4

The Impact of Information Structure on Strategic Behavior in Queueing Systems

Antonis ECONOMOU
Department of Mathematics, National and Kapodistrian University of Athens, Greece

The study of strategic customer behavior in service systems constitutes an important and dynamic trend in queueing theory. Indeed, the optimal design and control of service systems in real-life applications of queueing theory requires that the strategic dimension of customers is taken into account. Under this perspective, the customers are decision makers that aim to maximize their benefit, taking into account that the others have similar objectives. Therefore, the overall situation can be seen as a game among the customers and the administrator of the system. A central problem is how a social planner or a monopolist should act to incite customers to adopt a desirable behavior, one that increases the social welfare or the monopolist's revenue/profit, respectively. However, the intervention of a social planner and/or a monopolist should be indirect, as direct forcing of customers is considered unacceptable in the framework of a free market. To this end, several mechanisms have been proposed, e.g. pricing structures, priority systems and non-standard queueing disciplines. An important mechanism is the control of information that is provided to the customers. In this chapter, we will present several techniques for the control of information in a given system and their impact on strategic customer behavior, the throughput, the social welfare and a monopolist's revenue. These ideas will be explained in the simplest possible framework and then several extensions will be discussed. An overview of the corresponding literature is also included.

Queueing Theory 2,
coordinated by Vladimir ANISIMOV, Nikolaos LIMNIOS. © ISTE Editions 2020.

Queueing Theory 2: Advanced Trends,
First Edition. Vladimir Anisimov and Nikolaos Limnios.
© ISTE Ltd 2020. Published by ISTE Ltd and John Wiley & Sons, Inc.

4.1. Introduction

Classical queueing theory assumes that customers are passive entities that do not make decisions. Under this perspective, a very large number of queueing theoretic studies have been presented in the literature that deal with the performance evaluation, optimal design and optimal control of service systems that appear in a number of situations. However, an economic evaluation of a queueing system is not credible, unless we take into account the strategic nature of its customers. This point of view was assumed in the seminal paper of (Naor 1969), who studied the join-or-balk dilemma for the customers in the M/M/1 queue, when its queue length is observable. He also considered the problem of a social planner and a monopolist who optimize the social welfare and the revenue, respectively, taking into account customer strategic behavior. Edelson and Hildebrand (1975) complemented Naor's study by considering the same problems for the unobservable version of the system. In their model, the arriving customers are not allowed to observe the number of customers in the system and make their join-or-balk decisions relying solely on the knowledge of its operational and economic parameters, assuming that the system has reached stochastic steady-state. Since then, the literature on strategic behavior in queueing systems has grown considerably. Hassin and Haviv (2003) provided a nice overview of the basic methodology and the early results in this area. The monographs by (Stidham 2009) and (Hassin 2016) contain further material and overviews about models and methodologies in this subfield of queueing theory.

One recurrent theme in the strategic queueing literature is the appropriate use of mechanisms that will incentivize customers to behave according to a social planner's or a monopolist's objective. One such mechanism is the control of information that is provided to the customers before their decisions are made. The study of the effect of the level of information on the strategic customer behavior is an important theoretical issue per se, but it is also important because it raises several issues of practical relevance. The information provision mechanism is related to the design of the system, affects the psychology of the customers and may be costly. What is interesting is that the effect of the level of information is ambiguous. More information can benefit or hurt the customers and/or the service provider. The effect of more information may be negative or positive, depending on various parameters and structural assumptions of the underlying model. It should be emphasized here that the strategic interaction of the customers is the root cause of the phenomenon that the effect of more information might be negative, unlike simple optimization. The objective of this chapter is exactly to shed some light on these issues. Because of the complexity of the problem, we will unfold the presentation in the simplest possible framework, that is of the M/M/1 queue. Indeed, this is also the framework that most of the authors have adopted in the literature. In the last part of the chapter, we also discuss more involved models.

The rest of the chapter is organized as follows. In section 4.2, we present some basic concepts from game theory that will be used in the sequel. In sections 4.3 and 4.4, we present the analysis of the two "extreme" models regarding the information that is provided to the customers: the unobservable model of (Edelson and Hildebrand 1975) and the observable model of (Naor 1969). Subsequently, in section 4.5, we present briefly the main findings from the comparison of these models. Then, we describe three significant ideas that have appeared in the literature that bridge the observable and unobservable versions of a given model. More specifically, in sections 4.6–4.8, we discuss, respectively, partially observable models, heterogeneously observable models and observable-with-delay models. Apart from the descriptions of the main ideas behind each family of models, we present the basic results, the associated methodologies and the main qualitative findings. Section 4.9 is devoted to an overview of the various ideas that appear within the chapter. Moreover, it points to several important sources in the literature, some conclusions and topics for future research.

