Operations Research

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

Multilisting for Horizontally Differentiated Services: Implications for Throughput and Social Welfare

Zhou Chen, Yichuan Ding, Luyi Yang

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–20
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Contextual Areas

Multilisting for Horizontally Differentiated Services: Implications for Throughput and Social Welfare

Zhou Chen,[a] **Yichuan Ding,**[b,*] **Luyi Yang**[c]

[a] School of Economics and Management, Southeast University, Nanjing, Jiangsu 210096, China; [b] Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada; [c] Haas School of Business, University of California Berkeley, Berkeley, California 94720

*Corresponding author

chenzhou@seu.edu.cn (Z. Chen), daniel.ding@mcgill.ca (Y. Ding), luyiyang@haas.berkeley.edu (L. Yang)

https://orcid.org/0000-0003-4404-1763 (Z. Chen), https://orcid.org/0000-0003-3014-8973 (Y. Ding), https://orcid.org/0000-0002-5370-6926 (L. Yang)

Received: June 21, 2024
Revised: June 17, 2025; January 25, 2026
Accepted: July 13, 2026
Published Online in Articles in Advance: September 11, 2026

Area of Review: Operations and Supply Chain

https://doi.org/10.1287/opre.2024.1130

Abstract. This paper studies a queueing system with horizontally differentiated servers. Each customer decides whether to join the queues of multiple servers simultaneously ("multilist"), join the queue of her preferred server, or balk. Such a multilisting system contrasts with a single-listing system, which permits server choice but prohibits multilisting, and a pooling system, which precludes server choice by enforcing multilisting. We develop queueing-game-theoretic models for all three systems and compare their performance in terms of equilibrium throughput and social welfare. Although single-listing better matches customers to preferred servers and pooling improves load balancing, each may outperform the other depending on the performance metric. One might expect multilisting to dominate both designs by combining their respective advantages. Indeed, we find that multilisting consistently outperforms pooling in both throughput and social welfare and also achieves higher throughput than single-listing. However, it can surprisingly yield lower social welfare than single-listing because of customer misalignment in queue selection. Unlike equilibrium behavior, the socially optimal routing in the multilisting system can be asymmetric even when servers are ex ante symmetric. Customers underuse multilisting under light congestion but may overuse it under heavy congestion, relative to the social optimum. We show that nonnegative, asymmetric pricing can restore efficiency in equilibrium. The revenue maximization problem has a similar asymmetric structure in pricing and routing. Our results offer actionable insights for configuring multiserver service systems to balance customer preferences with operational efficiency.

Funding: This work was supported by National Natural Science Foundation of China [Grant 72401252 (Z. Chen)] and the Canadian Network for Research and Innovation in Machining Technology, Natural Sciences and Engineering Research Council of Canada [Grants RGPIN-2019-05539 and RGPIN-2025-05592 (Y. Ding)].

Supplemental Material: All supplemental materials, including the code, data, and files required to reproduce the results, are available at https://doi.org/10.1287/opre.2024.1130.

Keywords: queueing • redundancy • pooling • externality • process flexibility • social welfare

1. Introduction

Pooling is a foundational concept in operations. In service systems, pooling customer demand is known to balance load across servers and improve operational efficiency. Yet, pooling deprives customers of server choice and assigns them to available resources without regard to individual preferences. Hence, the operational efficiency of pooling comes at the expense of a potential mismatch between customers and resources. By contrast, in a (hybrid) *multilisting* system, customers decide for themselves whether to join the queues of multiple servers simultaneously for faster access (i.e., "multilisting"), or join the queue of their preferred server exclusively for a better fit (i.e., "single-listing"), or not join any queue at all (i.e., balking). Such a system can arise in various settings, especially in the public sector: For example, patients seeking medical services may sign up on the wait lists of one or multiple physicians or forgo service altogether; individuals considering vaccination during a pandemic may join the wait lists for one or multiple vaccines or remain unvaccinated; households who consider applying for public housing may be able to indicate interest in housing of multiple types or at multiple locations, or apply only for housing at a specific location, or look elsewhere. In most scenarios, registering on multiple wait lists either requires or implies a commitment to accepting the first available item or service.

Relative to a pooling system, which essentially forces all joining customers to multilist, a multilisting system allows customers to specify where to receive service, thereby reducing customer-server mismatch. Relative to a (pure) single-listing system, which requires customers to commit to a server of their choice at the outset, a multilisting system presents customers with the option to be flexible with where to receive service, thereby reducing those customers' waiting time. Thus, a multilisting system seems to have the best of both worlds and one might expect it to encourage participation, improve access, and benefit customers. This paper scrutinizes this intuition.

We build a queueing-game-theoretic model of strategic customers and two servers that are identical in service speed but differ in service features, so that customers may value them differently. On the demand side, we capture overall market potential with a single exogenous parameter: the total mass of customers who are interested in obtaining service. We refer to this parameter as the *demand size*. We consider the case when customers are heterogeneous in their tastes: each customer has an idiosyncratic preference over the two servers. We define *service differentiation* as the extent of this preference heterogeneity, captured by a single parameter.

We conduct a three-way comparison of the multilisting system, the pooling system, and the single-listing system in terms of both throughput and social welfare in equilibrium where individual customers act in their own self-interest. Our throughput comparison reveals that single-listing achieves higher throughput than pooling if and only if both the demand size is large and service differentiation is high. Yet, multilisting attains (weakly) higher throughput than both single-listing and pooling. Because multilisting provides customers with the most choice, it integrates the operational efficiency of pooling with the matching advantage of single-listing, thus luring the most customers among the three systems considered.

The throughput dominance of multilisting does not always translate into superior welfare performance. Although multilisting continues to outperform pooling in social welfare and, in fact, achieves both the highest throughput and highest social welfare among the three systems considered when the demand size is small, it nevertheless generates strictly lower social welfare than single-listing when the demand size is intermediate, even though single-listing is by no means guaranteed to dominate pooling in social welfare. As compared with pooling, the multilisting system brings the most value to those who strongly favor one server over another by letting them choose their preferred server instead of leaving it to chance where to get served. Thus, the addition of server choice in the multilisting system benefits customers. Relative to the single-listing system, the multilisting system helps balance the load across servers but also increases the overall load when there is considerable demand and thus congestion. Such an increased load exacerbates the wait time experienced by customers with extreme preferences who would only choose their preferred server regardless of whether the multilisting option is available, putting downward pressure on social welfare. The multilisting option does bring in more customers who would otherwise not join the single-listing system, yet these tend to be customers who are not extraordinarily excited about either server and, therefore, their utility from participation is close to nil when the congestion level is high. Hence, their participation adds little to social welfare, which implies that the additional choice of multilisting only hurts customers in such a case.

The potentially deleterious effect of multilisting on social welfare tells a cautionary tale against multilisting when customers make self-interested routing decisions. However, if a central planner could act on behalf of each individual customer to maximize social welfare, a multilisting system—with an additional lever to pull—would be at least as good as a single-listing system. Thus, it would be interesting to examine the socially optimal routing policy to see how the equilibrium deviates from the social optimum and how to recover it in a decentralized world.

We find that unlike the equilibrium, the socially optimal routing policy may be asymmetric even when servers are ex ante symmetric, meaning that the social optimum may assign different volumes of single-listing customers to each server, leading to unequal server utilization. To extract the most value from matching, the social planner will assign customers with strong preferences to their desired servers through single-listing while assigning customers with less extreme preferences to either multilisting or balking. Under symmetric allocation, all multilisting customers have the same likelihood of being served by each server and thus receive identical utility. Likewise, all balking customers receive the same utility of zero. Thus, symmetric allocation can be socially inefficient as it fails to account for customers' diverging preferences toward servers in deciding between multilisting and balking. In contrast, under asymmetric allocation, customers with different server preferences expect different utilities from multilisting. This differentiation allows the planner to assign customers who derive higher utility from multilisting to multilisting, while assigning remaining customers to balking, thereby improving matching and overall welfare. This reasoning drives the emergence of asymmetry in the socially optimal allocation.

Relative to the socially optimal outcome, customers in the equilibrium multilisting system under-choose multilisting when the demand size is small but may over-choose multilisting otherwise. When customers are left to their own devices, they do not internalize the externalities their actions impose on others, which causes them to deviate from the socially optimal behavior. The multilisting option has two externalities. On the one hand, as a substitute for single-listing, multilisting helps balance the load across servers, which generates a positive externality for other customers. On the other hand, as a substitute for balking, multilisting entices customers to join, which imposes a negative congestion externality. When the demand size is small, congestion is modest and hence multilisting primarily generates a positive externality. Therefore, not as many customers choose the multilisting option as socially desired. Yet, the reverse may be true when a large demand size leads to heavy congestion. We also show that a socially optimal price can be charged for each of the joining options to govern customer behavior in equilibrium so as to recover the socially optimal outcome. We find that multilisting need not be subsidized even when customers should be encouraged to choose multilisting more. We also study a revenue maximization problem and show that the revenue-maximizing prices can also be asymmetric despite server symmetry.

Finally, we extend our main model of two servers to a system with $n > 2$ servers. Each customer may either single-list with one of the n servers, or multilist across all servers, or balk. We characterize the symmetric equilibrium of the multilisting system and compare its throughput and social welfare with those of single-listing and pooling. The throughput comparison resembles that in the two-server case. The social welfare comparison also carries over when n is even, but when n is odd, we observe that the multilisting system can achieve the highest social welfare even at high demand levels. Moreover, we find that the socially optimal allocation is symmetric when n is a multiple of four, but can be asymmetric otherwise. These findings offer new perspectives that complement those in the two-server system.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the main model of two servers. Section 4 characterizes the equilibrium. Section 5 conducts throughput and welfare comparison. Section 6 studies the social optimum. Section 7 studies revenue maximization. Section 8 extends the model to a general number of servers. Section 9 concludes the paper.

2. Related Literature

Our paper contributes to the strategic queueing literature pioneered by Naor (1969), who studies the equilibrium queue-joining behavior and the socially optimal one. Our paper also studies both the equilibrium and the social optimum but of a setting with multiple servers and heterogeneous customer preferences. We refer the reader to Hassin and Haviv (2003) and Hassin (2016) for a comprehensive review of this literature. Particularly relevant for our paper is a body of works that study the question of pooling and load balancing in a multiserver queueing system. Although the operational benefit of pooling in reducing wait time is well-acknowledged, the literature has identified various strategic effects that can negate the operational benefit, including servers' strategic capacity decision or speed choice (Gilbert and Weng 1998, Cachon and Zhang 2007, Armony et al. 2021, Wang et al. 2023) and (homogeneous) customers' strategic joining decision or search behavior (Yang et al. 2019, Sunar et al. 2021). In particular, Sunar et al. (2021) show that when homogeneous customers observe the queue length and decide whether to join as in the model of Naor (1969), pooling incentivizes customers to join longer queues and may reduce social welfare. A key driver for their result is queue observability. By contrast, our paper shows that when customers have heterogeneous preferences, even without queue observability, pooling can still reduce social welfare as it denies customers' choice over servers. Further, the literature above restricts attention to a two-way comparison between separate queues and pooling. We take one step further by also considering a multilisting system that may outperform both separate queues and pooling.

In studying multilisting, our paper draws on the queueing literature on flexibility (Bassamboo et al. 2012, Tsitsiklis and Xu 2012, Visschers et al. 2012) and redundancy (Gardner et al. 2015, 2016, 2017b; Anton et al. 2021; Nageswaran and Scheller-Wolf 2022). One key takeaway from this literature is that even a little flexibility can significantly improve system performance. In particular, our cancel-upon-entry (CUE) model builds on the exact analysis of the queueing model in Visschers et al. (2012), where flexible customers can be served by the earliest available server. Our cancel-upon-service (CUS) model builds on that in Gardner et al. (2015, 2016), where some customers send redundant copies of their request to multiple queues so that their request is complete as soon as the first copy is complete. Although these papers focus on performance evaluation without incorporating strategic customer behavior, Ata et al. (2017) study a selfish-routing game that allows for multilisting in organ transplantation. Unlike our work, they abstract away from customers' heterogeneous preferences or balking decisions, rely on fluid and diffusion approximations rather than exact analysis, and restrict attention to overloaded queues. They demonstrate that multilisting improves geographical equity, whereas we caution that multilisting can reduce social welfare in equilibrium once customers' heterogeneous preferences and balking decisions are considered.

