Equilibrium behavior in tandem Markovian queues with heterogeneous delay-sensitive customers

Dimitrakopoulos Yiannis (✉ dimgiannhs@math.uoa.gr)
National and Kapodistrian University of Athens

Research Article

Keywords: queueing games, strategic customers, equilibrium strategies, tandem queues, join-or-balk dilemma

Posted Date: January 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2402714/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Operations Research Forum on October 14th, 2023. See the published version at https://doi.org/10.1007/s43069-023-00263-y.

Equilibrium behavior in tandem Markovian queues with heterogeneous delay-sensitive customers

Yiannis Dimitrakopoulos

Department of Mathematics, National and Kapodistrian University of Athens, Panepistimiopolis, Athens, 15124, Greece.

Contributing authors: dimgiannhs@math.uoa.gr;

Abstract

We consider a system of two unobservable Markovian queues in tandem with strategic customers, who are heterogeneous regarding their delay sensitivity. The customers decide upon arrival whether to balk or join the system and receive service in the first queue, or in both queues. We analyze their equilibrium strategic behavior which is specified by double threshold strategies regarding their delay sensitivity parameter (one threshold for each queue). Moreover, we compare the strategic behavior of the heterogeneous customer population with its homogeneous counterpart. We complement our theoretical results with numerical experiments and provide managerial insight into the optimal control of the system parameters.

Keywords: queueing games; strategic customers; equilibrium strategies; tandem queues; join-or-balk dilemma

1 Introduction

The stochastic modeling and analysis of a large number of realistic applications that include some kind of complementary services, i.e. when the total value of serving a customer depends on a set of smaller services, is carried out through the formation and study of networks of service systems. Such realistic applications occur either in service delivery systems, e.g. at the emergency department (ED) of a healthcare facility or in make-to-order production systems. The queueing networks that represent such systems correspond to a variety of topologies, from relatively simple ones such as stations in tandem (when the service stages are predetermined and identical for all customers) to very complex ones (when the customers receive complementary services through alternative routes). In addition, most of these applications, involve active customers who make individual decisions for maximizing their utility, taking into account the congestion of the system at any level of service and the available information. For example, patients who arrive at an ED and need to go through blood tests before a physical examination may balk due to excessive delays.

Therefore, the appropriate framework for the study of service systems with complementary services should involve the consideration of strategic customers. This subarea of Queueing deals with the analysis of customers' decentralized behavior from an economic perspective, and started with the pioneering work of Naor [1] who studied the join-or-balk dilemma when the customers observe the queue length before making their decisions (observable model). Subsequently, Edelson and Hilderbrand [2] studied the same problem, when the customers assess the expected congestion of the system only through its known parameters, but without observing the queue length (unobservable model). Since then, there is a growing body of literature that extends in various directions.

For example, there are a number of papers dealing with various types of decisions (entry/exit, arrival time selection, reneging time selection, e.g., [3]), different types of systems (e.g. $M/M^K/1$: [4], clearing systems: [5], M/M/1 with retrials: [6], M/M/1 with service vacations: [7], [8, 9]), system operation management (e.g. [10, 11]): dynamic control of service rate), or coordinating the system [12]. The corresponding studies on queueing networks have mostly focused on parallel queueing systems, i.e., on customer routing problems (e.g., [13]), or in telecommunications (e.g.,[14]). Other works emphasize that customers exhibit nonlinear aversion to delay, see e.g., [15, 16]. For a comprehensive survey of queueing games, we refer to Hassin and Haviv [17] and Hassin [18].

In many cases, customers differ in their evaluation of excessive delays, since different customers value time differently. In fact, some customers are more patient and can wait for a long time to obtain service, while other customers are more sensitive to system congestion and could leave after a short time of waiting. Therefore, customers' delay sensitivity is an important factor that plays an essential role not only in customers' joining decisions but also in the operation of the system, especially in the case of complementary services where customers may abandon the system after completing some of the stages of service collecting a smaller service reward, before concluding their complete service. In applications, we observe this effect on ride-hailing platforms which are designed to offer different types of services based on the needs of customers that are more sensitive to service congestion. The effect of customers' heterogeneity regarding delay sensitivity on their strategic behavior was studied in [19] for the M/M/1 queue under three different levels of available information upon customers' arrival: no, partial and full information. The authors identified special cases where more accurate delay information improves the performance of the system. On the other hand, Guo and Hassin [9] analyzed the case of heterogeneous customers facing the join/balk dilemma in a vacation queue and showed that may exist multiple equilibria in such a system.

The present work focuses on a thread of research that studies the strategic customer behavior in tandem queues, see e.g., [20], [21] and [22]. More specifically, in the present paper, we analyze strategic customer joining behavior in a Markovian tandem queueing system, where customers are heterogeneous in their delay sensitivity. Specifically, we consider two $M/M/1$ queues in series with strategic customers, who may enter the system only from the first queue and proceed to the second queue after completing their service in the first one. Customers are making joining decisions at the entrance of each queue having the option to leave permanently the system before entering the second queue, and collecting the reward by the first service. In addition, we assume that both queues are unobservable, thus customers' join/balk decisions depend only on the operational and economic parameters of the system and not on the knowledge of the system state at their arrival instants, assuming different operational parameters and service values for each queue. As in the model by Burnetas [20] who analyzed customers' strategic behavior for the join/balk dilemma considering a tandem system of N unobservable Markovian queues, customers' decisions must take into account not only the trade-off between the acquired service value and the waiting cost in the current queue, but also the possibility that this loss may be compensated by joining the subsequent queue. In [20], the author determined the unique subgame perfect equilibrium of the problem by a backward recursion scheme.

The crucial difference from the model of Burnetas [20], is that in the present paper we analyze the effect of customer heterogeneity in delay sensitivity on customers' strategic behavior, which leads to the existence of equilibrium threshold strategies on their delay cost parameter. Since the complexity of the analysis increases significantly due to customers' heterogeneity, the present model permits to obtain of comparative results between the cases of heterogeneous and homogeneous customers for several performance measures, and, especially for equilibrium arrival rates. Other significant works in this research thread are the papers in [21], [23], and [22].

D'Auria and Kanta [21] consider a tandem system with two queues where customers cannot balk in between, and they have to go through the whole system. The case where the arriving customers know the total number of customers in the system was studied by D'Auria and Kanta [23] who proved the existence of a unique strategic equilibrium. Finally, Ji, Roet-Green, and Snitkovsky [22] investigate a tandem system of two M/M/1 queues where queue-length information is available at customers' arrival and customers have the option to renege at any time. In the case where customers observe the state of the entire system and decide whether to join or not, they showed that customers' equilibrium strategy is not necessarily a function of the total number of customers in the system, and also that customers may balk in front of the second queue, but never renege from the first queue. In all cases, customers are considered to be homogeneous.

The contribution of this paper is to enrich the literature in tandem Markovian queueing systems with strategic customers examining the more realistic case where customers differ on delay sensitivity. We show that equilibrium strategies are threshold-based in delay sensitivity, and we derive explicitly the equilibrium for the whole range of the parameters. Moreover, we apply our solution in the case of a uniformly distributed delay sensitivity and perform several numerical experiments under different cases of the operational and economic parameters to analyze the effect of the variance in customers' equilibrium behavior. The main message of the present work is that heterogeneity matters and can be quantified using our approach.

Specifically, there exists a set of critical values for the parameters where customer heterogeneity disappears and customers behave as almost homogeneous. In addition, as the second queue becomes more valuable for the customers or works faster, then, in equilibrium, fewer customers will join it as the variance of delay sensitivity increases. On the other hand, for low values of the fraction of the service values or the service rates, higher variance results in higher arrival rates. Therefore, there exist ideal sets of parameter values for which the information asymmetry can be controlled by the administrator of the system either by setting the service rates appropriately or by imposing an admission fee or a subsidy which in turn controls the fraction of the service values.

The rest of the paper is organized as follows. In Section 2 we describe the model, whereas in Section 3 we set up the strategy space, the equilibrium strategies, and the corresponding benefit functions. In Section 4 we establish the form of the unique symmetric equilibrium threshold strategy. In Section 5, we present numerical results on the comparison with the homogeneous case considering the equilibrium throughput and the equilibrium customers' social welfare. Finally, in Section 6, we conclude and present possible extensions for future research. Some technical material and the notation are presented in Appendix.

2 Model Description

We consider two M/M/1 tandem queues, namely queues 1 and 2, where the service times are exponentially distributed with rates μ_1 and μ_2, respectively. Potential customers arrive at queue 1 according to a Poisson process with rate Λ. We assume that both queues have infinite waiting spaces, and the service discipline is FCFS.

Customers are strategic, delay-sensitive, and risk-neutral optimizers: Their objective is to maximize their individual expected net benefit from their service at the system, taking into account that all customers have similar objectives and the same level of information. Each customer first decides whether to join or balk at queue 1 upon arrival. If she joins, she makes a second join/balk decision upon her arrival at the second queue.

The service values at queues 1 and 2 for an arbitrary customer are R_1 and R_2, respectively. To avoid trivialities, we assume that $R_1, R_2 > 0$. A customer may choose either to balk upon arrival (in which case she receives 0 total service value), or to join the first queue and to depart after finishing her service there (in which case her total service value is R_1), or to go through both queues (in which case she receives $R_1 + R_2$ total service value).

Customers accumulate waiting costs as long as they stay in the system. Considering their delay sensitivity, we assume that they are heterogeneous in their valuations. Specifically, their valuation on the delay cost is linear on the waiting time with delay cost parameter C per unit time which is a customer-specific parameter indicating the importance of time. Furthermore, we assume that C is a non-negative continuous random variable with $H(\cdot)$ being its cumulative distribution having bounded support. Therefore $C \sim H(c),\ c \in [c_l, c_h]$, where $0 \leq c_l \leq c_h < \infty$, and c_l, c_h represent the lowest and the highest value of the delay sensitivity parameter, defined as $c_l = \max\{c : H(c) = 0\}$ and $c_h = \min\{c : H(c) = 1\}$. For a tagged customer, we denote the realization of her delay sensitivity parameter by c. This is considered private information, not known to the other customers.