4.2. Game-theoretical framework in queueing

In the framework of classical game theory, a game is specified by a set of players $\mathcal{N} = \{1, 2, \ldots, n\}$, sets of action plans $\mathcal{A}_i$, one for each player $i = 1, 2, \ldots, n$, and payoff (utility) functions $\mathcal{U}_i$, one for each player $i = 1, 2, \ldots, n$. The set $\mathcal{A}_i$ contains all available action plans for player i that specify what actions should be taken during the game, according to its state at every decision point. Every element of $\mathcal{A}_i$ is referred to as a pure strategy of i. A probability distribution on $\mathcal{A}_i$ is referred to as a mixed strategy of i. When player i uses a certain mixed strategy, the player chooses one of his/her pure strategies according to the probability distribution of the mixed strategy. The set of mixed strategies for of the player i is denoted by $\mathcal{S}_i$.

A strategy profile s is an ordered n-tuple of strategies, one for each player, i.e.

$$\mathbf{s} = (s_1, s_2, \ldots, s_n), \quad s_i \in \mathcal{S}_i, \ i = 1, 2, \ldots, n.$$

Given a strategy profile $\mathbf{s}$, we write $\mathbf{s} = (s_i, \mathbf{s}_{-i})$, to denote that the $(n-1)$-dimensional vector $\mathbf{s}_{-i}$ contains the strategies of $\mathbf{s}$ except the one that corresponds to player i. The payoff function $\mathcal{U}_i$ is defined on the set of all strategy profiles and takes real values. Its value $\mathcal{U}_i(\mathbf{s}) = \mathcal{U}_i(s_i, \mathbf{s}_{-i})$ corresponds to the payoff of player i, if the strategy profile $\mathbf{s}$ is adopted, i.e. if player i uses his/her strategy s_i and the other players the strategies in $\mathbf{s}_{-i}$. The function $\mathcal{U}_i(\mathbf{s}) = \mathcal{U}_i(s_i, \mathbf{s}_{-i})$ is linear with respect to s_i, i.e. if the strategy s_i mixes the strategies s_i^k, $k = 1, 2, \ldots, r$ with probabilities α_k, $k = 1, 2, \ldots, r$, respectively (with $\sum_{k=1}^r \alpha_k = 1$), then

$$\mathcal{U}_i(s_i, \mathbf{s}_{-i}) = \sum_{k=1}^{r} \alpha_k \mathcal{U}_i(s_i^k, \mathbf{s}_{-i}).$$

If s_i^1 and s_i^2 are strategies of player i, then we say that s_i^1 weakly dominates s_i^2, if for any strategy profile for the other players, $\mathbf{s}_{-i}$, then $\mathcal{U}_i(s_i^1, \mathbf{s}_{-i}) \geq \mathcal{U}_i(s_i^2, \mathbf{s}_{-i})$, with the inequality being strict for at least one strategy profile $\mathbf{s}_{-i}$. We say that s_i^1 strongly dominates s_i^2, if the inequality is strict for all strategy profiles $\mathbf{s}_{-i}$.

Consider, now, a player i. Given a strategy profile $\mathbf{s}_{-i}$ for the other players, the strategy s_i^* of i is said to be the best response against $\mathbf{s}_{-i}$, if for every strategy s_i of i we have $\mathcal{U}_i(s_i^*, \mathbf{s}_{-i}) \geq \mathcal{U}_i(s_i, \mathbf{s}_{-i})$, i.e. if s_i^* maximizes $f(s_i) = \mathcal{U}_i(s_i, \mathbf{s}_{-i})$. The set of best responses to $\mathbf{s}_{-i}$ is denoted by $BR(\mathbf{s}_{-i})$.

A strategy s_i of i is said to be weakly (respectively, strongly) dominant, if it weakly (respectively, strongly) dominates every other strategy of i.

A strategy profile $\mathbf{s}^e = (s_1^e, s_2^e, \ldots, s_n^e)$ is said to be an equilibrium profile, if for every player i, the strategy s_i^e is the best response against $\mathbf{s}_{-i}^e$. In other words, a strategy profile is an equilibrium if no player has an incentive to deviate from it unilaterally.

