UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 11-33 (25)
Statistics and Econometrics Series
October, 2011

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

EQUILIBRIUM STRATEGIES IN A TANDEM QUEUE UNDER VARIOUS LEVELS OF INFORMATION

Bernardo D'Auria[1] and Spyridula Kanta[1]

Abstract

We analyze from an economical point of view a tandem network with two nodes. We look at different situations, that is, when customers upon their arrival are no informed, partially informed or totally informed about the state of the system. For each case, we look for the strategy that optimizes the individual net benefit. In addition, for the totally unobservable case, we also study the strategy that would be socially optimal, i.e. maximizing the overall welfare.

Keywords: Tandem queues, Nash equilibrium, Optimization, Economical Strategies.

Acknowledgements: The first author is partially supported by the Spanish Ministry of Education and Science Grants MTM2010-16519, SEJ2007-64500 and RYC-2009-04671.

[1] Departamento de Estadística, Universidad Carlos III de Madrid,
Avda. Universidad 30, 28911 Leganes (Madrid), Spain;
emails: bernardo.dauria@uc3m.es, spyridoula.kanta@uc3m.es

Equilibrium strategies in a tandem queue under various levels of information

Bernardo D'Auria Spyridoula Kanta

Abstract

We analyze from an economical point of view a tandem network with two nodes. We look at different situations, that is, when customers upon their arrival are no informed, partially informed or totally informed about the state of the system. For each case, we look for the strategy that optimizes the individual net benefit. In addition, for the totally unobservable case, we also study the strategy that would be socially optimal, i.e. maximizing the overall welfare.

1 Introduction

In the last decades there is a tendency of studying queueing models from an economic viewpoint. More concretely a reward/cost structure is introduced and the objective is the optimization of the system. Each customer can take his own decision. Of course the customers act individually and independently one from the others in order to maximize their welfare. Inevitably, each customer's decision affects and is affected by the decisions of the other customers and the administrator of the system. The result is an equilibrium scheme where no one has incentive to deviate from this. Such kind of situations can be considered as a game between the customers or between the customers and the manager of the system. For that reason game-theoretic ideas are applied when a queueing system is analyzed under an economic framework.

The notions of supply and demand lie at the heart of the economic analysis. As it is straightforward, the price is used as the rationing device in most economic models. A closely related form of rationing that can be distinguished is the waiting line. Therefore, when an economic framework is considered, concepts of the economic theory should be introduced so that a queueing system can be studied. According to Martin and Smith (1999), except for the price, the waiting list and the waiting line are the most related to the price forms of rationing. The individuals join and remain in a queue to gain access to the good or service concerned. About fifty years ago, a discussion began regarding the question if an entrance fee imposed to the arriving customers at a service station is a rational measure. The situation is similar when a highway is considered and the imposition of a toll to control the traffic is discussed. A rational consumer is searching for low waiting time on one hand and on the other hand is his willingness of paying the corresponding price. A long line indicates an underpriced experience. By charging each individual, the demand can be rationed until the queue is winnowed down to something reasonable. Hence, pricing a queue can regulate the demand.

As stated before, a specific characteristic of a queueing system is the waiting time of the customers. The waiting time plays a significant role in the decision of the consumers and it forms part of the product quality (Zimmerman and Enell, 1993). Therefore, the waiting time can be considered as an inseparable part of a good or service. Greater waiting times imply larger costs on behalf of the consumers additionally to the price of the product or service. Rational consumers

should take into account the full price, that is the monetary price plus the waiting cost, when they are about to decide if they will join the queue or not for a good or service. In the concepts of cost or reward as used in queueing models, the opportunity costs, the lost time, the pleasure of receiving service or obtaining a good or any other benefit that provides utility are included. According to a basic principle of economics, a rational person will take a decision if and only if the marginal benefit of his action is greater than the corresponding marginal cost. Moreover a rational customer adopts a stopping rule in his search for low waiting time. He will join only those queues whose length is less than or equal to some critical value. The critical queue length depends on the distribution of the queue length, the value of the time, the cost of search and the availability of other suppliers (Vany, 1976). It also depends on the related information structure which has an important role in the decision process.

In general in these models the customers can take several types of decisions, for example if they will join the queue or they will balk, if they will join and then renege at some moment or they will stay until served, if they will buy priority so that to overpass some or all of the existing customers, etc. There are three different directions of studying these kind of queueing models under this game-theoretic/economic framework. The first interest of the study is to identify the equilibrium behavior of the customers. When all customers behave with the same way, then no one has incentive to deviate from that strategy. This equilibrium from the point of view of the society may not be optimal. A common concept in economics is the existence of externalities. In queueing systems this phenomenon also appears since the decision of the customers affects the congestion found by others. The difference between the individually and socially optimal behavior is due to the existence of the externalities that customers imply on later arrivals. At the time of the decision, the customers do not take into account these externalities. Hence a second task is to specify the socially optimal strategy that when followed by all customers, the social welfare is maximized. By social welfare, the total expected net benefit of the society including both customers and the manager of the system is considered. Finally, the problem that the administrator of the system is willing to solve is to identify the admission fee that he has to impose so that his own profit is maximized. Moreover a crucial point of consideration when studying any of the aforementioned directions is the level of information available to the customers at the time of taking their decision. The customers may or may not have at their disposal some information regarding the state of the system. Obviously their decision and consequently their behavior varies depending on the received information. Hassin and Haviv (2003) provides a rich bibliography where the basic models are analyzed with an extensive literature regarding this research area.

The simplest queueing system consists of a Poisson arrival process, exponential service times, one server and FCFS discipline. This is the well known M/M/1 queue that it was first studied under an economic framework by Naor (1969) and Edelson and Hildebrand (1975). In their pioneering works, they study the M/M/1 queue under two different levels of information, where the customers have to decide at the time that they arrive at the system if they will enter or not. The study of the utility function is the central idea for carrying out the analysis. They identified the equilibrium strategies as well as the social and profit maximizing ones. The difference between the two studies is the level of the information available to the customers at the time of their arrival. Naor (1969) considered this markovian model assuming a simple linear reward/cost structure, where the customers upon their arrival are informed about the exact number of customers already present in the system. This is referred as the observable model. He argues that in order to make the customers behave in a socially optimal way, the imposition of tolls is necessary. On the other hand, Edelson and Hildebrand (1975) studied the same model for the unobservable case, that is, under the assumption that the customers should take their decision without having any knowledge about the state of the system. The effect of the information is of interest when studying such models from the point of view of the manager as well as of the customers since their decisions vary.

Several papers refer to the value of information in the customers' strategies and the performance of the system, see for example Hassin (2007), Guo and Zipkin (2007).

After the pioneering works of Naor (1969) and Edelson and Hildebrand (1975), the interest on the study of queueing models in an economic context or use queueing theory to model the way a market functions, has become more intensive. Towards this direction, Parra-Frutos and Aranda-Gallego (1999) study a model where two classes of customers, that they differ in their tolerance to wait for service, arrive at a firm. They identify two thresholds that represent the highest number of customers that each class of consumers tolerates in the system, otherwise they decide to balk, that is, to leave the system without obtaining the product or service. More complicated or non-Markovian models are extensively studied by several authors. See for example Yechiali (1971, 1972), Mandelbaum and Shimkin (2000), Lin and Ross (2001), Armony and Haviv (2003), Economou and Kanta (2008). Queues with vacations of the server with linear rewards/costs were first analyzed by Burnetas and Economou (2007) under various levels of information and an extension to general service and vacation times appears in Economou et al. (2011).