At their arrival instants, customers are not informed about the state of the queues, i.e., we consider the unobservable case of this model, but they know the operational and economic parameters of the system.

For convenience in the presentation, we assume that $\Lambda < \min\{\mu_1, \mu_2\}$ in order to ensure the stability of the system even if all customers decide to go through both queues. In Figure 1, we illustrate the tandem system along with the customers' decisions and the operational and economic parameters.

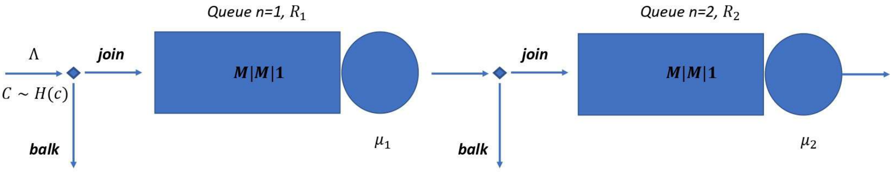

Fig. 1 Visualization of the setting

3 Game formulation

Since the expected benefit of an arriving customer at the system depends on the effective arrival rate at each queue, which in turn depends on the joining decisions of the population of customers, we have that the interaction among the customers can be viewed as a symmetric game (since they are a priori indistinguishable).

The system is unobservable, i.e., the customers are not aware of the system state at any queue upon their arrival instants. Therefore, a pure strategy of a customer with a given delay sensitivity c (in the following we refer to her as a customer of type$-c$) is of the form $\underline{x}(c) = (x_1(c), x_2(c))$ where $x_n(c) \in \{0,1\}$, and $x_n(c) = 1$ if and only if the customer joins queue n. Note that, if $x_1(c) = 0$, then $x_2(c) = 0$, because in case a customer balks after departing from queue 1, then at the same time she definitely departs from the system. Therefore, a customer's pure strategy is a function

$$X : [c_l, c_h] \to \{0,1\}^2 : c \to (x_1(c), x_2(c)), \tag{1}$$

with $X(c) = (x_1(c), x_2(c)) \in \{(0,0), (1,0), (1,1)\}$. Note that, $(0,0)$ represents the decision that a type$-c$ customer does not enter the system, $(1,0)$ refers to the decision of joining only the first queue and then leave the system, and, finally, $(1,1)$ stands for joining for both queues.

Furthermore, we consider an equivalent definition of X which mirrors customers' pure strategies to a partition of $[c_l, c_h]$, as follows

$$V_{00}(X) = \{c \in [c_l, c_h] : X(c) = (0,0)\}, \tag{2}$$
$$V_{10}(X) = \{c \in [c_l, c_h] : X(c) = (1,0)\}, \tag{3}$$
$$V_{11}(X) = \{c \in [c_l, c_h] : X(c) = (1,1)\}. \tag{4}$$

This partition corresponds to the notion of pure strategies employed by customers of type$-c$, since for any $c \in V_{00}(X)$ customers will balk the system, for any $c \in V_{10}(X)$ will join only the first queue, and for any $c \in V_{11}(X)$ will definitely join both queues. Note that, it would be equivalent to a threshold strategy if and only if there exist thresholds $c_1 \geq c_2 \in [c_l, c_h]$ with $c_1 \geq c_2$ where

$$V_{11}(X) = [c_l, c_2], V_{10}(X) = (c_2, c_1], V_{00}(X) = (c_1, c_h]. \tag{5}$$

In order to derive the effective arrival rates at each queue of the system, we note that $V_{10}(X) \cup V_{11}(X)$ refers to those type$-c$ customers who will join at least queue 1, whereas $V_{11}(X)$ corresponds to those customers that will join at both queues 1 and 2. Letting $V_1(X) := V_{10}(X) \cup V_{11}(X)$ and $V_2(X) := V_{11}(X)$, then the corresponding effective arrival rates denoted by $\lambda_1(X), \lambda_2(X)$ at queue 1 and 2, respectively, can be defined as functions of X, as

$$\lambda_1(X) = \Lambda P\left(V_1(X)\right) = \Lambda \int_{V_1(X)} dH(c) \tag{6}$$

$$\lambda_2(X) = \Lambda P\left(V_2(X)\right) = \Lambda \int_{V_2(X)} dH(c). \tag{7}$$

Note that $\lambda_1(X) \geq \lambda_2(X)$, since we consider that queues 1 and 2 are in series and arrivals occur only at queue 1, and, thus, the arrival rate at queue 2 cannot exceed the arrival rate of queue 1.

Therefore, under a given strategy X, the corresponding effective arrival rates at queues 1, 2 are equal to $\lambda_1(X)$ and $\lambda_2(X)$, respectively. Thus the expected sojourn time (delay) of any customer entering queue n is given by $W_n(X) = \frac{1}{\mu_n - \lambda_n(X)}$ with $\lambda_n(X) < \mu_n$.

Suppose that the population of customers follows a strategy X and let $\lambda_n(X)$, $n = 1, 2$, be the corresponding arrival rates. Then, the expected net benefit of a tagged customer with delay sensitivity parameter c' for joining queue n is

$$B_{n,c'}(X) = R_n - \frac{c'}{\mu_n - \lambda_n(X)}. \tag{8}$$

Therefore, the total expected utility of the tagged type$-c'$ customer from following the decision $Y(c') = \underline{y}(c') = (y_1(c'), y_2(c'))$, $y_1, y_2 \in \{0, 1\}$, dictated by her strategy Y, when all others follow a strategy X, is

$$U_{c'}(Y;\ X) = y_1(c')\left(R_1 - \frac{c'}{\mu_1 - \lambda_1(X)}\right) + y_2(c')\left(R_2 - \frac{c'}{\mu_2 - \lambda_2(X)}\right). \tag{9}$$

Let $Y^*(X)$ be her optimal response against X, which refers to the strategy that maximizes $U_{c'}(Y;\ X)$ given that $1 \geq y_1(c') \geq y_2(c') \geq 0$, then a symmetric Nash equilibrium is defined as a strategy X such that $Y^*(X) = X$, i.e., a best response against itself.

The best response $Y^*(X)$ is determined by solving for all c' the optimization problem

$$B_{1,c'}(X)y_1^* + B_{2,c'}(X)y_2^* = \max_{y_1, y_2 \in \{0,1\}} \{B_{1,c'}(X)y_1 + B_{2,c'}(X)y_2\}. \tag{10}$$

Note that the above maximization could be separated into the corresponding joining decisions at each queue considering the signs of $B_{n,c'}$ and $B_{1,c'} + B_{2,c'}$ and their sum, as follows,

i. customer of type$-c'$ joins queue 1 $\iff \max\{B_{1,c'},\ B_{1,c'} + B_{2,c'},\ 0\} > 0$,
ii. customer of type$-c'$ joins queue 2 $\iff \max\{B_{2,c'},\ 0\} > 0$.

In Table 1, we present the feasible best responses of a tagged customer, according to the relevant signs of the above quantities.

B_1	B_1+B_2	B_2	(y_1^*, y_2^*)
$+$	$+$	$+$	$(1,1)$
$+$	$+$	$-$	$(1,0)$
$+$	$-$	$+$	Not feasible
$+$	$-$	$-$	$(1,0)$
$-$	$+$	$+$	$(1,1)$
$-$	$+$	$-$	Not feasible
$-$	$-$	$+$	$(0,0)$
$-$	$-$	$-$	$(0,0)$

Table 1 Feasible best responses of tagged customer.

Therefore, the tagged customer's best response, $\underline{y}^*(c')$, assumes the form

$$(y_1^*(c'), y_2^*(c')) = \begin{cases} (0,0), & B_{1,c'}(X) \leq 0 \text{ and } B_{1,c'}(X) + B_{2,c'}(X) \leq 0 \\ (\{0,1\},0), & B_{1,c'}(X) = 0 \text{ and } B_{2,c'}(X) \leq 0 \\ (1,0), & B_{1,c'}(X) \geq 0 \text{ and } B_{2,c'}(X) \leq 0 \\ (1,\{0,1\}), & B_{1,c'}(X) \geq 0 \text{ and } B_{2,c'}(X) = 0 \\ (1,1), & B_{2,c'}(X) \geq 0 \text{ and } B_{1,c'}(X) + B_{2,c'}(X) \geq 0 \\ (\{0,1\},\{0,1\}), & B_{1,c'}(X) \leq 0 \text{ and } B_{1,c'}(X) + B_{2,c'}(X) = 0 \\ (\{0,1\},\{0,1\}), & B_{1,c'}(X) = B_{2,c'}(X) \\ & = B_{1,c'}(X) + B_{2,c'}(X) = 0. \end{cases} \tag{11}$$

Note that by (8), for any fixed X, $B_{n,c'}(X)$ for $n = 1,2$, and $B_{1,c'}(X) + B_{2,c'}(X)$ are strictly decreasing in c' . Therefore, by (11), when all others follow any fixed X, a customer of type$-c'$ won't enter the system if $c' > \max\{k_1, k_{12}\}$, will enter only queue 1 if $k_2 < c' < k_1$, and she will join both queues if $c' \leq \min\{k_2, k_{12}\}$, where

$$\begin{aligned} k_1 &= R_1\left[\mu_1 - \lambda_1(X)\right], \\ k_2 &= R_2\left[\mu_2 - \lambda_2(X)\right], \\ k_{12} &= (R_1 + R_2)\left[\frac{1}{\mu_1 - \lambda_1(X)} + \frac{1}{\mu_2 - \lambda_2(X)}\right]^{-1}. \end{aligned}$$

Therefore, depending on the values of the parameters, and the relative ordering of k_1, k_2 and k_{12}, it follows that there exist thresholds $c_1 \geq c_2$ such that the best response of the tagged customer to be

$$(y_1^*, y_2^*) = \begin{cases} (0,0), \ c' > c_1, \\ (1,0), \ c_2 < c' \leq c_1, \\ (1,1), \ c' \leq c_2, \end{cases}, \tag{12}$$

and, thus, $V_{11}(Y^*(X)) = [c_l, c_2]$, $V_{10}(Y^*(X)) = (c_2, c_1]$ and $V_{00}(Y^*(X)) = (c_1, c_h]$.