In the study of strategic customer behavior in queueing systems, the game-theoretical concepts that we described above are very useful. However, they cannot be applied immediately, because there are two fundamental problems. The first is the fact that the number of players is infinite, since the potential customers of the system are infinite. The second is that the customers–players do not make simultaneously their decisions since they arrive sequentially during an infinite time horizon. These problems are bypassed by defining analogous concepts and exploiting the homogeneity of the customers. For simplicity, we will assume that all customers are identical. However, the framework can be extended to allow heterogeneous customers, assuming that there are various classes with homogeneous customers within each class.

In the case of homogeneous customers, a game among them is specified by the set of their common strategies, $\mathcal{S}$, and from a payoff function $\mathcal{U}(s, s')$ that specifies the utility of a customer that uses strategy s, when all other customers follow s'. The function $\mathcal{U}(s, s')$ is linear in s, i.e. if a strategy s is the mixture of strategies s^k, $k = 1, 2, \ldots, r$ with corresponding probabilities α_k, $k = 1, 2, \ldots, r$, with $\sum_{k=1}^r \alpha_k = 1$, then

$$\mathcal{U}(s, s') = \sum_{k=1}^{r} \alpha_k \mathcal{U}(s^k, s').$$

If s^1 and s^2 are strategies of a player, then the strategy s^1 weakly dominates s^2, if for any strategy s' of the other players $\mathcal{U}(s^1, s') \geq \mathcal{U}(s^2, s')$. Moreover, we say that s^1 strongly dominates s^2, if the inequality is strict for all strategies s'.

Consider, now, a tagged customer and assume that the strategy s' is followed by the other customers. A strategy s^* of the tagged customer is said to be the best response against s', if for any strategy s of the tagged customer $\mathcal{U}(s^*, s') \geq \mathcal{U}(s, s')$, i.e. s^* maximizes $f(s) = \mathcal{U}(s, s')$. The set of best responses against s' is denoted by $BR(s')$.

A strategy s of a tagged player is said to be dominant if s is best response against any strategy of the others. A strategy s^e is said to be a (symmetric) equilibrium, if it is the best response against itself. In other words, the strategy s^e is equilibrium if

$$\mathcal{U}(s^e, s^e) \geq \mathcal{U}(s, s^e), \ s \in \mathcal{S},$$

or equivalently if $s^e \in BR(s^e)$. It should be noted here that because of the linearity of the payoff function with respect to its first argument, the indifference principle of a mixed equilibrium strategy holds: If an equilibrium mixed strategy assigns positive probability to some pure strategies, then all these pure strategies ensure the same payoff to a tagged customer given that the population of customers follow the equilibrium strategy.

A basic step for the study of strategic customer behavior concerns the computation of the payoff function $\mathcal{U}(s, s')$. The fundamental assumption for this computation is that if we consider a tagged customer who follows a strategy s, when all others follow a strategy s', then the tagged customer's strategy does not influence the performance measures of the system. The general behavior of the system and the corresponding performance measures are determined by the strategy s' that the other customers follow, since the impact of the tagged customer is negligible. Moreover, it is assumed that the system has reached stochastic steady-state. To determine the dominant and equilibrium customer strategies in a queueing system, a general methodology is applied using the following steps:

Step 1: the steady-state behavior of the system under an arbitrary strategy s' of the population of the customers is studied.

Step 2: the utility function $\mathcal{U}(s, s')$ of a tagged customer that follows strategy s, when all other customers follow strategy s', is computed.

Step 3: the best response of the tagged customer against an arbitrary strategy, s', of the population of the customers is computed:

$$BR(s') = \{s \in \mathcal{S} : \mathcal{U}(s, s') \geq \mathcal{U}(\hat{s}, s'), \hat{s} \in \mathcal{S}\}.$$

Step 4: all strategies with the property $s^e \in BR(s^e)$ are identified. These are exactly the equilibrium strategies. If an equilibrium strategy s^e satisfies the stronger condition $s^e \in BR(s)$, for all $s \in \mathcal{S}$, then it is a dominant strategy.

