Information cascades in a queueing system with parallel servers

Refael Hassin

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

Ricky Roet-Green

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

Abstract—When a customer makes an appointment at the doctor, call the plumber or reserve flight tickets, he joins a queue. By making the reservation, the customer can learn his position in the service provider's queue. If the customer is unsatisfied with his position, he might consider calling an alternative service provider, hoping to get a better position. In our model, the service is provided by n parallel servers. Upon arrival to the system, each customer randomly chooses one server and inspects it. Then, he either joins it or inspects another queue. If he inspects another queue, he can join the shorter queue, or continue his search and inspect another queue. We assume that each inspection is associated with a fixed cost. The solution of this model is not straightforward even when $n = 2$, and is characterized by cascades. In equilibrium, there exist isolated queue lengths (holes) at which customers inspect the other queue. In other queue lengths customers join the first queue. In some cases, there exist queue lengths at which customers adopt a mixed strategy.

I. INTRODUCTION

In many services, customers make an appointment in advance. For example, medical service providers as physicians, surgeons, psychologists and psychiatrists, require that the patient will make an appointment. Reservations are also necessary in other fields: for example, if someone needs the service of a plumber or an electrician, he would have to schedule the visit. Booking is also required in many travel services: flights, hotels, restaurants, shows etc.

When a customer makes an appointment at the doctor, calls the plumber or reserves flight tickets, he joins a queue. By making the reservation, the customer can learn where he is positioned in the service provider's queue. If the customer is unsatisfied with his position, he might consider calling an alternative service provider, to get a better position. The customer can continue inspecting other queues, until he is satisfied. But calls are not costless: apart from the cost of the call itself, there is effort that is associated with it (for example, if the customer needs to provide details as contact information in order to make the appointment).

In our model, we assume that customers have to consume a service (for example, visit the doctor). The service is provided by n servers. Upon arrival to the system, each customer randomly chooses one server and inspects it. Then, he decides whether to join it, or inspects another queue. If he inspects

another queue, he can join the shorter queue, or continue his search. We assume that each inspection is associated with a constant non-negative cost.

We focus on the case of two servers. The solution is not straightforward even in this case. In this system, the more customers inspect the other queue, the more is an individual inclined to avoid them and join the first queue he observed without inspecting the other queue. If a customer assumes with a high probability that the customer in front of him 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.

A threshold strategy $x = n+p, n \in N, p \in [0, 1)$, prescribes one action, say A_1, for every state $0 \le i \le n - 1$; another action, say A_2, for every state $i > n$; and when $i = n$, selects randomly between A_1 and A_2, assigning probability p to A_1 and $(1-p)$ to A_2. Equilibrium solution of threshold strategies is common in queueing systems (Hassin and Haviv [7], 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. For example, a long queue at a given time may indicate that other queues are probably also long at that time. In other cases, it may be an indicator that the server provides high quality service, or that it is a slower server. This information externality makes the analysis of such systems very interesting, and the solution might not be of the threshold type.

Several works had shown results of a similar nature. Whitt, [12], proved that joining the shortest line is not always optimal in a system of two parallel queues in front of two identical servers, and depend on the difference between the length of two queues. A symmetric Nash equilibrium which is not a threshold was also found in several works (Altman and Hassin [3], Haviv and Kerner [9], Hassin and Roet-Green [8]).

Positive externalities exist also in models where servers are different in their service quality, and it results in an involved structure of the equilibrium. A pioneering work is Banerjee's [1], who considered a sequential decision model in which each decision maker choose between several options, that only one of them is the correct option. He showed that the decision rule is characterized by herd behavior, meaning that people follow the crowd even if it is against their own information.

Some works showed equilibrium strategies that are characterized by cascades: Bikhchandani, Hirshleifer and Welch [2] considered models in which individuals start with some private information, obtain some information from predecessors, and then decide on a particular action. They found that there exists informational cascade, in which an individual takes an action that does not depend on his private information, and as a result of that no further information accumulates, and the following individuals adopt herd behavior.

Another solution that is characterized by cascades was found by Debo, Parlour and Rajan [4]. In their model, customers arrive to a single observable queue, and decide whether or not to join it. The decision relies on a private signal that indicates the quality of the service, while the queue length provides (positive) information externality. They showed that for customers with private signal that indicates bad service, there may exist equilibrium strategies with "holes": for example, such a customer will join the queue if its length is $3, 5$ or 6 but not when the queue length is 4. Other works on positive externalities due to service quality differences between parallel servers were written by Veeraraghavan and Debo [10],[11].