As we have noticed, due to the monotonicity of $B_{n,c'}$ and $B_{1,c'} + B_{2,c'}$ with respect to c', the best response of the tagged is specified by thresholds c_1, c_2 for joining each queue, formulating a threshold-based strategy of the form (c_1, c_2)

where a customer of type$-c$ will join queue n if and only if $c \leq c_n$. Therefore, the symmetric Nash equilibrium would also be a threshold strategy denoted by (c_1^e, c_2^e).

In the following, we proceed with the equilibrium analysis restricting to the following set of threshold joining strategies

$$TH = \{(c_1, c_2) : c_1, c_2 \in [c_l, c_h],\ c_1 \geq c_2\}.$$

4 Equilibrium analysis

We consider threshold joining strategies of the form (c_1, c_2) as defined in (5) where a customer of type c joins queue 1 if and only if $c \leq c_1$, and queue 2 if and only if $c \leq c_2$. Denoting C a random variable representing the delay sensitivity of an arbitrary customer, with distribution $H(c)$, the corresponding joining probabilities are $P(C \leq c_1)$ for joining queue 1 and $P(C \leq c_2, C \leq c_1)$ for joining queue 2, respectively. But, since only departures from queue 1 can be arrivals to queue 2, we can restrict our attention to the case where $c_2 \leq c_1$. Therefore, the system corresponds to two M/M/1 queues in tandem with arrival rates $\lambda_1 = \Lambda H(c_1)$ and $\lambda_2 = \Lambda H(c_2)$. Note that there is a one-one correspondence between $V_1(X), V_2(X)$ and the threshold strategies c_1, c_2. The illustration of the system when all customers follow a threshold strategy $\underline{c} = (c_1, c_2) \in TH$ can be seen in Figure 2.

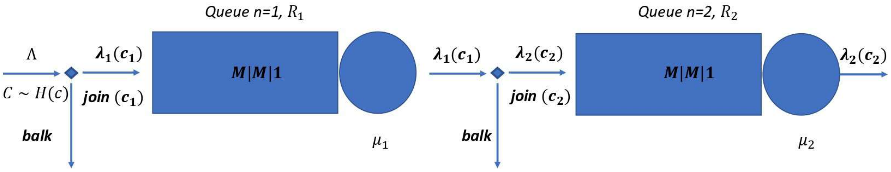

Fig. 2 The tandem system in the framework of customer joining threshold strategies

Therefore, letting $\underline{\tilde{c}} = (\tilde{c}_1, \tilde{c}_2)$ be the threshold strategy followed by a tagged type-c' customer, and $\underline{c} = (c_1, c_2)$ the one followed by the population of customers, her total expected net benefit in (10) can be written as:

$$U_{c'}(\underline{\tilde{c}};\ \underline{c}) = B_{1,c'}(c_1)\mathbb{1}_{\{c' \leq \tilde{c}_1\}} + B_{2,c'}(c_2)\mathbb{1}_{\{c' \leq \tilde{c}_2\}}, \tag{13}$$

and her optimal response is defined as the threshold strategy $\underline{\tilde{c}}^* = (\tilde{c}_1^*, \tilde{c}_2^*)$, that maximizes her expected net benefit $U_{c'}(\underline{\tilde{c}};\ \underline{c})$ when all others follow the threshold strategy $\underline{c} = (c_1, c_2)$. This can be written as follows,

$$(\tilde{c}_1^*, \tilde{c}_2^*) = \arg \max_{\tilde{c}_1, \tilde{c}_2 \in [c_l, c_h] : \tilde{c}_1 \geq \tilde{c}_2} \{B_{1,c'}(c_1)\mathbb{1}_{\{c' \leq \tilde{c}_1\}} + B_{2,c'}(c_2)\mathbb{1}_{\{c' \leq \tilde{c}_2\}}\}. \tag{14}$$

In correspondence with the previous analysis in Section 3, her optimal response is given by

$$(\tilde{c}_1^*, \tilde{c}_2^*) = \begin{cases} (c_l, c_l), & B_{1,c'}(c_1) \leq 0 \text{ and } B_{1,c'}(c_1) + B_{2,c'}(c_2) \leq 0 \\ (c_1, c_l), & B_{1,c'}(c_1) = 0 \text{ and } B_{2,c'}(c_2) \leq 0 \\ (c_h, c_l), & B_{1,c'}(c_1) \geq 0 \text{ and } B_{2,c'}(c_2) \leq 0 \\ (c_h, c_2), & B_{1,c'}(c_1) \geq 0 \text{ and } B_{2,c'}(c_2) = 0 \\ (c_h, c_h), & B_{2,c'}(c_2) \geq 0 \text{ and } B_{1,c'}(c_1) + B_{2,c'}(c_2) \geq 0 \\ (c_1, c_1), & B_{1,c'}(c_1) \leq 0 \text{ and } B_{1,c'}(c_1) + B_{2,c'}(c_2) = 0 \\ (c_1, c_2), & B_{1,c'} = B_{2,c'} = B_{1,c'} + B_{2,c'} = 0 \end{cases}, \tag{15}$$

where $\tilde{c}_1^*, \tilde{c}_2^*$ represent the optimal thresholds against the threshold strategy (c_1, c_2) visualized in Figure 3.

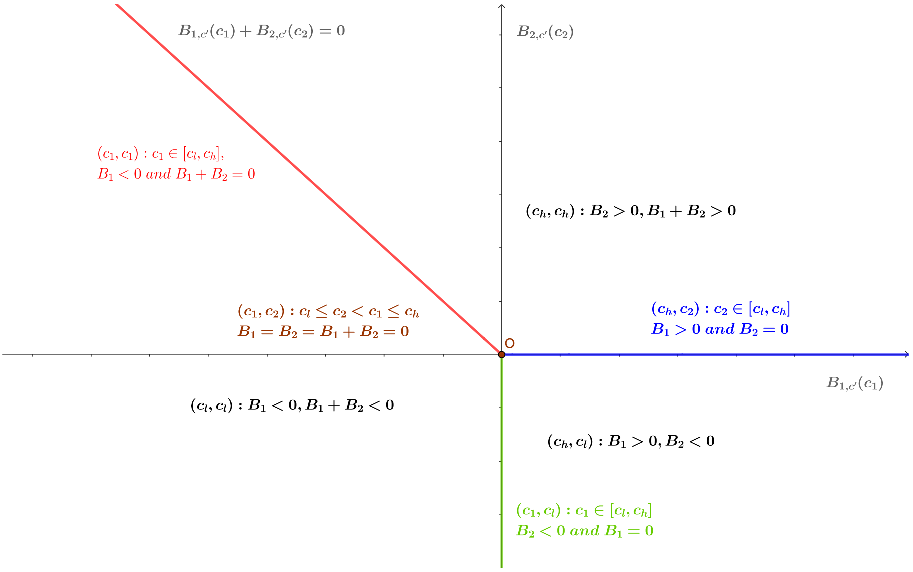
$B_{1,c'}(c_1) + B_{2,c'}(c_2) = 0$
$B_{2,c'}(c_2)$
$(c_1, c_1) : c_1 \in [c_l, c_h]$,
$B_1 < 0$ *and* $B_1 + B_2 = 0$
$(c_h, c_h) : B_2 > 0, B_1 + B_2 > 0$
$(c_1, c_2) : c_l \leq c_2 < c_1 \leq c_h$
$B_1 = B_2 = B_1 + B_2 = 0$
$(c_h, c_2) : c_2 \in [c_l, c_h]$
$B_1 > 0$ *and* $B_2 = 0$
O
$B_{1,c'}(c_1)$
$(c_l, c_l) : B_1 < 0, B_1 + B_2 < 0$
$(c_h, c_l) : B_1 > 0, B_2 < 0$
$(c_1, c_l) : c_1 \in [c_l, c_h]$
$B_2 < 0$ *and* $B_1 = 0$

Fig. 3 The best response of a type$-c'$ customer in the B_1B_2-plain

Consider that all customers follow the threshold strategy (c_1, c_2), and a tagged with delay cost parameter c' who joins queue n according to threshold $\tilde{c}_n$. Once again, the joining decision at each queue depends only on the corresponding thresholds, thus we simplify the notation, as follows. We let

$$B_{n,c'}(c_n) = R_n - \frac{c'}{\mu_n - \lambda(c_n)} \tag{16}$$

be the expected net benefit of the tagged type-c' customer, when she joins queue $n, n = 1, 2$, respectively. Finally, we let r_n be the unique roots of the equations $B_{n,r_n}(r_n) = 0 \iff R_n - \frac{r_n}{\mu_n - \lambda(r_n)} = 0$, or equivalently,

$$r_n = R_n[\mu_n - \Lambda H(r_n)]. \tag{17}$$

In the following Theorem 1, we characterize the equilibrium threshold strategies (c_1^e, c_2^e) for $c_l \leq c_2^e \leq c_1^e \leq c_h$ with respect to the values of the parameters. Each of the seven areas (I, II, III, IV, V, VIa, VIb) in Figure 4, corresponds to each case of the above characterization.