The closest work to ours is a recent paper by Nageswaran (2023), who employs queueing-game-theoretic models to study vaccine choice. Nageswaran's model of patients having a vaccine choice is similar to our single-listing system, whereas Nageswaran's model without a vaccine choice is similar to our pooling system in the CUS mode. Nageswaran (2023) analytically compares the throughput of these two systems and further numerically conducts a three-way throughput comparison of the choice option (single-listing), no-choice option (pooling), and a hybrid option, which is similar to the multilisting system under CUS in our model. Although Nageswaran's numerical three-way throughput comparison yields similar insights as our analytical throughput comparison, our investigation expands the scope of Nageswaran (2023) in a number of important directions. First, we consider not only the CUS mode but also the CUE mode, which allows us to cover a wider range of applications and compare the performance of the two modes. Second, we compare not only the throughput of the three systems but also social welfare. Third, we provide analytical results for the three-way throughput and social welfare comparison. Fourth, we adopt a canonical Hotelling model to accommodate customers' heterogeneous preferences for both servers and varying levels of service differentiation, whereas Nageswaran (2023) assumes heterogeneous customer preferences for one server but not for the other in the base model. Nageswaran's extension to heterogeneous preferences for both servers is limited to the case in which the level of service differentiation scales with the customer valuation of an ideal server; we do not impose such a restriction. Fifth, although Nageswaran (2023) focuses on the equilibrium without the service provider's intervention, we additionally study the socially optimal routing policy and the revenue maximization problem.

As such, our paper is related to the literature on optimal routing, which explores various tradeoffs such as waiting versus call resolution (Zhan and Ward 2014) and service speed versus quality (Zhan and Ward 2019). By contrast, our routing problem trades off matching value with waiting cost and considers not only where to allocate customers but also whether to admit them. In particular, we find that even for symmetric servers, the socially optimal routing policy can be asymmetric. This result connects our paper with recent work in the queueing literature that proposes optimal asymmetric dispatching policies for otherwise symmetric servers (Xie et al. 2024).

Besides, the fact that multilisting customers do not know where they will get served when they choose to multilist connects our paper to the literature on probabilistic goods and services whose identities are unknown to customers at the time of purchase (Fay and Xie 2008, Jerath et al. 2010). Most related to our work in this stream of literature is Xu et al. (2016), who study probabilistic services in a Hotelling queue using queue-game-theoretic models. In Xu et al. (2016), customers who choose a probabilistic service are randomly assigned to one of the queues, whereas, in our model, customers who choose to multilist will be served by the earliest available server. In Xu et al. (2016), offering a probabilistic service helps segment customers of different types but plays no role in load balancing, whereas in our model, offering the multilisting option is conducive to both customer segmentation and load balancing.

3. Model

We consider a queueing system with two servers, server A and server B. Both servers have an exponentially distributed service time of mean $1/\mu$. We refer to μ as the capacity of each server. The total capacity of the system is 2μ. Each server maintains a separate queue and processes customer requests on a first-come-first-served basis. Customers' service needs arise according to a Poisson process with rate Λ. We refer to Λ as the demand size, which represents the baseline arrival rate prior to any endogenous participation decisions. Customers incur a waiting cost of c per unit time.

To capture customers' heterogeneous preferences over the two servers, we employ a Hotelling model of horizontal differentiation (Hotelling 1929). Customers are uniformly distributed on the unit interval [0, 1]. A customer of type $x \in [0,1]$ obtains gross valuation $v_A(x) = v - tx$ if served by server A and $v_B(x) = v - t(1-x)$ if served by server B, where v is the valuation a customer receives from service by her ideal server, and the parameter t measures the degree of service differentiation, which can be driven by geographical distance, mismatch with a customer's preferences or needs, or other horizontally differentiated attributes. Larger t implies stronger preference heterogeneity: customers are more polarized and more strongly favor one server over the other. The variable x captures the customer's taste along the unit interval representing the spectrum of service attributes (e.g., spatial location, brand affinity, or specialization).

When a service need arises, a customer chooses one of the four options in her choice set $\{A, B, R, K\}$ to maximize her expected utility, where A stands for joining the queue of server A; B stands for joining the queue of server B; R stands for multilisting, that is, joining both queues; and K stands for balking, that is, not fulfilling her need. We normalize the utility of choosing option K to zero as the customer will not receive any service nor experience any waiting.

If a customer chooses to be multilisted (option R), she sends a copy of her request to each queue and her request is fulfilled as soon as one of the copies is complete. In terms of when a duplicate copy disappears from the system, we consider two modes, following the taxonomy of Nageswaran and Scheller-Wolf (2022):

(1) *Cancel-upon-service* (CUS): As soon as one copy is complete, the other copy is canceled.

(2) *Cancel-upon-entry* (CUE): As soon as one copy starts service, the other copy is canceled.

The CUS and CUE modes fit different applications. In the case of public housing, the capacity of a server corresponds to the rate at which a housing unit at a particular location becomes available. Thus, a multilisting applicant essentially indicates in her application that she is flexible with location and will take the first available housing unit from any location. Thus, once she receives a housing unit, she gets off the wait list and her request for housing is complete. Thus, the public housing example better fits the CUS protocol. In the case of medical services, as soon as a patient commences service with a provider, they would cancel their appointments with other providers to avoid penalty fees for no-shows or last-minute cancellations/rescheduling. Thus, the example of medical services better fits the CUE protocol.

Let λ_A, λ_B, λ_R be the (endogenous) arrival rates of customers who choose A, B, R, respectively. Given $(\lambda_A, \lambda_B, \lambda_R)$, let $W_i(\lambda_A, \lambda_B, \lambda_R)$ denote the expected steady-state response times (including both waiting time and service time) of a customer who chooses option $i \in \{A, B, R\}$. Let $p_A(\lambda_A, \lambda_B, \lambda_R)$ and $p_B(\lambda_A, \lambda_B, \lambda_R)$ be the steady-state probabilities that a customer who chooses option R ends up having her request completed by server A and server B, respectively, with $p_A(\lambda_A, \lambda_B, \lambda_R) + p_B(\lambda_A, \lambda_B, \lambda_R) = 1$. The expected utility of a customer of type $x \in [0, 1]$ choosing option A, $u_A(x)$, that of choosing option B, $u_B(x)$, and that of choosing option R, $u_R(x)$, are

$$
\begin{aligned}
u_A(x) &= v - tx - cW_A(\lambda_A, \lambda_B, \lambda_R),\\
u_B(x) &= v - t(1-x) - cW_B(\lambda_A, \lambda_B, \lambda_R); \qquad (1)\\
u_R(x) &= p_A(\lambda_A, \lambda_B, \lambda_R)(v - tx)\\
&\quad + p_B(\lambda_A, \lambda_B, \lambda_R)(v - t(1-x)) - cW_R(\lambda_A, \lambda_B, \lambda_R). \qquad (2)
\end{aligned}
$$

Given $(\lambda_A, \lambda_B, \lambda_R)$, let $q_i(x)$ be the probability that an expected-utility-maximizing customer of type x chooses option $i \in \{A, B, R\}$; the probability of balking is thus $1 - q_A(x) - q_B(x) - q_R(x)$:

$$
(q_A(x), q_B(x), q_R(x)) \in \underset{(q_A, q_B, q_R) \in [0,1]^3}{\arg\max} \; q_A u_A(x) + q_B u_B(x) + q_R u_R(x), \quad \text{s.t. } q_A + q_B + q_R \le 1. \qquad (3)
$$

Let $q_i(\lambda_A, \lambda_B, \lambda_R)$ be the proportion of customers who choose option $i \in \{A, B, R\}$ given $(\lambda_A, \lambda_B, \lambda_R)$. Thus, $q_i(\lambda_A, \lambda_B, \lambda_R) = \int_0^1 q_i(x)\,\mathrm{d}x$, $i \in \{A, B, R\}$. In equilibrium, the arrival-rate vector $(\lambda_A, \lambda_B, \lambda_R)$ solves the system of fixed-point equations: $\lambda_i = \Lambda q_i(\lambda_A, \lambda_B, \lambda_R)$, $i = \{A, B, R\}$.

Lemma 1 establishes structural properties of the equilibrium choice probabilities $q_A(x)$ and $q_B(x)$.

Lemma 1 (Symmetry in Equilibrium). *In equilibrium, there exists $x_m \in [0, 1/2]$ such that a customer of type x chooses queue A with probability one if and only if $0 \le x \le x_m$ and chooses queue B with probability one if and only if $1 - x_m \le x \le 1$, that is,*

$$
q_A(x) = \begin{cases} 1, & \text{for } x \in [0, x_m],\\ 0, & \text{for } x \in (x_m, 1]; \end{cases}
$$

$$
q_B(x) = \begin{cases} 1, & \text{for } x \in [1 - x_m, 1],\\ 0, & \text{for } x \in [0, 1 - x_m). \end{cases}
$$

As a result, $\lambda_A = \lambda_B$ in equilibrium.

Lemma 1 shows that customers with a strong preference (those with an extreme value of x) choose to be singly listed with their preferred server instead of multilisting or balking. Lemma 1 further establishes that there are just as many customers who exclusively choose server A as those who exclusively choose server B. Such symmetry of the equilibrium customer choice between the two servers is unsurprising because the servers are symmetric to begin with. Lemma 1 also implies that customers without a strong preference (those with an intermediate value of x) will not be singly listed and may only choose between multilisting and balking.

Given symmetry in equilibrium ($\lambda_A = \lambda_B$), a multilisted customer will have her request first completed by server A or server B with equal probability. that is, $p_A(\lambda_A, \lambda_B, \lambda_R) = p_B(\lambda_A, \lambda_B, \lambda_R) = 1/2$, and the expected response time of singly listed customers is the same between both servers, that is, $W_A(\lambda_A, \lambda_B, \lambda_R) = W_B(\lambda_A, \lambda_B, \lambda_R)$. Further, $p_A(\lambda_A, \lambda_B, \lambda_R) = p_B(\lambda_A, \lambda_B, \lambda_R) = 1/2$ implies that $u_R(x) = v - t/2 - cW_R(\lambda_A, \lambda_B, \lambda_R)$. Hence, customers who choose multilisting all have the same expected utility regardless of their type x. The expected response times $W_i(\lambda_A, \lambda_B, \lambda_R)$ for $i = \{A, B, R\}$ can be derived in closed form following the results of Visschers et al. (2012), Gardner et al. (2015), and Nageswaran and Scheller-Wolf (2022). We summarize the expressions of $W_i(\lambda_A, \lambda_B, \lambda_R)$, $i = \{A, B, R\}$ for both the CUS and CUE models in Table 1.

We make the following assumptions on the waiting tolerance c to avoid triviality.

Table 1. Expressions for the Expected Response Times When $\lambda_A = \lambda_B$

Mode	$W_R(\lambda_A, \lambda_B, \lambda_R)$	$W_A(\lambda_A, \lambda_B, \lambda_R) = W_B(\lambda_A, \lambda_B, \lambda_R)$
CUS	$\frac{1}{2\mu - 2\lambda_A - \lambda_R}$	$\frac{1}{2\mu - 2\lambda_A - \lambda_R} + \frac{1}{2\mu - 2\lambda_A}$
CUE	$\frac{1}{2\mu - 2\lambda_A - \lambda_R} + \frac{\mu - \lambda_A}{\mu(2\mu + \lambda_R)}$	$\frac{1}{2\mu - 2\lambda_A - \lambda_R} + \frac{\mu - \lambda_A}{\mu(2\mu + \lambda_R)} + \frac{(2\mu - \lambda_A)(\lambda_R + 2\lambda_A)}{\mu(2\mu - 2\lambda_A)(2\mu + \lambda_R)}$

Assumption 1. (i) $c < (v - t/2)\mu$; (ii) $c < t\mu$.

Assumption 1(i) ensures that a multilisted customer has a positive expected utility if no other customers are ahead of her. Hence, multilisting may arise in equilibrium. Assumption 1(ii) ensures a sufficient amount of horizontal differentiation between the two servers (i.e., a high-enough t) so that a customer with $x = 0$ ($x = 1$) has a strong-enough preference for exclusively joining queue A (B) over multilisting. Hence, at least some customers will be singly listed in equilibrium.

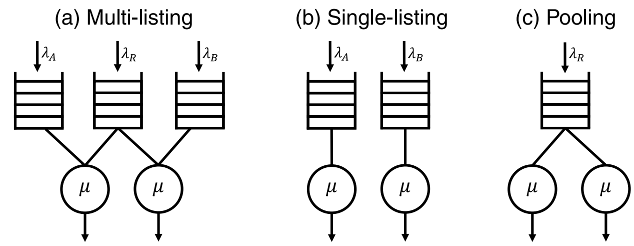
3.1. Two Benchmarks

We refer to the above system where customers choose to be singly listed or multilisted as a multilisting system for brevity. We also consider two benchmark systems: single-listing and pooling.

3.1.1. Single-Listing. In the single-listing benchmark, customers cannot multilist and must choose one of the two queues to join if they decide to join at all. That is, customers no longer have option R and their choice set becomes $\{A, B, K\}$. As defined earlier, customer x's expected utilities for joining queue A and queue B are $u_A(x)$ and $u_B(x)$, respectively. It is straightforward to extend Lemma 1 and show that the equilibrium will also be symmetric in this case, that is, $\lambda_A = \lambda_B$. Hence, each server operates an M/M/1 queue with an expected response time $1/(\mu - \lambda_i)$, $i \in \{A, B\}$. Note that this expected response time can also be obtained by letting $\lambda_R = 0$ in $W_A(\cdot)$ and $W_B(\cdot)$ in Table 1 (for either CUS or CUE).