In our model, when customers arrive to the system, the queues in front of the servers are unobservable to them. When a queue is inspected, it becomes observable to the customer. The question of acquiring information about queue lengths also receives significant attention in the literature. Hassin and Haviv [6] considered a model where customers arrive to a system with two identical parallel servers. An arriving customer can acquire the information about which queue is shorter by paying a fixed amount, and then join the shorter queue. A customer who does not purchase the information chooses one of the queues randomly. After joining, customers jockey without additional cost from one queue to another, when the difference between them reaches a given threshold of N. Hassin and Haviv computed the value of information and the equilibrium threshold strategies. As opposed to intuition, that the value of information is a decreasing function of the proportion of informed customers p (Avoid the crowd), they showed that under certain parameters this value increases with p (Follow the crowd).

In another paper, Hassin [5] considered a model of two queues, where all arriving customers observe the first queue, and decide upon arrival whether to join it or to jockey to the other queue which is unobservable to them. This decision is irrevocable, and balking is not an option. 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. Hassin found that the average arrival rate into the first queue is greater than the rate into the second one, and that customers' equilibrium strategy is of the threshold type: they join the first queue if it is shorter than a threshold which is the expected length of the other queue, but they join the other queue if its expected length is shorter, and adopt a mixed strategy if these values are equal.

Our model relates to these works in several aspects. The main difference between these models and the one that we present here is that in our model, customers' decision problem has to take into account the cost of inspecting the other unobserved queue, which was costless in the previous models.

It is not trivial to determine whether there exists an equilibrium of the threshold type in this case. Consider a threshold strategy where a customer decides to inspect the other queue if he observes at least i customers in his own first queue. Suppose that the customer observes $i+5$ customers. With a reasonably high probability, he knows that the last customer in the queue has recently inspected the other queue and concluded that it was not worth joining it. Hence, the current customer may use this information and avoid expending the amount required to verify it. This contradicts the assumption of a threshold equilibrium. Indeed, our investigation into this model reveals that there exists a unique equilibrium solution, and according to this solution a customers strategy has an involved structure that is characterized by cascades.

II. THE GENERAL MODEL

Consider a system of n parallel queues. The service time at the k-th server is exponentially distributed with parameter μ_k, $k = 1, \ldots, n$.

Customers' arrival process to the system is Poisson with parameter λ. When a customer arrives, he inspects the queue in front of server k with probability α_k, $k = 1, 2, \ldots, n$, and observes its length. We assume that $\sum_{k=1}^{n} \alpha_k = 1$. Then, arrival to queue k is Poisson with parameter $\alpha_k \lambda$. The utilization factor is:

$$\rho = \frac{\lambda}{\sum_{k=1}^{n} \mu_k} \, . \tag{1}$$

For stability, we will require that ρ will be strictly less than 1, since it is impossible to balk from the system.

After inspecting the first queue, each customer chooses between joining it, or inspecting another queue. If he decides to inspect another queue, he inspects queue $k' \neq k$ with probability $\frac{\alpha_{k'}}{1-\alpha_k}$. After inspecting the length of k', the customer

chooses whether to join one of the queues that he already inspected, or inspect another queue, and so on. We assume that the customers are rational agents in this game, and therefore, among the queues that they already inspected, they decide to join the queue that minimizes their expected waiting cost.

III. TWO 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. 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, he decides whether to join it, or to inspect the other queue. We assume that waiting has a cost of $C_W \geq 0$ per unit of time.

We use i for the state of Q1, and j for the state of Q2. By state we refer to the number of customers both in service and in queue.

To describe the birth and death process in this model, define the indicator function $\delta_{i,j}$ as:

$$\delta_{i,j} = \begin{cases} 1 & \frac{i+1}{\mu_1} \leq \frac{j+1}{\mu_2} \\ 0 & \frac{i+1}{\mu_1} > \frac{j+1}{\mu_2} \end{cases} \tag{2}$$

Consider a customer who inspected both queues, and observed the state (i, j). Then, if $\delta_{i,j} = 1$ he will choose Q1, and if $\delta_{i,j} = 0$ he will choose Q2.