Theorem 1 *The equilibrium joining threshold strategies for the system of the two $M/M/1$ queues in tandem are characterized as follows:*

i. If $R_1 \leq \frac{c_l}{\mu_1}$ and $R_1 + R_2 \leq \frac{c_l}{\mu_1} + \frac{c_l}{\mu_2}$, then the unique equilibrium threshold strategy is (c_l, c_l), which implies $\lambda_1^e = \lambda_2^e = 0$.

ii. If $R_1 \geq \frac{c_h}{\mu_1 - \Lambda}$ and $R_2 \leq \frac{c_l}{\mu_2}$, then the unique equilibrium threshold strategy is (c_h, c_l), which implies $\lambda_1^e = \Lambda$ and $\lambda_2^e = 0$.

iii. If $R_1 + R_2 \geq \frac{c_h}{\mu_1 - \Lambda} + \frac{c_h}{\mu_2 - \Lambda}$ and $R_2 \geq \frac{c_h}{\mu_2 - \Lambda}$, then the unique equilibrium threshold strategy is (c_h, c_h), which implies $\lambda_1^e = \lambda_2^e = \Lambda$.

iv. If $\frac{c_l}{\mu_1} < R_1 < \frac{c_h}{\mu_1 - \Lambda}$ and $R_2 \leq \frac{c_l}{\mu_2}$, then the unique equilibrium threshold strategy is (c_1^e, c_l), with $c_1^e = r_1$. In this case, $\lambda_1^e = \Lambda H(r_1)$ and $\lambda_2^e = 0$.

v. If $R_1 \geq \frac{c_h}{\mu_1 - \Lambda}$ and $\frac{c_l}{\mu_2} < R_2 < \frac{c_h}{\mu_2 - \Lambda}$, then the unique equilibrium threshold strategy is (c_h, c_2^e), with $c_2^e = r_2$. In this case, $\lambda_1^e = \Lambda$ and $\lambda_2^e = \Lambda H(r_2)$.

vi. If $\frac{c_l}{\mu_1} + \frac{c_l}{\mu_2} < R_1 + R_2 < \frac{c_h}{\mu_1 - \Lambda} + \frac{c_h}{\mu_2 - \Lambda}$, $R_1 < \frac{c_h}{\mu_1 - \Lambda}$, and $R_2 > \frac{c_l}{\mu_2}$ then

(a) If $R_2 < \frac{r_1(R_1)}{\mu_1 - \Lambda H(r_1(R_1))}$, the unique equilibrium threshold strategy is (c_1^e, c_2^e), with $c_1^e = r_1$ and $c_2^e = r_2$, if $r_1 > r_2$. In this case, $\lambda_1^e = \Lambda H(r_1)$ and $\lambda_2^e = \Lambda H(r_2)$.

(b) Otherwise, the unique equilibrium threshold strategy is (c_1^e, c_1^e), with $c_1^e = r$, where r is the solution of the following equation

$$R_1 - \frac{r}{\mu_1 - \Lambda H(r)} + R_2 - \frac{r}{\mu_2 - \Lambda H(r)} = 0. \tag{18}$$

In this case, $\lambda_1^e = \lambda_2^e = \Lambda H(r)$.

Proof i. Consider that nobody joins the system, then all follow the threshold strategy $c_1 = c_2 = c_l$. From the best response (c_1^*, c_2^*) of the tagged$-c'$ customer against (c_l, c_l), it follows that (c_l, c_l) would be an equilibrium if

$B_{1,c'}(c_l) \leq 0$ and $B_{1,c'}(c_l) + B_{2,c'}(c_l) \leq 0$ for all $c' \in [c_l, c_h]$. Since both are strictly decreasing in c', it suffices to assume that $B_{1,c_l}(c_l) \leq 0$ and $B_{1,c_l}(c_l)+B_{2,c_l}(c_l) \leq 0$. These imply by (16) that if $R_1 \leq \dfrac{c_l}{\mu_1}$ and $R_2+R_1 \leq \dfrac{c_l}{\mu_1}+\dfrac{c_l}{\mu_2}$, then the unique equilibrium threshold strategy is (c_l, c_l), and, thus, in this case, nobody joins the system.

ii. Consider that all join the first queue and balk, i.e., all follow the threshold strategy $c_1 = c_h,\ c_2 = c_l$. Once again by the best response function of a tagged$-c'$ customer against (c_h, c_l), it follows that in order (c_h, c_l) to be an equilibrium strategy, then $B_{1,c'}(c_h) \geq 0$ and $B_{2,c'}(c_l) \leq 0$ for all $c' \in [c_l, c_h]$. Following that, and since $B_{1,c'},\ B_{2,c'}$ are strictly decreasing in c', it suffices to assume that $B_{1,c_h}(c_h;\ c_h) \geq 0$ and $B_{2,c_l}(c_l;\ c_l) \leq 0$. By substitution from (16), it follows that if $R_1 \geq \dfrac{c_h}{\mu_1 - \Lambda}$ and $R_2 \leq \dfrac{c_l}{\mu_2}$, then the unique equilibrium strategy is (c_h, c_l).

iii. Consider that all join both queues, i.e., all follow the threshold strategy $c_1 = c_h,\ c_2 = c_h$. From the best response of the tagged$-c'$ customer against (c_h, c_h), it follows that (c_h, c_h) would be an equilibrium if $B_{2,c'}(c_h) \geq 0$ and $B_{1,c'}(c_h) + B_{2,c'}(c_h) \geq 0$ for all $c' \in [c_l, c_h]$. Once again, since both are strictly decreasing in c', it suffices to consider $B_{2,c_h}(c_h) \geq 0$ and $B_{1,c_h}(c_h) + B_{2,c_h}(c_h) \geq 0$. Substituting the expressions for B_i considering (16), it follows that if $R_2 \geq \dfrac{c_h}{\mu_1 - \Lambda}$ and $R_2+R_1 \geq \dfrac{c_h}{\mu_2 - \Lambda}+\dfrac{c_h}{\mu_1 - \Lambda}$, the unique equilibrium will be (c_h, c_h).

iv. Consider that a fraction of customers join the first queue and then balk, then all follow a threshold strategy (c_1, c_l), where $c_1 \in (c_l, c_h)$. By (15), (c_1, c_l) would be best response of the tagged$-c'$ customer against (c_1, c_l), when

$$B_{1,c'}(c_1) \geq 0 \text{ and } B_{2,c'}(c_l) \leq 0 \text{ for all } c' \leq c_1, \text{ and,}$$
$$B_{1,c'}(c_1) \leq 0 \text{ and } B_{1,c'}(c_1) + B_{2,c'}(c_l) \leq 0 \text{ for all } c' > c_1.$$

Since $B_{1,c'}$ is strictly decreasing in c', (c_1^e, c_l) with $c_1^e \in (c_l, c_h)$ is an equilibrium strategy if and only if $B_{1,c_1^e}(c_1^e) = 0$, and, also,

$$B_{2,c'}(c_l) \leq 0 \text{ for all } c' \leq c_1^e, \text{ and,}$$
$$B_{1,c'}(c_1^e) + B_{2,c'}(c_l) \leq 0 \text{ for all } c' \geq c_1^e.$$

Once again, due to the monotonicity of $B_{2,c'}$ and $B_{1,c'} + B_{2,c'}$ with respect to c', the above necessary conditions for equilibrium are equivalent to

$$B_{2,c_l}(c_l) \leq 0, \text{ and} B_{1,c_1^e}(c_1^e) + B_{2,c_1^e}(c_l) \leq 0, \tag{19}$$

where c_1^e is the solution of the equation

$$B_{1,c_1^e}(c_1^e) = 0 \Leftrightarrow R_1 - \frac{c_1^e}{\mu_1 - \Lambda H(c_1^e)} = 0 \text{ with } c_1^e \in (c_l, c_h). \tag{20}$$

The latter coincides with r_1 in (17). Substituting the expressions for B_n given in (16) in (19), it follows that $R_2 \leq \dfrac{c_l}{\mu_2}$ and $R_2 \leq \dfrac{c_1^e}{\mu_2}$. Since, c_1^e lies in the interval (c_l, c_h), we derive that $R_2 \leq \dfrac{c_l}{\mu_2}$. In addition, the right part of $R_1 = \dfrac{c_1^e}{\mu_1 - \Lambda H(c_1^e)}$ is increasing in $c_1^e \in (c_l, c_h)$, thus $\dfrac{c_l}{\mu_1} < R_1 < \dfrac{c_h}{\mu_1 - \Lambda}$. Therefore, if $R_2 \leq \dfrac{c_l}{\mu_2}$ and $\dfrac{c_l}{\mu_1} < R_1 < \dfrac{c_h}{\mu_1 - \Lambda}$, then the unique equilibrium strategy is (c_1^e, c_l), where $c_1^e = r_1$, the solution of (17) in (c_l, c_h).

v. Consider that a fraction of customers join the second queue, whereas all have joined queue 1 at first, then all customers follow the threshold strategy $c_1 = c_h,\ c_2 \in (c_l, c_h)$. By (15), (c_h, c_2) would be the best response of the tagged$-c'$ customer against (c_h, c_2), when

$$\begin{aligned} &B_{2,c'}(c_2) \geq 0 \text{ and } B_{1,c'}(c_h) + B_{2,c'}(c_2) \geq 0 \text{ for all } c' \leq c_2, \text{ and,} \\ &B_{1,c'}(c_h) \geq 0 \text{ and } B_{2,c'}(c_2) \leq 0 \text{ for all } c' > c_2. \end{aligned}$$

Since $B_{2,c'}$ is strictly decreasing in c', (c_h, c_2^e) with $c_2^e \in (c_l, c_h)$ is an equilibrium strategy if and only if $B_{2,c_2^e}(c_2^e) = 0$, and, also,

$$\begin{aligned} B_{1,c'}(c_h) + B_{2,c'}(c_2) \geq 0 \text{ for all } c' \leq c_2^e, \text{ and,} \\ B_{1,c'}(c_h) \geq 0 \text{ for all } c' \geq c_2^e. \end{aligned}$$

Once again, due to the monotonicity of $B_{1,c'}$ and $B_{1,c'} + B_{2,c'}$ with respect to c', the above equilibrium conditions are equivalent to

$$B_{1,c_2^e}(c_h) + B_{2,c_2^e}(c_2^e) \geq 0, \text{ and} B_{1,c_h}(c_h) \geq 0, \tag{21}$$

where c_2^e is the solution of the equation

$$B_{2,c_2^e}(c_2^e) = 0 \Leftrightarrow R_2 - \frac{c_2^e}{\mu_2 - \Lambda H(c_2^e)} = 0 \text{ with } c_2^e \in (c_l, c_h). \tag{22}$$