The literature on queueing networks is rich, most focused on the performance of these systems with many extensions and generalizations. To the best of our knowledge, there is scarce literature on the study of queueing networks from an economic viewpoint, some results could be found in Section 3.8 in Hassin and Haviv (2003) or in the recent paper by Burnetas (2011). A thorough analysis of pricing in communications networks is carried out in the book by Courcoubetis and Weber (2003) with many references in relative issues. The ways that a communications network should be priced, the necessity and the advantages of the pricing are presented in details as well as many applications on the Internet and telecommunications. The economic consequences or the problems of congestion that arise because of the lack of pricing a network are also analyzed. In everyday life the need of waiting at more than one successive queues is very frequent. The consumers arriving at a firm may have to wait to two different counters before obtaining the service or good considered. Some of them may need to pass from only one or more counters depending on their needs or obligations. A very simple but representative example is a health center where all customers have to pass by the secretary but then each person is assigned to the appropriate special doctor. In such cases the demand as expressed by the effective arrival rate is not stable since potential customers may or may not join the clinic if they observe too many people waiting on the secretary's desk, assuming that we are not in a monopoly market so they can choose to balk and go to another clinic. On the other hand they may decide to join the clinic if they know that the persons waiting for the same specialist are not too many. Consequently any information available regarding the state of the system, that is the number of customers already present, affects their joining/balking decision. The effective capacity of the clinic as expressed by the number of customers served per time unit is not infinite. Of course, the waiting space of the clinic is not infinite neither. The customers can not avoid waiting which is directly connected to a cost. Therefore, before taking the decision of joining or balking, they take into consideration the cost of waiting on one hand and the necessity of receiving the service, reclaiming of course the providing (if any) piece of information.

Motivated by such examples that can be found plenty of them in real life, we study a firm where customers after joining they have to visit sequentially two queues. This is what in queueing theory is called a network of two tandem queues. Regarding the level of information available to the customers, the manager may provide several information. In the present paper, we focus on the study of the case where the customers may observe the exact number of consumers in front of them in both queues. This is the full information case or fully observable model. If only the information of the number of customers in one of the queues, either the first or the second one, is provided we have the partial information cases or partially observable system. We also study the case where the information available is the exact number of customers present in the whole

system but without knowing exactly the situation in each one of the queues. We refer to this case as partially unobservable. Finally we have the fully unobservable model where the customers take their decision of joining or balking without having at their disposal any kind of information regarding the state of the system.

The structure of the paper is as follows. In Section 2 we introduce the probability model, then in Section 3 we study the optimal threshold policy of the customers when they have all information about the state of the system at their arrival epochs. In Section 4 we study the individual, social and administrator equilibrium policies when customers have no information about the system besides its structural parameters. Finally in Section 5 we study the optimal threshold policies when the customers have only partial information about the state of system upon their arrival, respectively the total number of customers in the tandem network (subsection 5.1), the number of customers at the first node (subsection 5.2) and the number of customers at the second node (subsection 5.3).

2 The model

We consider a service system that consists of two nodes. Customers arrive at the system according to a Poisson process at rate λ and begin their service at the first node. After completing their first service they join the line at the second node, and then after completing their second service they finally leave the system. We assume that both nodes have infinite capacity and that the two servers are heterogeneous, i.e. customers experience two different and independent service times. The service times are assumed to be exponentially distributed random variables with rates μ_i, $i = 1, 2$ at the respective node i. This model is generally known as an open tandem network and the state of the system at time t is given by a random vector $(Q_1(t), Q_2(t))$, where $Q_i(t)$ denotes the number of customers in queue i, $i = 1, 2$.

The goal of this work is to study the behavior of the customers when they can decide whether to join the system or balk upon their arrival. To this framework we introduce the reward/cost structure. We assume that each customer receives a reward of R units due to service completion and at same time suffer a waiting cost for the total amount of time that they spend in the system, either waiting or being served in any of the two queues. This cost is equal to C_i per time unit spent at queue i, $i = 1, 2$. Customers are assumed to be risk neutral, that is, they aim to maximize the expected value of their net benefit, by using all the information to them available at the arrival time. The amount of information we are going to consider includes all the structural parameters of the system, the reward/costs fees as well as some level of information about the current state of the system.

We explicitly notice that the decision of the customers has to be taken at the arrival moment and it is irrevocable, that is they can decide to join or to balk the queue, but no reneging of the joining customers or retrials of the balking customers is allowed.

In order to avoid trivialities we assume that

$$R > \frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}. \tag{1}$$

If this condition fails to hold, then even a customer that finds both queues empty suffers on average a negative benefit by joining. This implies that each customer will decide not to join the system and therefore the system will stay continuously empty.

The basic characteristic that affects the decision of the customers is the level of the information that they receive upon their arrival. We study three different levels of information. In the first case, referred to as the *fully observable*, the customers can observe the exact state of the system, that is, they know the exact number of customers at each node. In the second case, the *fully unobservable*,

the customers upon their arrival receive no information about the state of the system so that they take a random decision if they will join or not. Finally we study the cases, the *partially observable* ones, where the customers get informed only partially about the state of the system. By partial information we mean, the total number of customers in the system, only the number of customers at the first node, and finally, only the number of customers at the second node.

3 Fully observable case

With the *fully observable* case we refer to the case when customers upon their arrival can take decision about joining the tandem network also by observing the full state of the system, i.e. the number of customers present in each queue. In this case a pure threshold strategy is defined by a pair of integers (n, m), meaning that a customer decides to enter in the system as long as the number of customers in the first queue is at most n (himself included) and at most m in the second queue, otherwise he balks.

Let define by $S_i(n, m)$, $i = 1, 2$, the sojourns times, respectively in the first and the second queue, spent by a tagged customer that at the moment of his arrival finds $n - 1$ customers in the first queue and m customers in the second queue. Let $T_i(n, m) = E[S_i(n, m)]$ the corresponding expectation and define $T(n, m) = T_1(n, m) + T_2(n, m)$ as the total expected sojourn time in the system of the tagged customer. Under the assumption that the service time in the first node is distributed according to an exponential random variable with parameter μ_1 we have that

$$S_1(n, m) \sim \text{Erlang}(n, \mu_1) \ ,$$

and hence $T_1(n, m) = n/\mu_1$. The distribution of the sojourn time in the second queue is more complicated to compute, but for our purposes we only need its expectation, therefore in what follows we only provide an algorithm for the computation of the quantity $T(n, m)$. Then it is straightforward to compute the expected sojourn time at the second node as $T_2(n, m) = T(n, m) - T_1(n, m) = T(n, m) - n/\mu_1$.

Consider a tagged customer that upon arrival is getting informed that there are $n-1$ customers in the first node and m in the second one. If he decides to enter then he will be assigned to the $n - th$ position of the first queue and his net profit is given by

$$P(n, m) = R - C_1 \frac{n}{\mu_1} - C_2 T_2(n, m) = R - (C_1 - C_2)\frac{n}{\mu_1} - C_2 T(n, m) \ ,$$

while if he decides not to enter his net profit is zero. He will decide to enter if his profit is non negative. Of course, for such a customer, the decisions of the future arrivals will not affect his sojourn time.

Assuming that the customer has joined the system and denoting by (n, m) the state of the system, we have that for $n \geq 1$ and $m \geq 1$, the next event is either a service completion in the first queue, after an exponential (μ_1) time, or a service completion in the second queue, after an exponential (μ_2) time, whatever happens first. Thus the next event will occur after an exponential $(\mu_1 + \mu_2)$ time. With probability $\mu_1/(\mu_1 + \mu_2)$ the customer in service at the first queue completes his service and proceeds to the second queue. The remaining sojourn time of the tagged customer is then given by $T(n-1, m+1)$. On the other hand, with probability $\mu_2/(\mu_1+\mu_2)$ a service in the second queue is completed and the remaining sojourn time of the tagged customer is given by $T(n, m - 1)$. By first step analysis argument, we therefore get

$$T(n, m) = \frac{1}{\mu_1 + \mu_2} + \frac{\mu_1}{\mu_1 + \mu_2} T(n - 1, m + 1) + \frac{\mu_2}{\mu_1 + \mu_2} T(n, m - 1), \quad n, m > 0. \tag{2}$$