A strategy consists of two vectors P_I^1 and P_I^2, and a constant $0 \leq \alpha \leq 1$, where $P_I^k = [P_I^k(0), P_I^k(1), \ldots], k = 1, 2$, $\forall i : P_I^k(i) \in [0, 1]$, where $P_I^k(i)$ is the probability that the customer inspects the other queue if he observes $i \geq 1$ customers in Qk, and $1 - P_I^1(i)$ is the probability of joining Qk without inspecting the other queue.

Consider a given α. A customer arrives to the system, and inspects first Q1 with probability α. For any queue length of i customers in that queue, $i = 0, 1, 2, \ldots$, the customer joins Q1 with probability $(1 - P_I^1(i))\alpha$, or inspect Q2 with probability $P_I^1(i)\alpha$ and joins it if $\delta_{i,j} = 0$. Otherwise, he joins Q1. The customer inspects Q2 first with probability $1 - \alpha$. For any queue length of j customers in that queue, $j = 0, 1, 2, \ldots$, the customer joins Q2 with probability $(1 - P_I^2(j))(1 - \alpha)$, or inspects Q1 with probability $P_I^2(j)(1 - \alpha)$ and joins it if $\delta_{i,j} = 1$. Otherwise, he joins Q2.

Denote:

$$\begin{aligned} M_{i,j}^1 &= \left(\alpha + P_I^2(j)(1 - \alpha)\right) \delta_{i,j} \\ &+ (1 - P_I^1(i))\alpha(1 - \delta_{i,j}) \end{aligned} \tag{3}$$

and:

$$\begin{aligned} M_{i,j}^2 &= \left(P_I^1(i)\alpha + (1 - \alpha)\right)(1 - \delta_{i,j}) \\ &+ (1 - P_I^2(j))(1 - \alpha)\delta_{i,j} \end{aligned} \tag{4}$$

Then, given the state (i, j), $M_{i,j}^1$ is the probability that the system proceeds to state $(i+1, j)$, while $M_{i,j}^2$ is the probability that the system proceeds to state $(i, j + 1)$.

A. 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. If the customer observes Q1 first, let U_J^1 be the expected cost from joining Q1 without inspecting Q2, given state i in Q1. Then:

$$U_J^1(i) = \frac{C_W \cdot (i + 1)}{\mu_1} \tag{5}$$

Notice, that the expected cost from joining Q1 does not include the cost of inspecting Q1, C_1, since it has been spent already.

Let π_{ij} be the steady-state probability of state (i, j). Let $U_I^1(i)$ be the expected cost of inspecting Q2. Define auxiliary variable X_1 as:

$$X_1(i) = \left\lfloor \frac{(i+1)\mu_2}{\mu_1} \right\rfloor - 1 \tag{6}$$

When $j < X_1(i)$, then $\delta_{i,j} = 0$ and the informed customers join Q2. When $j \geq X_1(i)$, the informed customers join Q1. Therefore:

$$U_I^1(i) = C_2 + \frac{C_W}{\mu_2} \cdot \frac{\sum_{j=0}^{X_1(i)-1} \pi_{ij} \cdot (j + 1)}{\pi_i} + \frac{C_W}{\mu_1} \cdot \frac{\sum_{j=X_1(i)}^{\infty} \pi_{ij} \cdot (i + 1)}{\pi_i} \tag{7}$$

where:

$$\pi_i = \sum_{j=0}^{\infty} \pi_{ij}. \tag{8}$$

The expected cost of a customer who observes i customers in Q1, is:

$$U^1(i) = \min\{U_J^1(i), U_I^1(i)\}. \tag{9}$$

Similarly, if the customer inspects Q2 first, define $U_J^2(j)$ as the expected cost from joining Q2 without inspecting Q1, and $U_I^2(j)$ as the expected cost from joining Q2 without inspecting Q1, given state j in Q2. Then:

$$U_J^2(j) = \frac{C_W \cdot (j + 1)}{\mu_2} \tag{10}$$

Define auxiliary variable $X_2(j)$ as:

$$X_2(j) = \left\lfloor \frac{(j+1)\mu_1}{\mu_2} \right\rfloor - 1 \tag{11}$$

When $i < X_2(j)$, then $\delta_{i,j} = 1$ and informed customers join Q1. When $i \geq X_2(j)$, informed customers join Q2. Therefore:

$$U_I^2(j) = C_1 + \frac{C_W}{\mu_1} \cdot \frac{\sum_{i=0}^{X_2(j)-1} \pi_{ij} \cdot (i+1)}{\pi_j}$$
$$+ \frac{C_W}{\mu_2} \cdot \frac{\sum_{i=X_2(j)}^{\infty} \pi_{ij} \cdot (j+1)}{\pi_j} \quad (12)$$

where:

$$\pi_j = \sum_{i=0}^{\infty} \pi_{ij}. \quad (13)$$

The expected cost of a customer who observes j customers in Q2, is:

$$U^2(j) = \min\{U_J^2(j), U_I^2(j)\}. \quad (14)$$

Customers wish to minimize their cost, and therefore choose *best response strategies*. Since in this model the customers are homogeneous, we are interested in a *symmetric equilibrium*. A strategy profile is a *symmetric equilibrium profile* if it is a best response against itself.

Define as E_1 customers' expected cost from inspecting Q1 first. Then:

$$E_1 = C_1 + \sum_{i=0}^{\infty} \pi_i U^1(i) \quad (15)$$

Define as E_2 customers' expected cost from inspecting Q2 first. Then:

$$E_2 = C_2 + \sum_{j=0}^{\infty} \pi_j U^2(j) \quad (16)$$

In equilibrium:

$$\alpha = \begin{cases} 0 & E_1 > E_2 \\ 1 & E_1 < E_2 \\ \alpha \in [0,1] & E_1 = E_2 \end{cases} \quad (17)$$

For a given α, denote by (P_I^{1*}, P_I^{2*}) the best response strategy for a customer given that all others adopt strategy (P_I^1, P_I^2). Denote by $U_J^k(r), U_I^k(r)$ the expected costs from joining with and without inspecting the other queue when the length of the first inspected queue Qk is r. Then:

$$P_I^{k*}(r) = \begin{cases} P_I^k(r) = 0 & U_J(k) < U_I(k) \\ P_I^k(r) = 1 & U_J(k) > U_I(k) \\ P_I^k(r) = p \in [0,1] & U_J(k) = U_I(k) \end{cases} \quad (18)$$

The symmetric equilibrium strategy, $(P_I^k)^e$ is a best response against itself. In order to find equilibrium strategy, we find $(P_I^k)^e$ for a given α in general. Then, we substitute P_I^e into E_1, E_2, and find the appropriate α.

To avoid trivial solutions, we assume that $C_W > 0$ and $C_1 > 0$ or $C_2 > 0$.

B. Numerical Method

In our numerical study, we assumed that each server has a limited waiting room of N (including the customer in service). Therefore, each server has $N + 1$ states: $i, j = 0, 1, \ldots, N$. We assume that a customer who observes queue length of N customers inspects the other queue, and join it unless it is also full. If both queues are full, an arriving customer balks and never return to the system. We choose N big enough such as the probability that the queues fill up is negligible.

For a given α, we use the following algorithm to compute the equilibrium strategy:

1) Choose arbitrary strategy vectors $P_I^k, k = 1, 2$, and define a tolerance parameter ϵ.
2) Compute the steady states probabilities matrix $\pi_{i,j}$ using P_I^k.
3) Compute U_I^k, U_J^k using the $\pi_{i,j}$ matrix.
4) If $U_I^k(i) > U_J^k(i) + \epsilon$, set $P_I^{k*}(i) = 0$.
5) If $U_I^k(i) < U_J^k(i) + \epsilon$, set $P_I^{k*}(i) = 1$.
6) If $|U_I^k(i) - U_J^k(i)| < \epsilon$, set $P_I^{k*}(i) = P_I^k(i)$.
7) If $|P_I^{k*} - P_I^k| < \epsilon$, stop, and declare the equilibrium strategy as P_I^*.
8) If $|P_I^{k*} - P_I^k| \geq \epsilon$, define a "new" strategy $(P_I^k)^{\text{new}}$ as a convex combination of the "old" strategy P_I^k and its best response P_I^{k*}, using a random number $\gamma \in (0,1)$ as a weight. Continue from the second step and repeat the process, until you reach the stopping criteria.

We applied this algorithm with $\epsilon = 0.001$ in our numerical examples.

C. Numerical results

Numerical results of this model show that customers' behavior in equilibrium has a complex structure that characterized by cascades. To illustrate that, we look at the special case where the servers are identical in their rate of service $\mu_1 = \mu_2$, and also in their inspection cost $C_1 = C_2 > 0$. As a result, customers are indifferent between the servers, and therefore they choose which queue to inspect first randomly (meaning $\alpha = 0.5$).