The latter coincides with r_2 the solution of equation (17) for $n = 2$. Substituting the expressions for B_n in (21), it follows that $R_1 \leq \dfrac{c_2^e}{\mu_1 - \Lambda}$ and $R_1 \geq \dfrac{c_h}{\mu_1 - \Lambda}$. Since c_2^e lies in the interval (c_l, c_h), it follows that $R_1 \geq \dfrac{c_h}{\mu_1 - \Lambda}$. In addition, the right part of $R_2 = \dfrac{c_2^e}{\mu_2 - \Lambda H(c_2^e)}$ is increasing in $c_2^e \in (c_l, c_h)$, and, thus $\frac{c_l}{\mu_2} < R_2 < \frac{c_h}{\mu_2 - \Lambda}$. Therefore, if $R_1 \geq \frac{c_h}{\mu_1 - \Lambda}$ and $\frac{c_l}{\mu_2} < R_2 < \frac{c_h}{\mu_2 - \Lambda}$, then the unique equilibrium strategy is (c_h, c_2^e), where $c_2^e = r_2$, the solution of (17) in (c_l, c_h) for $n = 2$.

vi. Consider that a fraction of customers joins both queues, which means that all customers adopt the joining thresholds c_1, c_2 with $c_1 \geq c_2$ since $\lambda_1 \geq \lambda_2$. We consider the following two cases: (a) $c_1 > c_2$ and (b) $c_1 = c_2$.

a. For case (a), where $c_1 > c_2$, (c_1, c_2) would the best response of a tagged type$-c'$ customer against itself, when (15),

$$
\begin{aligned}
&B_{2,c'}(c_2) \geq 0 \text{ and } B_{1,c'}(c_1) + B_{2,c'}(c_2) \geq 0 \text{ for all } c' \leq c_2,\\
&B_{1,c'}(c_1) \geq 0 \text{ and } B_{2,c'}(c_2) \leq 0 \text{ for all } c_2 \leq c' \leq c_1.\\
&B_{1,c'}(c_1) \leq 0 \text{ and } B_{1,c'}(c_1) + B_{2,c'}(c_2) \leq 0 \text{ for all } c' \geq c_1.
\end{aligned}
$$

By the monotonicity of $B_{n,c'}$, i.e., $B_{1,c'}, B_{2,c'}$ are strictly decreasing in c', (c_1^e, c_2^e) is an equilibrium strategy with $c_l < c_2^e < c_1^e < c_h$ if and only if

$$
B_{2,c_2^e}(c_2^e) = 0 \text{ and } B_{1,c_1^e}(c_1^e) = 0 \text{ for } c_l < c_2^e < c_1^e < c_h, \tag{23}
$$

and, in addition,

$$
\begin{aligned}
B_{1,c'}(c_1) + B_{2,c'}(c_2) \geq 0 \text{ for all } c' \leq c_2^e, \text{ and,}\\
B_{1,c'}(c_1) + B_{2,c'}(c_2) \leq 0 \text{ for all } c' \geq c_1^e.
\end{aligned}
$$

Once again, due to the monotonicity of $B_{1,c'} + B_{2,c'}$ with respect to c', the above necessary equilibrium conditions are equivalent to

$$
B_{1,c_2^e}(c_1^e) + B_{2,c_2^e}(c_2^e) \geq 0, \text{ and } B_{1,c_1^e}(c_1^e) + B_{2,c_1^e}(c_2^e) \leq 0, \tag{24}
$$

where c_1^e, c_2^e satisfy the following equations

$$
B_{1,c_1^e}(c_1^e) = 0 \Leftrightarrow R_1 - \frac{c_1^e}{\mu_1 - \Lambda H(c_1^e)} = 0 \text{ with } c_1^e \in (c_2^e, c_h), \tag{25}
$$

$$
B_{2,c_2^e}(c_2^e) = 0 \Leftrightarrow R_2 - \frac{c_2^e}{\mu_2 - \Lambda H(c_2^e)} = 0 \text{ with } c_2^e \in (c_l, c_1^e), \tag{26}
$$

which correspond to the roots r_n of (17) for $n = 1, 2$, respectively. Since $B_{1,c_1^e}(c_1^e) = 0, B_{2,c_2^e}(c_2^e) = 0$, and by substitution of B_n in (24), it follows that

$$
B_{1,c_2^e}(c_1^e) \geq 0 \Leftrightarrow R_1 \geq \frac{c_2^e}{\mu_1 - \Lambda H(c_1^e)}, \tag{27}
$$

$$
B_{2,c_1^e}(c_2^e) \leq 0 \Leftrightarrow R_2 \leq \frac{c_1^e}{\mu_2 - \Lambda H(c_2^e)}, \tag{28}
$$

which both are true if $c_1^e = r_1 > r_2 = c_2^e$. Furthermore, by (25) and (26), it follows that $R_1 = \dfrac{c_1^e}{\mu_1 - \Lambda H(c_1^e)}$ for $c_1^e \in (c_2^e, c_h)$ and $R_2 = \dfrac{c_2^e}{\mu_2 - \Lambda H(c_2^e)}$ for $c_2^e \in (c_l, c_1^e)$. As in the proof of (v), the right part of R_1, R_2 is increasing in $c_2^e \in (c_l, c_1^e)$ and $c_1^e \in (c_2^e, c_h)$, respectively, thus $\dfrac{c_2^e}{\mu_1 - \Lambda H(c_2^e)} < R_1 < \dfrac{c_h}{\mu_1 - \Lambda}$ and $\dfrac{c_l}{\mu_2} < R_2 < \dfrac{c_1^e}{\mu_2 - \Lambda H(c_1^e)}$.
Taking into account the necessary conditions for equilibrium described by in (27) and (28), we derive that the point (R_1, R_2) must lie in the region

$$\left(\frac{c_2^e}{\mu_1 - \Lambda H(c_1^e)}, \frac{c_h}{\mu_1 - \Lambda}\right) \times \left(\frac{c_l}{\mu_2}, \frac{c_1^e}{\mu_2 - \Lambda H(c_2^e)}\right) \tag{29}$$

It is easy to show that the remaining area from the previous cases (i.e., VI in Figure 4) is created from the set of equations,

$$R_1+R_2 = \frac{c_l}{\mu_1}+\frac{c_l}{\mu_2}, R_1+R_2 = \frac{c_h}{\mu_1 - \Lambda}+\frac{c_h}{\mu_2 - \Lambda}, R_1 = \frac{c_h}{\mu_1 - \Lambda}, \text{ and } R_2 = \frac{c_l}{\mu_2}.$$

Clearly, (29) is a sub-area of VI which corresponds to the area VI.a in Figure 4 in which, the unique equilibrium strategy is (c_1^e, c_2^e), where $c_n^e = r_n$, the solution of (17) in (c_l, c_h) for $i = 1, 2$.

b. The remaining area of IV corresponds to the pairs of (R_1, R_2) which belong to VI but not in VI.a (the complement of VI.a). This takes place when (27) and (28) does not hold, i.e., $c_1^e \leq c_2^e$, which correspond to the equilibrium threshold strategy (c_1^e, c_1^e). By using similar arguments, and based on the optimal response of the tagged customer in (15), in the area of VI.b the equilibrium strategy (c_1^e, c_1^e) corresponds to (r, r) where r is the solution of $B_{1,r}(r) + B_{2,r}(r) = 0$, i.e., equation (18).
Finally, regarding the red line in Figure 4 that separates the areas VI.a and VI.b, we can derive its form as follows:
We define the r_1, r_2 solutions of equations (17) as functions of R_1 and R_2, respectively. Specifically,

$$r_1(R_1) = R_1\left[\mu_1 - \Lambda H(r_1(R_1))\right] \Leftrightarrow R_1 = \frac{r_1(R_1)}{\mu_1 - \Lambda H(r_1(R_1))}$$
$$r_2(R_2) = R_2\left[\mu_2 - \Lambda H(r_2(R_2))\right] \Leftrightarrow R_2 = \frac{r_2(R_2)}{\mu_2 - \Lambda H(r_2(R_2))}$$

Cases (a) and (b) are distinguished by the ordering of r_1, r_2, i.e. case (a) holds if $r_1 > r_2$, whereas case (b) holds if $r_1 \leq r_2$. In order to define this area explicitly with respect to the parameters R_1, R_2, we need to derive the line which is defined by the equation $r_1(R_1) = r_2(R_2)$. Substituting the expressions for $r_1(R_1)$ and $r_2(R_2)$, it follows that $R_2 = \dfrac{r_1(R_1)}{\mu_1 - \Lambda H(r_1(R_1))}$.

The latter coincides with the red curve in Figure 4 and its shape depends on the distribution function $H(c)$.

Thus, the case where $R_2 < \dfrac{r_1(R_1)}{\mu_1 - \Lambda H(r_1(R_1))}$, i.e., $r_1 > r_2$ corresponds to the VI.a area. Otherwise, (R_1, R_2) pairs lie in the supplementary area VI.b.

□

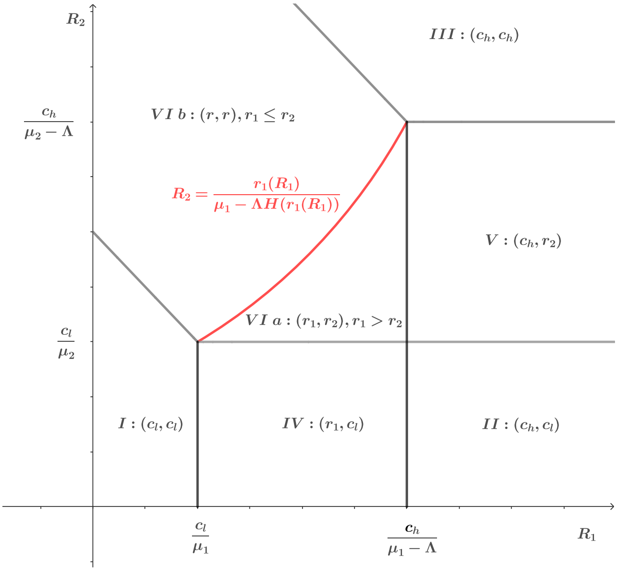
R_2
$\frac{c_h}{\mu_2 - \Lambda}$
$\frac{c_l}{\mu_2}$
$III : (c_h, c_h)$
$VI\ b : (r, r), r_1 \leq r_2$
$R_2 = \frac{r_1(R_1)}{\mu_1 - \Lambda H(r_1(R_1))}$
$V : (c_h, r_2)$
$VI\ a : (r_1, r_2), r_1 > r_2$
$I : (c_l, c_l)$
$IV : (r_1, c_l)$
$II : (c_h, c_l)$
$\frac{c_l}{\mu_1}$
$\frac{c_h}{\mu_1 - \Lambda}$
R_1

Fig. 4 Equilibrium thresholds in the R_2R_1-plane

5 The impact of heterogeneity on customers' strategic behavior: Numerical Experiments

As previously mentioned, the presence of customer heterogeneity regarding delay sensitivity is compatible with a number of situations that arise in several real-life applications, especially in the field of complementary services. These situations determine which parameters can be controlled by a decision-maker to improve the performance of a system. Our objective is to study the effect of customer heterogeneity with respect to delay sensitivity, considering several uniform distributions of different variances for the delay sensitivity C, and the case of homogeneous customers as a benchmark. The latter is approximated by considering a uniform distribution for C where its variance is almost 0.