If $n = 0$ means that the tagged customer is the last one in the second queue and hence, he only has to wait for the customers in front of him to be served plus himself. We have then

$$T(0, m) = \frac{m}{\mu_2}. \tag{3}$$

If at the time of arrival of the tagged customer the second queue is empty, i.e. $m = 0$, then the next event that will happen is a service completion in the first queue. Thus we obtain the equation

$$T(n, 0) = \frac{1}{\mu_1} + T(n-1, 1) \ , \tag{4}$$

valid for $n \geq 1$. The system of the equations (2), (3) and (4) give the sojourn times for any $n \geq 1$ and $m \geq 0$, i.e. for any possible value of (n, m). To simplify the solution of these equations we consider solutions of the form

$$T(n, m) = y(n, m) \left(\frac{\mu_2}{\mu_1 + \mu_2} \right)^m + \frac{n+m}{\mu_2}. \tag{5}$$

By plugging this expression in equations (2)–(4) we obtain the following system of equations:

$$y(n, m) = \frac{\mu_1 \, \mu_2}{(\mu_1 + \mu_2)^2} \, y(n-1, m+1) + y(n, m-1) \tag{6}$$

$$y(n, 0) = \frac{1}{\mu_1} + \frac{\mu_2}{\mu_1 + \mu_2} \, y(n-1, 1) \tag{7}$$

$$y(0, m) = 0 \tag{8}$$

where n, $m > 0$. By induction it is easy to check that the following equation can be used to compute the values of $y(n, m)$ for $m \geq 1$

$$y(n, m) = \frac{1}{\mu_1} + \frac{\mu_2}{\mu_1 + \mu_2} y(n-1, 1) + \frac{\mu_1 \, \mu_2}{(\mu_1 + \mu_2)^2} \sum_{k=0}^{m-1} y(n-1, k+2) \ , \tag{9}$$

that only makes use of the values of $y(n-1, k)$, with $1 \leq k \leq m+1$. By computing recursively the values of $y(n, m)$, one immediately gets the values of $T(n, m)$ in (5). This allows to compute exactly the individual net benefit of the tagged customer for any value of (n, m) and to get the equilibrium threshold strategy $m^*(n)$ by the relation

$$P(n, m^*(n)) \geq 0 \quad \text{and} \quad P(n, m^*(n)+1) < 0 \ , \tag{10}$$

where we have assumed $P(n, -1) = 0$. If $m^*(n) = -1$, then a tagged customer that finds n customers in the first queue decides to balk independently of the number of customers at the second queue.

Figure 1 below shows the two thresholds for a certain scenario for the parameters of the system. In the numerical example, we have used the following values for the system parameters: $R = 20$,$C_1 = 1$, $C_2 = 2$, $\mu_1 = 1.2$ and $\mu_2 = 0.7$. The graph show the function $(n, m^*(n))$ and, for instance, if a customer upon his arrival receives the information $(0, 6)$ meaning that the first queue is empty and that in the second there are 6 customers, his decision will be to enter the system. This is because, with himself included, the state of the system will be $(1, 6)$ which is a "permitted" strategy. His decision of course will be the same for any of the states $(0, m)$ for $m = 0, 1, \ldots, 6$. The polygonal line divides the plane in two regions: the region under the function $m^*(n)$ contains all the pairs of integers (n, m) where the decision is to join and the region above that consists of all the pairs that the decision is to balk.

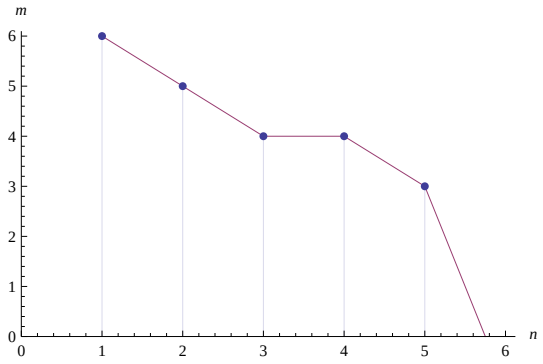


Figure 1: Thresholds (n, m)

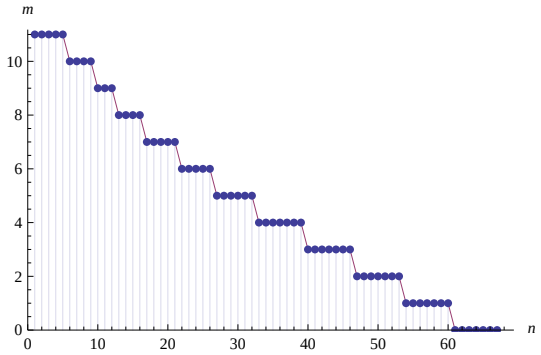


Figure 2: Thresholds (n, m)

In Figure 2, we computed the equilibrium strategy after having interchanged the values of the two service rates, i.e. now $\mu_1 = 0.7$ and $\mu_2 = 1.2$. We can see that the performance of the system changes significantly. Before we had a cheaper and faster first server in comparison with the second one. In this scenario we still have a cheaper first server but the second server is now faster than the first one. Even though the second server is more expensive and the reward R stays fixed, the maximum possible number of customers in both queues increases to a value that almost doubles the one in the previous scenario.

4 Fully Unobservable

In this section we study the case where the customers upon their arrival do not receive any information about the state of the system. The decision that they have to make is to join or to balk. We assume that all customers follow the same mixed strategy q, namely they join with probability q and they balk with probability $1-q$. Under this strategy, the system is an open Jackson network where the arrival process is Poisson (λq). For this system, the real arrival rate in each queue, taking into account the traffic equations, is λq. Denoting by T_i, $i = 1, 2$ the expected sojourn time of a customer at node i of the network, we have that $T_i = (\mu_i - \lambda q)^{-1}$.

4.1 Equilibrium behavior

Consider a customer that arrives at the system. If he decides to join his expected net benefit is $P(q) = R - C_1\, T_1 - C_2\, T_2$ which can be rewritten as

$$P(q) = R - \frac{C_1}{\mu_1 - \lambda q} - \frac{C_2}{\mu_2 - \lambda q}\ . \tag{11}$$

We assume that the parameters satisfy the following conditions

$$\mu_1 - \lambda > 0 \quad \text{and} \quad \mu_2 - \lambda > 0. \tag{12}$$

ensuring that the system is stable under any strategy q that customers may follow. Note that when these conditions hold then $\mu_1 + \mu_2 - 2\lambda > 0$.

Theorem 1. *Consider the fully unobservable system of two tandem queues where the conditions (1) and (12) hold. There is a unique equilibrium strategy "enter with probability q_e" where q_e is*

given by

$$q_e = \begin{cases} q^* & \text{if } R \in \left[\frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}, \frac{C_1}{\mu_1-\lambda} + \frac{C_2}{\mu_2-\lambda}\right) \\ 1 & \text{if } R \in \left[\frac{C_1}{\mu_1-\lambda} + \frac{C_2}{\mu_2-\lambda}, +\infty\right) \end{cases} \tag{13}$$

where

$$q^* = \frac{R(\mu_1+\mu_2) - (C_1+C_2) - \sqrt{[R(\mu_1-\mu_2) - (C_1-C_2)]^2 + 4C_1C_2}}{2\lambda R} \tag{14}$$

Proof. Consider a customer that arrives at the system when all other customers follow the same mixed strategy q, i.e. they enter with probability $q \in [0,1]$. His expected net profit, if he decides to enter, is given by (11). Obviously he decides to join as long as his benefit is non-negative. We distinguish three cases:

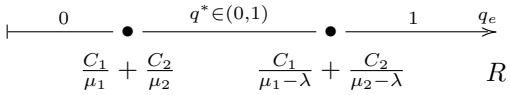
- If $R < \frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}$ then the benefit of the tagged customer is negative and the best decision for him is to balk. Balking is then a dominant strategy and the system stays continuously empty, as noticed before.

- If $R \in \left[\frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}, \frac{C_1}{\mu_1-\lambda} + \frac{C_2}{\mu_2-\lambda}\right)$, there is a unique value of q such that $P(q) = 0$. Indeed the equation $P(q) = 0$ is quadratic and admits two real roots but, one of which, under the condition (12), is always greater than 1 and therefore does not define a probability. So we have only one root which is q^* as given by (14). For this value of q, the tagged customer is indifferent between joining the system and balking so any joining probability $q \in [0,1]$ is a best response. The unique strategy that is a best response against itself is q^*. So $q_e = q^*$ is the equilibrium strategy and we obtain the first brunch of (13).