Once a customer arrives to this system, he inspects one of the queues. Given its length, the customer decides whether to inspect the other queue or not. Since the servers are identical, $P_I^1 = P_I^2$, and therefore we use P_I for customers' strategy vector. If the customer observe i customers in the first queue he inspected, he inspects the other queue with probability $P_I(i)$, or join the first queue without inspecting the other one with probability $1 - P_I(i)$. Customers' equilibrium strategy consists of the vector P_I, and is computed using $U_J(i) = U_J^1(i)$ (see Equation 5) and $U_I(i) = U_I^1(i)$ (see Equation 7). Figure 1 shows an example of customers' equilibrium strategy when $\mu_1 = \mu_2 = 3, \lambda = 7, C_1 = C_2 = 1, C_W = 2.5$. In the x-axis

there are the different feasible queue lengths, while in the y-axis there is the probability to inspect the other queue length, P_I.
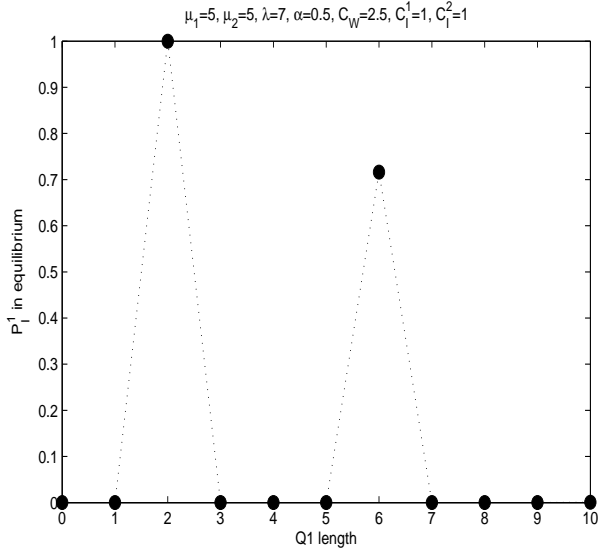

Fig. 1. "Holes" in the equilibrium strategy

From figure 1 one can learn that the equilibrium strategy is characterized by cascades: for $i \leq 1$ customers join Q1, for $i = 2$ they inspect Q2, for $3 \leq i \leq 5$ they join Q1. For $i = 6$ they adopt a mixed strategy: they join Q1 with probability $P_I \approx 0.7$, and inspect Q2 otherwise. For $i \geq 7$ they join Q1.

These cascades coincide with the following behavior: if a customer assumes with a high probability that the customer in front of him has already inspected the second queue and still chose to stay in the first queue (or joined from the other queue), then this serves as an indication that the present queue is shorter, or at least not much longer than the other queue. But if he assumes with a high probability that the customer in front of him did not inspect the second queue, he intends to check the other queue length. Thus, the actions of other customers influence the customer's behavior as positive externalities.

We compared different levels of queue congestion, in order to find out how it influence customers' behavior. We found that cascades appear as ρ increases, since the queue state is more unstable for larger ρ than for smaller ρ, and therefore it is more informative. As a result, a customer that inspects the queue can assume with high probability that the customer in front of him inspected the same queue length when he joined it, and therefore in equilibrium the queue length will indicate the action of the last customer who joined it. For small ρ, the queue is less informative, and customers choose to inspect the other queue if the cost of inspection is low. Figure III-C demonstrates this phenomenon.
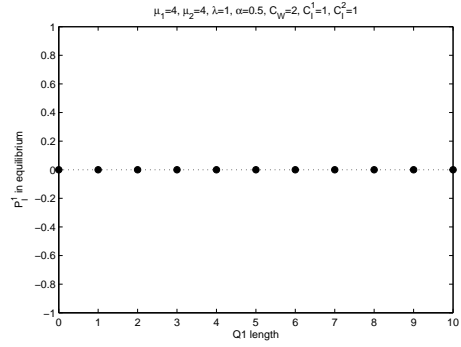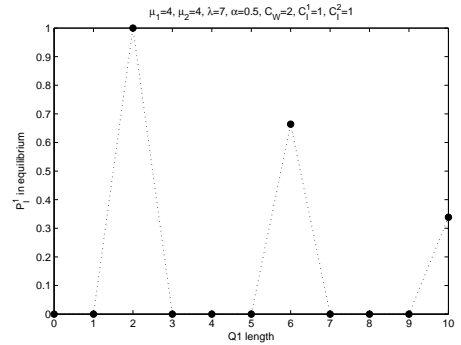