3.1.2. Pooling. In the pooling benchmark, customers cannot choose between queues and must multilist at both queues if they choose to join. That is, customers' choice set becomes $\{R, K\}$. As defined earlier, customer x's expected utility of joining is $u_R(x)$. Because the two servers have equal capacity, a joining customer is equally likely to have her request first completed by either server. Given joining rate λ_R, a joining customer's expected response time can be found by letting $\lambda_A = \lambda_B = 0$ in $W_R(\lambda_A, \lambda_B, \lambda_R)$ in Table 1. In the CUS case, the expected response time is $W_R(0, 0, \lambda_R) = 1/(2\mu - \lambda_R)$. The queueing system operates as if it were an M/M/1 queue with total capacity 2μ and arrival rate λ_R. The two servers effectively pool capacity so that a customer request is complete whenever the first server completes the request. In the CUE case, the expected time is $W_R(0, 0, \lambda_R) = \frac{1}{2\mu - \lambda_R} + \frac{1}{2\mu + \lambda_R}$. The queueing system operates as if it were an M/M/2 queue with total capacity 2μ and arrival rate λ_R. The two servers effectively pool capacity so that a customer is assigned to the first available server. Figure 1 illustrates the three system configurations.

Figure 1. Illustration of the Three System Configurations



3.2. Performance Measures

We consider two performance measures: system throughput (which is the total number of customers served per unit time) and social welfare (which is the total utility generated per unit time). Given the equilibrium choice probabilities $q_i(x)$ of option $i \in \{A, B, R\}$ and the corresponding expected utility $u_i(x)$, all as a function of customer type $x \in [0, 1]$, system throughput is defined as $\Lambda\int_0^1 [q_A(x) + q_B(x) + q_R(x)]\mathrm{d}x$, whereas social welfare is defined as $\Lambda\int_0^1 [q_A(x)u_A(x) + q_B(x)u_B(x) + q_R(x)u_R(x)]\mathrm{d}x$. That is, social welfare is the aggregation of the expected utility across all customers per unit time. It is also equivalent to the total value generated less the total expected waiting cost per unit time, consistent with the social welfare definition in the literature (Naor 1969, Hassin and Haviv 2003). Note again that the utility of balking is normalized to zero. We henceforth let $\mu = 1/2$ in our main model of two servers without loss of generality.

4. Equilibrium

This section characterizes the equilibrium for three different systems (multilisting, single-listing, and pooling) under two different modes (CUS and CUE), a total of $3 \times 2 = 6$ model variants and effectively five different ones because CUS and CUE are the same for single-listing. For each of the five variants, we present the equilibrium customer choice as a function of customer type $x \in [0, 1]$.

Proposition 1 (Multilisting Equilibrium). *In the equilibrium of the multilisting system, customers with $x \in [0, x_m]$ choose queue A; those with $x \in [1 - x_m, 1]$ choose queue B, that is,*

$$q_A(x) = \begin{cases} 1, & \text{for } x \in [0, x_m], \\ 0, & \text{for } x \in (x_m, 1]; \end{cases}$$

$$q_B(x) = \begin{cases} 1, & \text{for } x \in [1 - x_m, 1], \\ 0, & \text{for } x \in [0, 1 - x_m); \end{cases}$$

and those with $x \in (x_m, 1 - x_m)$ behave in one of the following ways:

(i) *all choose the multilisting option R, that is,* $q_R(x) = \begin{cases} 1, & \text{for } x \in (x_m, 1 - x_m), \\ 0, & \text{for } x \in [0, x_m] \cup [1 - x_m, 1], \end{cases}$ *if* $\Lambda < \Lambda_1$;

(ii) *all balk, that is, $q_R(x) = 0$ for $x \in [0, 1]$, if $\Lambda \geq \Lambda_2$ and $t > \hat{t}$;*

(iii) *choose option R with probability* κ *and balk with probability* $1-\kappa$, *that is,* $q_R(x) = \begin{cases} \kappa, & \text{for } x \in (x_m, 1-x_m), \\ 0, & \text{for } x \in [0, x_m] \cup [1-x_m, 1], \end{cases}$ *if* $\Lambda \geq \Lambda_2$ *and* $t \leq \hat{t}$ *or* $\Lambda \in [\Lambda_1, \Lambda_2)$, *where expressions of* $(x_m, \Lambda_1, \Lambda_2, \hat{t}, \kappa)$ *for the CUS and CUE modes are given in the proof in Section* EC.2 *in the Online Appendix.*[1]

Proposition 1 characterizes the equilibrium structure in the multilisting system. Customers with a strong preference (those with an extreme value of x) choose to be singly listed with their preferred server (as indicated in Lemma 1). Moreover, customers without a strong preference (those with an intermediate value of x) do not care much about where to receive service and may therefore choose multilisting in exchange for faster access. These customers also do not have a high valuation of either server and are among the first to balk if the system is too congested. How these intermediate customers choose between balking and multilisting depends on two model parameters: demand size Λ and the unit misfit cost t. When Λ is sufficiently low and thus congestion is negligible, no customers balk and all customers in the intermediate segment choose multilisting. When Λ is high, congestion drives balking. In this case, if t is small, then both servers are a good fit and very much undifferentiated in the eyes of most customers. Thus, most customers fall into the intermediate segment and they play a mixed strategy by randomizing between multilisting and balking. However, when both Λ and t are large, the two servers are clearly differentiated and most customers have a strong preference and choose single-listing; only a small fraction of customers do not have a strong preference, but they also dislike both (owing to a high misfit cost t) and choose to balk (owing to congestion caused by a large demand size Λ). Hence, in this case, multilisting does not arise in equilibrium. It is noteworthy that in all cases, the equilibrium is intuitively symmetric in that $q_A(x) = q_B(1-x)$ and $q_R(x) = q_R(1-x)$ for $x \in [0,1]$.

Proposition 2 (Single-Listing Equilibrium). *In the equilibrium of the single-listing system, customers with* $x \in [0, x_s]$ *choose queue A; those with* $x \in [1-x_s, 1]$ *choose queue B; those with* $x \in (x_s, 1-x_s)$ *balk, where* $x_s = \frac{t/2+\Lambda v-\sqrt{(t/2-\Lambda v)^2+4\Lambda ct}}{2\Lambda t} \in (0, 1/2]$. *In particular,* $x_s = 1/2$ *(i.e., no customers balk) if and only if* $\Lambda \leq \Lambda_s := 1 - \frac{2c}{v-t/2}$.

Proposition 2 shows that in the single-listing system, when the demand size is small, all customers join the queue of their preferred server. Otherwise, customers with an intermediate x balk.

Proposition 3 (Pooling Equilibrium). *In the equilibrium of the pooling system,*

(i) *all customers join the queue if* $\Lambda \leq \Lambda_p$;

(ii) *customers join the queue with probability* Λ_p/Λ *and balk with probability* $1-\Lambda_p/\Lambda$ *if* $\Lambda > \Lambda_p$, *where threshold* $\Lambda_p = \Lambda_p^c := 1 - \frac{c}{v-t/2}$ *under CUS and threshold* $\Lambda_p = \Lambda_p^e := \sqrt{1-\frac{2c}{v-t/2}}$ *under CUE.*

Proposition 3 shows that in the pooling system, when the demand size is small, all customers join. Otherwise, all customers play a mixed strategy and randomize between joining and balking.

The equilibrium characterizations in Propositions 1, 2, and 3 enable us to derive the system throughput and social welfare for the multilisting, single-listing, and pooling systems, as summarized in Corollaries EC.1, EC.2, and EC.3 in Section EC.2 in the Online Appendix.

5. Throughput and Welfare Comparison

In this section, we compare the throughput and social welfare across the three different systems (multilisting, single-listing, and pooling) under two different modes (CUS and CUE). Section 5.1 conducts throughput comparison, and Section 5.2 conducts social welfare comparison.

5.1. Throughput Comparison

Theorem 1 presents the result of the throughput comparison in the CUS mode.

Theorem 1 (Throughput Comparison Under CUS). *In the CUS mode, multilisting achieves higher throughput than single-listing and pooling, that is,* $Th_m = \max\{Th_s, Th_p\}$. *Between single-listing and pooling, the former achieves higher throughput* ($Th_s > Th_p$) *if and only if both* Λ *and* t *are high. More specific throughput comparison is summarized in Table 2.*

When the demand size is small ($\Lambda \leq \Lambda_s$), no customers balk irrespective of the system design because of negligible waiting times across the board. Hence, the equilibrium throughput rate equals the demand size Λ in all three systems. When the demand size is not small ($\Lambda > \Lambda_s$), balking kicks in, and the throughputs of the three systems diverge. Two factors drive balking: long waiting times and mismatching between customers and servers. Relative to single-listing, pooling reduces waiting times because of its well-acknowledged operational advantage of eliminating the coexistence of idleness and queues, which effectively balances the load across servers. However, pooling falls short of single-listing in matching as it denies customers the choice of their preferred server. When the unit misfit cost t is

Table 2. Throughput Comparison Under CUS

Condition	Comparison
$\Lambda \leq \Lambda_s$	$Th_m = Th_p = Th_s$
$\Lambda > \Lambda_s, t \leq v$ or $\Lambda_s < \Lambda < \Lambda_2^c, t > v$	$Th_m = Th_p > Th_s$
$\Lambda > \Lambda_2^c, t > v$	$Th_m = Th_s > Th_p$

low ($t \le v$), customer preferences for servers are of secondary concern and, therefore, the pooling system attracts more customers with its operational advantage and thus outperforms single-listing in throughput. Yet, in the presence of significant service differentiation ($t > v$), the customer-server match becomes crucial, allowing the single-listing system to overtake the pooling system in throughput when the system is highly congested ($\Lambda > \Lambda_2^c$).

The multilisting system has the best of both worlds. It combines the shorter waiting times of pooling with the improved matching of single-listing. Hence, it entices just as many customers as the better-performing system of the two (in terms of throughput). Notably, Theorem 1 shows that multilisting and pooling have equivalent throughput as long as some customers choose to multilist in the former system. This result arises because the expected response time of multilisting under CUS only depends on the total number of joining customers per unit time (i.e., throughput), but not how joining customers are split between single-listing and multilisting. This intriguing property was also highlighted in Gardner et al. (2015), who note that for a given throughput, multilisting customers are "immune" to the pain of more customers choosing multilisting.

Next, we turn to the throughput comparison in the CUE mode and state the result in Theorem 2. Denote $\Lambda_{sp}^e := \frac{\sqrt{\frac{1}{4}-\frac{c}{2v-t}}}{\frac{v}{t}-\frac{c}{\frac{t}{2}-t\sqrt{\frac{1}{4}-\frac{c}{2v-t}}}}$.

Theorem 2 (Throughput Comparison Under CUE). *In the CUE mode, multilisting achieves weakly higher throughput than single-listing and pooling, that is, $Th_m \ge \max\{Th_s, Th_p\}$. Between single-listing and pooling, the former achieves higher throughput ($Th_s > Th_p$) if and only if both Λ and t are high. More specific throughput comparison is summarized in Table 3.*

Although the throughput comparison of the CUE case largely mirrors that of the CUS case, one notable difference is that in the CUE mode, multilisting can yield *strictly* higher throughput than both single-listing and pooling. In particular, under CUE, the multilisting system can strictly outperform the pooling system in throughput even when multilisting arises in the former, whereas under CUS, the two systems would have identical throughput in such a case. Unlike the CUS mode, multilisting customers in the CUE mode are adversely affected by more customers choosing to be multilisted for a given throughput. The higher the proportion of joining customers who multilist, the longer the expected waiting time is for these customers. Because the pooling system essentially forces all joining customers to multilist, it has a longer expected waiting time than that experienced by multilisting customers (who account for only a fraction of joining customers) in the multilisting system for a given throughput. This further implies that the pooling system cannot attract as many customers as the multilisting system in equilibrium when the demand size is large. Altogether, a multilisting system is always the optimal choice for maximizing the equilibrium throughput (among the three system designs considered), regardless of whether the system operates in the CUS or CUE mode.

Table 3. Throughput Comparison Under CUE

Condition	Comparison
$\Lambda \le \Lambda_s$	$Th_m^e = Th_p^e = Th_s$
$\Lambda_s < \Lambda \le \Lambda_p^e$	$Th_m^e = Th_p^e > Th_s$
$\Lambda > \Lambda_p^e, t \le 2v - \frac{v^2}{v-c}$ or $\Lambda_p^e < \Lambda < \Lambda_{sp}^e, t > 2v - \frac{v^2}{v-c}$	$Th_m^e > Th_p^e > Th_s$
$\Lambda > \Lambda_{sp}^e, 2v - \frac{v^2}{v-c} < t \le v - \frac{4c^2}{v}$ or $\Lambda_{sp}^e < \Lambda < \Lambda_2^e$, $t > v - \frac{4c^2}{v}$	$Th_m^e > Th_s > Th_p^e$
$\Lambda \ge \Lambda_2^e, t > v - \frac{4c^2}{v}$	$Th_m^e = Th_s > Th_p^e$

We next compare the equilibrium throughput between CUS and CUE in Proposition 4.