- If $R \geq \frac{C_1}{\mu_1-\lambda} + \frac{C_2}{\mu_2-\lambda}$, then even in the case that all customers decide to enter ($q = 1$) the net benefit of the tagged customer is nonnegative, that is $P(q) \geq 0, \forall q \in [0,1]$. So the best response for him is to join (with probability 1). Hence joining is a dominant strategy.

□

The results of Theorem 1 can be graphically summarized by the following diagram



Remark 2. *The function $P(q)$ is a decreasing function with respect to the strategy q followed by the customers. Suppose that all join with probability q, then the utility of a tagged customer that joins with probability s is equal to $s\,P(q)$. We denote by $s^*(q)$ the best response of the tagged customer against the strategy q chosen by the other customers. It is easy to see that $s^*(q) = 0$ when $q > q_e$ and $s^*(q) = 1$ when $q < q_e$. When $q = q_e < 1$ all strategies are best responses against q_e and q_e is therefore the unique best response against itself, i.e. it is a (symmetric) equilibrium strategy. It follows that the best response function is not increasing with respect to q, that implies that the model is of* Avoid The Crowd*(ATC) type. It is well known that this kind of models admits a unique equilibrium point.*

4.2 Social Maximization

In this section we focus on solving the social optimization problem, that is we look for a strategy q_s that, when followed by all customers, maximizes the overall welfare. Recall that the customers receive no information about the state of the system, that is the number of customers in any of the two nodes. When all customers follow the same strategy q, the arrival process is Poisson(λq) and the social net profit per time unit is given as

$$P_s(q) = \lambda q \left(R - \frac{C_1}{\mu_1 - \lambda q} - \frac{C_2}{\mu_2 - \lambda q} \right) \tag{15}$$

We aim at identifying the probability $q \in [0,1]$ that maximizes the above function. Towards this direction, we have the following theorem.

Theorem 3. *Consider the fully unobservable system of two-queues tandem network where the conditions (12) and (1) hold. There is a unique strategy that maximizes the social benefit per time unit 'enter with probability q_s', where q_s is given by*

$$q_s = \begin{cases} q_s^*, & \text{if } R \in \left(\frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}, \frac{\mu_1 C_1}{(\mu_1 - \lambda)^2} + \frac{\mu_2 C_2}{(\mu_2 - \lambda)^2} \right) \\ 1, & \text{if } R \in \left[\frac{\mu_1 C_1}{(\mu_1 - \lambda)^2} + \frac{\mu_2 C_2}{(\mu_2 - \lambda)^2}, +\infty \right) \end{cases} \tag{16}$$

where q_s^* is the unique solution in (0,1) of the equation

$$R = \frac{\mu_1 C_1}{(\mu_1 - \lambda q)^2} + \frac{\mu_2 C_2}{(\mu_2 - \lambda q)^2}. \tag{17}$$

Proof. In order to identify the strategy q that maximizes the social profit per time unit we consider the first order condition. The first derivative of (15) is given by

$$P_s'(q) = \lambda \left(R - \frac{\mu_1 C_1}{(\mu_1 - \lambda q)^2} - \frac{\mu_2 C_2}{(\mu_2 - \lambda q)^2} \right) . \tag{18}$$

We are looking for the roots of the equation $P_s'(q) = 0$ (if any) in $[0,1]$. Note that $P_s'(q)$ is clearly a decreasing of q in the interval $[0,1]$, $P_s'(0) = \lambda(R - C_1/\mu_1 - C_2/\mu_2)$ and $P_s'(1) = \lambda(R - \mu_1 C_1/(\mu_1 - \lambda)^2 - \mu_2 C_2/(\mu_2 - \lambda)^2)$. We distinguish the following three cases:

- If $R \leq \frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}$, the above condition ensures that $P_s'(0) \leq 0$ and therefore $P_s'(1) < 0$. Consequently the function $P_s(q)$ is decreasing in $[0,1]$ and attains its maximum value for $q_s = 0$. This case is meaningless since such a case would result in a behavior on behalf of the customers that they decide never to join and the system would be continuously empty. This trivial case never appears under the condition (1) that we have assumed in the theorem.

- If $R \geq \frac{\mu_1 C_1}{(\mu_1 - \lambda)^2} + \frac{\mu_2 C_2}{(\mu_2 - \lambda)^2}$, $P_s'(1) \geq 0$ and therefore $P_s'(0) > 0$. We conclude then that the function $P_s(q)$ is increasing in $[0,1]$ and consequently obtains its maximum value at $q_s = 1$. Under this strategy it is socially optimal that all customers join the system. This case corresponds to the second brunch of (16).

- If $R \in \left(\frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}, \frac{\mu_1 C_1}{(\mu_1 - \lambda)^2} + \frac{\mu_2 C_2}{(\mu_2 - \lambda)^2} \right)$, no one of the two cases above hold, then we conclude that $P_s'(0) > 0$ and $P_s'(1) < 0$. Therefore there is a unique (since P_s' is decreasing) root of the function P_s' in $(0,1)$. We denote this root by q_s^*, and we have that $P_s'(q_s^*) = 0$. The monotonicity of P_s' implies that the function $P_s(q)$ is a concave function with respect to q in

$(0,1)$ and attains its maximum at the point q_s^*. As explained, the value of q_s^* can be found as the solution of the equation $P_s'(q) = 0$ in $(0,1)$, that is the solution of the equation (17) in $(0,1)$. This case corresponds to the first brunch of (16) and completes the proof of the theorem.

□

Finally the following result shows the relation between the optimal individual and social behavior.

Theorem 4. *For any values of the parameters, under the stability condition $\mu_1, \mu_2 > \lambda$ we have that*

$$q_s \leq q_e \ . \tag{19}$$

Proof. First note that $\mu_i/(\mu_i - \lambda) > 1$, $i = 1, 2$, so that we have to study the inequality (19) in the following three intervals:

- If $R \in \left(\frac{C_1}{\mu_1} + \frac{C_2}{\mu_2}, \frac{C_1}{\mu_1 - \lambda} + \frac{C_2}{\mu_2 - \lambda}\right)$, according to Theorem 1 we have that $q_e = q^*$, with q^* given in (14), and, by Theorem 3, $q_s = q_s^*$, with q_s^* being the probabilistic solution of (17). From the definition of the equilibrium strategy, q_e, we have $P(q_e) = 0$ and $P(q) \leq 0$ for any $q \in [q_e, 1]$. Having that $P_s(q) = \lambda q\, P(q)$, it follows also that $P_s(q) \leq 0$ for any $q \in [q_e, 1]$. Since in the interval we are considering $P_s(0+) > 0$ and q_s is where the function $P_s(q)$ reaches its maximum we have that $P_s(q_s) > 0$ and it follows that $q_s < q_e$.

- If $R \in \left[\frac{C_1}{\mu_1 - \lambda} + \frac{C_2}{\mu_2 - \lambda}, \frac{C_1\mu_1}{(\mu_1 - \lambda)^2} + \frac{C_2\mu_2}{(\mu_2 - \lambda)^2}\right)$, according to Theorem 1, $q_e = 1$, while by Theorem 3, $q_s = q_s^* \in (0,1)$ and therefore the result is straightforward.

- If $R \geq \frac{C_1\mu_1}{(\mu_1 - \lambda)^2} + \frac{C_2\mu_2}{(\mu_2 - \lambda)^2}$, both optimal joining probabilities are equal to 1 and the result holds as equality.

□

The ordering relation between the individual and the social strategies, as commented in the introduction, is due to the negative externalities that customers impose on later arrivals. Each customer, when thinking individually, takes into account only his own benefit without bothering for the benefit of the total welfare and so ignoring the negative effect that his decision implies on the customers that arrive to the system after him.

4.3 Profit Maximization

In this section we consider the profit maximization problem, that is, we assume that there is an administrator of the system that look for a strategy that maximizes his profit. The strategy consists in imposing an admission fee to every customer entering the system.