A related problem from a social planner's point of view is the maximization of the social welfare per customer given that a symmetric strategy s is followed by the population of customers. This is defined to be the quantity $\mathcal{U}(s,s)$. Then, a socially optimal symmetric strategy s_{soc} is such that $\mathcal{U}(s_{soc},s_{soc}) \geq \mathcal{U}(s,s)$, for $s \in \mathcal{S}$. To determine such strategies we must solve the optimization problem $\max_{s\in\mathcal{S}} \mathcal{U}(s,s)$.

4.3. The unobservable model

We consider an M/M/1 queue, where customers arrive according to a Poisson process at rate λ and the service times are exponentially distributed with rate μ. Each customer receives mean reward R for his/her service completion, whereas he/she accumulates waiting costs at rate C, during his/her sojourn time in the system (assuming that the cost is accumulated with the same rate whether the customer stays in the waiting space or receives service). The dilemma of the customers is whether to join or balk. In this section, we consider the unobservable version of the model, i.e. we assume that the customers make their join-or-balk decisions, without observing the queue length in the system. However, the various operational and economic parameters of the system, λ, μ, R and C are common knowledge to the customers.

In this case, the pure strategies of a customer are two: Join (1) or balk (0). A mixed strategy is specified by a join probability $q \in [0,1]$. Edelson and Hildebrand (1975) studied the equilibrium customer strategies for the unobservable M/M/1 queue regarding the join-or-balk dilemma and the associated social optimization and revenue maximization problems of the administrator of the system. We summarize their main findings, using the 4-step methodology that applies as follows: Suppose that the population of the customers follow a join probability q. Then, because of the thinning property of the Poisson process, the system becomes an M/M/1 queue with arrival rate λq and service rate μ; hence the mean sojourn time of a customer who decides to join is $\frac{1}{\mu-\lambda q}$ as long as $\lambda q < \mu$ and 0 otherwise (see, e.g. (Hassin and Haviv 2003), section 1.4). Consider, now, a tagged customer who joins with probability q', when the others join with probability q. Then, his/her expected utility is

$$\mathcal{U}(q',q) = (1-q')\cdot 0 + q'\left(R - \frac{C}{\mu-\lambda q}\right).$$

Therefore, to find his/her best response against q, the tagged customer has to solve the problem $\max_{q'\in[0,1]} \mathcal{U}(q',q)$. However, the function $\mathcal{U}(q',q)$ is linear with respect to q', so the tagged customer bases his/her decision on the sign of the quantity

$$S_{ind}^{(un)}(q) = R - \frac{C}{\mu-\lambda q}.$$

The superscript "(un)" in the notation of the function $S_{ind}^{(un)}(q)$ indicates that we refer to the unobservable model, whereas the subscript "ind" refers to the individualistic point of view that we consider here. Similar notational conventions will be used in the sequel. Let

$$\bar{q}_e = \frac{1}{\lambda}\left(\mu - \frac{C}{R}\right).$$

be the root of $S_{ind}^{(un)}(q)$. Then, the set of best responses against q, $BR(q)$, is (as far as $q \in [0,1]$ and $\lambda q < \mu$)

$$BR(q) = \begin{cases} \{0\}, & \text{if } q > \bar{q}_e, \\ [0,1], & \text{if } q = \bar{q}_e, \\ \{1\}, & \text{if } q < \bar{q}_e. \end{cases}$$

We can now proceed to the computation of the equilibrium strategies:

The strategy of "always balk" ($q_e = 0$) is equilibrium strategy, if and only if $0 \in BR(0)$, i.e. $0 \geq \bar{q}_e$, which reduces to $R \leq \frac{C}{\mu}$.

A strategy $q_e \in (0,1)$ is equilibrium strategy, if and only if $q_e \in BR(q_e)$, i.e. $q_e = \bar{q}_e$, which reduces to $q_e = \frac{1}{\lambda}\left(\mu - \frac{C}{R}\right)$. This is valid as far as $\bar{q}_e \in (0,1)$, which occurs if and only if $\frac{C}{\mu} < R < \frac{C}{\mu-\lambda}$.

Finally, the "always join" ($q_e = 1$) is equilibrium strategy, if and only if $1 \in BR(1)$, i.e. $1 \leq \bar{q}_e$, which reduces to $R \geq \frac{C}{\mu-\lambda}$.