Proposition 4 (CUS vs. CUE in Throughput). *In a multilisting system, CUS has a weakly higher throughput than CUE ($Th_m \ge Th_m^e$).*

The result of Proposition 4 aligns with intuition: by yielding shorter waiting times across all three queues (as implied by Table 1), CUS is more attractive to customers than CUE, leading to more system participation and a higher throughput rate. Further, it is straightforward to verify that the same comparison holds in the pooling system (whereas the single-listing system is unaffected by CUS or CUE).

5.2. Social Welfare Comparison

Theorem 3 presents the social welfare comparison in the CUS mode.

Theorem 3 (Social Welfare Comparison Under CUS). *In the CUS mode, multilisting always achieves higher social welfare than pooling ($SW_m > SW_p$) but achieves weakly lower social welfare than single-listing ($SW_m \le SW_s$) when Λ is high. More specific social welfare comparison is summarized in Table 4.*

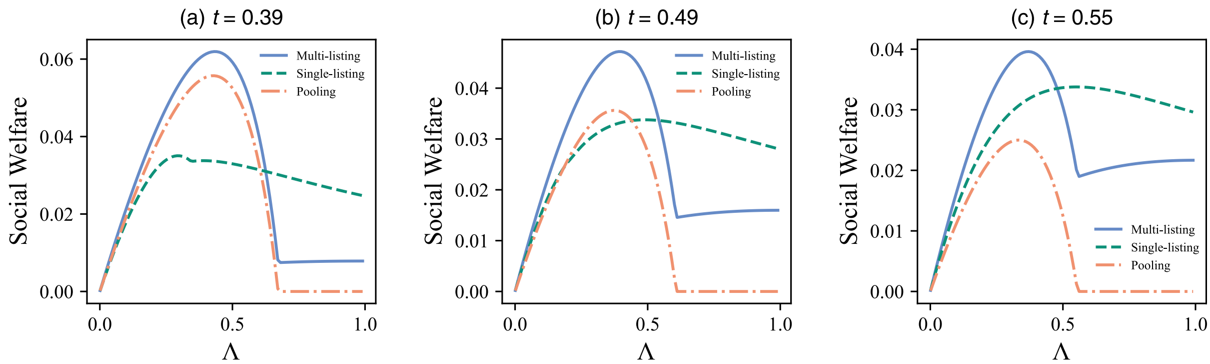
Figure 2 provides a graphical illustration of Theorem 3.

5.2.1. Pooling vs. Single-Listing. We first provide some intuition for the social welfare comparison

Table 4. Social Welfare Comparison Under CUS

Condition	Comparison
$\Lambda \le \Lambda_s, v < 4c$ or $\Lambda \le \Lambda_s, t \le 4c < v$ or $1 - 4c/t < \Lambda \le \Lambda_s, 4c < t \le v$	$SW_m > SW_p > SW_s$
$\Lambda < 1 - 4c/t, 4c < t \le v$ or $\Lambda \le \Lambda_s, t > v > 4c$	$SW_m > SW_s > SW_p$
$\Lambda > \Lambda_1^c, t \le v$ or $\Lambda_1^c < \Lambda < \Lambda_2^c, t > v$	$SW_s > SW_m > SW_p$
$\Lambda \ge \Lambda_2^c, t > v$	$SW_s = SW_m > SW_p$

Figure 2. (Color online) Social Welfare Comparison Under CUS



Note. $v = 0.5, c = 0.1$.

between pooling and single-listing. Similar to the throughput comparison, the comparison of social welfare is driven by two factors: the cost of waiting and the value of customer-server matching. Pooling excels at reducing the former, whereas single-listing excels at increasing the latter. When the unit misfit cost t is low, pooling achieves higher social welfare than single-listing for small demand Λ, but the reverse is true for large demand, as demonstrated in Figure 2(a). When demand is small, almost no customers balk, making the throughput comparable between the two systems. When t is low, service differentiation is low, and the expected waiting time is a determining factor of social welfare. Pooling reduces the expected waiting time (for a given throughput) because of its operational efficiency and thus delivers more value to customers than single-listing. However, when demand is large, customers without a strong preference balk in the single-listing system, which alleviates congestion and brings value to customers with a strong preference as they join their preferred server. By contrast, all customers in the pooling system suffer from a long wait without a server choice. Thus, in the single-listing system, customers with a strong preference benefit from better matching and reduced congestion, and customers without a strong preference are not much worse off by choosing to balk because they would still experience a long wait even if they join the pooling system. Therefore, in this case, single-listing achieves higher social welfare than pooling.

When the unit misfit cost t is intermediate, single-listing achieves higher social welfare than pooling not only when the demand size is large but also when the demand size is small; it nevertheless underperforms pooling in social welfare when the demand size is intermediate, as demonstrated in Figure 2(b). When t is not too low, service differentiation becomes more significant, which favors single-listing. When the demand size is small, the amount of congestion is also limited. Hence, the advantage of pooling in reducing waiting time is overshadowed by the disadvantage of pooling in accommodating diverging customer preferences for servers, causing the pooling system to fall short of the single-listing system in social welfare. However, when the demand size is intermediate, congestion becomes more significant and the advantage of pooling in reducing waiting time becomes the dominant force, enabling it to outperform single-listing. However, when the unit misfit cost t is sufficiently high, service differentiation becomes so pronounced that pooling underperforms single-listing for any demand size, as demonstrated in Figure 2(c).

5.2.2. Pros and Cons of Multilisting. Among the three systems considered, the multilisting system provides customers with the most choices. As such, it marries the advantage of pooling in wait-time reduction with the advantage of single-listing in matching. Customers with a strong preference can meet their need by joining their favored server, whereas customers without a strong preference can choose to be multilisted and thus flexible with where to receive service in exchange for faster access. Such flexibility balances the load across the servers and also benefits other customers. Indeed, these strengths enable the multilisting system to achieve higher social welfare than the two other systems when the demand size Λ is small (and thus almost all customers join regardless of the system design). This implies that a multilisting system is the best design that maximizes both social welfare and throughput when the system congestion is light.

Nevertheless, the multilisting system surprisingly achieves lower social welfare than the single-listing system when Λ is large. This result contrasts with that of the throughput comparison where multilisting is always the best design regardless of the demand size. To see why, note that when Λ is large, customers with an intermediate x choose to balk in the single-listing system because of heavy congestion but may choose to participate in the multilisting system by joining both queues to expedite access, although doing so still generates a low expected utility owing to congestion and these intermediate customers' inherent lukewarm interest in either server. Hence, despite their participation, these multilisting customers contribute little to social welfare, but their presence in the multilisting system increases the waiting time of customers with a more extreme x who choose to be singly listed with their preferred server, thereby adversely affecting social welfare. In essence, although the additional option to multilist attracts more customers (and thus increases throughput), it also exacerbates congestion and ultimately harms social welfare when Λ is large.

Although the multilisting system is not always as socially desirable as the single-listing system, it is guaranteed to beat the pooling system in social welfare. Recall from our earlier discussion that under CUS, customers who choose to multilist in the multilisting system have the same expected response time and thus the same expected utility as customers who join the pooling system (when the throughput of the two systems is the same). However, customers who choose to be singly listed with their preferred server in the multilisting system derive a higher expected utility than their counterparts who join the pooling system without the guarantee of being matched to their preferred server. Therefore, multilisting achieves higher social welfare than pooling when the two systems enjoy the same throughput. When the two systems differ in throughput, balking arises in both systems, in which case social welfare is zero in the pooling system (because all customers are indifferent between joining and balking), whereas social welfare is positive in the multilisting system (thanks to the positive expected utility enjoyed by customers who choose to be singly listed). Hence, the multilisting system still outperforms pooling in social welfare.

Next, Theorem 4 presents the social welfare comparison in the CUE mode.

Theorem 4 (Social Welfare Comparison Under CUE). *In the CUE mode, multilisting always achieves higher social welfare than pooling ($SW_m^e > SW_p^e$) but achieves weakly lower social welfare than single-listing ($SW_m^e \le SW_s$) when Λ is high. More specific social welfare comparison is summarized in Table 5.*

Table 5. Social Welfare Comparison Under CUE

Condition	Comparison
$\Lambda < \frac{\sqrt{16c^2+t^2}-4c}{t}, t < \frac{3v-3c}{2} - \frac{\sqrt{v^2-2vc+9c^2}}{2}$ or $\Lambda < \Lambda_s, t \ge \frac{3v-3c}{2} - \frac{\sqrt{v^2-2vc+9c^2}}{2}$	$SW_m^e > SW_s > SW_p^e$
$\frac{\sqrt{16c^2+t^2}-4c}{t} < \Lambda < \Lambda_s, t < \frac{3v-3c}{2} - \frac{\sqrt{v^2-2vc+9c^2}}{2}$	$SW_m^e > SW_p^e > SW_s$
$\Lambda > \Lambda_1^e, t \le v - \frac{4c^2}{v}$ or $\Lambda_1^e < \Lambda < \Lambda_2^e, t > v - \frac{4c^2}{v}$	$SW_s > SW_m^e > SW_p^e$
$\Lambda \ge \Lambda_2^e, t > v - \frac{4c^2}{v}$	$SW_s = SW_m^e > SW_p^e$

Figure 3 illustrates Theorem 4. The comparison of social welfare in the CUE mode largely resembles that in the CUS mode. One notable difference is that in the CUS mode, when demand size Λ is sufficiently small, pooling outperforms single-listing in social welfare if the unit misfit cost t is small, but in the CUE mode, when demand size Λ is sufficiently small, single-listing beats pooling in social welfare regardless of t. The single-listing system is the same across the two modes, but the pooling system under CUE has a longer expected waiting time than that under CUS (for a given throughput) because the former requires duplicate copies to be canceled earlier and thus the level of resource pooling is not as thorough. This undermines the operational efficiency of pooling (relative to the CUS mode), causing the pooling system in the CUE mode to fall short of the single-listing system in social welfare when the demand size is small.

We next compare social welfare between CUS and CUE in Proposition 5. One might expect CUS to yield higher social welfare because of its shorter expected response times. Indeed, it is straightforward to show that in the pooling system, CUS has higher social welfare than CUE. However, as Proposition 5 shows, this intuition does not always hold in the multilisting system.

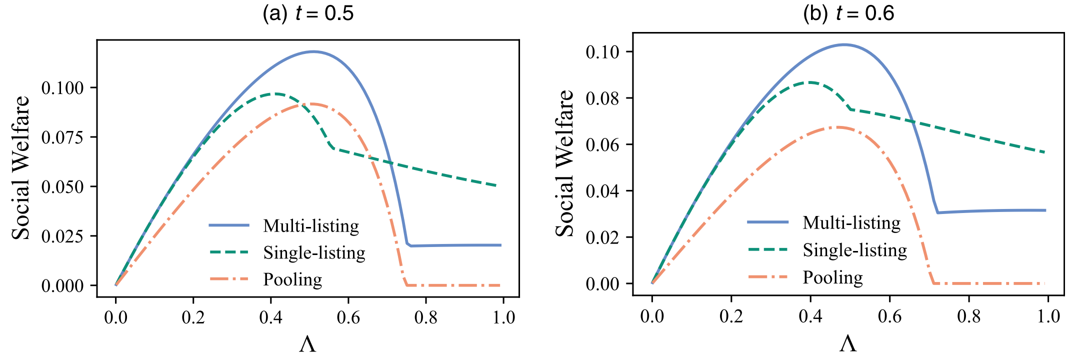
Proposition 5 (CUS vs. CUE in Social Welfare). *In a multilisting system, CUE has weakly higher social welfare than CUS ($SW_m^e \ge SW_m$) when $\Lambda > \Lambda_1^c$.*

Proposition 5 indicates that CUS can, in fact, lead to lower social welfare than CUE when demand is high. In this case, high demand triggers balking in both modes and implies that customers who choose multilisting receive zero surplus. It is only single-listing customers who contribute a positive surplus to social welfare. Under CUE, the increase in expected response time is more pronounced for the multilisting queue R than for the single-listing queues A and B. As a result, the proportion of customers who join the single-listing queues tends to be larger under CUE, resulting in higher social welfare under CUE. Figure 4 illustrates Proposition 5.

6. Social Optimum

In this section, we take the perspective of a social planner and study the optimal routing policy that maximizes social welfare. We characterize the structure of the social optimum. Then, we compare the social optimum with the equilibrium outcome. Finally, we identify the optimal price the social planner would charge to induce the socially optimal outcome in equilibrium.

Figure 3. (Color online) Social Welfare Comparison Under CUE



Note. $v = 0.7, c = 0.1$.