When an additional admission fee r is imposed, the reward of each customer is reduced from R to $R - r$. Therefore assuming that they follow their equilibrium strategy, the customers adjust their strategy to the new equilibrium point. Hence they decide to join the system with a probability that it is defined by (11). More concretely, substituting $R - r$ in (11) instead of R and solving the equation $P(q) = 0$ with respect to r we obtain

$$r = R - \frac{C_1}{\mu_1 - \lambda q} - \frac{C_2}{\mu_2 - \lambda q} \ . \tag{20}$$

Each toll r imposed, induces on behalf of the customers a different equilibrium strategy q. The administrator's profit per time unit is $P_{ad}(q) = \lambda q\, r$ which taking into account (20) obtains the form

$$P_{ad}(q) = \lambda q \left(R - \frac{C_1}{\mu_1 - \lambda q} - \frac{C_2}{\mu_2 - \lambda q} \right) . \tag{21}$$

The problem of identifying the fee that maximizes the administration's problem reduces to the computation of the strategy q (the joining probability) that maximizes the above function. Observe that the function (21) of the administrator's profit is the same as (15) that gives the social net profit per time unit. That is the two functions coincide. This is a common phenomenon that appears in the models where no information is available (see Hassin and Haviv(2003)). The objectives of a profit maximizer and the society coincide. In these cases, a monopoly does not leave a positive surplus to the customers.

5 Partially observable case

In this section we study the optimal threshold policies when the customers have only partial information about the state of system upon their arrival. In particular we look at the three cases when customers know, in order of presentation, only the total number of customers in the tandem network, only the number of customers at the first node and finally only the number of customers at the second node.

In order to describe in a more conveniently form the results in the last two cases, we introduce here some matrix convention. The identity and the zero matrices are denoted respectively by $\mathbb{I}$ and $\mathbb{O}$. The transposition operator is denoted by $\mathtt{t}$, diag$[\boldsymbol{x}]$ denotes a diagonal matrix with the main diagonal given by the vector $\boldsymbol{x}$. and $\boldsymbol{e}_i$ is the $i+1$th vector of the canonical base in $\mathbb{R}^n$.

In addition we define the left and right shift $K+1$-square matrices by U_L and U_R. Their elements are given by $(U_L)_{i,j} = \delta_{i-1,j}$ and $(U_R)_{i,j} = \delta_{i+1,j}$ for $0 \le i,j \le K$ where $\delta_{i,j}$ denotes the Kronecker delta. Finally we define the projection matrix onto the $k-th$ coordinate, with $0 \le k \le K$ by U_k whose elements are given by $(U_k)_{i,j} = \delta_{i,k}\delta_{j,k}$.

Notice that we are using the convention that the first element of a matrix has position $(0,0)$. The sizes of the defined square matrices will be clear by the context.

5.1 Total number of customers in the network

In this section we study the case that an arriving customer is informed only about the total number of customers in the system, i.e. $n+m$, but does not how they are distributed in the two nodes. We assume that all customers follow the same threshold strategy K, that is they join as long as the total number of customers is less than K. As a consequence of this strategy, the total capacity of the system as well as the length of each queue can not exceed this number K. We define by $Q^*_{K,i}$ the stationary random number of customers at node i, $i = 1,2$, under the strategy K, and $Q^*_K = Q^*_{K,1} + Q^*_{K,2}$. The stationary distribution is given by

$$\pi_K(n,m) = \mathbb{P}(Q^*_{K,1} = n, Q^*_{K,2} = m) = c_K\, \rho_1^n \rho_2^m, \quad n+m \le K$$

where $c_K^{-1} = \sum_{n+m\le K} \rho_1^n \rho_2^m$ is calculated by the normalization equation.

We tag a customer that just arrives at the system and receives the information k, that is upon his arrival there are $k \le K$ customers in the system. We define also $T_{K,i}(k) = \mathbb{E}[S_i | Q^*_K = k]$ and $T_K(k) = T_{K,1}(k) + T_{K,2}(k)$. The expected net benefit of the tagged customer if he decides to join is

$$P_{tot}(k) = R - (C_1 - C_2)\, T_{K,1}(k) - C_2\, T_K(k) . \tag{22}$$

To compute $T_{K,1}(k)$ we have that

$$T_{K,1}(k) = \sum_{n=0}^{k} T_1(n, k-n)\, \mathbb{P}(Q^*_{K,1} = n | Q^*_K = k) = \frac{1}{\mathbb{P}(Q^*_K = k)} \sum_{n=0}^{k} \frac{n+1}{\mu_1}\, \pi_K(n, k-n)\ , \quad (23)$$

where we recall that $T_1(n,m) = \mathbb{E}[S_1(n,m)]$. Using the stationary distribution we can easily calculate the probability $\mathbb{P}(Q^*_K = k)$:

$$\mathbb{P}(Q^*_K = k) = \sum_{n=0}^{k} c_K\, \rho_1^n\, \rho_2^{k-n} = c_K\, \frac{\rho_2^{k+1} - \rho_1^{k+1}}{\rho_2 - \rho_1}\ . \quad (24)$$

Substituting (24) in (23) we obtain the following expression for the expected sojourn time of the tagged customer in the first queue given that his information is that upon his arrival there are k customers in the system:

$$\begin{aligned} T_{K,1}(k) &= \frac{\rho_2 - \rho_1}{\mu_1(\rho_2^{k+1} - \rho_1^{k+1})} \sum_{n=0}^{k} (n+1)\rho_1^n \rho_2^{k-n} = \frac{(\rho_2 - \rho_1)\rho_2^k}{\mu_1(\rho_2^{k+1} - \rho_1^{k+1})} \sum_{n=0}^{k} (n+1) \left(\frac{\rho_1}{\rho_2}\right)^n \\ &= \frac{(\rho_2 - \rho_1)\rho_2^k}{\mu_1(\rho_2^{k+1} - \rho_1^{k+1})} \frac{\rho_2^{k+2} - (k+2)\rho_2\rho_1^{k+1} + (k+1)\rho_1^{k+2}}{\rho_2^k(\rho_2 - \rho_1)^2}\ . \end{aligned} \quad (25)$$

Taking into account that $\rho_i = \lambda/\mu_i$, $i = 1, 2$, last equation can be simplified as follows:

$$T_{K,1}(k) = \frac{\rho_2^{k+2} - (k+2)\rho_2\rho_1^{k+1} + (k+1)\rho_1^{k+2}}{\mu_1(\rho_2 - \rho_1)(\rho_2^{k+1} - \rho_1^{k+1})} = \frac{1}{\mu_1 - \mu_2} - \frac{k+1}{\mu_1}\ \frac{\mu_2^{k+1}}{\mu_1^{k+1} - \mu_2^{k+1}}. \quad (26)$$

As for $T_K(k)$, we have similarly:

$$\begin{aligned} T_K(k) &= \sum_{n=0}^{k} T(n+1, k-n)\, \mathbb{P}(Q^*_{K,1} = n | Q^*_K = k) \\ &= (1 - \frac{\mu_2}{\mu_1})\ \frac{\mu_1^{k+1}}{\mu_1^{k+1} - \mu_2^{k+1}} \sum_{n=0}^{k} T(n+1, k-n) \left(\frac{\mu_2}{\mu_1}\right)^n \end{aligned} \quad (27)$$

where we recall that $T(n+1, m) = \mathbb{E}[S(n+1, m)]$ is the expected sojourn time of a customer that finds n customers in the first queue and m customers in the second queue. It can be calculated through the algorithm that we described in Section 4 about the fully observable case.

Substituting (26) and (27) in (22), we obtain that the individual net benefit of the tagged customer can be expressed in terms of the parameters of the system and the available information through the following equation:

$$\begin{aligned} P_{tot}(k) &= R - (C_1 - C_2) \left(\frac{1}{\mu_1 - \mu_2} - \frac{k+1}{\mu_1}\ \frac{\mu_2^{k+1}}{\mu_1^{k+1} - \mu_2^{k+1}} \right) \\ &\quad - C_2 \left((1 - \frac{\mu_2}{\mu_1})\ \frac{\mu_1^{k+1}}{\mu_1^{k+1} - \mu_2^{k+1}} \sum_{n=0}^{k} T(n+1, k-n) \left(\frac{\mu_2}{\mu_1}\right)^n \right) \end{aligned} \quad (28)$$

The tagged customer decides to enter in the system if and only if his benefit is non negative. More specifically, he will join the system for any information k about the total number of customers in the system that the above expression is non negative.