Specifically, we perform extensive numerical experiments to obtain further insights on the effective arrival rates at each queue $(\lambda_1^e, \lambda_2^e)$, the corresponding joining double threshold strategy (c_1^e, c_2^e), as well as the total expected customers' benefit S^e, in equilibrium, under different scenarios. In all numerical experiments, we apply the uniform distribution for the delay sensitivity C, i.e., customers' waiting cost per unit time, in the interval $[c_l, c_h]$ for given values of c_l, c_h. Note that, all these uniform distributions are of the same mean $\frac{c_l+c_h}{2}$, but we vary their variance $Var(C) = \frac{(c_h-c_l)^2}{12}$.

The customers' expected benefit from joining in equilibrium is given by the formula

$$\begin{aligned} S^e &= \Lambda \int_{c_l}^{c_h} \mathbb{1}_{\{c \leq c_1^e\}} B_{1,c}(c_1^e) + \mathbb{1}_{\{c \leq c_2^e\}} B_{2,c}(c_2^e)\, dH(c) \\ &= \Lambda \left(\int_{c_l}^{c_2^e} (B_{1,c}(c_1^e) + B_{2,c}(c_2^e))\, dH(c) + \int_{c_2^e}^{c_1^e} B_{1,c}(c_1^e)\, dH(c) \right). \end{aligned} \tag{30}$$

In the following figures, we consider the terms almost zero, medium, and large variance, as a characteristic of each uniform distribution that we consider for the waiting cost rate C, since all have the same mean value, i.e., $\mathbb{E}(C) = \dfrac{c_l + c_h}{2} = 4.5$. As mentioned above, the almost zero variance which corresponds to approximately homogeneous customers, i.e., $Var(C) \approx 0$, refers to $C \sim U[c_l = 4.45, c_h = 4.55]$, the medium one with $Var(C) \approx 0.75$, refers to $C \sim U[c_l = 3, c_h = 6]$, and the large one with $Var(C) \approx 4.08$, refers to $C \sim U[c_l = 1, c_h = 8]$. In the following figures, these cases are represented through the dots, the dashes, and the continuous lines, respectively. Moreover, the black and red colors stand for the customers' strategic behavior in queue 1 and queue 2, respectively, plotting the equilibrium arrival rates λ_1^e, λ_2^e in the left panels, and the corresponding equilibrium threshold strategy (c_1^e, c_2^e) in the middle panels, whereas the blue color is used only for the plots of S^e in the right panels of the following figures.

5.1 Ratio of the service values

First, we study the effect on strategic customer behavior of the fraction of service valuation at each queueing system, R_2/R_1, for different stochastic valuations of the customers' delay sensitivity C, as we have analyzed above. We perform two numerical experiments considering different service speeds for the two queues, where queue 1 is faster with $\mu_1 = 8$ than queue 2 where $\mu_2 = 6$ in Figure 5, and vice versa, in Figure 6. This way, we have a clearer view of not only how customers differ in the evaluation of their delay, but also how their behavior is affected by the operation of the whole system. For both experiments, we fix $\Lambda = 5$ and $R_1 = 1$. Therefore, as we increase R_2 with respect to R_1, we cover the following cases

1. Queue 1 is faster and more valuable for customers than queue 2, i.e., Figure 5 for $R_2/R_1 < 1$.

2. Queue 1 is faster than queue 2 but less valuable for the customers, i.e., Figure 5 for $R_2/R_1 > 1$.
3. Queue 1 is slower than queue 2 but more valuable for the customers, i.e., Figure 6 for $R_2/R_1 < 1$.
4. Queue 1 is slower than queue 2 and less valuable for the customers, i.e., Figure 5 for $R_2/R_1 > 1$.

Our objective is to study the impact of customer heterogeneity along the four cases, and whether there exists an ideal fraction of service values for which the information asymmetry can be controlled by the administrator of the system either by setting the service rates appropriately or by imposing an appropriate service fee. Note that, even though the above model formulation does not consider service fees explicitly, the service administrator could impose a service fee that decreases the corresponding service values at the desired levels.

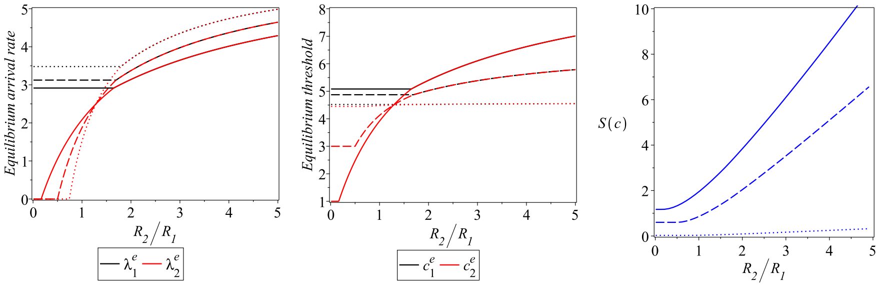

Fig. 5 Equilibrium arrival rate, threshold, and total benefit with respect to service values ratio when $\mu_1 > \mu_2$.

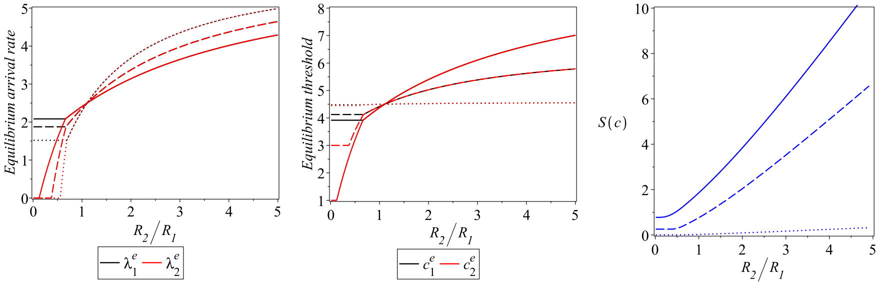

Fig. 6 Equilibrium arrival rate, threshold, and total benefit with respect to service values ratio when $\mu_1 < \mu_2$.

Considering both Figures 5 and 6, we observe that both effective arrival rates, the corresponding thresholds at each queue, and the customers' total benefit, in equilibrium, are non-decreasing in R_2/R_1. Therefore, the more valuable the service in the second queue becomes the more customers will join queue 2, as expected. A faster queue 2 will be more attractive for customers to join it even for smaller values of R_2/R_1, since all plots in Figure 6 are moved to the left with respect to those in Figure 5. Also, in both Figures, we identify a critical value of the fraction of the service values, $R_2/R_1 \approx 1.7$ in Figure 5 and $R_2/R_1 \approx 0.65$ in Figure 6, where the equilibrium threshold strategy changes from two distinct thresholds, i.e., $c_1^e > c_2^e$, to equal ones. This means that as the value of R_2/R_1 increases, customers tend to join either both queues or none, since the service at the second queue will cover potential losses from the expected delays. Note that this critical point is lower when the service speed in queue 2 is higher. Therefore, there is a trade-off between service speed, and the value of service. Indeed, when the second queue works faster, customers will adopt the same threshold for both queues if their service value R_2 is above the 60% of R_1. Finally, customers' overall benefit will be slightly higher if queue 1 works faster.

Considering the effect of customers' heterogeneity with respect to their delay sensitivity on the performance measures in equilibrium, we observe in Figure 5 where $\mu_1 > \mu_2$, that lower variance always leads to a higher arrival rate at queue 1, while the opposite occurs for the second queue when R_2/R_1 is low. Specifically, there is a critical value of $R_2/R_1 \approx 1.3$, where all plots of λ_2^e intersect and the lower variance induces a higher arrival rate at queue 2 afterward. A similar behavior of λ_2^e occurs if the second queue is faster (see Figure 6), but the critical point now moves to a lower value, i.e., when $R_2/R_1 \approx 1.1$, while for low values of R_2/R_1, the arrival rate at the first queue is higher as the variance of C increases. This fact changes after a critical value of R_2/R_1: As in the previous case, lower variance induces a higher arrival rate for both queues, since $\lambda_1^e = \lambda_2^e$. Therefore, in both cases, when the fraction R_2/R_1 is high, the effect of heterogeneity is more detrimental regarding the resulting equilibrium arrival rates, while for low values of R_2/R_1 this happens only if $\mu_1 > \mu_2$. Finally, in both numerical experiments, this critical value also indicates a relative value of the corresponding rewards, where customer heterogeneity disappears and customers behave as homogeneous in the evaluation of their delays.

Regarding the threshold strategies in Figure 5, when the variance is large, we observe that the corresponding threshold is higher, while the arrival rate is lower. Although it seems counter-intuitive, it is quite reasonable in some cases, since the difference $c_h - c_l$ is increasing in the variance. Therefore, it leads to a lower arrival rate ($\lambda_1^e = \lambda H(c_1^e)$, where $H(c) = \frac{c-c_l}{c_h-c_l}$). In this example, for the large variance case, we obtain $c_1^e = 5.1$; thus we have $\lambda_1^e = 5\frac{5-1}{8-1} \approx 2.85$. On the other hand, for the medium variance case, we have that $c_1^e = 4.9$ which implies that $\lambda_1^e = 5\frac{4.9-3}{6-3} \approx 3.16$. Another impact of the variance concerns the threshold strategies. The numerical experiments suggest that the equilibrium threshold strategy at queue 1, is always higher for a larger variance. This does not happen in queue 2, since for low values of R_2/R_1, the lower variance leads to a higher threshold up to a point where $R_2/R_1 \approx 1.3$, and the opposite takes place afterward. In both cases (medium and large variance) when the fraction becomes sufficiently large, the two thresholds c_1^e and c_2^e coincide, similarly to the effective arrival rates. In Figure 6, in contrast to the previous case, the equilibrium threshold strategy for queue 1 is higher for low variance, whereas for higher values of R_2/R_1, the two threshold strategies coincide again and the higher variance leads to a higher threshold, independently of which queue is faster.