In summary, we have the following result:

THEOREM 4.1.– For the join-or-balk customer dilemma in the unobservable M/M/1 queue, a unique equilibrium strategy q_e exists, given by the formula

$$q_e = \begin{cases} 0, & R \leq \frac{C}{\mu}, \\ \frac{1}{\lambda}\left(\mu - \frac{C}{R}\right), & \frac{C}{\mu} < R < \frac{C}{\mu-\lambda}, \\ 1, & R \geq \frac{C}{\mu-\lambda}. \end{cases}$$

Therefore, unless $R > \frac{C}{\mu-\lambda}$, the social welfare under the equilibrium strategy is 0. We now consider the problem of the system administrator that can act as a social planner who wants to induce a socially optimal strategy, q_{soc}, to maximize the social welfare per time unit. To this end, the administrator of the system can impose an admission fee (entrance or service price) p that will change the service reward from

R to $R - p$. Note that p may be even negative, corresponding to a subsidy for service. Using this price mechanism, the administrator may induce whatever join probability, q, he desires. Then, the system behaves as an M/M/1 queue with arrival rate λq and service rate μ. The revenue of the administrator per time unit will be λqp, whereas the total customer utility will be $\lambda q\left(R - p - \frac{C}{\mu-\lambda q}\right)$. Therefore, the social welfare per time unit is given by

$$S_{soc}^{(un)}(q) = \lambda q\left(R - p - \frac{C}{\mu - \lambda q}\right) + \lambda qp = \lambda q\left(R - \frac{C}{\mu - \lambda q}\right),$$

which is independent of p (since the transfer payments do not appear in the social welfare function). The subscript "soc" in the notation of the function $S_{soc}^{(un)}(q)$ shows that we adopt here the point of view of a social planner.

Note, now, that

$$\begin{aligned}\frac{d}{dq}S_{soc}^{(un)}(q) &= \lambda\left(R - \frac{C\mu}{(\mu - \lambda q)^2}\right),\\ \frac{d^2}{dq^2}S_{soc}^{(un)}(q) &= -\frac{2C\mu\lambda^2}{(\mu - \lambda q)^3} < 0.\end{aligned}$$

Therefore, the function $S_{soc}^{(un)}(q)$ is concave for $q \in [0, \frac{\mu}{\lambda})$ and attains its maximum at the root $\bar{q}_{soc}$ of $\frac{d}{dq}S_{soc}^{(un)}(q)$, which is given by the formula

$$\bar{q}_{soc} = \frac{1}{\lambda}\left(\mu - \sqrt{\frac{C\mu}{R}}\right). \qquad [4.1]$$

When $\bar{q}_{soc} \in [0, 1]$, we deduce that the socially optimal strategy (i.e. admission probability) is given by the formula [4.1], otherwise the maximum of the social welfare is attained at 0 or 1. More concretely, we have the following result:

THEOREM 4.2.– For the social planner's admission problem in the unobservable M/M/1 queue, a unique socially optimal strategy exists, given by the formula

$$q_{soc} = \begin{cases} 0, & R \leq \frac{C}{\mu}, \\ \frac{1}{\lambda}\left(\mu - \sqrt{\frac{C\mu}{R}}\right), & \frac{C}{\mu} < R < \frac{C\mu}{(\mu-\lambda)^2}, \\ 1, & R \geq \frac{C\mu}{(\mu-\lambda)^2}. \end{cases}$$

We can easily see that $q_{soc} \leq q_e$, for all parameter values λ, μ, R and C. Indeed, for a mixed strategy $q \in (0,1)$, the individual condition of optimality is $R = \frac{C}{\mu-\lambda q}$, whereas the first-order condition for social optimality is $R = \frac{C\mu}{(\mu-\lambda q)^2}$. The quadratic term in the latter appears because of the negative external effects of joining (for details, see the excellent paper of (Haviv and Oz 2018)). Therefore, we observe that, without an admission fee, the customers tend to use the system more than what is socially desirable.

We now consider the monopolist's pricing problem, when the objective of the system administrator is the maximization of the revenue per time unit. By imposing an admission fee, p, the customers adopt the corresponding equilibrium strategy q_e that we described above that corresponds to the service reward $R - p$. For inducing join probability $q \in (0,1)$, the administrator should impose an admission fee $p = R - \frac{C}{\mu-\lambda q}$. For inducing $q = 1$, he should impose the maximum possible price that allow all customers to join, which is $p = R - \frac{C}{\mu-\lambda}$. When he induces join probability $q = 0$, by imposing a large admission fee, his profit will be 0. Therefore, we see that in any case, the function of the revenue per time unit is

$$S_{prof}^{(un)}(q) = \lambda q p = \lambda q \left(R - \frac{C}{\mu - \lambda q} \right),$$

i.e. it is identical to the function of the social welfare per time unit.