The expression (28) depends on the values of the function $T(n,m)$ and as such we cannot expect to be able to give a closed formula for the equilibrium threshold strategy. However using the algorithm developed in Section 4 we can always compute it numerically.

Note that the profit of the customer does not depend neither on the arrival rate nor on the strategy of the other customers (K does not appear since it is inside the constant c_K of the stationary distribution that cancels). This means that the strategy that we find using the function $P_{tot}(k)$ is a dominant strategy.

5.2 Number of customers at the first queue

In this section we study the case that an arriving customer is informed only about the number of customers at the first node. We assume that all customers follow the same threshold strategy N, that is they join as long as the number of customers in the first queue is less than N. As a consequence of this strategy, the number of customers in the first queue can not exceed this number N.

In this case the system substantially differs from the previous cases, as the tandem queues do not form anymore a Jackson network with finite or infinite capacity. To analyze the system we use the results in Kroese et al. (2004) that we briefly resume in the proposition below. We tag a customer that just arrives at the system and receives the information n, that is upon his arrival there are $n \leq N$ customers at the first node. We define by $Q^*_{N,i}$ the stationary random number of customers at node i, $i = 1, 2$, under the strategy N, and $Q^*_N = Q^*_{N,1} + Q^*_{N,2}$. Again we let $T_{N,i}(n) = \mathbb{E}[S_i | Q^*_{N,1} = n]$, and $T_N(n) = T_{N,1}(n) + T_{N,2}(n)$ the *row* vector $\boldsymbol{\pi}_N(m) = (\pi_N(0,m), \ldots, \pi_N(N,m))$ with $\pi_N(n,m) = \mathbb{P}(Q^*_{N,1} = n, Q^*_{N,2} = m)$.

Proposition 5 (in Kroese et al. (2004))**.** *The stationary distribution $\boldsymbol{\pi}_N(m)$ is given by*

$$\boldsymbol{\pi}_N(m) = \boldsymbol{\pi}_N(0)\, H^m \quad m \geq 0 \ , \tag{29}$$

where H is the minimal non-negative solution of the equation

$$A_0 + H\, A_1 + H^2\, A_2 = \mathbb{O} \ , \tag{30}$$

where $A_0 = \mu_1 U_R$, $A_1 - (\lambda + \mu_1 + \mu_2)\, \mathbb{I} + \mu_1\, U_0 + \lambda(U_N + U_L)$ and $A_2 = \mu_2\, \mathbb{I}$ are square matrices of size $(N+1)$.

The value of $\boldsymbol{\pi}_N(0)$ is given by

$$\boldsymbol{\pi}_N(0) = \frac{\boldsymbol{y}}{\boldsymbol{y}\, \boldsymbol{v}^t} \ ,$$

where $\boldsymbol{v}^t = (\mathbb{I} - H)^{-1}\, \mathbf{1}^t$ and $\boldsymbol{y}$ is a probability vector, satisfying the condition $\boldsymbol{y}\, \boldsymbol{v}^t < \infty$, that solves the equation

$$\boldsymbol{y}\, (A_1 + A_2 + H\, A_2) = \mathbf{0} \ .$$

We notice that H depends on the level N, but we decided to omit this dependence in the notation to keep it simpler.

Let $P_{1st,N}(n)$ be the expected net benefit of an entering customer that is going to take position n in the first queue after assuming that all other customers decide not to enter if the first queue has more than N customers in the system. It is given by

$$P_{1st,N}(n) = R - (C_1 - C_2)\, T_{N,1}(n) - C_2\, T_N(n) \tag{31}$$

We have that

$$T_{N,1}(n) = \frac{n}{\mu_1}$$

and

$$T_N(n) = \sum_{m=0}^{\infty} T(n,m)\,\mathbb{P}(Q^*_{N,2} = m | Q^*_{N,1} = n-1) = \sum_{m=0}^{\infty} T(n,m)\,\frac{\pi_N(n-1,m)}{\pi_N(n-1,\cdot)} \tag{32}$$

with $1 \le n \le N+1$ and $\pi_N(n,\cdot) = \mathbb{P}(Q^*_{N,1} = n) = \sum_m \pi_N(n,m)$.

As it shown in the formula above, the problem in computing $T_N(n)$ follows by the need of summing an infinite number of terms. To overcome this difficulty we introduce the partial generating function of $T(n,m)$, $\phi(n,z)$, defined as

$$\phi(n,z) = \sum_{m=0}^{\infty} T(n,m)\,z^m \ . \tag{33}$$

Substituting equation (29) in (32) we get

$$T_N(n) = \frac{\boldsymbol{\pi}_N(0)\,(\sum_{m=0}^{\infty} T(n,m)\,H^m)\,\boldsymbol{e}_{n-1}^{\mathtt{t}}}{\boldsymbol{\pi}_N(0)\,(\sum_{m=0}^{\infty} H^m)\,\boldsymbol{e}_{n-1}^{\mathtt{t}}} = \frac{\boldsymbol{\pi}_N(0)\,\phi(n,H)\,\boldsymbol{e}_{n-1}^{\mathtt{t}}}{\boldsymbol{\pi}_N(0)\,(\mathbb{I}-H)^{-1}\,\boldsymbol{e}_{n-1}^{\mathtt{t}}} \ . \tag{34}$$

We remind that given a scalar function $f(x)$, its evaluation at the square matrix point X is given by

$$f(X) = \sum_{k=0}^{\infty} f^{(k)}(0) X^k \ ,$$

see Gohberg et al. (1982). In particular if X is diagonal with the main diagonal given by the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ then $f(X) = \mathrm{diag}[f(x_i),\, 1 \le i \le n]$, and if the matrix X admits the diagonal form $X = A^{-1}\,D\,A$ then $f(X) = A^{-1} f(D) A$.

Finally, defining

$$\psi(n,z) = \sum_{m=0}^{\infty} y(n,m)\,z^m \ , \tag{35}$$

we get by (33) and (5) the following expression

$$\phi(n,z) = \sum_{m=0}^{\infty} y(n,m)\,\left(\frac{\mu_2\,z}{\mu_1+\mu_2}\right)^m + \frac{z + n\,(1-z)}{\mu_2\,(1-z)^2} = \psi\left(n, \frac{\mu_2\,z}{\mu_1+\mu_2}\right) + \frac{z+n\,(1-z)}{\mu_2\,(1-z)^2} \ , \tag{36}$$

and to be able to compute $T_N(n)$ in (34) we only are left with computing the function $\psi(n,z)$. We show how to do it in the following section.

5.2.1 Computing $\psi(n,z)$

Multiplying (6) by z^m and summing over $m \ge 0$ we get

$$\frac{1}{z}\sum_{m=1}^{\infty} y(n+1,m)\,z^m = \frac{\mu_2\,\mu_1}{z^2\,(\mu_1+\mu_2)^2}\sum_{m=2}^{\infty} y(n,m)\,z^m + \sum_{m=0}^{\infty} y(n+1,m)\,z^m \tag{37}$$

and adding and subtracting terms we have that

$$\frac{1}{z}\Big(\psi(n+1,z) - y(n+1,0)\Big) = \frac{\mu_2\,\mu_1}{z^2\,(\mu_1+\mu_2)^2}\Big(\psi(n,z) - y(n,0) - z\,y(n,1)\Big) + \psi(n+1,z) \ . \tag{38}$$

By rearranging terms we finally get

$$\psi(n+1,z) = \frac{\mu_2\,\mu_1}{(\mu_1+\mu_2)^2}\frac{\psi(n,z)}{z-z^2} - \frac{\mu_2\,\mu_1}{(\mu_1+\mu_2)^2}\frac{y(n,0)}{z-z^2} - \frac{\mu_2\,\mu_1}{(\mu_1+\mu_2)^2}\frac{y(n,1)}{1-z} + \frac{y(n+1,0)}{1-z} \ , \tag{39}$$

that allows to compute recursively $\psi(n,z)$ using only the values of $\psi(k,z)$ for $k < n$, starting with $\psi(0,z) = 0$, and the values of $y(k,0)$ and $y(k,1)$ for $k \leq n$ that have already been computed previously in the fully observable case.