Regarding the customers' overall benefit in equilibrium, we observe in Figure 5 that it is always higher when the variance is high and non-decreasing at a high rate. Obviously, for the case of almost zero variance which refers to the approximately homogeneous case, it tends to be zero, which coincides with the case where queue 2 is the faster one, as in Figure 6. The latter indicates that in this case, the relative service speed between the two queues doesn't play any role regarding the effect of the variance on customers' overall benefit.

In short, Figures 5 and 6 illustrate the following:

- The equilibrium arrival rates $(\lambda_1^e, \lambda_2^e)$, the corresponding thresholds (c_1^e, c_2^e), and the customers' overall benefit S^e are non-decreasing in R_2/R_1.
- The equilibrium arrival rates coincide above a critical value of R_2/R_1, since the equilibrium thresholds become equal, and customers decide either to enter both queues or balk from the beginning. A faster queue 2 moves this point towards lower values of R_2/R_1.
- A faster queue 2 will impose a greater effective arrival rate at queue 1 as the variance increases, but lower for the same variance in the opposite case, whereas λ_2^e increases with respect to the variance for low values of R_2/R_1 and decreases for greater ones. Equilibrium thresholds have the opposite behavior as the variance increases. There is a second critical point of R_2/R_1, where the effective arrival rate at queue 2 becomes equal for any variance, and the heterogeneity effect seems to disappear. A higher variance will lead to greater customers' overall benefit in equilibrium.

5.2 Ratio of the service rates

In the second set of experiments, we perform a sensitivity analysis with respect to the ratio μ_2/μ_1 for $\Lambda = 0.8, \mu_1 = 1$, when the service reward is higher on the first queue than the second one, i.e., $(R_1, R_2) = (9, 1)$ in Figure 7, and vice versa where $(R_1, R_2) = (1, 9)$ in Figure 8. Once again, as in Subsection 5.1, we consider the same three different uniform distributions for customers' delay sensitivity, presented at the beginning of Section 5 above.

We consider the four similar cases discussed in the previous subsection but now as a function of the ratio μ_2/μ_1.

1. Queue 1 is faster and more valuable for customers than queue 2, i.e., Figure 7 for $\mu_2/\mu_1 < 1$.
2. Queue 1 is slower than queue 2 but more valuable for the customers, i.e., Figure 7 for $\mu_2/\mu_1 > 1$.
3. Queue 1 is faster than queue 2 but less valuable for the customers, i.e., Figure 8 for $\mu_2/\mu_1 < 1$.

4. Queue 1 is slower than queue 2 and less valuable for the customers, i.e., Figure 8 for $\mu_2/\mu_1 > 1$.

Once again, our objective is to study the impact of customer heterogeneity along the four cases, and whether there exists an ideal fraction of service rates for which the information asymmetry can be controlled by the administrator of the system either by setting the service rates appropriately or by imposing an appropriate service fee.

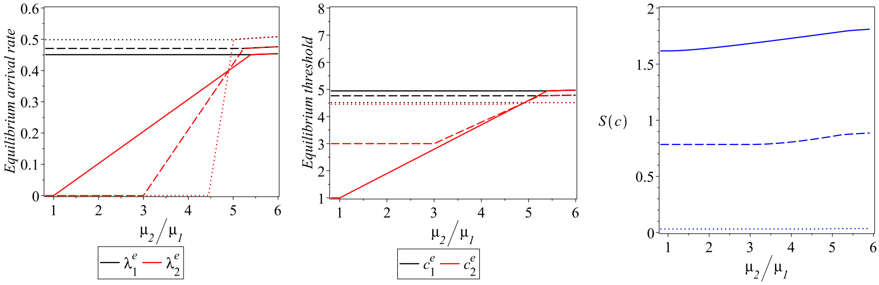

Fig. 7 Equilibrium arrival rate, threshold and total benefit with respect to service rates ratio when $R_1 > R_2$.

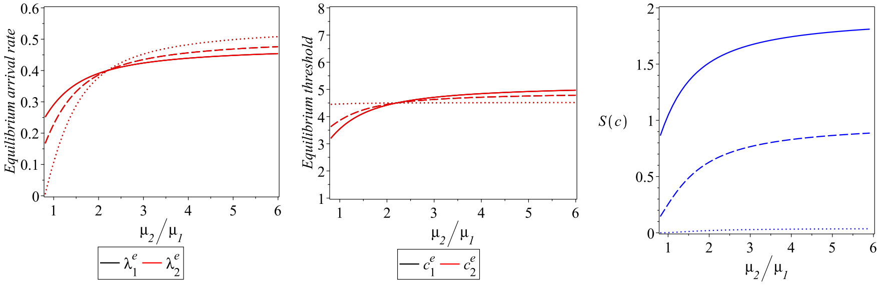

Fig. 8 Equilibrium arrival rate, threshold and total benefit with respect to service rates ratio when $R_1 < R_2$.

In Figure 6 where the service reward for the first queue is very high, while is very low for the second queue, a large fraction of customers join the first queue, and in order to continue to the second queue, it requires a high increase in its service rate. Even in this case, only a very small portion of customers will continue to the second queue. The effect of the variance is very intense in the first queue since it leads fewer customers to decide to join when it is high. As in the previous experiments, for high values of μ_2/μ_1, the equilibrium effective arrival rates for both queues coincide, since the equilibrium thresholds become equal, and customers adopt a threshold to either join both queues or to balk. Also, with respect to the variance, the equilibrium arrival rates at queue 2 intersect at a certain value, and the effect of customer heterogeneity seems to disappear, making customers to be considered almost homogeneous in delay sensitivity. Furthermore, for lower values of μ_2/μ_1 greater variance results in a higher effective arrival rate but in a lower threshold, while greater values of the fraction of service rates lead the arrival rate to become increasing as the variance also increases. Finally, in terms of customers' overall benefit, the higher variance reflects a higher overall benefit but at a very low increasing rate (almost constant) due to the devaluation of the second queue, and customers mainly benefit from the reward collected by the first queue.

On the other hand, in Figure 8 where the service rewards are exchanged, making the service value of the first queue very low, we observe that customers will adopt the same threshold strategy at both queues, i.e., $c_1^e = c_2^e$, since they are mainly interested to collect the high service reward from the second queue, and thus the arrival rates for both queues coincide. Note that, again, there is a critical point of the ratio μ_2/μ_1. For values of μ_2/μ_1 lower than this critical value, the arrival rate increases, and the threshold decreases as the variance increases. For values of μ_2/μ_1 that are higher than the critical value, the effect is reversed. Regarding the customers' overall expected benefit, we observe again that regardless of the order of the service values, the higher variance induces a higher benefit, but at a greater rate now.

In short, Figures 7 and 8 illustrate the following:

- The equilibrium arrival rates $(\lambda_1^e, \lambda_2^e)$, the corresponding thresholds (c_1^e, c_2^e), and the customers' overall benefit S^e are non-decreasing in μ_2/μ_1.
- The equilibrium arrival rates coincide above a critical value of μ_2/μ_1 when $R_1 > R_2$. In the other case, where $R_1 = 1 < 9 = R_2$, the high difference in the service values forces customers to adopt a single threshold for either joining both queues or balking. In a more profitable queue 1, the almost homogeneous customers will impose a higher effective arrival rate which decreases as the corresponding variance increases.
- Once again, there is a critical value of μ_2/μ_1, where the effective arrival rates at queue 2, become equal for any variance, and the heterogeneity effect seems to disappear. A higher variance will lead to greater customers' overall benefit in equilibrium at a much higher rate when queue 2 is more profitable.

6 Conclusion

In this paper, we considered a Markovian tandem system of two unobservable M/M/1 queues in series with strategic customers who are heterogeneous in their delay sensitivity when they make join/balk decisions when arriving in front of each queue. We identified the unique symmetric Nash equilibrium strategy which is threshold-based, and depending on the values of the parameters it can dictate to customers either to join each queue under a different threshold or to use the same joining threshold for both queues. We performed a number of numerical experiments to study the impact of customer heterogeneity regarding delay sensitivity on customers' strategic behavior, and we also compared it with the case of almost homogeneous customers.

The considered model is the first step towards a more thorough study of the impact of customers' delay sensitivity heterogeneity on their strategic joining behavior in a network with queues in series. The current work can be extended in several directions to study the impact of customer heterogeneity in customer equilibrium behavior either considering a finite number of queues in series as in [20] or more general networks with feedback loops. Another approach in this direction could incorporate aspects of customers' risk aversion who adopt a non-linear utility. Finally, studying the impact of customer heterogeneity in combination with the information disclosure of the number of customers in the system at any decision instant will provide managerial insights referring to even more realistic applications.

Acknowledgments. This research is co-financed by Greece and the European Union (European Social Fund - ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).

Declarations

- **Funding** This research is co-financed by Greece and the European Union (European Social Fund - ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY).
- **Conflicts of interest/Competing interests** The author has no relevant financial or non-financial interests to disclose.
- **Authors' contributions** Not applicable.

Appendices

A Notation

In Tables 2 and 3, we present the notation throughout the paper and the different cases with the corresponding parameters for the numerical section.