We conclude that the optimal join probability for the system administrator when he acts as a monopolist is the socially optimal join probability. This occurs because of symmetric information (i.e. all information known to the customers is also known to the monopolist) and customer homogeneity that allow the monopolist to extract all utility surplus from the customers. Then, the objective functions of the social welfare and the monopolist's revenue coincide. Therefore, the administrator of the system imposes an admission fee

$$p_{prof} = R - \frac{C}{\mu - \lambda q_{soc}}$$

which maximizes the social welfare, and collects all of this revenue for himself. Substituting the socially optimal join probability into the formula for p_{prof} yields

$$p_{prof} = \begin{cases} R - \frac{C}{\mu-\lambda}, & \text{if } \lambda < \mu - \sqrt{\frac{C\mu}{R}} \\ R - \sqrt{\frac{RC}{\mu}}, & \text{if } \lambda \geq \mu - \sqrt{\frac{C\mu}{R}}, \end{cases}$$

which is a decreasing and ultimately constant function of λ. This may seem a paradox at first glance, since λ can be interpreted as the demand for service and therefore we expect that an increase in the demand will make the monopolist raise the price. However, an increase in the demand induces a significant decrease in the “quality” of the service, because of increasing delays. Therefore, the customers become more reluctant to buy the service and the monopolist cannot increase the price.

We further notice that pricing is not the only way for the regulation of the unobservable M/M/1 queue. Indeed, (Haviv and Oz 2018) describe eight different regulation schemes for this model.

4.4. The observable model

Consider, now, the join-or-balk customer dilemma in the M/M/1 queue with the same operational and economic parameters as in section 4.3, assuming that each customer makes his/her decision after observing the number of present customers upon arrival. In this case, a mixed customer strategy is specified by a sequence $\mathbf{q} = (q_0, q_1, q_2, \ldots)$, where $q_n \in [0, 1]$, $n = 0, 1, 2, \ldots$, is the join probability when an arriving customer finds n customers in the system (without counting herself/himself). Naor (1969) studied the equilibrium customer strategies for this case and the associated social optimization and revenue maximization problems. We will present the corresponding findings in this section.

First, we focus on the equilibrium strategies of the customers. When the population of the customers follow a strategy $\mathbf{q}$, the system behaves as an M/M/1 queue with state-dependent arrival rates $\lambda_n = \lambda q_n$, and any performance measure can be easily computed using standard results from birth–death processes. More importantly, the conditional mean sojourn time of a tagged customer, given that he/she finds n customers in the system, does not depend on the strategy $\mathbf{q}$ employed by the population of customers. Because of the FCFS queueing discipline and the memoryless property of the exponential service times, we can easily argue that his/her conditional mean sojourn time is $\frac{n+1}{\mu}$.

Suppose that the tagged customer follows strategy $\mathbf{q}' = (q'_0, q'_1, q'_2, \ldots)$, when the population of the others follows strategy $\mathbf{q} = (q_0, q_1, q_2, \ldots)$. If the tagged customer finds n customers upon arrival, then his/her expected utility is

$$\mathcal{U}(\mathbf{q}', \mathbf{q}|n) = (1 - q'_n) \cdot 0 + q'_n \left(R - \frac{C(n+1)}{\mu} \right).$$

Therefore, his/her best response depends on the sign of the quantity

$$S_{ind}^{(obs)}(n) = R - \frac{C(n+1)}{\mu}.$$

The set of best responses $BR(\mathbf{q}|n)$ of the tagged customer against a strategy $\mathbf{q}$ of the others, when he/she finds n customers in the system is

$$BR(\mathbf{q}|n) = \begin{cases} \{0\}, & \text{if } \frac{R\mu}{C} - 1 < n, \\ [0,1]\,, & \text{if } \frac{R\mu}{C} - 1 = n, \\ \{1\}, & \text{if } \frac{R\mu}{C} - 1 > n, \end{cases}$$

which clearly does not depend on $\mathbf{q}$. Therefore, we have the following result:

THEOREM 4.3.– For the join-or-balk customer dilemma in the observable M/M/1 queue, the n_e-threshold strategy with

$$n_e = \left\lfloor \frac{R\mu}{C} \right\rfloor, \qquad [4.2]$$

that prescribes a customer to join the system as long as the number of customers in the system including his/her is at most n_e is the dominant strategy (and therefore the equilibrium strategy).