5.3 Number of customers at the second queue

In this section we study the complementary case of the previous section, i.e. an arriving customer is informed only about the number of customers at the second node. We assume that all customers follow the same threshold strategy M, that is they join as long as the number of customers in the second queue is less than M. One main difference with the previous case, is that in this case the state is the same as the one for the open network, that is at a generic time we can find any number of customers at both nodes of the tandem networks.

Similarly to the notation introduced in the previous section, we define $Q^*_{M,i}$ as the stationary random number of customers at node i, $i = 1, 2$, under the strategy M, $Q^*_M = Q^*_{M,1} + Q^*_{M,2}$, $T_{M,i}(m) = \mathbb{E}[S_i | Q^*_{M,2} = m]$, and $T_M(m) = T_{M,1}(m) + T_{M,2}(m)$. In addition we define the probability *row* vector $\boldsymbol{\pi}_M(n) = (\pi_M(n,0), \ldots, \pi_M(n,M))$ with $\pi_M(n,m) = \mathbb{P}(Q^*_{M,1} = n, Q^*_{M,2} = m)$. The following result, proved in Leskelä and Resing (2007), gives how to determine the stationary distribution of the customers in the network given the strategy M.

Proposition 6 (in Leskelä and Resing (2007))**.** *The stationary distribution $\boldsymbol{\pi}_M(n)$ is given by*

$$\boldsymbol{\pi}_M(n) = \boldsymbol{\pi}_M(0)\, H^n \quad n \geq 0 \ , \tag{40}$$

where H is the minimal non-negative solution of

$$\sum_{n=0}^{\infty} H^n\, A_n = \mathbb{O} \ , \tag{41}$$

where the $(M+1)$-square matrices A_n are given by $A_0 = \lambda \mathbb{I}$, $A_1 = \mu_2\, U_L - (\lambda + \mu 1 + \mu 2)\mathbb{I} + \mu_2\, U_0$, $A_2 = \mu_1\, (U_R + \kappa_1\, U_M)$ and $A_{n+1} = \mu_1\, \kappa_n\, U_M$, with $\kappa_n = 1/(n+1)\, \binom{2n}{n} (\mu_1/(\mu_1+\mu_2))^{n-1} (\mu_2/(\mu_1+\mu_2))^n$.

The value of $\boldsymbol{\pi}_M(0)$ is given by

$$\boldsymbol{\pi}_M(0) = \frac{\mu_2}{\lambda + \mu_2\, \boldsymbol{x}\, (\mathbb{I} - H)^{-1}\, \boldsymbol{e}_0^t}\, \boldsymbol{x} \ ,$$

where the row vector $\boldsymbol{x}$ is the unique positive solution of

$$\boldsymbol{x} \sum_{n=0}^{\infty} H^n\, B_n = \mathbb{O} \ ,$$

satisfying $\boldsymbol{x}\, \boldsymbol{v}^t = 1$ with $\boldsymbol{v}^t = (\mathbb{I} - H)^{-1}\, \mathbf{1}^t$. The $(M+1)$-square matrices B_n are given by $B_0 = \mu_2\, U_L - (\lambda + \mu 2)\mathbb{I} + \mu_2\, U_0$, $B_1 = \mu_1\, (U_R + U_M)$ and $B_{n+1} = \mu_1\, (1 - \kappa_1 - \cdots - \kappa_n)\, U_M$.

Again to keep notation simpler we omit the dependence of H to the strategy level M.

We denote by $P_{2nd,M}(m)$ the profit of an entering customer that is going to enter the first queue knowing that in the second queue there are m customers and assuming that all other customers decide not to enter if the second queue has more than M customers in the system. It follows that

$$P_{2nd,M}(m) = R - (C_1 - C_2)\, T_{M,1}(m) - C_2\, T_M(m) \tag{42}$$

We have that

$$T_{M,1}(m) = \sum_{n=0}^{\infty} T_1(n+1,m)\,\mathbb{P}(Q^*_{M,1}=n|Q^*_{M,2}=m) = \sum_{n=0}^{\infty} \frac{n+1}{\mu_1}\,\frac{\pi_M(n,m)}{\pi_M(\cdot,m)}$$

and

$$T_M(m) = \sum_{n=0}^{\infty} T(n+1,m)\,\mathbb{P}(Q^*_{M,1}=n|Q^*_{M,2}=m) = \sum_{n=0}^{\infty} T(n+1,m)\,\frac{\pi_M(n,m)}{\pi_M(\cdot,m)} \tag{43}$$

valid for any $m \geq 0$, with $\pi_M(\cdot,m) = \mathbb{P}(Q^*_{M,2}=m) = \sum_n \pi_M(n,m)$.

Defining

$$\phi(z,m) = \sum_{n=0}^{\infty} T(n+1,m)\,z^n \ . \tag{44}$$

after substituting equation (40) in (43) we get

$$T_{M,1}(m) = \frac{1}{\mu_1}\frac{\boldsymbol{\pi}_M(0)\,\sum_{n=0}^{\infty}(n+1)\,H^n\,\boldsymbol{e}_m^{\mathtt{t}}}{\boldsymbol{\pi}_M(0)\,\sum_{n=0}^{\infty}H^n\,\boldsymbol{e}_m^{\mathtt{t}}} = \frac{1}{\mu_1}\frac{\boldsymbol{\pi}_M(0)\,(\mathbb{I}-H)^{-2}\,\boldsymbol{e}_m^{\mathtt{t}}}{\boldsymbol{\pi}_M(0)\,(\mathbb{I}-H)^{-1}\,\boldsymbol{e}_m^{\mathtt{t}}} \tag{45}$$

$$T_M(m) = \frac{\boldsymbol{\pi}_M(0)\,\sum_{n=0}^{\infty}T(n+1,m)\,H^n\,\boldsymbol{e}_m^{\mathtt{t}}}{\boldsymbol{\pi}_M(0)\,\sum_{n=0}^{\infty}H^n\,\boldsymbol{e}_m^{\mathtt{t}}} = \frac{\boldsymbol{\pi}_M(0)\,\phi(H,m)\,\boldsymbol{e}_m^{\mathtt{t}}}{\boldsymbol{\pi}_M(0)\,(\mathbb{I}-H)^{-1}\,\boldsymbol{e}_m^{\mathtt{t}}} \tag{46}$$

Finally, defining

$$\psi(z,m) = \sum_{n=0}^{\infty} y(n+1,m)\,z^n \ , \tag{47}$$

from (44) and (5) we get

$$\begin{aligned}\phi(z,m) &= \left(\frac{\mu_2}{\mu_1+\mu_2}\right)^m \sum_{n=0}^{\infty} y(n+1,m)\,z^n + \frac{1+m\,(1-z)}{\mu_2\,(1-z)^2} \\ &= \left(\frac{\mu_2}{\mu_1+\mu_2}\right)^m \psi(z,m) + \frac{1+m\,(1-z)}{\mu_2\,(1-z)^2}\end{aligned} \tag{48}$$

and to be able to compute $T_{M,1}(m)$ and $T_M(n)$ in (45) and (46) we are only left with computing the function $\psi(z,m)$. We show how to do it in the following section.

