edu/marketing/design-services.

U.S. Department of the Treasury (2017) Shared services. Accessed July 8, 2018, https://www.fiscal.treasury.gov/fsservices/gov/fit/fit_fssp.htm.

van Dijk N, van der Sluis E (2009) Pooling is not the answer. *Eur. J. Oper. Res.* 197(1):415–421.

van Dijk NM, van der Sluis E (2008) To pool or not to pool in call centers. *Production Oper. Management* 17(3):296–305.

Wang J, Zhou YP (2017) Impact of queue configuration on service time: Evidence from a supermarket. *Management Sci.* 64(7):3055–3075.

Xerox (2013) Contact Center Consolidation: A Best Practices Blueprint. Accessed July 1, 2018, https://www.xerox.com/downloads/services/brochure/contact-center-consolidation.pdf.

Yang L, Debo L, Gupta V (2017) Trading time in a congested environment. *Management Sci.* 63(7):2377–2395.

Yu Y, Benjaafar S, Gerchak Y (2015) Capacity sharing and cost allocation among independent firms with congestion. *Production Oper. Management* 24(8):1285–1310.