We now turn to the problem of the administrator of the system, when he acts as a social planner, aiming to induce a socially optimal strategy $\mathbf{q}_{soc}$,to maximize the social welfare per time unit, i.e. the quantity

$$S_{soc}^{(obs)}(\mathbf{q}) = S_{soc}^{(obs)}(q_0, q_1, \ldots) = \lambda_e^{(obs)}(\mathbf{q})R - CE_{\mathbf{q}}[Q], \qquad [4.3]$$

where $\lambda_e^{(obs)}(\mathbf{q})$ is the steady-state throughput, under strategy $\mathbf{q}$, and $E_{\mathbf{q}}[Q]$ the corresponding expected steady-state number of customers in the system. The classical approach for solving this problem is the use of stochastic dynamic programming (Stidham 1985), where it is shown that the optimal strategy is of threshold type. Indeed, the social optimization problem can be solved by considering an appropriate Markov decision process. Since the state of this process is fully observable, there always exists a non-randomized optimal policy that can be seen to be a pure threshold optimal policy. However, in what follows, we will focus on the version of this problem where the social planner optimizes the social welfare per time unit by imposing a common admission fee (price) to all arriving customers. To this end, we suppose that the administrator charges an admission fee p. Then, the customers follow a threshold strategy $n = \left\lfloor \frac{(R-p)\mu}{C} \right\rfloor$ and the system reduces to an M/M/1/n queue. We limit the study, for simplicity, to the case $\rho < 1$ (but the case $\rho \geq 1$ can be treated similarly). Then, using standard formulas for the M/M/1/n queue regarding its throughput and mean queue length (see, e.g., section 7.3.2 in

(Kulkarni 2010)), and substituting them into [4.3], we conclude that the social welfare, when the n-threshold strategy has been imposed is given by

$$S_{soc}^{(obs)}(n) = \lambda R \frac{1-\rho^n}{1-\rho^{n+1}} - C\left[\frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}\right].$$

After a bit of algebraic manipulation, we see that

$$S_{soc}^{(obs)}(n) - S_{soc}^{(obs)}(n-1) = \frac{\lambda R(1-\rho)^2\rho^{n-1}}{(1-\rho^{n+1})(1-\rho^n)} + \frac{C((n+1)\rho - \rho^{n+1} - n)\rho^n}{(1-\rho^{n+1})(1-\rho^n)}.$$

For $\rho < 1$,

$$\begin{aligned} S_{soc}^{(obs)}(n) - S_{soc}^{(obs)}(n-1) \geq 0 \Leftrightarrow \lambda R(1-\rho)^2 & \\ \geq C\rho(n + \rho^{n+1} - (n+1)\rho) & \\ \Leftrightarrow \frac{R\mu}{C} \geq \frac{n + \rho^{n+1} - (n+1)\rho}{(1-\rho)^2}. & \end{aligned}$$

Let $g(n)$ be the quantity in the right-hand side of the last inequality. Then,

$$\begin{aligned} g(n) &= \frac{n + \rho^{n+1} - (n+1)\rho}{(1-\rho)^2} \\ &= \frac{1}{(1-\rho)^2}\left(n(1-\rho) - \rho(1-\rho^n)\right) \\ &= \frac{1}{1-\rho}\left(n - \sum_{k=1}^{n} \rho^k\right) \\ &= \frac{1}{1-\rho}\sum_{k=1}^{n}(1-\rho^k), \end{aligned} \qquad [4.4]$$

which is obviously strictly increasing in n. Moreover, $g(0) = 0$ and $\lim_{n\to\infty} g(n) = \infty$. Therefore, a unique number n_{soc} exists such that $g(n) \leq \frac{R\mu}{C}$, for $n \leq n_{soc}$, whereas $g(n) > \frac{R\mu}{C}$, for $n > n_{soc}$. Hence, $S_{soc}^{(obs)}(n) - S_{soc}^{(obs)}(n-1) \geq 0$ for $n \leq n_{soc}$, whereas $S_{soc}^{(obs)}(n) - S_{soc}^{(obs)}(n-1) < 0$ for $n > n_{soc}$. We conclude that $S_{soc}^{(obs)}(n)$ is unimodal with a maximum at n_{soc}. In a nutshell, we have the following result:

THEOREM 4.4.– For the social planner's admission problem in the observable M/M/1 queue, the n_{soc}-threshold strategy with

$$n_{soc} = \max\left\{n : g(n) \leq \frac{R\mu}{C}\right\},$$

where the function $g(n)$ is given by [4.4], is the socially optimal strategy. The threshold n_{soc} is induced by the administrator of the system, by imposing an admission fee p_{soc}, such that $n_{soc} = \left\lfloor \frac{(R-p_{soc})\mu}{C} \right\rfloor$, i.e. by imposing any $p_{soc} \in \left(R - \frac{C(n_{soc}-1)}{\mu}, R - \frac{Cn_{soc}}{\mu}\right]$.

