The impact of inspection costs on equilibrium in a queueing system with parallel servers

Refael Hassin[**] and Ricky Roet-Green[**]

[**]Department of Statistics and Operations Research , School of Mathematical Sciences, Tel Aviv University, Israel

February 20, 2015

Abstract

When time-sensitive customers arrive to a system of parallel servers, they search for the least congested queue. Customers do not always have full information of the system state, since the search is associated with a cost. We present a model of parallel servers that provide the same service. Upon arrival, each customer inspects the queue in front one server. Then, she either joins it or inspects another queue. After each inspection, the customer can end the sequential process by joining an inspected queue that minimizes her sojourn time. The solution of this model is not straightforward even when the system contains only two servers, and the equilibrium is not always a threshold strategy. We prove that in many cases, there exists a unique equilibrium strategy that contains cascades: customers choose one action (join or inspect) when they observe i and i+2 customers in the first observed queue, and the other action when they observe i+1 customers in the first observed queue. We find cascade equilibrium strategies even when the servers are identical with respect to service rate or inspection cost, and when the buffer size is finite or infinite.

1 Introduction

In many service systems, customers search among servers in order to minimize their expected waiting time. Marylandï£¡s Motor Vehicle Administration (MVA) is an example of such a system. Vehicle owners in Maryland state are required periodically to bring their vehicle to an emissions inspection. A vehicle owner can search among several nearby centers for the location with the shortest queue

length. To do so, she can use the MVA website, which provides online information about the local service centers, including their current waiting times.

But search is not costless. Inspection cost is a fundamental cause for the generating of queues. Haddock and Mcchesney (1994) argue that queues arise when customers are not informed of the system condition. If inspection costs where zero, customers would have joined non-congested queues, and congestion would have been avoided, or at least dramatically reduced. Apart from the cost of the inspection itself (for example, the cost associated with making a phone call), there is an effort associated with it. Even when the information is provided online, customers are required to pre-register to the website or download the application, and to acquire the relevant equipment (such as smartphones, tablets or similar) which has internet access.

Another aspect of inspection cost is the price of privacy that the customers pay when they search for such information, especially when inspecting health care providers. In many settings customers are required to reveal private information as a part of the inspection process. As argued in the literature, customers consider sharing personal information as a cost that they are not always willing to pay (for examples, see Miyazaki and Fernandez (2001), Sheehan (2002), Hann et al. (2002), and Huang and Van Mieghem (2013)).

Similar to the MVA setting, we consider a Markovian queueing system that contains n parallel servers. Each customer that arrives to the system chooses a server, and inspects its queue length. Inspection is associated with a fixed cost. Given the information about the first observed queue, the customer decides whether to join it, or to inspect another queue at additional cost. In the latter case, she can join the queue that minimizes her expected sojourn time in the system, or continue the sequential search until she decides to join one of the inspected queues. We assume that customers valuation of the service is very high, and balking is not allowed.

We assume that customers are homogenous with respect to their waiting costs per unit time. We also assume that customers are time-sensitive and strategic, and therefore they choose the action that minimizes their expected cost. Our goal is to find the customers' behavior in equilibrium, and as a solution concept we seek for a symmetric Nash equilibrium.

Our model is innovative in combining few aspects:

- Strategic customers: We assume that a central controller does not exist. Instead, customers are strategic in the sense of making there own decisions based on their own information.
- Costly search: We assume that inspection is costly, and that the strategic customer takes this cost into consideration.
- Sequential search: We assume that customers search among servers sequentially, meaning that after every inspection, there is an "exit point", where the customer can end the search process by joining one of the observed queues.
- Dependent queue lengths: We assume that the number of servers is finite, and therefore the lengths of the queues are dependent.
- Equilibrium solution of threshold strategies is common in queueing systems (Hassin and Haviv (2003), p.7-9). Our work reveals a non-threshold strategy, which contains *cascades*.

We start by reviewing the relevant literature of each aspect that was discussed above.

- **Sequential search, strategic customers and dependent queue lengths**

  Our study is related to *the supermarket model* of Mitzenmacher (2001), which is a dynamic version of *the load balancing model* of Azar et al.(1999). Consider a Markovian queueing system of n servers. Customers arrive to the system due to Poisson process. Each customer chooses independently, uniformly and at random a fixed number of servers, and joins the one that is less congested at that time.

  While both the static and the dynamic supermarket models assume that inspection is costless, Breitgand et al.(2006) suggest an extended model in which inspection has a cost, and this cost is incorporated into the decision of how many servers should the customer inspect prior to joining. They show that the efficiency of the system is affected by the tradeoff from the reduction of the average waiting time due to increasing of management information and the cost of its maintenance.

  Xu and Hajek (2012) look at the supermarket game with strategic customers, who wish to minimize the sum of inspection and waiting costs. The authors prove the existence of a symmetric equilibrium strategy when the number of servers goes to infinity.

  Unlike theses works, we assume that the search is sequential. The customer does not decide

prior to arrival how many servers to inspect. Instead, after every inspection, the customer considers the information that she has already gathered, and decides whether to continue the search or stop it and join one of the queues that she has already inspected.

Davidson (1988) also analyzes sequential costly search among competing servers. Unlike in our model, Davidson considers that the number of queues is so large, that their lengths are independent. Davidson shows that in equilibrium, all servers select the same price, and face the same arrival rate. Our model deals with a more complicated problem, since the number of servers is small, and therefore their queue lengths are dependent.

- **Costly search**

  The classical model of customers decision making in an unobservable queue assumes that it is either too costly to acquire the queue length information (Edelson and Hildebrand, 1975), or that inspection is free of charge (Naor, 1969). Later models consider queuing systems with two or more queues, where "joining the shortest queue" is an optimal customers' policy (but not always, see Whitt (1986) for counterexamples). Technology nowadays makes information accessible more than ever. Yet, as we argued above, the search for information is not costless. Hassin and Haviv (1994) also assumed that inspection has a cost. They considered a model where customers arrive to a system of two identical parallel servers. An arriving customer can acquire the information about which queue is shorter, and then join the shorter queue. A customer who does not purchase the information chooses one of the queues randomly. After joining, customers jockey from one queue to another.

  Hassin (1996) considered a model of two queues, where all arriving customers observe the first queue, and decide whether to join it or the other unobservable queue. A motivation for this model is the example of two gas stations that are located one after the other on a main road. Drivers make their decision by comparing their expected waiting cost at the first station, to the conditional expected cost at the second one.

  Our model relates to these works in several aspects. We solve two cases: In the first case, the queues have identical service time distribution and inspection cost. Therefore, customers randomly decide upon arrival which queue to inspect first, similarly to Hassin and Haviv's

model (1994), but with sequential search. In the second case, the inspection of one of the queues is free, while the inspection of the other queue is associated with positive cost. In that case, all customers first observe the free-of-charge queue, and then decide whether to inspect the other queue or not, similar to Hassin's gas stations model (1996), but with the option of joining the first queue after observing both queues.

The question of a costly search in queueing systems is also analyzed by Hassin and Roet-Green (2012). They consider an M/M/1 queueing model, where customers of an unobservable queue choose among three options: join the queue, balk, or inspect with a cost and then decide whether or not to join. Introducing the customers with this third option creates a model that bridges the classical queueing models of the observable and unobservable queue. They prove the existence and uniqueness of the equilibrium, and show that the monopoly firm can increase the throughput by adding costly search.

- **Non-threshold strategies**

  A threshold strategy $x = n + p, n \in \mathbb{N}, p \in [0, 1)$, prescribes one action, say A_1, for every state $0 \leq i \leq n - 1$; another action, say A_2, for every state $i > n$; and when $i = n$, it randomly selects A_1 or A_2, assigning probability p to A_1 and $(1 - p)$ to A_2.

  Equilibrium solution with a threshold strategy is common in queueing systems (Hassin and Haviv (2003), p.7-9). But in many queueing systems, a customer's choice between alternative servers is based on partial information about these queues. Since customers' decisions interact, a customer may infer about the state of a particular queue from the information available about the other queue. In other cases, it may be an indicator that the server provides high quality service, or that it is a slower server. Customers decision is influenced by this information, which makes the analysis of such systems very interesting, and the solution might not be of the threshold type.

  The general model that is presented here is too complex to be solved. Therefore, we focus on the case of two queues. The solution is not straightforward even in this case. Indeed, our investigation reveals that the equilibrium strategy has an involved structure that is often characterized by *cascades*: the customer inspects the other queue (or joins the first observed

queue) when she observes i or $i+2$ customers in the first observed queue, and joins the first observed queue (or, respectively, inspect the other queue) when she observes $i+1$ customers in the first observed queue.

A symmetric non-threshold Nash equilibrium has been found in several works (see Whitt (1986) for a model of two parallel queues in front of two identical servers, Altman and Hassin (2002), Haviv and Kerner (2007) and Kerner (2011) for M/G/1 queue, and Haviv, Kella and Kerner (2010) for M/M/N/N loss system).

The intuition behind the non-threshold strategies in our model is as follows. The more customers inspect the other queue, the more is an individual inclined to join the first queue she observed without inspecting the other queue. If a customer assumes with a high probability that the customer in front of her has already inspected the other queue and nevertheless chose to stay in that queue, then this serves as an indication that the present queue is shorter, or at least not much longer than the other queue. Thus, the actions of other customers also serve as signals, rendering the search associated with positive externalities.

Search externalities also exist in models where servers are heterogeneous in their service quality, and result in involved structures of the equilibrium (see Banerjee (1992) and Bikhchandani, Hirshleifer and Welch (1998) for static models, and Debo and Veeraraghavan (2014) for M/M/1 queue with service rate and quality as random variables).

A solution that contains cascades was also found by Debo, Parlour and Rajan (2012). In their model, customers arrive to a single observable queue, and decide whether or not to join it. The decision is based on a private signal that indicates the quality of the service, while the queue length provides (positive) information externality. They show that for customers with a private signal that indicates bad service, there may exist equilibrium strategies with "holes". Other works on positive externalities due to service quality differences between parallel servers were written by Veeraraghavan and Debo (2008, 2009).