Variable/Parameters	**Definition**
Λ	potential customers arrival rate
R_n	customer service valuation at queue n, $n = 1, 2$
μ_n	service rate at queue n, $n = 1, 2$
c	customer-type delay sensitivity
$C \sim H(c)$	continuous random variable C and CDF
c_l, c_h	lower and higher value of c
$\underline{x}(c) = (x_1(c), x_2(c))$	pure strategy of a customer given c
X	function of c with image the pairs of pure strategies
$V_{00}(X), V_{10}(X), V_{11}(X)$	partition of pure strategies by type$-c$ customers
$\lambda_n(X)$	effective arrival rate at queue n
$W_n(X)$	expected customer delay from joining queue n
$B_{n,c'}(X)$	expected type c' customer benefit from joining queue n
$\underline{y}(c') = (y_1(c'), y_2(c'))$	pure strategy of a tagged type c' customer
Y	function of c' with image the pairs of pure strategies
(y_1^*, y_2^*)	optimal response against $\underline{x}$
TH	set of threshold joining strategies
$\lambda_n(x_n)$	effective arrival rate at queue $n = 1, 2$
$W_n(x_n)$	expected customer delay from joining queue n
$U_{c'}(\underline{y};\ \underline{x})$	total expected net benefit of a tagged customer
$\underline{y}^*(\underline{x})$	optimal response against $\underline{x}$
TH	set of threshold joining strategies
$\underline{c} = (c_1, c_2)$	threshold joining strategies
$\underline{\tilde{c}} = (\tilde{c}_1, \tilde{c}_2)$	threshold joining strategies of a type c' tagged customer
$(\tilde{c}_1^*, \tilde{c}_2^*)$	optimal response against $(\tilde{c}_1, \tilde{c}_2)$
$\lambda_n = \Lambda H(c_n)$	effective arrival rate at queue n
$U_{c'}(\underline{\tilde{c}};\ \underline{c})$	total expected net benefit of a tagged customer
r_1, r_2, r	solutions of $B_{n,r_n} = 0, n = 1, 2$ and $B_{1,r} + B_{2,r} = 0$

Table 2 Basic notation.

Parameters	**Case**
$(c_l, c_h) = (4.45, 4.55)$	Almost Homogeneous (dots)
$(c_l, c_h) = (3, 6)$	Medium Variation (dashes)
$(c_l, c_h) = (1, 8)$	Large Variation (lines)
$R_1 = 1, \Lambda = 5, \mu_1 = 8, \mu_2 = 6, R_2$ *variable*	Figure 5 (Fast service at Queue 1)
$R_1 = 1, \Lambda = 5, \mu_1 = 6, \mu_2 = 8, R_2$ *variable*	Figure 6 (Fast service at Queue 2)
$\mu_1 = 1, \Lambda = 0.8, R_1 = 9, R_2 = 1, \mu_2$ *variable*	Figure 7 (High Valuation at Queue 1)
$\mu_1 = 1, \Lambda = 0.8, R_1 = 1, R_2 = 9, \mu_2$ *variable*	Figure 8 (High Valuation at Queue 2)

Table 3 Cases for the numerical section.

B The case of uniformly distributed delay sensitivity parameter

In this appendix, we summarize how the results are applied in the case where the delay sensitivity parameter C follows a uniform distribution on the interval $[c_l, c_h]$. These results are used to perform our numerical experiments in Section 5. For the case of a uniform distribution, the cumulative distribution function on $[c_l, c_h]$ is equal to $H(r_n) = \frac{r_n - c_l}{c_h - c_l}$, and the solution in (17) can be easily seen to be equal to

$$r_n = \frac{R_n\left(\mu_n(c_h - c_l) + \Lambda c_l\right)}{R_n\Lambda + c_h - c_l}. \tag{31}$$

Following that, the red curve in Figure 4, that characterizes (a) and (b) in case iv. of Theorem 1 can be also obtained by equating $r_1(R_1) = r_2(R_2)$, and solving with respect to R_2. After some algebra, we derive that

$$R_2 = \frac{A}{\Phi\left(1 + \frac{A}{R_1}\right) - 1} \text{ with } A = \frac{c_h - c_l}{\Lambda} \text{ and } \Phi = \frac{(c_h - c_l)\mu_2 + \Lambda c_l}{(c_h - c_l)\mu_1 + \Lambda c_l}.$$

Therefore, for $\frac{c_l}{\mu_1} + \frac{c_l}{\mu_2} - R_2 < R_1 < \frac{c_h}{\mu_1 - \Lambda}$ and

$$\frac{c_l}{\mu_2} < R_2 < \min\left\{\frac{c_h}{\mu_1 - \Lambda} + \frac{c_h}{\mu_2 - \Lambda} - R_1, \frac{A}{\Phi\left(1 + \frac{A}{R_1}\right) - 1}\right\},$$

the equilibrium threshold is

$$(c_1^e, c_2^e) = \left(\frac{R_1\left(\mu_1(c_h - c_l) + \Lambda c_l\right)}{R_1\Lambda + c_h - c_l}, \frac{R_2\left(\mu_2(c_h - c_l) + \Lambda c_l\right)}{R_2\Lambda + c_h - c_l}\right),$$

otherwise there exists the same joining threshold in both queues, $c_1^e = r$, the solution of (18), which can be simplified to the following quadratic equation:

$$(R_1 + R_2)\left(\mu_1 - \Lambda\tilde{r}\right)\left(\mu_2 - \Lambda\tilde{r}\right) = \left((c_h - c_l)\tilde{r} + c_l\right)\left(\mu_1 + \mu_2 - 2\Lambda\tilde{r}\right), \tag{32}$$

with $\tilde{r} = \frac{r - c_l}{c_h - c_l}$.

References

[1] Naor, P.: The regulation of queue size by levying tolls. Econometrica **37**, 15–24 (1969). https://doi.org/10.2307/1909200

[2] Edelson, N.M., Hilderbrand, D.K.: Congestion tolls for poisson queuing processes. Econometrica **43**(1), 81–92 (1975). https://doi.org/10.2307/1913415

[3] Economou, A., Logothetis, D., Manou, A.: The value of reneging for strategic customers in queueing systems with server vacations/failures. Eur. J. Oper. Res. **299**(3), 960–976 (2022). https://doi.org/10.1016/j.ejor.2022.01.010

[4] Bountali, O., Economou, A.: Equilibrium joining strategies in batch service queueing systems. Eur. J. Oper. Res. **260**(3), 1142–1151 (2017). https://doi.org/10.1016/j.ejor.2017.01.024

[5] Manou, A., Economou, A., Karaesmen, F.: Strategic customers in a transportation station: When is it optimal to wait? Oper. Res. **64**(4), 910–925 (2014). https://doi.org/10.1287/opre.2014.1280

[6] Economou, A., Kanta, S.: Equilibrium customer strategies and social–profit maximization in the single-server constant retrial queue. Nav. Res. Logist. **58**(2), 107–122 (2011). https://doi.org/10.1002/nav.20444

[7] Burnetas, A., Economou, A.: Equilibrium customer strategies in a single server markovian queue with setup times. Queueing Syst **56**(3–4), 213–228 (2007). https://doi.org/10.1007/s11134-007-9036-7

[8] Guo, P., Hassin, R.: Strategic behavior and social optimization in markovian vacation queues. Oper. Res. **59**(4), 986–997 (2011). https://doi.org/10.1287/opre.1100.0907

[9] Guo, P., Hassin, R.: Strategic behavior and social optimization in markovian vacation queues: The case of heterogeneous customers. Eur. J. Oper. Res. **222**(2), 278–286 (2012). https://doi.org/10.1016/j.ejor.2012.05.026

[10] Dimitrakopoulos, Y., Burnetas, A.N.: Customer equilibrium and optimal strategies in an M/M/1 queue with dynamic service control. Eur. J. Oper. Res. **252**(2), 477–486 (2016). https://doi.org/10.1016/j.ejor.2015.12.029

[11] Dimitrakopoulos, Y., Burnetas, A.: The value of service rate flexibility in an m/m/1 queue with admission control. IISE Transactions **49**(6), 603–621 (2017). https://doi.org/10.1080/24725854.2016.1269976

[12] Haviv, M., Oz, B.: Self-regulation of an unobservable queue. Manage Sci **64**(5), 2380–2389 (2018). https://doi.org/10.1287/mnsc.2017.2728

[13] Bell, C.E., Stidham Jr, S.: Individual versus social optimization in the allocation of customers to alternative servers. Manage Sci **29**(7), 831–939 (1983)

[14] Altman, E., Boulogne, T., El-Azouzi, R., Jiménez, T., Wynter, L.: A survey on networking games in telecommunications. Comput Oper Res and Oper. Res. **33**(2), 286–311 (2006). https://doi.org/10.1016/j.cor.2004.06.005

[15] Benioudakis, M., Burnetas, A., Ioannou, G.: Lead-time quotations in unobservable make-to-order systems with strategic customers: Risk aversion, load control and profit maximization. Eur. J. Oper. Res. **289**(1), 165–176 (2021). https://doi.org/10.1016/j.ejor.2020.06.047

[16] Benioudakis, M., Burnetas, A., Ioannou, G.: Single versus dynamic

lead-time quotations in make-to-order systems with delay-averse customers. Ann. Oper. Res. **318**, 33–65 (2022). https://doi.org/10.1007/s10479-022-04802-4

[17] Hassin, R., Haviv, M.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems. Springer, New York, NY (2003). https://doi.org/10.1007/978-1-4615-0359-0

[18] Hassin, R.: Rational Queueing. Chapman & Hall book. CRC press, New York, NY (2016). https://doi.org/10.1201/b20014

[19] Guo, P., Zipkin, P.: Analysis and comparison of queues with different levels of delay information. Manage Sci **53**(6), 962–970 (2007). https://doi.org/10.1287/mnsc.1060.0686

[20] Burnetas, A.N.: Customer equilibrium and optimal strategies in markovian queues in series. Ann. Oper. Res. **208**(1), 515–529 (2013). https://doi.org/10.1007/s10479-011-1010-4

[21] D'Auria, B., Kanta, S.: Equilibrium strategies in a tandem queue under various levels of information (2011). Preprint at: https://researchportal.uc3m.es/display/act391234

[22] Ji, J., Roet-Green, R., Snitkovsky, R.I.: Foresee the Next Line: On Information Disclosure in Tandem Queues. Available at SSRN: https://ssrn.com/abstract=3728894 (2022)

[23] D'Auria, B., Kanta, S.: Pure threshold strategies for a two-node tandem network under partial information. Oper. Res. Lett. **43**(5), 467–470 (2015). https://doi.org/10.1016/j.orl.2015.06.014