6.1. Social Welfare Maximization Problem

We first define the social planner's routing problem. The social planner maximizes social welfare by dictating what each arriving customer chooses based on their type $x \in [0,1]$. That is, the social planner determines $G_i(x)$, the probability that a customer with type x is assigned to option $i \in \{A,B,R\}$. Thus, $1 - G_A(x) - G_B(x) - G_R(x)$ is the probability that the social planner requests customer x to balk. The social planner solves the following problem:

$$\begin{aligned} \max_{G_A(\cdot),G_B(\cdot),G_R(\cdot)} \quad & \Lambda \int_0^1 [G_A(x)u_A(x) + G_B(x)u_B(x) \\ & \quad + G_R(x)u_R(x)]\,dx, \\ \text{s.t.} \quad & G_A(x) \ge 0,\ G_B(x) \ge 0,\ G_R(x) \ge 0, \quad \forall x \in [0,1], \\ & G_A(x) + G_B(x) + G_R(x) \le 1, \qquad \forall x \in [0,1], \end{aligned} \tag{4}$$

where $u_i(x)$ is defined in (1)–(2) with $\lambda_i = \Lambda \int_0^1 G_i(x)\,dx$ for $i \in \{A,B,R\}$.

The social planner's routing problem is a functional optimization problem that determines the routing probabilities as a function of customer types, characterized by the three functions $G_A(\cdot)$, $G_B(\cdot)$, and $G_R(\cdot)$. Alternatively, the social planner's problem can be viewed as an infinite-dimensional optimization problem because there is a continuum of (and thus infinitely many) customer types. Note that the equilibrium choice probabilities for the three systems (multilisting, single-listing, pooling) characterized in Section 4 are all feasible solutions to this optimization problem. We refer to the optimal $G_A(\cdot)$, $G_B(\cdot)$, $G_R(\cdot)$ as the socially optimal routing policy. Proposition 6 characterizes the structural properties of a socially optimal routing policy.

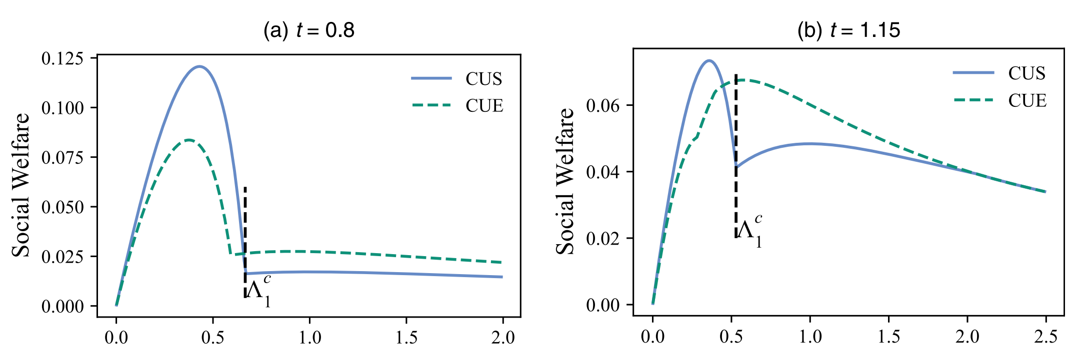
Proposition 6 (Structure of Social Optimum). *There exists a socially optimal routing policy in which customers with $x \in [0, x_A]$ are routed to queue A; those with $x \in [1 - x_B, 1]$ are routed to queue B such that $x_B \le x_A$; those with $x \in (1 - x_B - x_R, 1 - x_B)$ are multilisted; and those with $x \in (x_A, 1 - x_B - x_R]$ balk for some $x_A, x_B, x_R \ge 0$ and $x_A + x_B + x_R \le 1$, that is,*

$$G_A(x) = \begin{cases} 1, & \text{for } x \in [0, x_A], \\ 0, & \text{for } x \in (x_A, 1]; \end{cases}$$

$$G_B(x) = \begin{cases} 1, & \text{for } x \in [1 - x_B, 1], \\ 0, & \text{for } x \in [0, 1 - x_B); \end{cases}$$

$$G_R(x) = \begin{cases} 1, & \text{for } x \in (1 - x_B - x_R, 1 - x_B), \\ 0, & \text{for } x \in [0, 1 - x_B - x_R] \cup [1 - x_B, 1]. \end{cases}$$

Figure 4. (Color online) Social Welfare Comparison Between CUS and CUE of the Multilisting System



Note. $v = 1, c = 0.2$.

Without loss of generality, we assume that $x_B \le x_A$ such that the arrival rate of queue B is no less than queue A. Then either $x_A = x_B$, in which case the optimal routing follows a symmetric structure, or $x_A > x_B$, in which case the optimal routing is no longer symmetric, that is, $G_A(x)$ does not necessarily equal $G_B(1-x)$ and $G_R(x)$ does not necessarily equal $G_R(1-x)$. Moreover, unlike the equilibrium routing, where customers in the intermediate segment may randomize between multilisting and balking, the socially optimal policy presented does not involve probabilistic routing. See Figure 5 for an illustration of the optimal routing policy characterized in Proposition 6.

To build intuition toward possible asymmetry of the optimal routing policy, it is useful to rationalize how the social planner would assign customers of different types to each option when the arrival rate for each option is fixed. To maximize social welfare, the social planner would like to extract the most matching value by first routing customers with an extreme preference to their preferred server (through single-listing). Customers in the middle without an extreme preference would incur a high misfit cost with either server and would be best assigned to either the multilisting or balking option, depending on the arrival rate of single-listing customers at the two servers. Whichever server has a (weakly) lower single-listing arrival rate (the "less busy" server) would be (weakly) more likely to serve a multilisting customer.

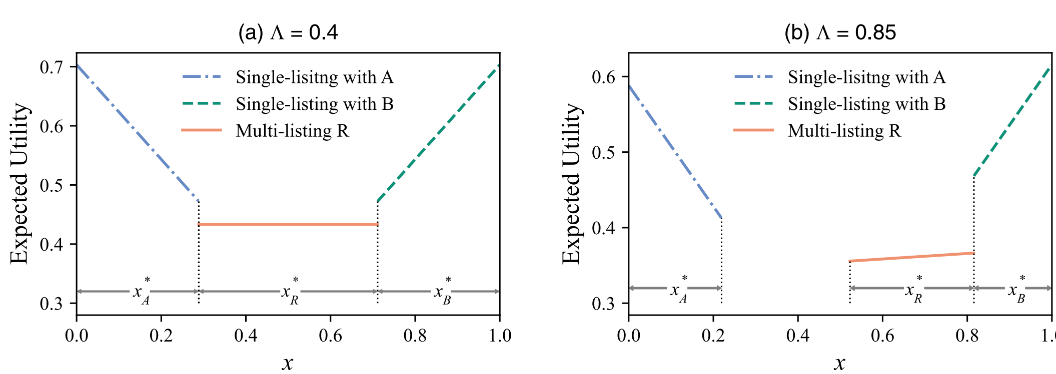
Thus, out of customers in the middle that are yet to be assigned, the total misfit will be minimized if customers who prefer the "less busy" server are assigned to the multilisting option and customers who dislike the "less busy" server the most are assigned to balking. Such an arrangement gives rise to the type-dependent routing policy presented in Proposition 6. This policy subsumes various symmetric equilibrium structures characterized in Proposition 1 as special cases, including a symmetric policy without customer balking ($x_A = x_B$ and $x_A + x_B + x_R = 1$) and a symmetric policy without multilisting ($x_A = x_B$ and $x_R = 0$). More importantly, it can at least match any symmetric probabilistic allocation where customers in the middle randomize between multilisting and balking because it can always assign one subset of customers to multilisting and another subset of customers to balking while achieving the same expected utilities for all customers as the probabilistic allocation. Unlike the symmetric probabilistic allocation, where customers in the middle randomly choose between multilisting and balking without regard to their type, the socially optimal policy differentiates customers based on their preferences to the extent possible and can be asymmetric when all four options are assigned to a nonempty segment of customers. The possible optimality of asymmetric routing is striking in a setting with symmetric servers.

Note that $x_A = \int_0^1 G_A(x)\,dx$, $x_B = \int_0^1 G_B(x)\,dx$, $x_R = \int_0^1 G_R(x)\,dx$. Hence, x_i can be interpreted as the proportion of customers who are required to take option $i \in \{A, B, R\}$. Thus, the arrival rate of customers assigned to option i is $\lambda_i = \Lambda x_i$, $i \in \{A, B, R\}$. Proposition 6 enables us to translate the infinite-dimensional (functional) optimization problem in (4) into a three-dimensional nonlinear optimization problem over (x_A, x_B, x_R), as formulated in (5):

$$\begin{aligned}
\max_{x_A, x_B, x_R} \quad & \Lambda x_A \left[v - \frac{t x_A}{2} - cW_A(\Lambda x_A, \Lambda x_B, \Lambda x_R) \right] \\
& + \Lambda x_B \left[v - \frac{t x_B}{2} - cW_B(\Lambda x_A, \Lambda x_B, \Lambda x_R) \right] \\
& + \Lambda x_R \Bigg[v - \Bigg[p_A(\Lambda x_A, \Lambda x_B, \Lambda x_R)(1 - 2x_B - x_R) \\
& \qquad + \left(x_B + \frac{x_R}{2} \right) \Bigg] t - cW_R(\Lambda x_A, \Lambda x_B, \Lambda x_R) \Bigg], \\
\text{s.t.} \quad & x_A + x_B + x_R \le 1, \; x_A \ge x_B \ge 0, x_R \ge 0.
\end{aligned} \tag{5}$$

Next, we focus our analysis on the CUS model. The analysis of the CUE model yields similar results and

Figure 5. (Color online) Expected Utility as a Function of Customer Type in the Social Optimum Under CUS



Note. The vertical axis is $G_A(x)u_A(x) + G_B(x)u_B(x) + G_R(x)u_R(x)$. $v = 1$, $c = 0.1$, $t = 0.8$.

will be provided in Section EC.5 in the Online Appendix. We will substitute the corresponding expressions of $W_i(\cdot)$ for $i \in \{A, B, R\}$ and $p_A(\cdot)$ into the social welfare maximization problem to obtain more specific results about the structure of the social optimum (e.g., when asymmetry arises, comparison with the equilibrium outcome) in Section 6.2 and characterize the socially optimal prices in Section 6.3.

6.2. Characterization of Social Optimum

In the CUS model, Gardner et al. (2015) derive the expected response time for choosing option $i \in \{A, B, R\}$, $W_i(\lambda_A, \lambda_B, \lambda_R)$, as a function of $(\lambda_A, \lambda_B, \lambda_R)$:

$$\begin{aligned} W_i(\lambda_A, \lambda_B, \lambda_R) &= \frac{1}{2\mu - \lambda_A - \lambda_B - \lambda_R} + \frac{1}{\mu - \lambda_i} \\ &\quad - \frac{1}{2\mu - \lambda_A - \lambda_B}, \quad i \in \{A, B\}, \\ W_R(\lambda_A, \lambda_B, \lambda_R) &= \frac{1}{2\mu - \lambda_A - \lambda_B - \lambda_R}. \end{aligned}$$

The expressions above generalize those in Table 1 by allowing $\lambda_A \neq \lambda_B$. When $\lambda_A \neq \lambda_B$, a multilisting customer is not equally likely to be served by either server. Building on Gardner et al. (2015), we derive the probability of a multilisting customer being served by a given server.

Lemma 2. *In the CUS mode of the multilisting system, given $(\lambda_A, \lambda_B, \lambda_R)$, a multilisting customer ends up having her request completed by server A with probability $p_A(\lambda_A, \lambda_B, \lambda_R) = \frac{\mu - \lambda_A}{2\mu - \lambda_A - \lambda_B}$.*

Now that we have the expressions of $W_i(\cdot)$ for $i \in \{A, B, R\}$ and $p_A(\cdot)$, we are ready to characterize the social optimum (i.e., the optimal solution to Problem (5), denoted by (x_A^*, x_B^*, x_R^*)) under CUS in more detail. Proposition 7 presents the result.