The remainder of this paper is structured as follows: In section 2, we present the general model, and the mathematical model for a system of two servers. In section 3, we solve the case of identical servers. We prove the existence and uniqueness of a symmetric Nash equilibrium strategy for

systems of two identical servers with buffers of three or four slots. We prove the existence of an equilibrium cascade strategy for the case of four slots. We also show that in the latter case, the probability of inspecting the other queue is not always monotonic in congestion and in cost. For both cases, we show that the probability that a customer arrives to the system and finds it at full capacity is non-decreasing in congestion and in cost, and that the effect of a change in the cost on that probability is relatively small. We also consider the case of two identical servers with infinite buffers. We show the existence of equilibrium strategies that contain multiple cascades. We find that as cost increases, customers tend to inspect the other queue less. However, for a given observed queue length, customers may inspect the other queue more as cost increases. In section 4, we consider the case of heterogeneous servers, with respect to service rate, and with respect to inspection cost. For each case, we calculate the symmetric equilibrium strategy. Through numerical analysis we show that the equilibrium strategy is unique, and that it may contain cascades. In section 5 we summarize our results and discuss future work.

2 The General Model

Consider a system of n parallel queues, denoted by Q_k, $k = 1, \ldots, n$. The service time at the k-th server is exponentially distributed with parameter μ_k. Customers' arrival process to the system is Poisson with parameter λ. When a customer arrives, she inspects one of the queues, and observes its length. Let α_k be the probability that the inspected queue is the one in front of server k. Thus, the arrival process to queue k is Poisson with parameter $\alpha_k\lambda$. Inspecting Q_k costs $C_k \geq 0$, and waiting costs $C_W \geq 0$ per unit time.

After inspecting the first queue, the customer chooses among three options: joining it, inspecting another queue, or balking from the system. If she decides to inspect another queue, she inspects queue $k' \neq k$ with probability $\frac{\alpha_{k'}}{1-\alpha_k}$. After inspecting queues with lengths $l_1, \ldots, l_j, j < n$, the customer uses the information gathered so far, and chooses among joining the queue with the minimum expected waiting time, inspecting another queue, or balking. If $j = n$, the customer is left with only two relevant actions: joining one of the queues or balking.

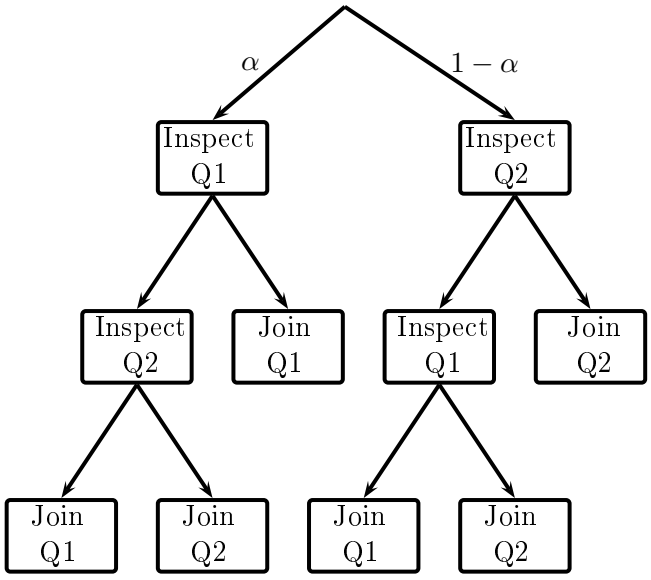
We also assume that each queue has a buffer. Let $N_k \in 1, 2, \ldots, \infty$ be the size of the buffer

of Q_k. If a customer observes N_k customers in Q_k (including the one in service), she can either inspect another queue or balk, but she cannot join it. If a customer observed all queues and found that they are all full, then she is rejected from the system. For the rest of the paper we will focus on the case of a system with two queues.

2.1 A system of two parallel queues

Consider a system of two parallel queues in front of two servers. We refer to them as Q1 and Q2. Inspecting a queue is associated with a cost: $C_1 \geq 0$ for Q1 and $C_2 \geq 0$ for Q2. To avoid trivial solutions, we assume that at least one of the queues has a positive inspection cost: $C_1 > 0$ or $C_2 > 0$. We assume that the reward from service completion is very high. As a result, customers do not balk from the system unless it is full, and they act to minimize their expected costs.

A customer who arrives to the system inspects at first one of the queues: Q1 with probability α, and Q2 with probability $1 - \alpha$. Then, she decides whether to join it, or to inspect the other queue. The following flowchart demonstrates the customers decision procedure, when the buffer sizes are infinite.



We use i for the state of Q1, and j for the state of Q2, where state refers to the total number of customers in service and in queue. We assume that the queues have buffers of the size N_1 and N_2,

respectively. Note, that if the customer arrives to a queue with full buffer capacity, then she may continue her search and join other queues, but she cannot join a full queue. If all queues are at full buffer capacity, an arriving customer is rejected (blocked), leaves the system and never returns.

To describe the birth-and-death process in this model, for $i < N_1$ and $j < N_2$ we define the indicator function $\delta_{i,j}$ as:

$$\delta_{i,j} = \begin{cases} 1 & \left(\frac{i+1}{\mu_1} < \frac{j+1}{\mu_2}\right) \\ 0.5 & \left(\frac{i+1}{\mu_1} = \frac{j+1}{\mu_2}\right) \\ 0 & \left(\frac{i+1}{\mu_1} > \frac{j+1}{\mu_2}\right) \end{cases} \tag{1}$$

Consider a customer who inspected both queues, and observed the state (i, j), where $i < N_1$ and $j < N_2$. Then, she will join Q1 with probability $\delta_{i,j}$ and Q2 with probability $1 - \delta_{i,j}$.

A strategy is consists of vectors P_I^1 and P_I^2, and a constant $0 \leq \alpha \leq 1$, where:

$$P_I^k = [P_I^k(0), P_I^k(1), \ldots, P_I^k(N_k - 1)], \quad k = 1, 2. \tag{2}$$

Consider a given α. A customer who arrives to the system inspects Q1 first with probability α. Given $i < N_1$ customers in Q1, the customer joins Q1 with probability $(1 - P_I^1(i))$ and inspects Q2 with probability $P_I^1(i)$. If she inspects Q2 and observes j customers there, she joins it with probability $1 - \delta_{i,j}$. Otherwise, she joins Q1. If she observes $i = N_1$ customers in Q1, she inspects Q2 with probability 1, and if she finds $j < N_2$ in Q2 she joins it, otherwise she balks from the system. Respectively, the same decision process takes place if she inspects Q2 first.

Suppose that the system is of state (i, j) when a new customer arrives. Let $Z_{i,j}^1$ be the probability that the system proceeds to state $(i + 1, j)$. Then

$$Z_{i,j}^1 = \alpha \left[(1 - P_I^1(i)) + P_I^1(i)\delta_{i,j}\right] + (1 - \alpha)P_I^2(j)\delta_{i,j} \qquad i = 0, \ldots, N_1 - 1, j = 0, \ldots, N_2 - 1. \tag{3}$$

$Z_{i,j}^1$ is the sum of three probabilities:

1. the probability of arriving to Q1 first and joining without inspecting Q2;

2. the probability of arriving to Q1 first, inspecting Q2 and joining Q1 with probability $\delta_{i,j}$;
3. the probability of arriving to Q2 first, inspecting Q1 and joining Q1 with probability $\delta_{i,j}$.

In the same way, let $Z^2_{i,j}$ be the probability that the system proceeds to state $(i, j+1)$. Then

$$Z^2_{i,j} = (1-\alpha)\left[(1 - P^2_I(j)) + P^2_I(j)(1 - \delta_{i,j})\right] + \alpha P^1_I(i)(1-\delta_{i,j}) \qquad i = 0, \ldots, N_1-1, j = 0, \ldots, N_2-1. \tag{4}$$

Let π_{ij} be the steady-state probability of state (i, j). We use $Z^1_{i,j}$ and $Z^2_{i,j}$ to write the balance equations, from which we calculate π_{ij}, and we use π_{ij} to calculate the equilibrium strategy, as we describe below.

2.2 Equilibrium

A customer observes the length of one of the queues and compares the expected cost from joining this queue with the conditional expected cost from inspecting the other queue and joining the shorter one. We distinguish between two complement scenarios: (a) the customer observed Q1 first, or (b) the customer observed Q2 first. Assume that the customer inspected Q1 first. Let $K^(_J i)$ be the expected cost from joining Q1 without inspecting Q2, given state i at Q1. Then

$$K^1_J(i) = C_W \frac{i+1}{\mu_1}. \tag{5}$$

If Q1 is not full, meaning $i < N_1$, let $K^1_I(i)$ be the expected cost associated with inspecting Q2. Then $K^1_I(i)$ is the sum of the following:

1. C_2 which is the cost of inspecting Q2.
2. The cost of waiting in Q2 if the customer finds that Q2 is shorter than i.

$$C_W \sum_{j=0}^{i-1} \frac{\pi_{ij}}{\pi_i} \cdot \frac{j+1}{\mu_1}, \tag{6}$$

3. The cost of waiting in Q1 if the customer finds that Q2 is not longer than i,

$$C_W \sum_{j=i}^{N_2} \frac{\pi_{ij}}{\pi_i} \cdot \frac{i+1}{\mu_1}. \tag{7}$$

Using the indicator function that we defined in (1), we can write $K_I^1(i)$ as:

$$K_I^1(i) = C_2 + C_W \sum_{j=0}^{N_2} \frac{\pi_{ij}}{\pi_i} \left[\frac{i+1}{\mu_1} \delta_{i,j} + \frac{j+1}{\mu_2}(1 - \delta_{i,j}) \right], \tag{8}$$

where $\pi_i = \sum_{j=0}^{N_2} \pi_{ij}$.

The expected cost of a customer who observes $i < N_1$ customers in Q1, is

$$K^1(i) = \min\{K_J^1(i), K_I^1(i)\}. \tag{9}$$

If Q1 is full, meaning $i = N_1$, then the customer would inspect Q2, and therefore her expected cost equals to:

$$K_I^1(N_1) = C_2 + C_W \sum_{j=0}^{N_2-1} \frac{\pi_{ij}}{\pi_i} \left[\frac{i+1}{\mu_1} \right]. \tag{10}$$

Similarly, we define $K^2(j)$ as the expected cost of a customer that inspected Q2 first.

We assume that the customers are homogeneous. Therefore, we seek for a *symmetric equilibrium*. A strategy profile is a *symmetric equilibrium profile* if it is a best response against itself. Define the best response for $i < N_1$:

$$\begin{cases} P_I^1(i) = 1 & K_I^1(i) < K_J^1(i) \\ P_I^1(i) = 0 & K_I^1(i) > K_J^1(i) \\ 0 \le P_I^1(i) \le 1 & K_I^1(i) = K_J^1(i), \end{cases} \tag{11}$$

and for $j < N_2$:

$$\begin{cases} P_I^2(j) = 1 & K_I^2(j) < K_J^2(j) \\ P_I^2(j) = 0 & K_I^2(j) > K_J^2(j) \\ 0 \leq P_I^2(j) \leq 1 & K_I^2(j) = K_J^2(j). \end{cases} \tag{12}$$

A best response strategy is any strategy that satisfies conditions (11) and (12)).