5.3.1 Computing $\psi(z,m)$

Multiplying (6) by z^n and summing over $n \geq 0$ we get

$$\sum_{n=0}^{\infty} y(n+1,m+1)\,z^n = \frac{\mu_2\mu_1\,z}{(\mu_1+\mu_2)^2}\sum_{n=0}^{\infty} y(n+1,m+2)\,z^n + \sum_{n=0}^{\infty} y(n+1,m)\,z^n \tag{49}$$

where in the second series we used (8). It follows that

$$\psi(z,m+1) = \frac{\mu_2\mu_1\,z}{(\mu_1+\mu_2)^2}\psi(z,m+2) + \psi(z,m) \quad m \geq 0 \ . \tag{50}$$

Doing the same with equation (7) we have that

$$\psi(z,0) = \frac{1}{\mu_1\,(1-z)} + \frac{\mu_2\,z}{\mu_1+\mu_2}\psi(z,1) \ . \tag{51}$$

Lemma 7. *The following inequalities hold*

$$0 \leq y(n,m) \leq \frac{n}{\mu_1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{-m} \tag{52}$$

for n, $m \geq 0$.

Proof. The lower bound is an immediate consequence of equation (9), so we are left with proving the upper bound. Note that

$$T(n,m) \leq n\left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right)+\frac{m}{\mu_2}\ .$$

Hence using (5),

$$y(n,m)\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{m}+\frac{n+m}{\mu_2} \leq \frac{n}{\mu_1}+\frac{n+m}{\mu_2}\ ,$$

we get that the inequality holds. □

Let $a_z(m) = \psi(z,m)$ and $c_z = z\,(\mu_1\,\mu_2)/(\mu_1+\mu_2)^2$, equation (50) can be rewritten as

$$c_z\,a_z(m+2) - a_z(m+1) + a_z(m) = 0 \quad m \geq 0, \tag{53}$$

that is a homogeneous difference equation whose solutions are given by

$$a_z(m) = c_{z,-}\,a_{z,-}^{-m} + c_{z,+}\,a_{z,+}^{-m} \quad m \geq 0, \tag{54}$$

with $c_{z,\pm}$ are constants with respect to m and $a_{z,\pm}$ are solutions of the following second order equation

$$a_z^2 - a_z + c_z = 0\ , \tag{55}$$

that are equal to $a_{z,\pm} = 1/2 \pm \sqrt{1/4 - c_z}$. Note that for $0 \leq z \leq 1$, we have that $0 \leq c_z \leq 1/4$ and therefore the two roots are both real, $0 \leq a_{z,-} \leq 1/2$ and $1/2 \leq a_{z,+} \leq 1$.

Lemma 8. *The constant* $c_{z,-}$ *is equal to zero.*

Proof. We have that $\psi(z,m) \geq 0$ for $m \geq 0$ and $0 \leq z < 1$, therefore

$$c_{z,-}\,a_{z,-}^{-m} + c_{z,+}\,a_{z,+}^{-m} \geq 0\ .$$

Dividing by $a_{z,-}^{-m}$ we get

$$c_{z,-} + c_{z,+}\left(\frac{a_{z,+}}{a_{z,-}}\right)^{-m} \geq 0\ ,$$

where $a_{z,+}/a_{z,-} > 1$, if $z > 0$. Taking the limit as $m \to \infty$ we get that $c_{z,-} \geq 0$.

Having that $\mu_1, \mu_2 > 0$ and $0 < z < 1$, it is easy to prove that

$$a_{z,-} < b < a_{z,+}\ ,$$

where $b = \mu_2/(\mu_1+\mu_2)$. Using equation (47) and the inequalities (52) we get

$$0 \leq \psi(z,m) \leq \frac{1}{\mu_1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{-m}\frac{1}{(1-z)^2}$$

for n, $m \geq 0$ and $0 \leq z < 1$, that is equivalent to

$$c_{z,-}\,a_{z,-}^{-m} + c_{z,+}\,a_{z,+}^{-m} \leq k_z\,b^{-m}\ ,$$

where $1/k_z = \mu_1 (1-z)^2$, and dividing both sides by b^{-m} we get

$$c_{z,-} \left(\frac{b}{a_{z,-}}\right)^m + c_{z,+} \left(\frac{b}{a_{z,+}}\right)^m \leq k_z \ .$$

Taking the limit as $m \to \infty$ the left side is unbounded if $c_{z,-} > 0$ and therefore the result holds true. □

Theorem 9. *The function $\psi(z,m)$ has the following expression*

$$\psi(z,m) = \frac{1}{\mu_1 (1-z)} \left(1 - \frac{\mu_2}{\mu_1+\mu_2} \frac{z}{a(z)}\right)^{-1} a^{-m}(z) \tag{56}$$

where $a(z) = 1/2 + \sqrt{1/4 - z\,(\mu_1\,\mu_2)/(\mu_1+\mu_2)^2}$.

Proof. Using the results of Lemma 8 we can write the expression of $\psi(z,m)$ in the following way

$$\psi(z,m) = c(z)\, a^{-m}(z)$$

and then using (51) we get

$$c(z) = \frac{1}{\mu_1 (1-z)} \left(1 - \frac{\mu_2}{\mu_1+\mu_2} \frac{z}{a(z)}\right)^{-1} \ .$$

□

Acknowledgements

The first author is partially supported by the Spanish Ministry of Education and Science Grants MTM2010-16519, SEJ2007-64500 and RYC-2009-04671.

References

M. Armony and M. Haviv (2003): *Price and delay competition between two service providers*. *European J.Oper.Res.* 147:32–50.

A. Burnetas (2011): *Customer Equilibrium and Optimal Strategies in Markovian Queues in Series*. *Annals of Operations Research* to appear.

A. Burnetas and A. Economou (2007): *Equilibrium customer strategies in a single server Markovian queue with setup times*. *Queueing Syst.* 56:213–228.

C. Courcoubetis and R. Weber (2003): *Pricing communication networks, Economics, Technology and Modelling*. Wiley, England.

A. Economou, A. Gómez-Corral and S. Kanta (2011): *Optimal balking strategies in single-server queues with general service and vacation times*. *Performance Evaluation* 68:967–982.

A. Economou and S. Kanta (2008): *Optimal balking strategies and pricing for the single server Markovian queue with compartmented waiting space*. *Queueing Syst.* 59:237–269.

N. M. Edelson and K. Hildebrand (1975): *Congestion tolls for Poisson queueing processes*. *Econometrica* 43:81–92.

I. Gohberg, P. Lancaster and L. Rodman (1982): *Matrix Polynomials*. Academic Press, New York.

P. Guo and P. Zipkin (2007): *Analysis and comparison of queues with different levels of delay information*. *Managm. Sci.* 53:962–970.

R. Hassin (2007): *Information and uncertainty in a queuing system*. *Prob.Eng.Inf.Sci.* 21:361–380.

R. Hassin and M. Haviv (2003): *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer, Boston.

D. Kroese, W. Scheinhardt and P. Taylor (2004): *Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process*. *The Annals of Applied Probability* 14.

L. LeskelÄ and J. Resing (2007): A tandem queueing network with feedback admission control. In *Network Control and Optimization - Proceedings of the First EuroFGI International Conference (NET-COOP)*, T. Chahed and B. Tuffin, editors, volume 4465 of *Lecture Notes in Computer Science*. Springer-Verlag, pp. 129–137.

K. Lin and S. Ross (2001): *Admission control with incomplete information of a queueing system*. *Oper. Res.* 51:645–654.

A. Mandelbaum and N. Shimkin (2000): *A model for rational abandonments from invisible queues*. *Queueing Syst.* 36:141–173.

S. Martin and P. Smith (1999): *Rationing by waiting lists: an empirical investigation*. *Journal of Public Economics* 71:141–164.

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

I. Parra-Frutos and J. Aranda-Gallego (1999): *Multiproduct monopoly: a queueing approach*. *Applied Economics* 31:565–576.

A. D. Vany (1976): *Uncertainty, waiting time and capacity utilization: a stochastic theory of product quality*. *Journal of Political Economy* 84:523–541.

U. Yechiali (1971): *On optimal balking rules and toll charges in the GI/M/1 queue*. *Oper. Res.* 19:349–370.

U. Yechiali (1972): *Customers' optimal joining rules and toll charges in the GI/M/s queue*. *Managm. Sci.* 18:434–443.

C. D. Zimmerman and J. Enell (1993): *Empresas de servicios*, volume II of *Manual de Control de Calidad*. McGraw–Hill / Interamericana de España, 4th edition.