Proposition 7 (Social Optimum Under CUS). *In the CUS mode, there exists a socially optimal solution (x_A^*, x_B^*, x_R^*) with the following properties:*

(i) *Some customers are singly listed, that is, $x_A^* + x_B^* > 0$;*

(ii) *When $t \le v$, some customers are multilisted, that is, $x_R^* > 0$;*

(ii-a) *When $\Lambda \le 1 - \sqrt{\frac{c}{v - t/2}}$, the social optimum is symmetric without balking, that is, $x_A^* = x_B^*$ and $x_A^* + x_B^* + x_R^* = 1$;*

(ii-b) *When $\Lambda > 1 - \sqrt{\frac{c}{v - t/2}}$, the social optimum is asymmetric with balking, that is, $x_A^* \neq x_B^*$ and $x_A^* + x_B^* + x_R^* < 1$;*

(iii) *When $t > v$,*

(iii-a) *When $\Lambda \le 1 - \sqrt{\frac{c}{v - t/2}}$, the social optimum is symmetric without balking;*

(iii-b) *When $\Lambda > 1 - \sqrt{\frac{c}{v - t/2}}$, the social optimum involves balking, that is, $x_A^* + x_B^* + x_R^* < 1$, and is asymmetric if and only if some customers are multilisted, that is, $x_A^* \neq x_B^*$ if and only if $x_R^* > 0$.*

Proposition 7 identifies conditions under which the socially optimal routing policy is symmetric or asymmetric. When the demand size is small ($\Lambda \le 1 - \sqrt{\frac{c}{v - t/2}}$), all customers join and the social optimum is symmetric, as illustrated by Figure 5(a). Although such symmetry is also seen in equilibrium, the social optimum does not coincide with the equilibrium. For instance, in Figure 5(a), the expected utility is discontinuous in customer type in the social optimum, whereas it would be continuous in equilibrium. This suggests that when left to their own devices, customers in equilibrium would not always choose the option they are assigned to in the social optimum despite symmetry being a structural feature shared by the equilibrium and the social optimum. When the demand size is large ($\Lambda > 1 - \sqrt{\frac{c}{v - t/2}}$), not all customers join and the social optimum can be asymmetric (and thus structurally different from the equilibrium), as illustrated by Figure 5(b). In such an asymmetric routing policy, the proportion of customers who are singly listed with server A differs from the proportion of customers singly listed with server B despite the ex ante symmetry between the two servers. When the unit misfit cost t is low ($t \le v$), the multilisting segment is nonempty and such asymmetry ($x_A^* \neq x_B^*$) holds for any large-enough demand size, as illustrated by Figure 6(a). However, when the unit misfit cost t is high ($t > v$), the multilisting segment can be either fully covered (when demand size is small) or empty (when demand size is large). In either case, the social optimum is symmetric, as illustrated by Figure 6(b). Altogether, when $t > v$, the socially optimal routing policy is asymmetric for an intermediate demand size, but symmetric for a demand size that is either small or large.

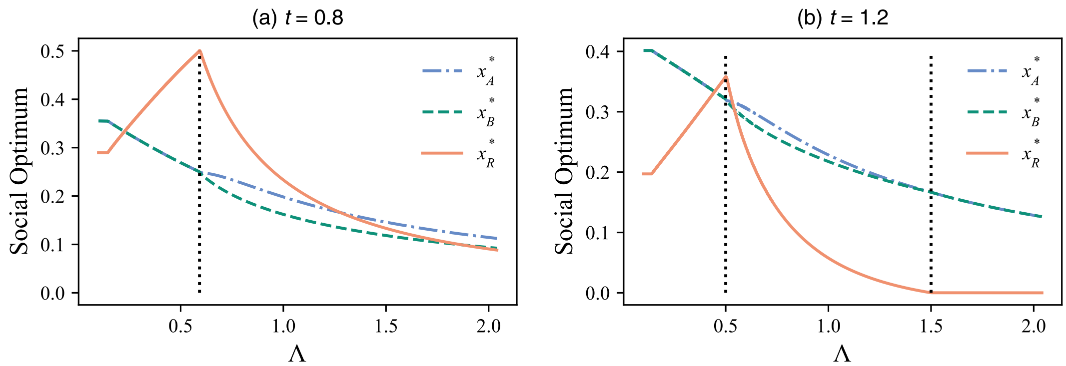
Next, we compare the social optimum with the equilibrium outcome in Theorem 5. Specifically, we compare the proportion of customers who take option $i \in \{A, B, R\}$ in the social optimum, denoted by x_i^* (the optimal solution to (5)) with its counterpart in equilibrium, denoted by q_i^{eq}.

Theorem 5 (Social Optimum vs. Equilibrium). *In the multilisting system, under CUS,*

(i) *When $\Lambda \le 1 - \sqrt{\frac{c}{v - t/2}}$, customers under-choose multilisting in equilibrium, that is, $q_R^{eq} < x_R^*$, and all join in both equilibrium and the social optimum, that is, $q_A^{eq} + q_B^{eq} + q_R^{eq} = x_A^* + x_B^* + x_R^* = 1$;*

(ii) *When $\Lambda > 1 - \sqrt{\frac{c}{v - t/2}}$ and $t \le v$, customers overjoin in equilibrium, that is, $q_A^{eq} + q_B^{eq} + q_R^{eq} > x_A^* + x_B^* + x_R^*$;*

Figure 6. (Color online) Socially Optimal Routing Under CUS



Note. $v = 1, c = 0.1$.

(iii) *Customers always over-choose single-listing in equilibrium, that is,* $q_A^{eq} + q_B^{eq} > x_A^* + x_B^*$.

Theorem 5 shows that customers do not multilist enough in equilibrium relative to what they should in the social optimum when the demand size is small. In this case, balking does not arise in either the equilibrium or the social optimum, and hence, routing is symmetric in both instances. Relative to the choice of single-listing, multilisting is a choice conducive to load balancing (as a form of partial pooling) and thus benefits other customers. Hence, it generates positive externalities that self-interested customers do not take into account when choosing between different options. Therefore, customers under-choose multilisting in equilibrium when the demand size is small. However, when the demand size is large, some balking is inevitable in both the equilibrium and social optimum. On the one hand, as a substitute for single-listing, multilisting generates positive externalities through the load-balancing effect. On the other hand, as a substitute for balking, multilisting attracts customers who have lukewarm interest in either service. The presence of these customers imposes negative congestion externalities on other customers. Thus, multilisting has both positive and negative externalities self-interested customers do not internalize, and thus, depending on the relative strength of these two competing forces, customers may under-choose or over-choose multilisting in equilibrium, as suggested by the numerical example in Figure 7.

As for the throughput comparison between the equilibrium and social optimum, our analytical result in Theorem 5 combined with our numerical observation indicates that the throughput in equilibrium is always weakly higher, that is, customers overjoin in equilibrium (although our analytical result cannot comment on the case in which both demand size Λ and unit misfit cost t are high, our numerical studies confirm such an insight). This result is well-aligned with that of the classical one in the strategic queueing literature (see, e.g., Hassin and Haviv 2003) and is driven by the negative congestion externalities of joining.

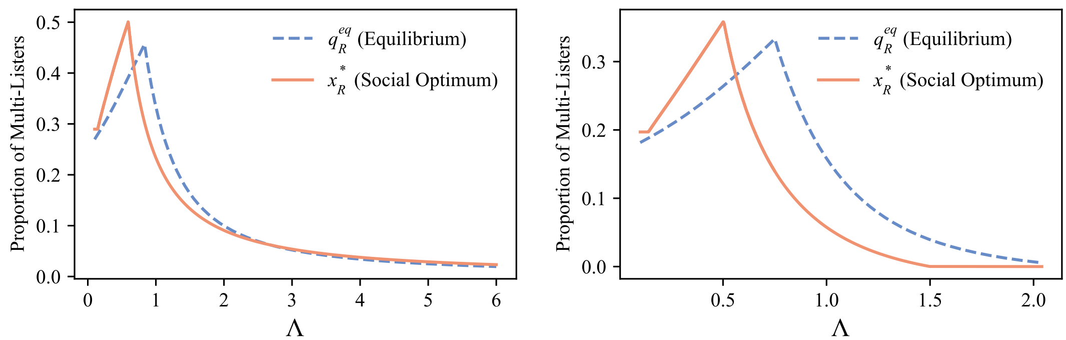
6.3. Socially Optimal Prices

The deviation of the equilibrium outcome from the social optimum begs the question of how to induce the social optimum through price control. That is, the social planner can charge customers price P_i for choosing option $i \in \{A, B, R\}$. To be general, these prices can be either positive or negative. In practice, a positive price can be implemented by an application fee (in the case of public housing) or treating healthcare reimbursement as a taxable income (in the case of medical services), whereas a negative price can be in the form of a subsidy or a tax credit. Consistent with the strategic queueing literature, these prices are treated as internal transfers as far as social welfare is concerned. Proposition 8 identifies the socially optimal prices (P_A, P_B, P_R) under which the equilibrium choice probabilities $q_i(x|P_A, P_B, P_R)$ match the socially optimal routing probabilities $G_i(x)$ (derived from the optimal solution to Problem (5)) for $x \in [0,1]$ and for $i \in \{A, B, K\}$.

Proposition 8 (Socially Optimal Prices). *In the multilisting system, under CUS, there exist nonnegative prices* (P_A, P_B, P_R) *that induce the socially optimal behavior in equilibrium, that is,* $q_i(x|P_A, P_B, P_R) = G_i(x)$ *for* $x \in [0,1]$ *and for* $i \in \{A, B, K\}$. *In particular,* $P_R \le \min\{P_A, P_B\}$ *for* $\Lambda \le 1 - \sqrt{\frac{c}{v - \frac{t}{2}}}$ *and* $P_R \le \max\{P_A, P_B\}$ *for any* Λ.

Because customers in equilibrium under-choose multilisting when the demand size is small (per Theorem 5), one might expect the social planner to subsidize multilisting to encourage more customers to take this option. Proposition 8 shows that this is not necessary because the same effect (inducing more multilisting) can be achieved by charging a lower price for multilisting than for single-listing (which is indeed confirmed by Proposition 8 for a small demand size). In fact, because

Figure 7. (Color online) Social Optimum vs. Equilibrium Under CUS: Proportion of Multilisting Customers



Note. $v = 1, c = 0.1$.

customers also overjoin in equilibrium, taxing multilisting helps deter customers from joining, thereby regulating congestion and improving social welfare. When the demand size is large, customers may overchoose both multilisting and single-listing, making it less clear how the socially optimal price for each option compares. Figure 8 explores this question numerically. Figure 8(a) shows an instance where the socially optimal price for multilisting is the lowest for any demand size. By contrast, Figure 8(b) shows an instance where the socially optimal price for multilisting is higher than one of the single-listing prices when the demand size is large. Finally, when demand size is large, the optimal pricing is asymmetric, that is, $P_B > P_A$, as shown in Figure 8. This is aligned with the asymmetry of the optimal routing policy as characterized in Proposition 7(ii-b).

7. Revenue Maximization

In this section, we study the problem where the planner aims to maximize revenue collected from customers, rather than social welfare, by charging customers price P_i for choosing option $i \in \{A, B, R\}$. The revenue maximization problem is formulated as

$$\max_{P_A, P_B, P_R} \quad \Lambda \int_0^1 [q_A(x)P_A + q_B(x)P_B + q_R(x)P_R]\, dx, \tag{6}$$

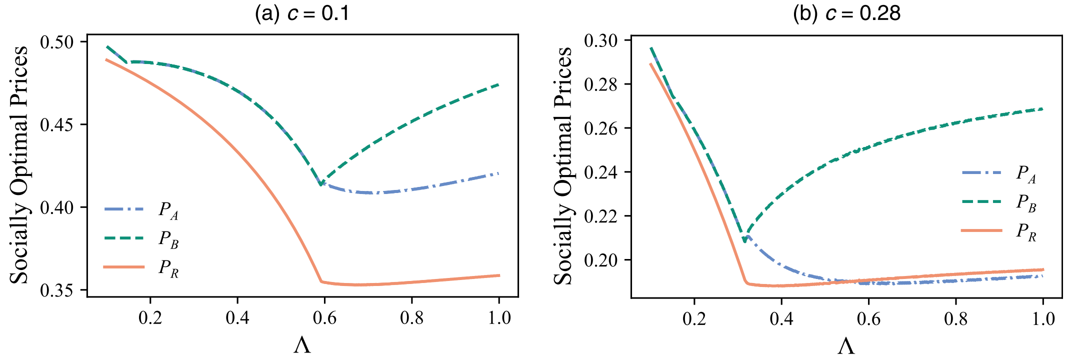
$$\text{s.t.} \quad (q_A(x), q_B(x), q_R(x)) \in \underset{\substack{(q_A, q_B, q_R) \in [0,1]^3 \\ q_A + q_B + q_R \le 1}}{\arg\max} \; q_A \tilde{u}_A(x) + q_B \tilde{u}_B(x) + q_R \tilde{u}_R(x), \quad \forall x \in [0,1], \tag{7}$$

where $\tilde{u}_i(x) = u_i(x) - P_i$ with $u_i(x)$ defined in (1)–(2) and $\lambda_i = \Lambda \int_0^1 q_i(x)\, dx$ for $i \in \{A, B, R\}$. In the revenue maximization formulation, the objective (6) gives the expected revenue generation rate, whereas (7) specifies customers' equilibrium choice probabilities for given prices, similar to (3).

Next, we focus our analysis on the CUS model. Let P_A^{RM}, P_B^{RM}, and P_R^{RM} be the revenue-maximizing prices and λ_R^{RM} be the resulting arrival rate of multilisting customers. Proposition 9 characterizes when the revenue-maximizing prices are symmetric ($P_A^{RM} = P_B^{RM}$) and when they are asymmetric ($P_A^{RM} \neq P_B^{RM}$).

Proposition 9 (Revenue Maximization). *In the CUS mode,*
(i) *If $t \le v$, some customers are multilisted, that is, $\lambda_R^{RM} > 0$;*

Figure 8. (Color online) Socially Optimal Prices Under CUS



Note. $v = 1, t = 0.8$.