Let $E_1(\alpha)$ be the customers' expected cost from inspecting Q1 first. Then:

$$E_1(\alpha) = C_1 + \sum_{i=0}^{N_1} K^1(i) \sum_{j=0}^{N_2} \pi_{i,j}. \tag{13}$$

Let $E_2(\alpha)$ be the customers' expected cost from inspecting Q2 first. Then:

$$E_2(\alpha) = C_2 + \sum_{j=0}^{N_2} K^2(j) \sum_{i=0}^{N_1} \pi_{i,j}. \tag{14}$$

Note, that $K^1(i)$ and $K^2(i)$ are also functions of α. In equilibrium:

$$\alpha = \begin{cases} 0 & E_1(0) > E_2(0) \\ 1 & E_1(1) < E_2(1) \\ \alpha \in [0,1] & E_1(\alpha) = E_2(\alpha) \end{cases} \tag{15}$$

The symmetric equilibrium strategy, $({P_I}^k)^e$ is satisfied if a best response against itself, meaning that it satisfies conditions (11)-(15). This strategy is a best response of a player, when all other players use $({P_I}^k)^e$.

The system is characterized by two normalized parameters:

$$\rho = \frac{\lambda}{\mu} \tag{16}$$

which is the congestion parameter, and

$$\kappa = \frac{\mu C_I}{C_W} \tag{17}$$

which is the normalized inspection cost parameter. We use these normalized parameters in the numerical analysis.

3 Identical servers

We assume first that the servers are identical, both in their service rate ($\mu_1 = \mu_2 = \mu$), in their buffer sizes ($N_1 = N_2 = N$) and in their inspection cost ($C_1 = C_2 = C_I$). Therefore, customers are indifferent when choosing which queue to inspect first. As a result, an arriving customer inspects Q1 with probability 0.5, and Q2 with probability 0.5. We found that for a buffer size larger than 3, customers' behavior is not a threshold strategy.

Note that for stability, we require that $\rho < 1$ for the case of infinite buffers. This assumption is not required for the case of finite buffers, because the queue lengths are bounded by the size of the buffers.

Since the servers are identical, $P_I^1 = P_I^2$, and therefore we use P_I for customers' strategy vector. The definitions of the expected cost from each action are the same as in the previous section.

Theorem 1 *(Existence of equilibrium) In a system of two identical servers, for each set of parameters ρ, κ, there exists a symmetric Nash equilibrium strategy.*

Proof: Existence of an equilibrium in this model follows from using a fixed-point theorem. This is a game of countably many players. A strategy in this game consists of a vector $P_I = [P_I(0), P_I(1), \ldots, P_I(N)]$, where $P_I(i)$ is the probability to inspect the other queue after observing i customers in the first observed queue, and N is the size of the buffer of each server. N can be either finite or infinite. Let X be the space of all mixed strategy vectors: $X = \Big\{ [P_I(0), P_I(1), \ldots, P_I(N)] : \forall i = 0, 1, \ldots, P_I(i) \in [0, 1] \Big\}$.

A strategy vector induces the steady state probabilities. If the buffer size N is finite, then the number of possible states, $(N+1)^2$, is also finite. For infinite buffer sizes, there are countably many possible states. In that case, for a given ρ, the steady state probability of each state (i, j) is bounded by the probability that (i, j) customers are in the system, which is $L_{i,j} = (1-\rho)^2 \rho^i \cdot \rho^j, \quad i, j = 0, 1, \ldots$.

Following the assumption that in this case $\rho < 1$, we get $\lim_{i\to\infty} L_{i,j} = \lim_{j\to\infty} L_{i,j} = 0$. Therefore, the number of possible states is numerically bounded, and X is the N-dimensional cube $X = [0,1]^N$. Therefore, X is a compact space.

Let $F : X \longrightarrow X$ be the function that generates the best response strategy: $F(x) = \left\{ y \in X : y = P_I^*(x) \right\}$, where $P_I^*(x) = [P_I^*(0), P_I^*(1), \ldots, P_I^*(N)] : P_I^*(i) \in \{0,1\}$ is the best response vector strategy, as was defined in conditions (11)-(12).

Let $y_1, y_2 \in F(x)$. Let $y_3 = \omega y_1 + (1-\omega)y_2$, where $\omega \in (0,1)$, be a point on the straight line segment that joins y_1 and y_2. If $y_1 = y_2$ then it is clear that $y_3 \in F(x)$. If $y_1 \neq y_2$, then for every component i for which $y_1(i) \neq y_2(i)$, the customer is indifferent between inspection and joining, and therefore for every ω we get $y_3 \in F(x)$. Therefore, F is convex.

Given a symmetric strategy, the steady-state probabilities are derived from the linear balance equations, which are continuous for any symmetric strategy vector. The cost function (Equation (9)) is also continuous as a minimum of two continuous functions. The function that assigns the best response to each steady-state probabilities (Equation (11)) is continuous, and F is continuous as the composition of the two. Therefore the graph of F, $\left\{ \{x,y\} \in X \times X : y \in F(x) \right\}$, is a closed set.

By Kakutani's fixed point theorem, the best response correspondence F has a fixed point P_I^e. This strategy is a best response of a player, when all other players use P_I^e, which defines a symmetric Nash equilibrium. ■

We wish to characterize all the feasible types of equilibrium strategies in this model. To do so, we define *cascade strategy* as follows:

Definition 1 *Consider a vector strategy $P = P(i)$, where i is the queue state. A cascade is a state $i \geq 1$ such that $P(i-1) = P(i+1) \in \{0,1\}$ and $P(i) = 1 - P(i-1)$. We say that P is a cascade strategy if it contains a cascade.*

We wish to prove the existence of an equilibrium cascade strategy. The general model is too complicated to be fully analyzed, and therefore we solve simpler cases. In the first case, each server has a buffer with three slots. This is the minimum buffer size that allows the appearance of a cascade

strategy. We prove however that the equilibrium strategy in that case is of the threshold type. Then, we solve the case of four slots at each buffer, in which we demonstrate the existence of a cascade equilibrium strategy. To complete the case of two identical servers, we consider infinite buffers and analyze the appearance of the cascades as a function of the problem's normalized parameters.

3.1 Three slots at each buffer

We now consider a two-servers loss system, where each server has a buffer with three slots. A customer who arrives to a queue inspects the other queue with probability $P_I^e(i)$, when $i = 0, 1, 2, 3$. Since the servers are identical, when a customer arrives to an empty queue, there is no advantage in inspecting the other queue. Therefore, $P_I^e(0) = 0$. As we assumed, a customer who arrives to a full queue, inspects the other queue with probability 1, meaning $P_I^e(3) = 1$.

It is left to calculate $P_I(1)$ and $P_I(2)$. To do so, we need to calculate the expected cost from joining and inspecting when $i = 1, 2$. When $i = 1$, the expected cost from joining is (see (5)):

$$K_J(1) = 2\frac{C_W}{\mu}, \tag{18}$$

and the expected cost from inspecting the other queue is (see (8)):

$$\begin{aligned} K_I(1) &= C_I + \frac{C_W}{\mu}\frac{(\pi_{1,0} + 2\pi_{1,1} + 2\pi_{1,2} + 2\pi_{1,3})}{\pi_1} \\ &\overset{*}{=} C_I + \frac{C_W}{\mu}\frac{(2\pi_1 - \pi_{1,0})}{\pi_1} = C_I + \frac{C_W}{\mu}\left(2 - \frac{\pi_{1,0}}{\pi_1}\right). \end{aligned} \tag{19}$$

In (*) we used the definition $\pi_i = \sum_{j=0}^{3} \pi_{i,j}$, $i = 1, 2$. Therefore $K_J(1) = K_I(1)$ when $C_I = \frac{C_W}{\mu}\frac{\pi_{1,0}}{\pi_1}$, or equivalently, when $\kappa = \frac{\pi_{1,0}}{\pi_1}$. Substituting this into condition (11), we get:

$$\begin{cases} P_I(1)^e = 1 & \kappa < \frac{\pi_{1,0}}{\pi_1} \\ P_I(1)^e = 0 & \kappa > \frac{\pi_{1,0}}{\pi_1} \\ 0 \le P_I(1)^e \le 1 & \kappa = \frac{\pi_{1,0}}{\pi_1}. \end{cases} \tag{20}$$

By the same way, we calculate $P_I(2)$:

$$\begin{cases} P_I(2)^e = 1 & \kappa < 2 \cdot \frac{\pi_{2,0}}{\pi_2} + \frac{\pi_{2,1}}{\pi_2} \\ P_I(2)^e = 0 & \kappa > 2 \cdot \frac{\pi_{2,0}}{\pi_2} + \frac{\pi_{2,1}}{\pi_2} \\ 0 \leq P_I(2)^e \leq 1 & \kappa = 2 \cdot \frac{\pi_{2,0}}{\pi_2} + \frac{\pi_{2,1}}{\pi_2}. \end{cases} \tag{21}$$

For any given ρ and κ, we can now calculate the equilibrium strategy $[0, P_I(1), P_I(2), 1]$ in the following procedure:

1. Given ρ, we calculate the transaction probability matrix $Z_{i,j}$.
2. Given $Z_{i,j}$, we calculate the steady state probability matrix $\pi_{i,j}$
3. Given κ and the steady state probability matrix $\pi_{i,j}$, we calculate the equilibrium strategy vector $[0, P_I(1), P_I(2), 1]$.

We distinguish between four possible types of pure equilibrium strategies:

I. $[0, 0, 0, 1]$. II. $[0, 0, 1, 1]$. III. $[0, 1, 1, 1]$. IV. $[0, 1, 0, 1]$.

The first three types of pure equilibrium strategies (strategies I - III) represent threshold equilibrium strategies: in case I, customers inspect the other queue only if they observe three customers in the queue, i.e., their behavior in equilibrium corresponds to a threshold strategy with a threshold 3. In case II, the threshold is 2, while in case III the threshold is 1. However, the fourth strategy is a cascade strategy, where customers do not inspect the other queue if the first observed queue has 0 or 2 customers in it, but inspects it if the length of the observed queue is 1 or 3.

Figure 1 shows a map of all equilibrium strategies of this model for $0 < \rho < 2$ and $0 < \kappa < 2$. For each ρ, we calculated the values of κ that satisfy the equilibrium conditions 20 and 21.