(a) $\rho = \frac{1}{8}$
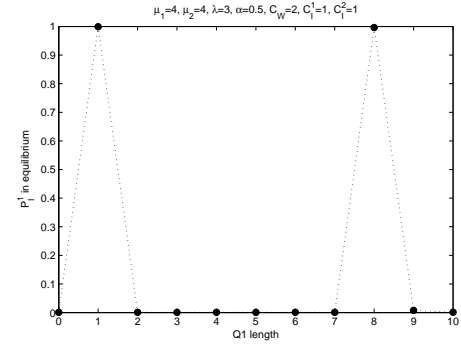
(b) $\rho = \frac{3}{8}$
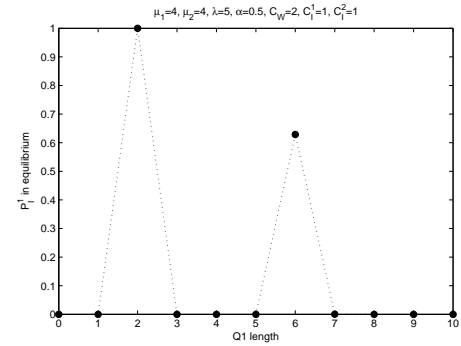
(c) $\rho = \frac{5}{8}$

(d) $\rho = \frac{7}{8}$

Fig. 2. Cascades multiplies as ρ increases

Notice, that the dashed line in all the figures in this paper is for graphical help - the strategy in all cases is discrete.

In Figure III-C, we fixed the following parameters: $\mu_1 = \mu_2 = 4, C_1 = C_2 = 1, C_W = 2$. We changed the customers' arrival rate, λ. In 2(a), $\lambda = 1$ and therefore $\rho = \frac{1}{8}$, and customers' strategy in equilibrium is to join Q1 without inspecting Q2, due to relatively high cost of inspection. As ρ increases to $\frac{3}{8}$ in Figure 2(b) due to increase in λ, cascades appear. As ρ continues to grow, cascades appear for earlier (see Figure 2(d) where $\rho = \frac{5}{8}$), and multiply (see Figure 2(c) where $\rho = \frac{7}{8}$).

We also looked into the changes in the equilibrium when the cost of waiting is changing with respect to the cost of inspection. The numerical results show, that when the ratio $\frac{C_W}{C_I} \to 0$, customers' dominant strategy is to join the first queue that they observed, no matter what is its current length. As the ratio $\frac{C_W}{C_I}$ increases, cascades appear, and customers inspect the other queue for isolated observable queue lengths. Cascades increase with $\frac{C_W}{C_I}$, and inspecting the other queue becomes a dominant strategy in equilibrium. Figure III-C is an example for that.

In Figure III-C, we fixed the following parameters: $\mu_1 = \mu_2 = 4, \lambda = 7.5, C_1 = C_2 = 1$. We changed the cost of waiting, C_W. In Figure 3(a), $C_W = 1$, and customers' strategy in equilibrium has one cascade when $i = 2$. As C_W increases, cascades increase (see Figure 3(b) for $C_W = 3$). In Figure 3(d), $C_W = 9$, and we get more cascades and a mixed strategy when $i = 8$. In Figure 3(c), $C_W = 25$, and it shows that in equilibrium, customers tend to inspect the other queue for many observable queue lengths.

IV. CONCLUSION

Our main conclusion is that when customers search for a server in a multiple servers system, their equilibrium behavior is not easily described by a simple strategy. In particular, it is not a threshold strategy. We focused on the case of two servers, and found that the equilibrium strategy is characterized by cascades. There exist queue lengths, for which customers join the first queue they inspect, for a longer queue they inspect the other queue, and for even longer queues they again join the first queue.

This behavior is a result of the positive externalities in this model: when a customer assumes with high probability that the customer in front of him inspected the other queue and yet decided to join this queue, it serves as a signal and influences his decision whether or not to inspect the other queue.

In the numerical analysis, we assumed that the servers are identical in their service rate and in their cost of inspection. We found that:

1) Cascades appear when ρ is small, and tend to multiply as ρ increases.