(i-a) *When* $\Lambda \le 1 - \sqrt{\frac{c}{v-t/2}}$, *the revenue-maximizing prices are symmetric, that is,* $P_A^{\mathrm{RM}} = P_B^{\mathrm{RM}}$;

(i-b) *When* $\Lambda > 2$, *the revenue-maximizing prices are asymmetric, that is,* $P_A^{\mathrm{RM}} \neq P_B^{\mathrm{RM}}$;

(ii) *If* $t > v$,

(ii-a) *When* $\Lambda \le 1 - \sqrt{\frac{c}{v-t/2}}$, *the revenue-maximizing prices are symmetric, that is,* $P_A^{\mathrm{RM}} = P_B^{\mathrm{RM}}$;

(ii-b) *When* $\Lambda > 2$, *the revenue-maximizing prices are asymmetric if and only if some customers are multilisted, that is,* $P_A^{\mathrm{RM}} \neq P_B^{\mathrm{RM}}$ *if and only if* $\lambda_R^{\mathrm{RM}} > 0$.

Figure 9 illustrates Proposition 9 by plotting the revenue-maximizing prices. Together, they show that the revenue-maximizing prices are asymmetric for small t and large Λ or large t and intermediate Λ. Such price asymmetry will also result in asymmetry in the volume of single-listing customers across the servers. The conditions under which asymmetry arises are analogous to those for the social optimum (cf. Proposition 7 and Figure 6), and the rationale is indeed similar. The socially optimal policy uses asymmetry to create more surplus for consumers, whereas the revenue-optimal pricing, in turn, extracts the higher surplus created by asymmetry.

For the CUE model, characterization of the revenue-maximizing equilibrium is analytically intractable. A numerical study is presented in Section EC.5.2 of the Online Appendix.

8. A General Number of Servers

This section studies an extension to $n > 2$ servers, each operating at an equal service rate of $\mu = 1/n$. Upon arrival, each customer chooses among $n+2$ options: (i) single-list with server i ($i = 1, \ldots, n$), (ii) multilist with all the n servers, or (iii) balk. To capture customers' heterogeneous preferences toward these n servers, we adopt the Salop Circular Model (Salop 1979), in which a unit mass of customers is uniformly distributed around a circle with perimeter one. The n servers are located equidistant around the circle, so that the distance between two adjacent servers is $1/n$, as illustrated in Figure 10. Let server i's coordinate be $\frac{i-1}{n}$. A customer with coordinate $x \in [0,1)$ derives a utility of $v - td_i(x)$ from being served by server i, where $d_i(x)$ denotes the (shorter) arc distance between customer x and server i, that is, $d_i(x) = \min\{|\frac{i-1}{n} - x|, 1 - |\frac{i-1}{n} - x|\}$.

Our analysis focuses on the CUS model, where closed-form expressions of the expected response times are available for general n because of Gardner et al. (2015, 2017a, 2019), who show that under CUS, given the arrival rate of customers who single-list with server i, λ_i for $i = 1, \ldots, n$, and the arrival rate of multilisting customers λ_R, the expected response time for multilisting customers, W_R, and that for customers who single-list with server i, W_i, are

$$W_R(\{\lambda_i\}_{i=1}^n, \lambda_R) = \frac{1}{n\mu - \sum_{i=1}^n \lambda_i - \lambda_R},$$

$$W_i(\{\lambda_i\}_{i=1}^n, \lambda_R) = \frac{1}{\mu - \lambda_i} + \frac{1}{n\mu - \sum_{i=1}^n \lambda_i - \lambda_R} - \frac{1}{n\mu - \sum_{i=1}^n \lambda_i}, \quad i = 1, \ldots, n.$$

Given $(\{\lambda_i\}_{i=1}^n, \lambda_R)$, for a customer with coordinate x, the expected utility of multilisting, $U_R(x)$, and that of single-listing with server i, $U_i(x)$, are given by

$$U_R(x) = v - t\sum_{i=1}^n \frac{\mu - \lambda_i}{n\mu - \sum_{i=1}^n \lambda_i} d_i(x) - cW_R(\{\lambda_i\}_{i=1}^n, \lambda_R),$$

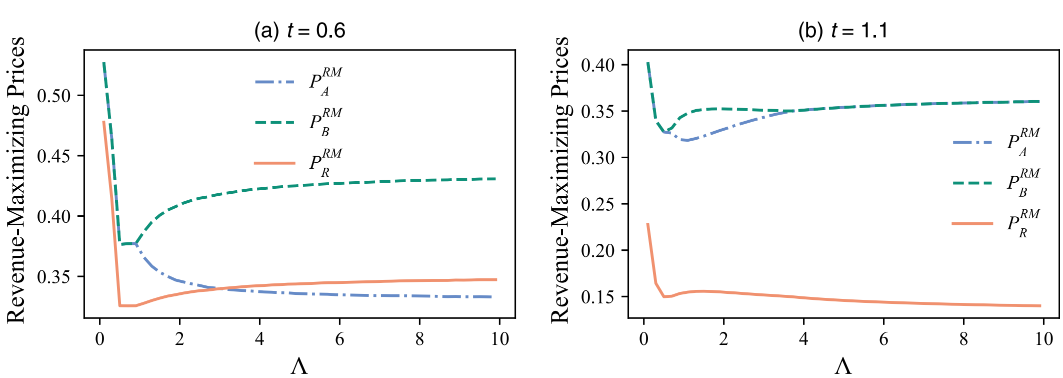
$$U_i(x) = v - td_i(x) - cW_i(\{\lambda_i\}_{i=1}^n, \lambda_R), \quad i = 1, \ldots, n.$$

8.1. Equilibrium Analysis

To proceed with the equilibrium analysis, we impose the following parametric assumptions.

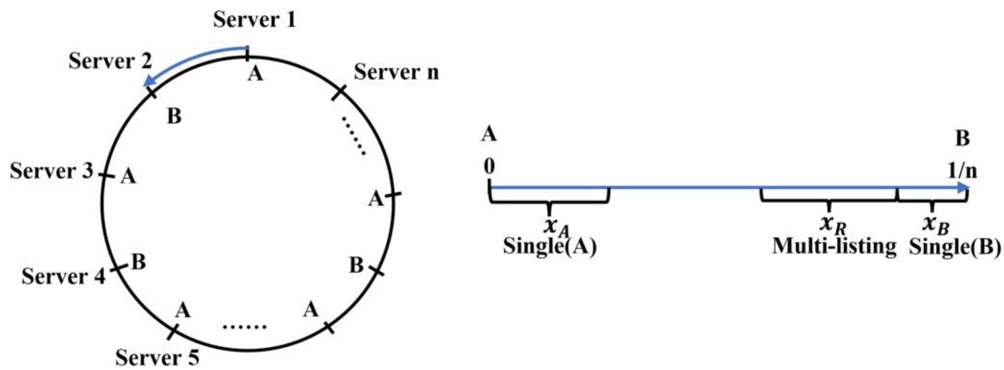
Assumption 2 (Even n). (i) $c < v - \frac{t}{4}$; (ii) $c < \frac{t}{4(n-1)}$; (iii) *if* $n \ge 4$, $c > \frac{t(1-2/n)}{4(n-1)}$.

Figure 9. (Color online) Revenue-Maximizing Prices Under CUS



Note. $v = 1, c = 0.2$.

Figure 10. (Color online) Illustration of the Salop Circular Model



Assumption 3 (Odd n). (i) $c < v - \frac{t}{4}(1 - \frac{1}{n^2})$; (ii) $c < \frac{t(n+1)}{4n^2}$; (iii) *if* $n \geq 3$, $c > \frac{t(n-1)}{4n^2}$.

For both even and odd values of n, conditions (i) and (ii) of Assumptions 2 and 3 play similar roles to their counterparts in Assumption 1 for the two-server case. Condition (iii), which is unique to the $n > 2$ setting, ensures that a positive fraction of customers prefer multilisting to single-listing. Proposition 10 characterizes the strategy of any customer $x \in [0, 1)$ in a symmetric equilibrium where each server receives an identical volume of single-listing customers, that is, $\lambda_1 = \lambda_2 = \cdots = \lambda_n$.

Proposition 10 (Multilisting Equilibrium for $n > 2$). *Under CUS, in a multilisting system with $n > 2$ servers, there exists a threshold $x_m^n \in [0, \frac{1}{2n}]$ such that in equilibrium, for $i = 1, \ldots, n$, customers with $x \in [\frac{i-1}{n}, \frac{i-1}{n} + x_m^n]$ choose to single-list with server i, those with $x \in [\frac{i}{n} - x_m^n, \frac{i}{n})$ choose to single-list with server $i+1$, and those with $x \in (\frac{i-1}{n} + x_m^n, \frac{i}{n} - x_m^n)$ behave as follows:*

- *For even n, they* (i) *all multilist if $\Lambda < \Lambda_{n_1}$,* (ii) *all balk if $\Lambda \geq \Lambda_{n_2}$ and $t > \frac{4(n-1)}{n}v$, and* (iii) *multilist with probability κ_n and balk with probability $1 - \kappa_n$ if $\Lambda \geq \Lambda_{n_2}$ and $t \leq \frac{4(n-1)}{n}v$ or $\Lambda_{n_1} \leq \Lambda < \Lambda_{n_2}$;*
- *For odd n, they* (i) *all multilist if $t \leq \frac{4n}{n+1}v$ and $\Lambda \leq 1 - \frac{c}{v - \frac{t}{4}(1+\frac{1}{n^2})}$;* (ii) *all balk if $t > \frac{4n}{n+1}v$; and* (iii) *multilist for $x \in (\frac{i-1}{n} + x_m^n, \frac{i-1}{n} + x_{R0}^n] \cup [\frac{i}{n} - x_{R0}^n, \frac{i}{n} - x_m^n)$ and balk for $x \in (\frac{i-1}{n} + x_{R0}^n, \frac{i}{n} - x_{R0}^n)$ if $t \leq \frac{4n}{n+1}v$ and $\Lambda > 1 - \frac{c}{v - \frac{t}{4}(1+\frac{1}{n^2})}$,*

where $(x_m^n, \Lambda_{n_1}, \Lambda_{n_2}, \kappa_n, x_{R0}^n)$ are given in the proof in Section EC.7 *in the Online Appendix.*

Proposition 10 specifies customer strategies for each of the n contiguous intervals $x \in [\frac{i-1}{n}, \frac{i}{n})$. $i = 1, \ldots, n$. The equilibrium structure when n is even generalizes that of the two-server case in Proposition 1. Importantly, customers who choose multilisting expect the same utility regardless of their server preferences (i.e., their location on the circle). By contrast, when n is odd, customers' expected utility from multilisting depends on their server preferences. Under odd n, the midpoint between two adjacent servers is directly opposite a server on the circle, which implies that multilisting yields a lower expected utility for customers located near such a midpoint than for those located further away. Consequently, unlike the case of an even n, in which customers may randomize between multilisting and balking, an odd n can give rise to distinct customer segments of multilisting and balking, with those near the midpoint of two adjacent servers choosing to balk, those near either server choosing to single-list with their respective preferred server, and those in between choosing to multilist while maintaining positive surplus.

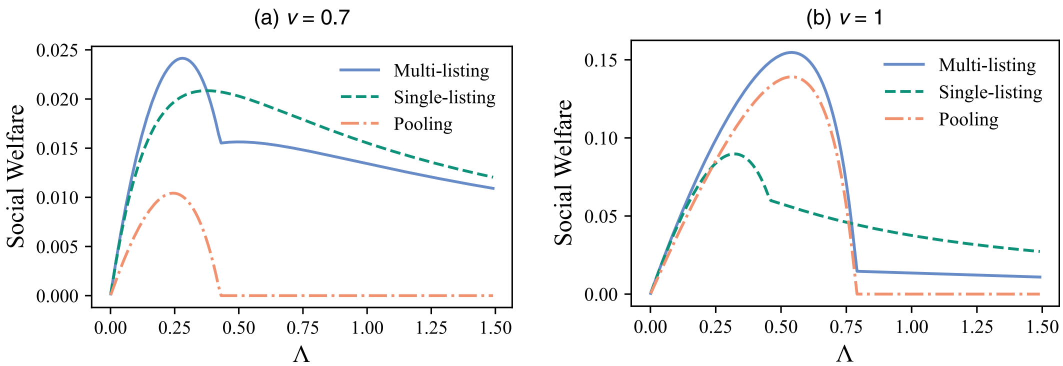
Following a similar approach, we characterize in Propositions EC.1 and EC.2 in the Online Appendix the equilibria for two benchmark systems: a pooling system (where a single queue feeds into n servers) and a single-listing system (where each of the n servers operates a separate queue). Let Th_m^n, Th_s^n, and Th_p^n denote the equilibrium throughput in the multilisting, single-listing, and pooling systems, respectively, and let SW_m^n, SW_s^n, and SW_p^n denote the corresponding social welfare levels. Theorem 6 conducts throughput comparison across the three systems, and Theorem 7 compares social welfare.