The figure is divided into three main regions, that distinguish between the three types of pure equilibrium strategies (strategies I - III). The regions in between relate to parameter values for which the equilibrium is a mixed strategy. Note, that there are no parameters for which the equilibrium strategy is of type IV. Proposition 1 states the uniqueness of a symmetric threshold equilibrium strategy in this case.

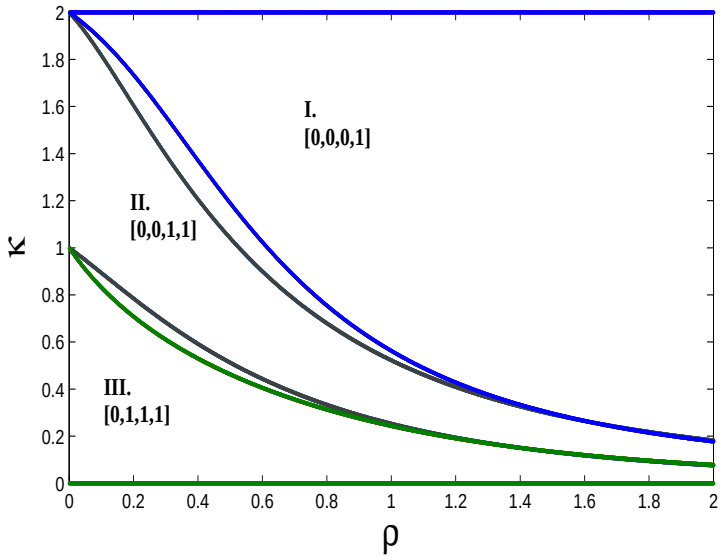


Figure 1: Map of equilibrium strategies for three slots system

Proposition 1 *(Uniqueness of symmetric equilibrium of the threshold type) In a system of two identical servers with buffers of three slots, for each set of parameters ρ, κ, there exists a unique symmetric Nash equilibrium strategy. Moreover, it is a threshold strategy.*

Proof: The numerical analysis that is provided in Figure 3.1 show that for each pair ρ, κ there exists only one equilibrium strategy. In Lemma 2 we prove that cascade strategy cannot exist in the three-slots case. Therefore, an equilibrium strategy must be a threshold strategy in this case. ■

Lemma 2 *In a system of two identical servers with buffers of three slots, a cascade strategy does not exist.*

Proof: Assume that a cascade strategy of type IV exists. Then, by condition (20), there exists a value of κ such that $\kappa < \frac{\pi_{0,1}}{\pi_1}$, and by condition (21) $\kappa > 2 \cdot \frac{\pi_{0,2}}{\pi_2} + \frac{\pi_{1,2}}{\pi_2}$. Define $\Delta = \frac{\pi_{0,1}}{\pi_1} - 2 \cdot \frac{\pi_{0,2}}{\pi_2} + \frac{\pi_{1,2}}{\pi_2}$. Such κ exists only if $\Delta > 0$.

We used the balance equations to calculate Δ. We substitute: $N = 3, P_1 = 1, P_2 = 0$ and find the steady state probabilities as a function of ρ:

$$\pi_{0,0} = \frac{(4+5\rho)(1+2\rho+2\rho^2)}{T} \qquad \pi_{0,1} = \pi_{1,0} = \frac{(4\rho+5\rho^2)(1+2\rho+2\rho^2)}{T}$$
$$\pi_{0,2} = \pi_{2,0} = \frac{2\rho^3(2+5\rho+4\rho^2)}{T} \qquad \pi_{0,3} = \pi_{3,0} = \frac{2\rho^4(1+\rho)(3+2\rho)}{T}$$
$$\pi_{1,1} = \frac{2\rho^2(1+\rho)(4+7\rho+6\rho^2)}{T} \qquad \pi_{1,2} = \pi_{2,1} = \frac{2\rho^3(2+6\rho+3\rho^2+6\rho^3)}{T} \tag{22}$$

$$\pi_{1,3} = \pi_{3,1} = \frac{2\rho^4(1+6\rho+8\rho^2+4\rho^3)}{T} \qquad \pi_{2,2} = \frac{4\rho^4(1+\rho)(1+2\rho+4\rho^2)}{T}$$
$$\pi_{2,3} = \pi_{3,2} = \frac{2\rho^5(1+2\rho)(2+5\rho+4\rho^2)}{T} \qquad \pi_{1,3} = \pi_{3,1} = \frac{4\rho^6(1+2\rho)(2+5\rho+4\rho^2)}{T} \tag{23}$$

where $T = 4 + 21\rho + 52\rho^2 + 84\rho^3 + 110\rho^4 + 128\rho^5 + 132\rho^6 + 124\rho^7 + 88\rho^8 + 32\rho^9$.

Therefore:

$$\begin{aligned} \pi_1 &= \frac{4 + 21\rho + 44\rho^2 + 50\rho^3 + 42\rho^4 + 28\rho^5 + 8\rho^6}{T} \\ \pi_2 &= \frac{8\rho^3 + 26\rho^4 + 42\rho^5 + 54\rho^6 + 44\rho^7 + 16\rho^8}{T} \end{aligned} \tag{24}$$

We substitute the probabilities into Δ and get:

$$\Delta = -2\rho^4 \cdot \frac{8 + 86\rho + 343\rho^2 + 730\rho^3 + 979\rho^4 + 940\rho^5 + 740\rho^6 + 492\rho^7 + 224\rho^8 + 48\rho^9}{(4 + 21\rho + 44\rho^2 + 50\rho^3 + 42\rho^4 + 28\rho^5 + 8\rho^6)(8\rho^3 + 26\rho^4 + 42\rho^5 + 54\rho^6 + 44\rho^7 + 16\rho^8)} \tag{25}$$

and since $\rho > 0$, we get $\Delta < 0$, which is a contradiction. Therefore, there is no cascade equilibrium strategy in this system. ■

The following observations are derived from Figure 1:

Observation 3 *In a system of two identical servers with buffers of three slots, in equilibrium, both $P_I(1)^e$ and $P_I(2)^e$ are monotonically decreasing in ρ and in κ.*

As illustrated in Figure 1, we found numerically that as κ increases, both $P_I(1)^e$ and $P_I(2)^e$ decreases monotonically from 1 to 0. The same happens as ρ increases.

Observation 4 *In a system of two identical servers with buffers of three slots, an equilibrium strategy contains at most one mixed component.*

Observe from Figure 1, that there is no pair (ρ, κ) for which customers in equilibrium are indifferent between joining their first observed queue and inspecting the other queue for both queue lengths 1 and 2. In other words, $0 < P_I(1)^e < 1$ and $0 < P_I(2)^e < 1$ do not appear simultaneously.

3.2 Four slots at each buffer

Next, we consider buffers of four slots. Here, we demonstrate the existence of an equilibrium cascade strategy.

The customer in this case has three states in which she chooses her action: when she observes queue length of one, two or three customers. In the other states (0 or 4) her action is determined as before: $P_I(0) = 0, P_I(4) = 1$. We use the same procedure as in the previous case to find the equilibrium strategy. For $i = 1, 2$ the equilibrium conditions are similar to those in conditions (20)-(21), only with one difference: in the current case $\pi_i = \sum_{j=0}^{4} \pi_{i,j}$. For $i = 3$ the equilibrium condition is:

$$\begin{cases} P_I(3)^e = 1 & \kappa < \frac{3\pi_{3,0}+2\pi_{3,1}+\pi_{3,2}}{\pi_3} \\ P_I(3)^e = 0 & \kappa > \frac{3\pi_{3,0}+2\pi_{3,1}+\pi_{3,2}}{\pi_3} \\ 0 \leq P_I(3)^e \leq 1 & \kappa = \frac{3\pi_{3,0}+2\pi_{3,1}+\pi_{3,2}}{\pi_3}. \end{cases} \tag{26}$$

We find eight types of pure equilibrium strategies:

I. $[0,0,0,0,1]$.	III. $[0,0,1,1,1]$.	V. $[0,0,1,0,1]$.	VII. $[0,1,1,0,1]$.
II. $[0,0,0,1,1]$.	IV. $[0,1,1,1,1]$.	VI. $[0,1,0,0,1]$.	VIII. $[0,1,0,1,1]$.

The first four types (strategies I - IV) represent threshold equilibrium strategies with thresholds 4,3,2 and 1 respectively. However, strategies V - VIII are pure cascade strategies. For example, in strategy V customers do not inspect the other queue if the first observed queue has $0, 1$ or 3 customers in it, but do inspect it if the length of the observed queue is 2 or 4.

For each pair (ρ, κ), we calculate the equilibrium strategy vector $P_I{}^e = [0, P_I(1)^e, P_I(2)^e, P_I(3)^e, 1]$. For each ρ, we calculated the values of κ that satisfy the equilibrium conditions. Figure 2 shows a map of all equilibrium strategies for $0 < \rho < 2$ and $0 < \kappa < 3$. The figure is divided into five main areas, corresponding to strategies I - V. The areas in between show mixed equilibrium strategies. There are no parameters for which the equilibrium strategy is of type VI - VIII.

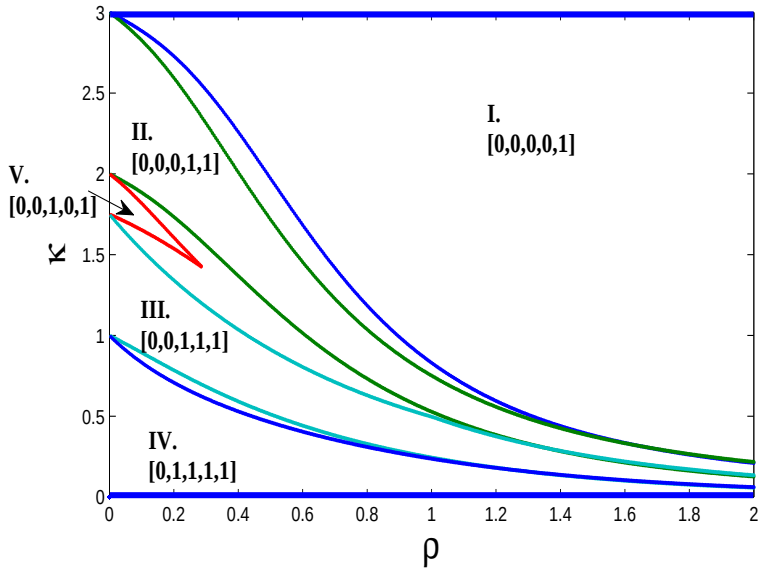


Figure 2: Map of equilibrium strategies for 4-slots system

The following statements are derived from Figure 2:

Proposition 5 *In a system of two identical servers with four-slots buffers, for each pair of parameters ρ, κ, there exists a unique equilibrium strategy. Moreover, there exists parameters ρ and κ for which customers' strategy in equilibrium is characterized by cascades.*

Proof: The uniqueness of the equilibrium strategy follows from the equilibrium map that is presented in Figure 2. The existence of a cascade strategy is also derived from this Figure. In particular, for $0 < \rho \leq 0.2877$, the graph show that there exists values of κ for which the unique equilibrium strategy is $P_I = [0, 0, 1, 0, 1]$. ■

Observation 6 *In a system of two identical servers with four-slots buffers, $P_I(1)^e$ and $P_I(2)^e$ are monotonically decreasing in ρ and in κ, while $P_I(3)^e$ is non-monotonic in ρ and in κ.*

An illustration of Observation 5-6 is shown in Appendix A.