(a) $\frac{C_W}{C_I} = 1$

(b) $\frac{C_W}{C_I} = 3$
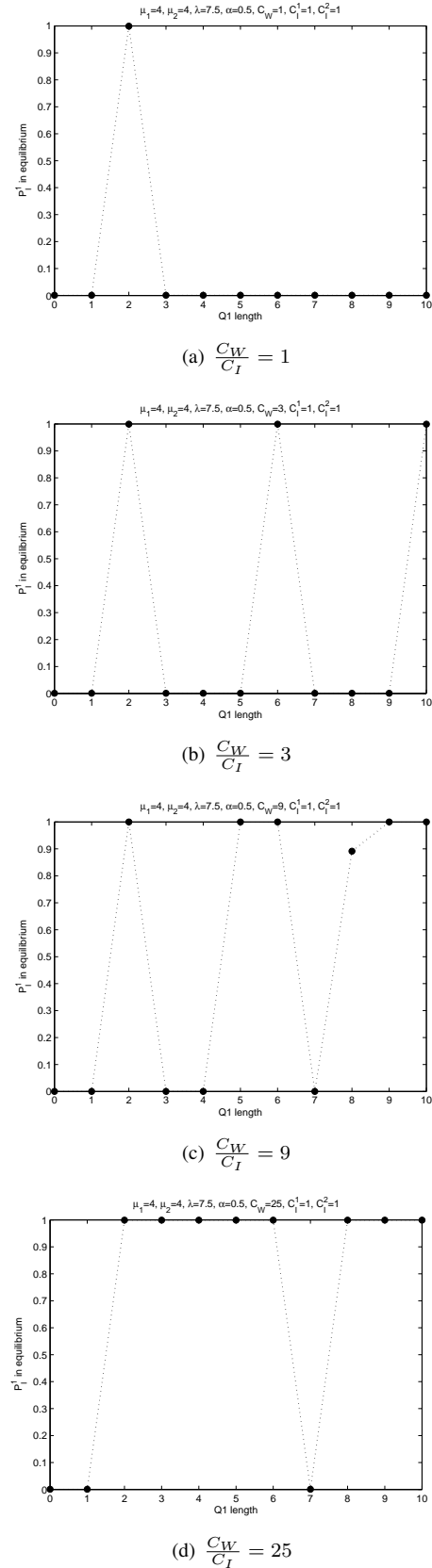

(c) $\frac{C_W}{C_I} = 9$

(d) $\frac{C_W}{C_I} = 25$

Fig. 3. Cascades multiplies as the ratio $\frac{C_W}{C_I}$ increases

2) When the cost of waiting is low comparing to the cost of inspection, customers' strategy in equilibrium is to join the first queue they inspect. As the ratio $\frac{C_W}{C_I}$ increases, cascades appear. For higher values of $\frac{C_W}{C_I}$, customers' strategy is to inspect the other queue.

How can a planner of future queueing systems take this behavior into account? Should competing servers control their cost of inspection in order to affect customers inspection policy? How does advertising affect customers' decision of which queue to inspect first? These questions and others arise from this model, and can be used as a basis for future research.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] A.V. Banerjee (1992), *A simple model of herd behavior*, The Quarterly Journal of Economics CVII, 797-817.

[3] 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.

[4] 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.

[5] L. Debo, C. Parlour, U. Rajan (2007), *Inferring quality from a queue*, Services and Operations Management (Forthcoming).

[6] R. Hassin (1995), *On the advantage of being the first server*, Management Science 42, 618-623.

[7] 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.

[8] R. Hassin and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*, Kluwer Academic Publishers, Norwell, Massachusetts, USA.

[9] R. Hassin and R. Roet-Green (2012), *Equilibrium in a two dimensional queueing game: When inspecting the queue is costly*.

[10] M. Haviv and Y. Kerner (2007), *On balking from an empty queue*, Queueing systems 55, 239-249.

[11] S. Veeraraghavan and L. Debo (2008a), *Joining longer queues: Information externalities in queue choice*, Services and Operations Management (Forthcoming).

[12] S. Veeraraghavan and L. Debo (2008b), *Is it worth the wait? Service choice when waiting is expensive*.

[13] W. Whitt (1986), *Deciding which queue to join: some counterexamples*, Operations Research 34, 55-62.