Theorem 6 (Throughput Comparison Under CUS for $n > 2$ Servers). *Under CUS, $Th_m^n = \max\{Th_s^n, Th_p^n\}$. Between single-listing and pooling, when n is even, $Th_s^n > Th_p^n$ if and only if $\Lambda > \Lambda_{n_2}$ and $t > \frac{4(n-1)}{n}v$; when n is odd, $Th_s^n > Th_p^n$ if and only if $t > \frac{4n}{n+1}v$.*

Theorem 7 (Social Welfare Comparison Under CUS for $n > 2$ Servers). *Under CUS, $SW_m^n > SW_p^n$; $SW_m^n > SW_s^n$ when Λ is low but $SW_m^n \leq SW_s^n$ when Λ is high and n is even. More specific social welfare comparison is summarized in Tables* EC.5 *and* EC.6 *in the Online Appendix.*

Consistent with the two-server case (Theorem 1), Theorem 6 shows that the multilisting system always

Figure 11. (Color online) Social Welfare Comparison Under CUS with an Even Number of Servers



Note. $c = 0.1, t = 2.1, n = 4$.

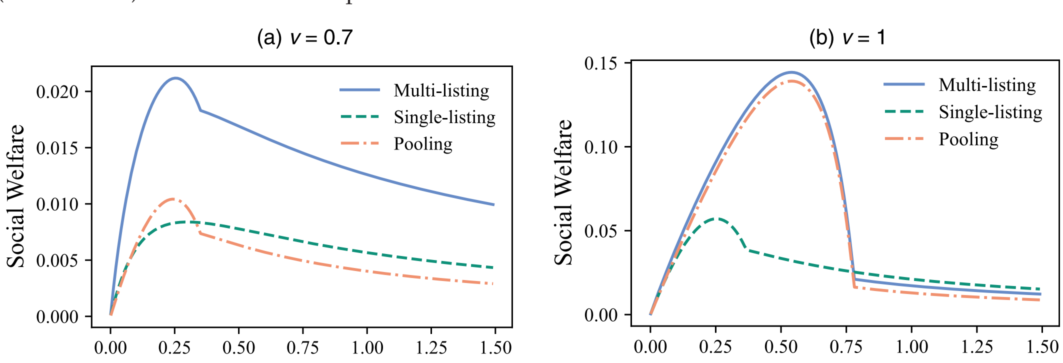
yields the highest throughput among the three designs. As for social welfare comparison, Theorem 7 shows that the multilisting system always outperforms pooling and also outperforms the single-listing system under low demand. However, it is outperformed by single-listing when demand is high and n is even. These insights are consistent with our findings in the base model (Theorem 3) and are illustrated in Figure 11. However, for odd n and high demand, we observe that although multilisting still falls short of single-listing in some scenarios, as illustrated in Figure 12(b), it can consistently outperform single-listing (and pooling) even at high demand levels, as illustrated in Figure 12(a). This improvement stems from the fact that under odd n, the option of multilisting helps create more refined customer segmentation (as explained after Proposition 10), allowing customers who multilist to enjoy positive surplus even under high demand (rather than expecting zero utility as they would under even n), which ultimately leads to higher social welfare.

8.2. Social Optimum

We study the social optimum of multilisting in the n-server system, with a particular focus on whether the socially optimal routing is symmetric or asymmetric. Given this focus, we restrict attention to periodic routing policies. Our proposed policy for even n generalizes the socially optimal routing policy characterized in Proposition 6 and is similarly specified by three parameters $(x_A, x_B, x_R) \in [0, \frac{1}{2n}]$ with $x_A + x_B + x_R \le \frac{1}{n}$. For customers with $x \in [\frac{i-1}{n}, \frac{i}{n})$, $i = 1, \ldots, n$, if i is odd, then customers with $x \in [\frac{i-1}{n}, \frac{i-1}{n} + x_A]$ single-list with server i, customers with $x \in (\frac{i-1}{n} + x_A, \frac{i}{n} - x_B - x_R]$ balk, customers with $x \in (\frac{i}{n} - x_B - x_R, \frac{i}{n} - x_B)$ multilist, and customers with $x \in [\frac{i}{n} - x_B, \frac{i}{n}]$ single-list with server $i+1$. If i is even, then customers with $x \in [\frac{i-1}{n}, \frac{i-1}{n} + x_B]$ single-list with server i, customers with $x \in (\frac{i-1}{n} + x_B, \frac{i-1}{n} + x_B + x_R)$ multilist, customers with $x \in [\frac{i-1}{n} + x_B + x_R, \frac{i}{n} - x_A)$ balk, and customers with $x \in [\frac{i}{n} - x_A, \frac{i}{n}]$ single-list with server $i+1$. See Figure 10 for an illustration of the proposed periodic routing policy.

The social planner determines (x_A, x_B, x_R) to maximize social welfare. Given (x_A, x_B, x_R), the arrival rate of single-listing customers at an odd-indexed (even-indexed) server is $2\Lambda x_A$ $(2\Lambda x_B)$. As in the two-server model, we say the social optimum (x_A^*, x_B^*, x_R^*) is symmetric if $x_A^* = x_B^*$, as all n servers have equal single-listing arrival rates. Otherwise $(x_A^* \neq x_B^*)$, the social

Figure 12. (Color online) Social Welfare Comparison Under CUS with an Odd Number of Servers



Note. $c = 0.1, t = 2.1, n = 5$.

optimum is asymmetric as odd-indexed servers differ from even-indexed servers in their single-listing arrival rates. Theorem 8 characterizes when the social optimum is symmetric/asymmetric.

Theorem 8 (Social Optimum for Even n). *Under CUS, if n is even:*

(i) *If n is a multiple of four, the social optimum is symmetric, that is,* $x_A^* = x_B^*$.

(ii) *If n is not a multiple of four, the social optimum is asymmetric if and only if some customers are multilisted and some balk, that is,* $x_A^* \neq x_B^*$ *if and only if* $x_R^* > 0$ *and* $x_A^* + x_B^* + x_R^* < \frac{1}{n}$.

If n is a multiple of four, then multilisting customers will have a constant expected utility independent of their server preferences, even when odd-indexed and even-indexed servers receive different volumes of single-listing customers. This negates the value of asymmetric routing, leading to a symmetric social optimum. By contrast, if n is even but not a multiple of four, asymmetric routing will cause multilisting customers' expected utility to vary with their type and hence is socially optimal when both multilisting and balking are active, as in the base model (cf. Proposition 7).

When n is odd, the social optimum can be symmetric or asymmetric, depending on the values of the parameters. We present examples for $n = 3$ in Section EC.7.4 in the Online Appendix.

9. Conclusion

We study the throughput and welfare implications of multilisting in service systems where customers have heterogeneous preferences toward different servers and make strategic decisions. We find that in equilibrium, the multilisting system (where customers choose among multilisting, single-listing, and balking) attracts more customers than the single-listing system (where customers choose between single-listing and balking) and the pooling system (where customers choose between multilisting and balking) and hence achieves higher throughput than both. As for social welfare, multilisting still outperforms pooling but nevertheless falls short of single-listing when the demand size is large. This implies that providing customers with an additional multilisting option, albeit attractive, may only hurt the interest of customers. This welfare loss is driven by self-interested customers' failure to internalize externalities on others when they are left to their own devices.

This social inefficiency in equilibrium motivates us to investigate the socially optimal routing policy in the multilisting system. When servers are symmetric, the equilibrium is also intuitively symmetric, with each server receiving an equal volume of single-listing customers (besides multilisting customers). By contrast, the socially optimal routing policy can be asymmetric. Further, relative to the socially optimal outcome, customers in equilibrium under-choose multilisting when the demand size is small but may over-choose multilisting otherwise. We identify socially optimal (asymmetric) prices to coordinate customer behavior in equilibrium and restore social efficiency. We then investigate the revenue maximization problem and show that the revenue-optimal prices can also be asymmetric. Finally, we extend our two-server analysis to a general number of servers.

In our model, we treat the capacity of servers as given, unaffected by the queue configuration. However, in some settings such as medical services, servers can be strategic and respond to the queue configuration. The empirical literature has documented evidence for the competition effect and the social loafing effect of pooling strategic servers (e.g., Wang and Zhou 2018). Future research can investigate how multilisting moderates such strategic interaction among servers.

Endnote

[1] We denote the expressions of (Λ_1, Λ_2) for the CUS mode by $(\Lambda_1^c, \Lambda_2^c)$ and those for the CUE mode by $(\Lambda_1^e, \Lambda_2^e)$.

References

Anton E, Ayesta U, Jonckheere M, Verloop IM (2021) On the stability of redundancy models. *Oper. Res.* 69(5):1540–1565.

Armony M, Roels G, Song H (2021) Pooling queues with strategic servers: The effects of customer ownership. *Oper. Res.* 69(1):13–29.

Ata B, Skaro A, Tayur S (2017) OrganJet: Overcoming geographical disparities in access to deceased donor kidneys in the United States. *Management Sci.* 63(9):2776–2794.

Bassamboo A, Randhawa RS, Van Mieghem JA (2012) A little flexibility is all you need: On the asymptotic value of flexible capacity in parallel queueing systems. *Oper. Res.* 60(6):1423–1435.

Cachon GP, Zhang F (2007) Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Sci.* 53(3):408–420.

Fay S, Xie J (2008) Probabilistic goods: A creative way of selling products and services. *Marketing Sci.* 27(4):674–690.

Gardner K, Hyytiä E, Righter R (2019) A little redundancy goes a long way: Convexity in redundancy systems. *Performance Evaluation* 131:22–42.

Gardner K, Harchol-Balter M, Hyytiä E, Righter R (2017a) Scheduling for efficiency and fairness in systems with redundancy. *Performance Evaluation* 116:1–25.

Gardner K, Harchol-Balter M, Scheller-Wolf A, Velednitsky M, Zbarsky S (2017b) Redundancy-d: The power of d choices for redundancy. *Oper. Res.* 65(4):1078–1094.

Gardner K, Zbarsky S, Doroudi S, Harchol-Balter M, Hyytia E (2015) Reducing latency via redundant requests: Exact analysis. *ACM SIGMETRICS Performance Evaluation Rev.* 43(1): 347–360.

Gardner K, Zbarsky S, Doroudi S, Harchol-Balter M, Hyytiä E, Scheller-Wolf A (2016) Queueing with redundant requests: Exact analysis. *Queueing Systems* 83(3–4):227–259.

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.

Hassin R (2016) *Rational Queueing* (CRC Press, Boca Raton, FL).

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, International Series in Operations

Research and Management Science, vol. 59 (Springer Science & Business Media, New York).

Hotelling H (1929) Stability in competition. *Econom. J.* 39(153): 41–57.

Jerath K, Netessine S, Veeraraghavan SK (2010) Revenue management with strategic customers: Last-minute selling and opaque selling. *Management Sci.* 56(3):430–448.

Nageswaran L (2023) Implications of vaccine shopping during pandemic. *Production Oper. Management* 32(4):1133–1149.

Nageswaran L, Scheller-Wolf A (2022) Queues with redundancy: Is waiting in multiple lines fair? *Manufacturing Service Oper. Management* 24(4):1959–1976.

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

Salop SC (1979) Monopolistic competition with outside goods. *Bell J. Econom.* 10(1):141–156.

Sunar N, Tu Y, Ziya S (2021) Pooled vs. dedicated queues when customers are delay-sensitive. *Management Sci.* 67(6): 3785–3802.

Tsitsiklis J, Xu K (2012) On the power of (even a little) resource pooling. *Stochastic Systems* 2(1):1–66.

Visschers J, Adan I, Weiss G (2012) A product form solution to a system with multi-type jobs and multi-type servers. *Queueing Syst.* 70(3):269–298.

Wang J, Zhou YP (2018) Impact of queue configuration on service time: Evidence from a supermarket. *Management Sci.* 64(7): 3055–3075.

Wang Z, Yang L, Cui S, Ülkü S, Zhou YP (2023) Pooling agents for customer-intensive services. *Oper. Res.* 71(3):860–875.

Xie R, Grosof I, Scully Z (2024) Heavy-traffic optimal size- and state-aware dispatching. *Proc. ACM Measurement Anal. Comput. Systems* 8(1):1–36.

Xu X, Lian Z, Li X, Guo P (2016) A hotelling queue model with probabilistic service. *Oper. Res. Lett.* 44(5):592–597.

Yang L, Debo LG, Gupta V (2019) Search among queues under quality differentiation. *Management Sci.* 65(8):3605–3623.

Zhan D, Ward AR (2014) Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing Service Oper. Management* 16(2):220–237.

Zhan D, Ward AR (2019) Staffing, routing, and payment to trade off speed and quality in large service systems. *Oper. Res.* 67(6):1738–1751.

Zhou Chen is an associate professor at the School of Economics and Management, Southeast University. Her research interests include service operations, queueing games, mechanism design, and decision making under uncertainty.

Yichuan Ding is an associate professor, Desautels Faculty Scholar, and academic director of the Global Manufacturing and Supply Chain Management program at McGill University's Desautels Faculty of Management. His research applies operations research and artificial intelligence to improve the efficiency and equity of healthcare delivery systems.

Luyi Yang is an associate professor in the Operations and Information Technology Management Group at the University of California, Berkeley's Haas School of Business. His research interests include service operations, digital marketplaces, sustainability, and operations-marketing interface.