3.3 The blocking probability

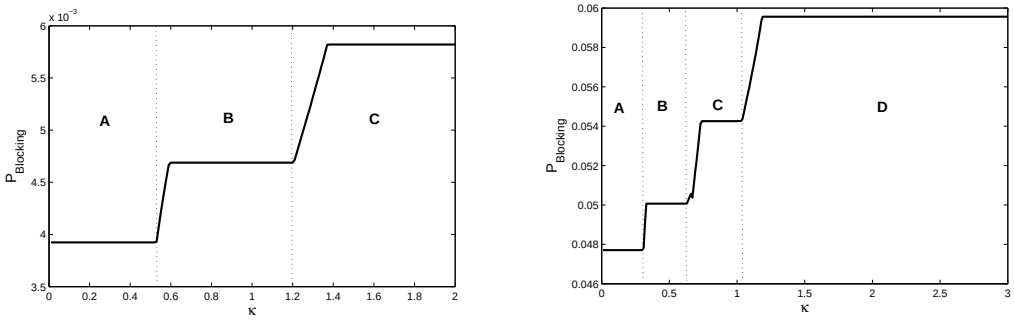
Customers who arrive to a loss system with a finite buffer size, are rejected and forced to leave if they find at arrival that the system is at full capacity. Denote $P_B = \pi_{N,N}$ is the steady-state probability of finding the system at full capacity, or the *Blocking probability*. We now analyze the effect of the system parameters on P_B. It is intuitively expected that for fixed costs (fixed κ), a growth in the congestion (ρ) will result in increasing of the probability to arrive to a full system (P_B). But it is not trivial to determine the changes in P_B for fixed ρ as κ increases. We study this question numerically, and the results are summarized in the following observations.

Observation 7 *For fixed ρ, P_B is monotonically increasing in κ. Yet, the effect of the change in κ on P_B is relatively small.*

We analyze numerically the change in P_B as a function of κ. For each value of ρ, $\rho = 0.1, 0.2, \ldots, 3$, we find that P_B is a non-decreasing function of κ, which is constant at intervals of κ that correspond to one of the possible pure equilibrium strategies.

For example, Figure 3 shows the increasing of P_B in the three-slots and four-slots cases. The x-axis shows κ while the y-axis shows P_B. In Figure 3(a), the buffer size is 3 and $\rho = 0.4$: in region A the equilibrium strategy is $[0, 0, 0, 1]$, in region B $[0, 0, 1, 1]$, and in region C $[0, 1, 1, 1]$. In Figure 3(b) the buffer size is 4 and $\rho = 0.8$: in region A where the equilibrium strategy is [0, 0, 0, 0, 1], in region B is [0, 0, 0, 1, 1], in region C is [0, 0, 1, 1, 1] and in region D is [0, 1, 1, 1, 1]. The cascade equilibrium strategy [0, 0, 1, 0, 1] is represented in a slight increase in P_B at the end of region B.

To show that the change in P_B as a function of κ is small, we calculated the minimum and the maximum values of P_B for constants values of ρ. For the three-slots case, the minimum value was calculated when the equilibrium strategy is $[0, 0, 0, 1]$ and the maximum when the equilibrium strategy is $[0, 1, 1, 1]$. For the four-slots case the equilibrium strategies were $[0, 0, 0, 0, 1]$ and $[0, 1, 1, 1, 1]$,



(a) $N = 3, \rho = 0.4$ (b) $N = 4, \rho = 0.8$

Figure 3: P_B increases with κ

respectively.

Table 1 shows the results for a system with buffers of three slots (see Table 3.3) and with buffers of four slots (see Table 3.3). For example, when $\rho = 0.6$, the minimum P_B in the three-slots case is 0.0269 and the maximum P_B is 0.0346, while in the four-slots case the minimum P_B is 0.0095 while the maximum P_B is 0.0141.

(a) Buffers of three slots

ρ	$\min(P_B)$	$\max(P_B)$
0.1	1.71×10^{-6}	3.19×10^{-6}
0.2	0.00009	0.00016
0.4	0.0039	0.0058
0.6	0.0269	0.0346
0.8	0.0827	0.0957
1.0	0.1633	0.1770
1.2	0.2502	0.2616
1.4	0.3308	0.3393
1.6	0.4029	0.4968
1.8	0.4601	0.4642
2.0	0.5100	0.5128
2.5	0.6038	0.6050
3.0	0.6684	0.6690

(b) Buffers of four slots

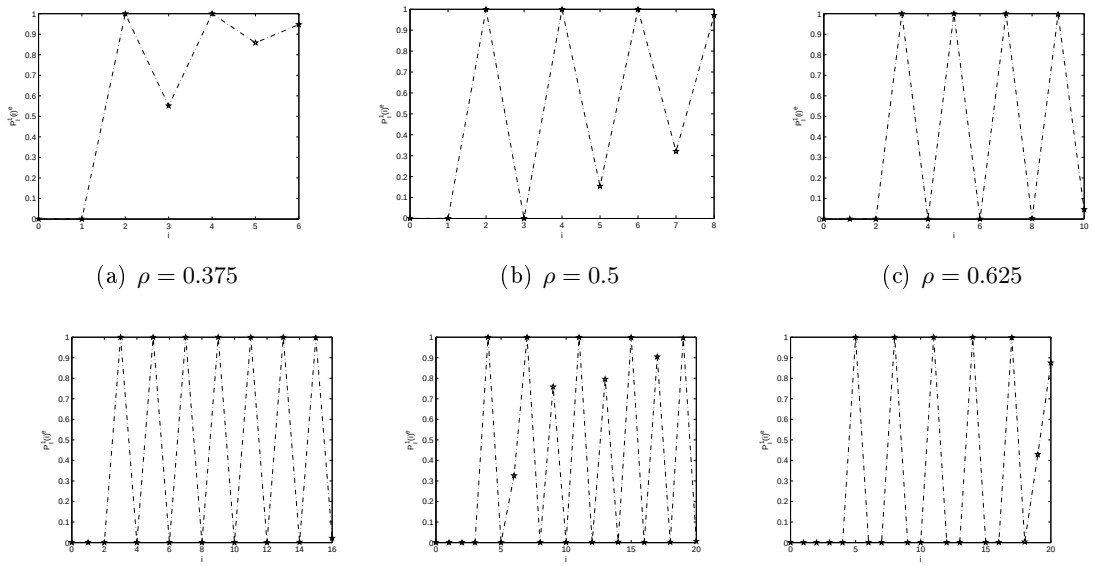
ρ	$\min(P_B)$	$\max(P_B)$
0.1	1.71×10^{-8}	4.004×10^{-8}
0.2	3.68×10^{-6}	8.004×10^{-6}
0.4	0.00063	0.0011
0.6	0.0095	0.0141
0.8	0.0477	0.0596
1.0	0.1240	0.1385
1.2	0.2181	0.2297
1.4	0.3082	0.3159
1.6	0.3856	0.3903
1.8	0.4498	0.4527
2.0	0.5029	0.5047
2.5	0.6008	0.6014
3.0	0.6670	0.6672

Table 1: Changes in the blocking probability P_B

3.4 Sensitivity analysis for a system of two identical queues with infinite buffers

We investigate the change of customers' behavior in equilibrium when the buffer of each queue is infinite. For all the numerical examples in this section, the x-axis represents i, the length of the first observed queue. The y-axis represents $P_I(i)^e$. Note, that the dashed line in all the figures in this paper is for graphical help - the queue length i in all cases is discrete.

We compare different levels of queue congestion. We find that more cascades appear as ρ increases. Examples are shown in Figure 4, where $\kappa = 1$.



(a) $\rho = 0.375$ (b) $\rho = 0.5$ (c) $\rho = 0.625$

(d) $\rho = 0.75$ (e) $\rho = 0.875$ (f) $\rho = 0.975$

Figure 4: Cascades multiplies as ρ increases

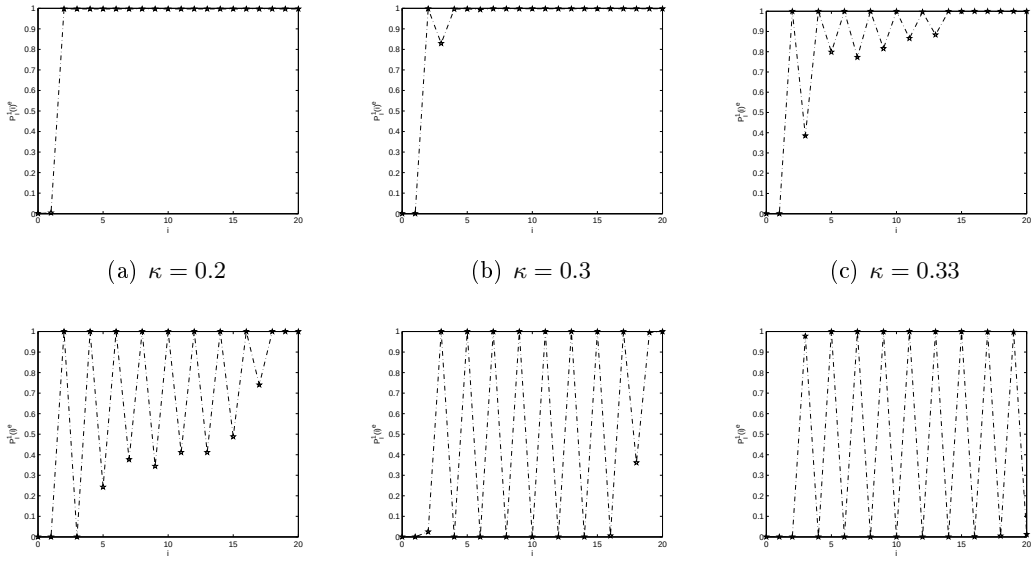
We found that when $\rho = 0.25$, and customers' strategy in equilibrium has a threshold of 2. As ρ increases, customers start to mix between joining and inspecting (e.g., Figure 4(a) for $\rho = 0.375$). As ρ continues to increase, cascades appear (e.g., Figure 4(b) for $\rho = 0.5$). As ρ continues to grow, cascades multiply (e.g., Figures 4(c) and 4(d) where $\rho = 0.625$ and $\rho = 0.75$, respectively). For higher values of ρ, customers tend to use a mixed strategy in the states that lie between the cascades (e.g., Figure 4(e) for $\rho = 0.875$), and as ρ continues to grow, gaps occur between the cascades (e.g., Figure 4(f)).

Observation 8 *For a given κ, as ρ increases, customers tend to join without inspecting the other*

queue for longer observed queue lengths, and the appearance of cascades is delayed respectively.

The intuition behind Observation 8 is that for a given κ, as ρ increases, the probability that the other queue is empty decreases, and as a result customers tend to join without inspecting the other queue when they observe longer queues. As a result, the appearance of cascades is delayed respectively.

We also look into the changes in the equilibrium as a function of κ. The numerical results show, that when κ is small, customers tend to inspect the other queue prior to joining. As κ increases, the number of cascades increase. When the value of κ is large, customers tend to join their first observed queue without inspecting the other queue, and the appearance of cascades decreases. The intuition behind this phenomenon is the customers' motivation to inspect the other queue is inversely proportional to cost. Figure 5 shows an example.



(a) $\kappa = 0.2$ (b) $\kappa = 0.3$ (c) $\kappa = 0.33$

(d) $\kappa = 0.4$ (e) $\kappa = 0.5$ (f) $\kappa = 0.75$

Figure 5: Changes in equilibrium strategy as κ increases

In Figure 5, we fixed $\rho = 0.9$. When κ is low, customers have a threshold strategy (e.g., Figure 5(a) for $\kappa = 0.2$). As κ increases, customers start to use a mixed strategy (e.g., Figure 5(b) and Figure 5(c) for $\kappa = 0.3$ and $\kappa = 0.4$, respectively). As κ continues to increase, cascades appear (e.g., Figure 5(d) for $\kappa = 0.4$), and multiply (e.g., Figure 5(e) and Figure 5(f) for $\kappa = 0.5$ and $\kappa = 0.75$, respectively).

The meaning of the multiple cascades is that customers inspect the other queue less as κ increases. Note, that this is not true when we compare strategies at a specific state: for example, as shown in Figure 5, when a customer observe $i = 5$ customers, she inspects the other queue with probability 1 when κ is low (e.g., Figure 5(a) and 5(b)), and also when κ is high (e.g., Figure 5(e) and 5(f)), and uses a mixed strategy in between (e.g., Figure 5(c) and 5(d)). Precisely, when $\kappa = 0.33$ we get $P_I^e(5) = 0.7982$, when $\kappa = 0.4$ we get $P_I^e(5) = 0.2426$, when $\kappa = 0.44$ we get $P_I^e(5) = 0.0013$, but then for $\kappa = 0.45$ we get $P_I^e(5) = 1$. Observation 9 summarizes these results.

Observation 9 *For a given ρ, as κ increases, customers tend to inspect the other queue less. However, for a given state, customers may inspect the other queue more as κ increases.*

4 Heterogeneous servers

We now abandon the assumption that the servers are identical in all aspects. The servers may differ in three parameters: the service rate μ, the inspection cost C_I, and the buffer size N. We analyze two scenarios: first, we assume that the servers differ in their service rate ($\mu_1 \neq \mu_2$), but identical in all other parameters. Second, we assume that the servers differ in their inspection costs ($C_1 \neq C_2$). Specifically, we analyze the case where $C_1 = 0$ while $C_2 > 0$. In each case, we calculate the symmetric equilibrium strategy.

4.1 When $\mu_1 \neq \mu_2$

In this case, we cannot calculate α directly from the model's assumptions. Instead, we calculate it numerically along with ${P_I}^e$. For $\alpha = 0, 0.01 \ldots, 1$ we calculate ${P_I}^e$. We substitute it into $E_1(\alpha)$ and $E_2(\alpha)$ (equations (13) and (14), respectively), and find the best response strategy α^*. Finally, we find α^e which is a fixed point in the graph of $\alpha*$.

Observation 10
1. *The best response strategy against α is to avoid the crowd (ATC).*
2. *The equilibrium α^e is unique.*
3. *α^e is continuous in μ_1 (and in μ_2). Furthermore, α^e is monotonically increasing in μ_1.*

The intuition behind the observation is as follows. If all customers tend to inspect Q1 first, then the congestion of Q1 increases while the congestion of Q2 decreases, and as a result the best response would be to inspect Q2 first. The numerical results support that observation: the best response strategy α^* is a non-increasing function of α.

Uniqueness follows directly from the ATC property. Since the best response strategy α^* is a non-increasing function of α, it has a unique fixed point which is the equilibrium α^e.

To explain monotonicity, we look at the ratio between the service rates. When $\mu_1 << \mu_2$, the first server is significantly slower than the second one. Therefore, even if both queues are observed, customers join Q2 no matter what is the length of Q1, and we get $\alpha^e = 0$. As μ_1 increases, customers start to mix between joining Q1 and Q2. As μ_1 approaches μ_2, α^e increases. When $\mu_1 = \mu_2$, the queues are identical and therefore we get $\alpha^e = 0.5$. When $\mu_1 > \mu_2$, symmetric results are derived. The continuity of α^e is derived from the numerical analysis.

Numerical results for cascade equilibrium strategies when $N \geq 4$ are shown in Appendix B.

4.2 Inspect one queue for free

Consider a system of two servers with infinite buffers. The queues may have different service rates. We wish to find customers' strategy in equilibrium, when inspecting one of the servers is free.

To illustrate that, we assume that inspecting Q1 is costless, $C_1 = 0$, while $C_2 > 0$. In this case, all customers inspect Q1 first, meaning $\alpha^e = 1$. Then, after observing its length, they decide whether to join it, or inspect Q2. Therefore, customers' equilibrium strategy consists of one vector P_I, which is equivalent to P_I^1, and is computed using $K_J(i) = K_J^1(i)$ and $K_I(i) = K_I^1(i)$ (see Equation 8 in section 4.1).

For a fixed arrival rate λ, the symmetric equilibrium profile is influenced by the ratio $\frac{\mu_2}{\mu_1}$.

When $\frac{\mu_2}{\mu_1}$ is very small, Q1 is significantly faster than Q2, and therefore customers join it without inspecting Q2. As $\frac{\mu_1}{\mu_2}$ increases, customers start to inspect Q2, and the equilibrium strategy involve cascades and mixed strategy. When $\frac{\mu_1}{\mu_2}$ is high, customers strategy becomes a threshold strategy: join Q1 when its state i is below a threshold n, inspect Q2 when $i > n$, and mix between joining and inspecting when $i = n$.

Numerical results are shown in Figures 6 and 7. Since it is a system with unbounded buffers and customers do not balk, we choose μ_1 and μ_2 such as the utilization factor satisfies $\frac{\lambda}{\mu_1+\mu_2} < 1$. The x-axis represents i, the number of customers in Q1, while the y-axis represents the equilibrium strategy $P_I(i)^e$. Note that the dashed line in all the figures is for graphical help - the queue length i in all cases is discrete.

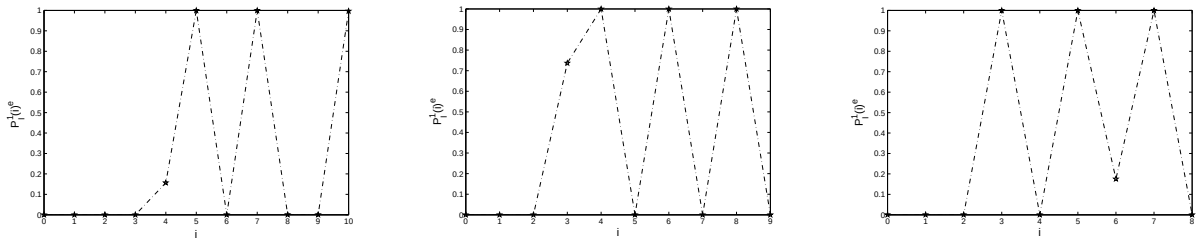
In Figure 6, $\kappa = 0.5$ and $\lambda = 2$. We find that for $\frac{\mu_2}{\mu_1} \leq 0.162$, customers join Q1 without inspecting Q2, no matter how long Q1 is. For $0.162 < \frac{\mu_2}{\mu_1} < 1.2$, customers have a non-threshold equilibrium strategy, which contains cascades (e.g. Figures 6(b) and 6(c)) and/or mixed strategies (e.g. Figures 6(d) and 6(e)). For $\frac{\mu_2}{\mu_1} \geq 1.2$ customers adopt a threshold strategy (e.g. Figures 6(f)).

In Figure 7, $\kappa = 1.5$ and $\lambda = 2$. We find that for $\frac{\mu_2}{\mu_1} \leq 0.18$, customers join Q1 without inspecting Q2, no matter how long Q1 is. For $0.18 < \frac{\mu_2}{\mu_1} < 1.56$, customers have non-threshold equilibrium strategy, which contains cascades (e.g. Figures 7(a) and 7(b)) and/or mixed strategies (e.g. Figures 7(c) and 7(d)). For $\frac{\mu_2}{\mu_1} \geq 1.56$ customers adopt a threshold strategy (e.g. Figures 7(e)).

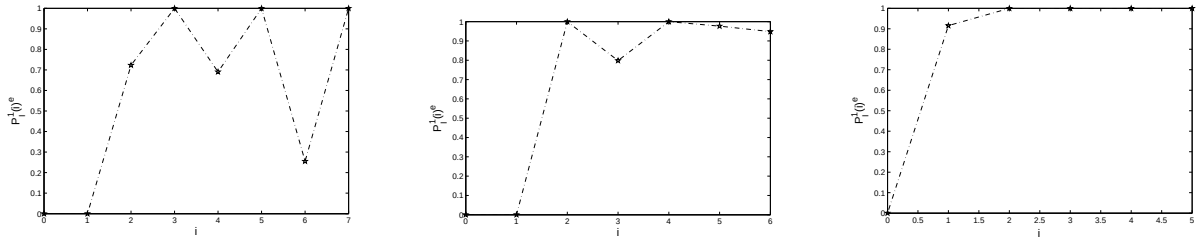
The numerical results are summarized in the following observation.

Observation 11 *In a system of two servers with infinite buffers, when the inspection of one queue is costless and the inspection of the second queue has a positive cost, customers' equilibrium strategy dependents on the ratio between the service rate of the costly-inspected queue and the service rate of the free-inspected queue:*

1. *When the ratio is low, customers join the free-inspected queue without inspecting the other queue.*
2. *When the ratio is high, customers adopt a threshold strategy, in which they inspect the other queue if the free-inspected queue is relatively long.*
3. *In between, customers adopt a non-threshold strategy which involve cascades and mixed strategies.*

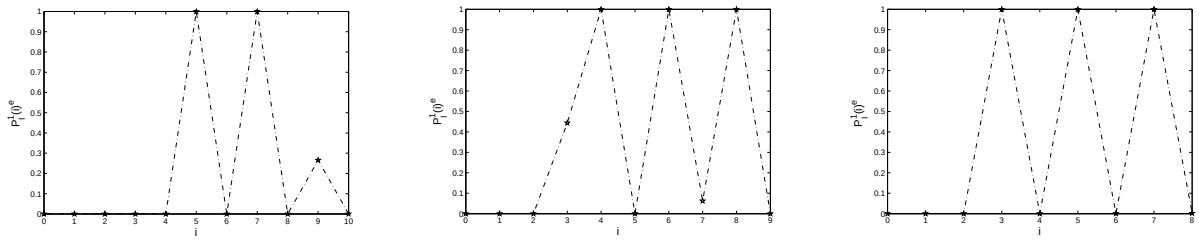


(a) $\frac{\mu_2}{\mu_1} = 0.3$ (b) $\frac{\mu_2}{\mu_1} = 0.5$ (c) $\frac{\mu_2}{\mu_1} = 0.6$

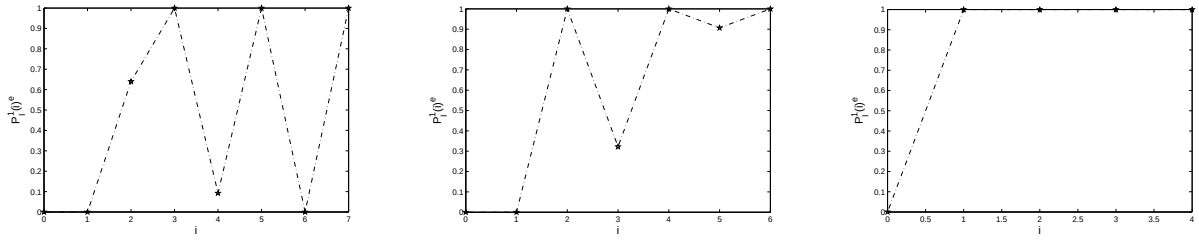


(d) $\frac{\mu_2}{\mu_1} = 0.75$ (e) $\frac{\mu_2}{\mu_1} = 1$ (f) $\frac{\mu_2}{\mu_1} = 1.5$

Figure 6: Customers' strategy when $\kappa = 0.5$ and $C_I^1 = 0$



(a) $\frac{\mu_2}{\mu_1} = 0.33$ (b) $\frac{\mu_2}{\mu_1} = 0.55$ (c) $\frac{\mu_2}{\mu_1} = 0.7$



(d) $\frac{\mu_2}{\mu_1} = 0.9$ (e) $\frac{\mu_2}{\mu_1} = 1.2$ (f) $\frac{\mu_2}{\mu_1} = 2.25$

Figure 7: Customers' strategy when $\kappa = 1.5$ and $C_I^1 = 0$

5 Concluding remarks

When customers search for a server in a multiple servers system, their equilibrium strategy is not necessary a threshold strategy. This is our main conclusion in this paper. In a system of two servers, we show that when the buffers are greater than 3, the equilibrium strategy is often characterized by cascades.

The cascade equilibrium strategy is a result of positive externalities that are induced by the customers. One can define this phenomenon as a *conditional ATC* behavior: given the first observed queue length, the customer tends to avoid the action that was made by former customers.

The model that we present here arises many questions that can serve as a basis for future research. Most queueing models assume that equilibrium strategies are of the threshold type. The appearance of cascade strategies arises the question is this assumption valid? If not, how can a planner of future queueing systems take this behavior into consideration?

Our model deals with parallel servers. Alternatively, one can assume that the servers are competing, trying to maximize their revenue by increasing their throughput on behalf of the other servers. What is the search cost that a competing server should fix in order to achieve the desirable customers' behavior? Would a slower server benefit from lowering its cost of inspection? Analysing such a different model requires a paper in itself and therefore we leave it for future research.

References

1. E. Altman and R. Hassin (2002), *Non threshold equilibrium for customers joining an $M/G/1$ queue*, In: Proceedings of the 10th international symposium of dynamic games.
2. Y. Azar, A.Z. Broder, A.R. Karlin and E. Upfal (1999), *Balanced Allocations*, SIAM Journal of Computer Science, Vol. 29, No.1, 180-200.
3. A.V. Banerjee (1992), *A simple model of herd behavior*, The Quarterly Journal of Economics CVII, 797-817.
4. S. Bikhchandani, D. Hirshleifer and I. Welch (1998), *Learning from the behavior of others: Conformity, fads, and informational cascades*, Journal of Economic Perspectives 12, 151-170.
5. D. Breitgand, A. Nahir and D. Raz (2006), *To know or not to know: on the needed amount of management information*, IBM research report.
6. S. Callander and J. Horner (2009), *The wisdom of the minority*, Journal of Economic Theory, 144, 1421ï£¡1439.

7. C. Davidson (1988), *Equilibrium in servicing industries: an economic application of queueing theory*, Journal of Business 61, 347-367.
8. L. Debo, C. Parlour, U. Rajan (2012), *Signaling quality via queues*, Management Science, 58, 876-891.
9. L. Debo and S. Veeraraghavan (2014), *Equilibrium in queues under unknown service times and service value*, Operations research 62, 38-57.
10. D. Haddock and F. Mcchesney (1994), *Why do firms contrive shortages? The economics of intentional mispricing*, Economic Inquiry 32/4, 562-581.
11. I.H. Hann, K.L. Hui, T.S. Lee, I.P.L. Png (2002), "Online information privacy: measuring the cost-benefit trade-off", Proceedings of the 23rd International Conference of Information Systems, `www.comp.nus.edu.sg/~ipng/research/privacy_icis.pdf`.
12. R. Hassin (1996), *On the advantage of being the first server*, Management Science 42, 618-623.
13. R. Hassin and M. Haviv (1994), *Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying*, Communications in statistics, Stochastic Models 10, 415-436.
14. R. Hassin and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer Academic Publishers, Norwell, Massachusetts, USA.
15. R. Hassin and R. Roet-Green (2012), *Equilibrium in a two dimensional queueing game: When inspecting the queue is costly*, submitted to a journal.
16. M. Haviv and Y. Kerner (2007), *On balking from an empty queue*, Queueing systems 55, 239-249.
17. M. Haviv, O. Kella and Y. Kerner (2010), *Equilibrium strategies in queues based on time or index of arrival*, Probability in the Engineering and Informational Sciences, 24, 13-25.
18. T. Huang and J.A. Van Mieghem (2013), *The promise of strategic customer behavior: on the value of click tracking*, Production and Operations mangements 22, 489-502.
19. A.D. Miyazaki and A. Fernandez (2001), "Consumer Perceptions of Privacy and Security Risks for Online Shopping", *Journal of Consumer Affairs*, Vol. 35/1, 27-44.
20. S. Kakutani (1941), *A generalization of Brouwerï£¡s fixed point theorem*, Duke Mathematics Journal, 8, 3, 457-459.
21. J.L. Kelley (1995), *General Topology*, Springer, New York, USA.
22. Y. Kerner (2011), *Equilibrium joining probabilities for an M/G/1 queue*, Games and Economic Behavior , 71, 2, 512-526.
23. M. Mitzenmacher, (2001), *The power of two choices in randomized load balancing*, IEEE Transactions on parallel and distributed systems, Vol. 12, No. 10, 1094-1104.
24. K.B. Sheehan (2002), "Toward a Typology of Internet Users and Online Privacy Concerns", *The Information Society: An International Journal*, 18/1, 21-32.
25. S. Veeraraghavan and L. Debo (2011), *Herding in queues with waiting costs: rationality and regret*, Manufacturing and Service management 13, 329-346.
26. S. Veeraraghavan and L. Debo (2009), *Joining longer queues: Information externalities in queue choice*, Manufacturing and service operations management, 11, 543-562.
27. W. Whitt (1986), *Deciding which queue to join: some counterexamples*, Operations Research 34, 55-62.
28. J. Xu and B. Hajek (2012), *The Supermarket Game*, 2012 IEEE International Symposium on Information Theory Proceedings, 2511-2515.

Appendices

A Numerical results of identical servers with four slots

An illustration of Observations 5-6 is shown in Figure 8. In region A, when κ is low, the threshold is 1. As κ increases, in region B, the threshold increases too and becomes 2. Note that between the pure strategies in region A and B there is a mixed strategy in $P_I(1)^e$. In region C we get a cascade strategy as was described in Observation 5. At the beginning of region D the equilibrium strategy is mixed in $P_I(2)^e$, until $P_I(2)^e$ drops to 0 and we get a new threshold of 3. As κ continues to increase, at region E we get another threshold strategy with a threshold 4.

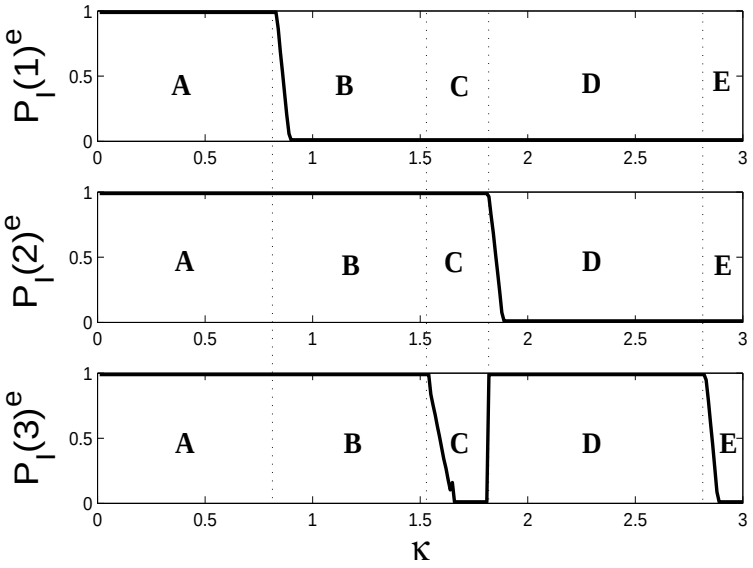
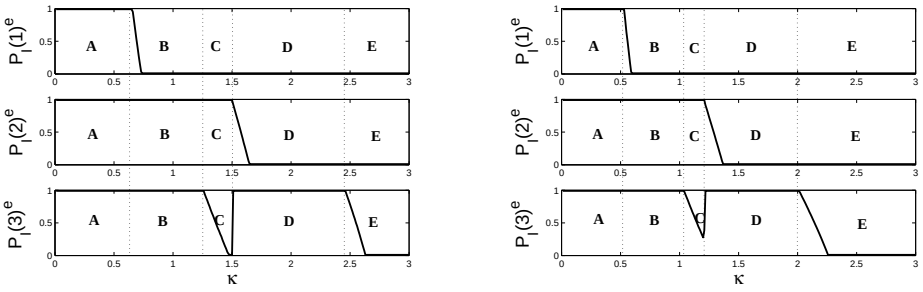


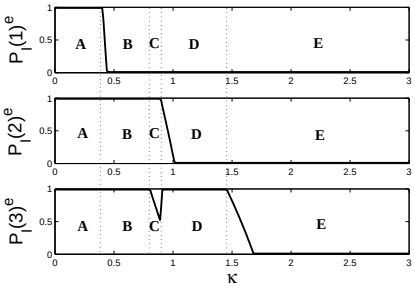
Figure 8: A cross section at $\rho = 0.1$

To present the changes in equilibrium strategy as a function of ρ and as a function of κ, we look at other vertical cross sections of Figure 8.

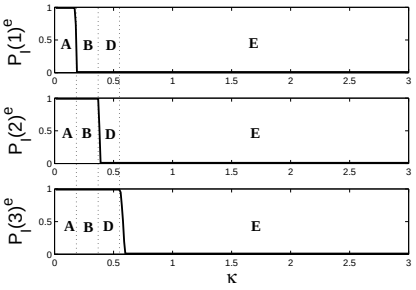


(a) $\rho = 0.25$

(b) $\rho = 0.4$



(c) $\rho = 0.6$



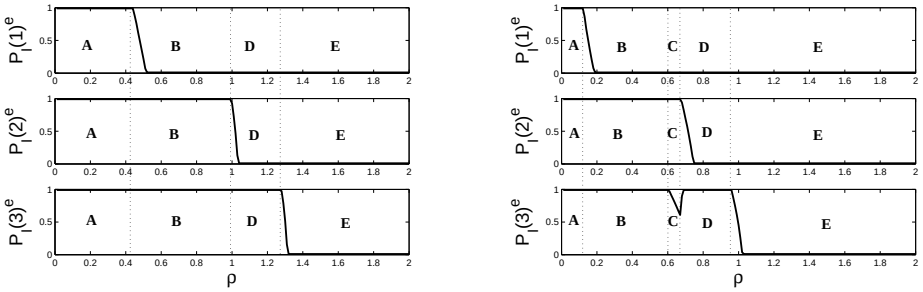
(d) $\rho = 1.2$

Figure 9: Changes in vertical cross sections as ρ increases

Figure 9 shows the change in equilibrium strategy for fixed values of ρ, as κ changes. Figure 9(a) shows the equilibrium strategy where $\rho = 0.25$. A cascade strategy occurs in region C, where the probability $P_I(3)^e$ drops to 0. In Figure 9(b), where $\rho = 0.4$, the cascade in region C is of a mixed strategy, where the probability $P_I(3)^e$ drops to ≈ 0.276. For larger values of ρ, we get a non-threshold strategy instead of pure cascade strategy, as shown in region C of Figure 9(c). As ρ increases, the cascade disappears, as is shown in Figure 9(d). Note, that Figures 9(a) - 9(c) present the non-monotonicity of $P_I(3)^e$, as was described in Observation 6.

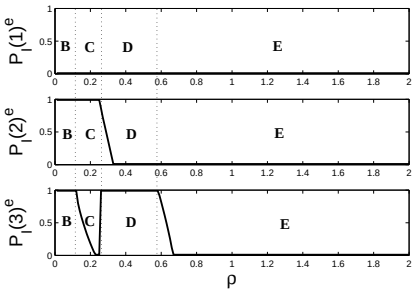
Figure 10 shows the change in equilibrium strategy for fixed values of κ, as ρ changes. Again, the cascade occurs in region C of each figure. Note, that as κ increases, equilibria with threshold of 1 (region A) and of 2 (region B) are gradually disappearing, while the cascade at region C is expanding. Here, the cascade occurs when ρ is relatively small, as captured in region A of each subgraph. Figure 10(a) shows the equilibrium strategy when $\kappa = 0.3$. Note that as ρ increases,

$P_I(1)^e$ decreases while $P_I(2)^e$ increases. In Figure 10(b), at the beginning of region B, we get a mixed strategy in both probabilities. Figure 10(d) and Figure 10(d) capture the non-monotonicity of $P_I(3)^e$.

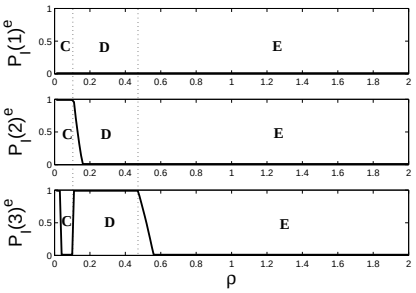


(a) $\kappa = 0.5$

(b) $\kappa = 0.8$



(c) $\kappa = 1.5$



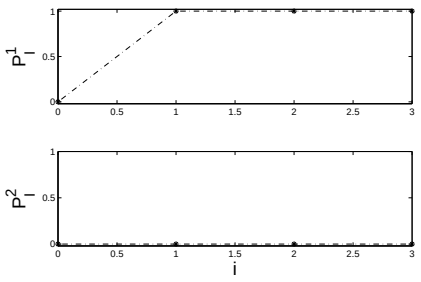
(d) $\kappa = 1.8$

Figure 10: Changes in horizonal cross sections as κ increases

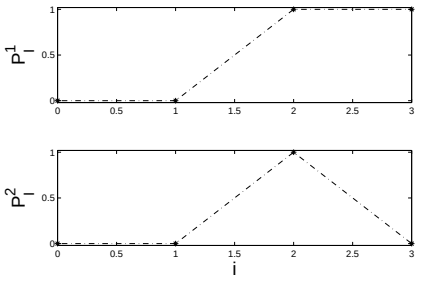
B Numerical results of cascade equilibrium strategies for heterogeneous servers

Figure 11 shows examples of non-threshold equilibrium strategy when $N = 4$, and $\mu_2 = 2, \lambda = 1, C_W = 1$ and $C_1 = C_2 = 0.5$.

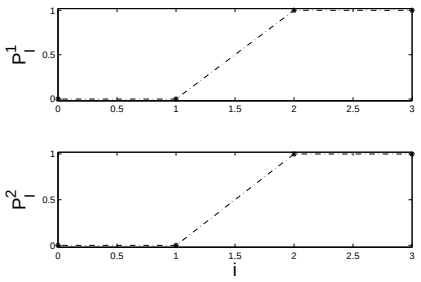
Each graph is divided into two subgraphs: the top one shows P_I^1 as a function of i, the observed number of customers in Q1. The bottom one shows P_I^2 as a function of i, the observed number of customers in Q2. Since $N = 4$, $i = 0, 1, 2, 3$. From the model assumptions, for $i = 4$, $P_I^1(4) = P_I^2(4) = 1$. Each graph calculates the equilibrium for different (increasing) value of μ, and therefore α^e is different (increases from one graph to the following one).



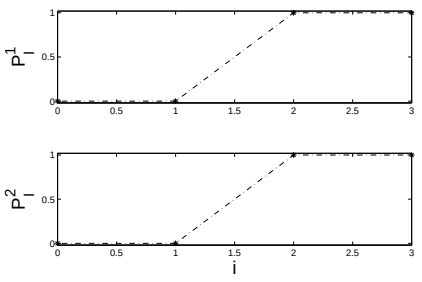
(a) $\mu_1 = 1.4, \alpha^e = 0$



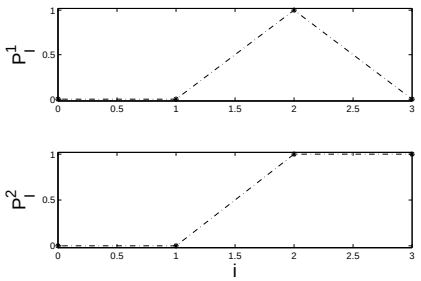
(b) $\mu_1 = 1.5, \alpha^e = 0.075$



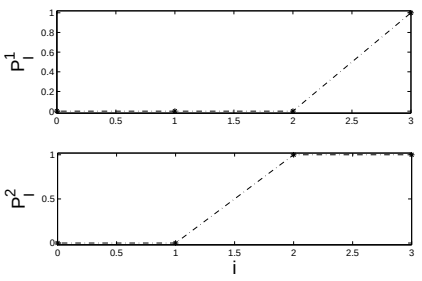
(c) $\mu_1 = 1.8, \alpha^e = 0.335$



(d) $\mu_1 = 2, \alpha^e = 0.5$



(e) $\mu_1 = 2.5, \alpha^e = 0.865$



(f) $\mu_1 = 3, \alpha^e = 1$

Figure 11: Changes in horizonal cross sections as κ increases

In Figure 11(a), P_I^1 is a threshold strategy with threshold 1, and P_I^2 is a threshold strategy with threshold 4. In Figure 11(b), P_I^1 is a threshold strategy with threshold 2, but P_I^2 is a non-threshold strategy with cascades: customers inspect Q1 with probability 1 when Q2 length is 2 or 4 and join Q2 without inspecting Q1 when the length is 0,1 or 3. Figure 11(c) and Figure 11(d) show the same strategy components for P_I^1, P_I^2, which are both having a threshold of 2. Yet they differ in α^e. Figure 11(e) is a symmetric picture of 11(b) regarding P_I^1, P_I^2. In Figure 11(f), both P_I^1 and P_I^2 are threshold strategies with thresholds 3 and 2 respectively.

Acknowledgment

This research was supported by the Israel Science Foundation grant no.1015/11. The authors thank Prof. Eilon Solan for his assistance in formalizing the proof of Theorem 1).