In addition, it is easy to see that $n_{soc} \leq n_e$, i.e. in the absence of an admission fee, the individually optimal threshold is greater than or equal to the socially optimal threshold. Indeed, we have

$$\begin{aligned} g(n) - n &= \frac{1}{1-\rho}\sum_{k=1}^{n}(1-\rho^k) - \frac{1}{1-\rho}n(1-\rho) \\ &= \frac{\rho}{1-\rho}\sum_{k=1}^{n}(1-\rho^{k-1}) \geq 0, \end{aligned}$$

whence $g(n_{soc}) \geq n_{soc}$. But $g(n_{soc}) \leq \frac{R\mu}{C}$, by the very definition of n_{soc}, so $n_{soc} \leq \frac{R\mu}{C}$ that yields $n_{soc} \leq \left\lfloor \frac{R\mu}{C} \right\rfloor = n_e$.

Therefore, we see that, without imposing an admission fee, the customers overuse the system. This occurs because they neglect the negative externalities of their joining decisions on future customers. The same phenomenon was also observed in the unobservable counterpart of the model.

We now consider the monopolist's problem. In this case, the administrator of the system imposes an admission fee, aiming to the maximization of his own revenue. If he imposes admission fee p, then the customers adopt the corresponding threshold strategy and his revenue becomes $\lambda_e^{(obs)}p$, where $\lambda_e^{(obs)}$ stands for the corresponding equilibrium throughput. To determine the threshold n_{prof} that maximizes the revenue, we express the revenue as a function of the imposed threshold n. For inducing a threshold n to the customers, the administrator should impose an admission fee p such that $\left\lfloor \frac{(R-p)\mu}{C} \right\rfloor = n$. In the monopolist's problem, the administrator benefits from imposing the maximum price that induces the threshold, i.e. he should impose $p = R - \frac{Cn}{\mu}$. Then, the monopolist's revenue is

$$S_{prof}^{(obs)}(n) = \lambda\frac{1-\rho^n}{1-\rho^{n+1}}\left(R - \frac{Cn}{\mu}\right) = \lambda R\frac{1-\rho^n}{1-\rho^{n+1}}\left(1 - \frac{n}{\nu_e}\right),$$

where $\nu_e = \frac{R\mu}{C}$. This equation shows that for $n > n_e = \lfloor \frac{R\mu}{C} \rfloor$ we have $S_{prof}(n) < 0$, since the administrator should set a negative admission fee (i.e. he should subsidized the entrance) to induce a threshold, which is greater than n_e. Therefore, we conclude that $n_{prof} \leq n_e$.

To study the monotonicity behavior of the function $S^{(obs)}_{prof}(n)$, we consider the ratio $S^{(obs)}_{prof}(n)/S^{(obs)}_{prof}(n-1)$. We present again only the case where $\rho < 1$.

$$\frac{S^{(obs)}_{prof}(n)}{S^{(obs)}_{prof}(n-1)} = \frac{(1-\rho^n)^2(\nu_e - n)}{(1-\rho^{n+1})(1-\rho^{n-1})(\nu_e - n + 1)}.$$

Then,

$$\begin{aligned}\frac{S^{(obs)}_{prof}(n)}{S^{(obs)}_{prof}(n-1)} \geq 1 &\Leftrightarrow \frac{1-\rho^n}{1-\rho^{n+1}}(\nu_e - n) \geq \frac{1-\rho^{n-1}}{1-\rho^n}(\nu_e - n + 1) \\ &\Leftrightarrow \frac{(1-\rho^n)^2 - (1-\rho^{n-1})(1-\rho^{n+1})}{(1-\rho^{n+1})(1-\rho^n)}(\nu_e - n) \geq \frac{1-\rho^{n-1}}{1-\rho^n} \\ &\Leftrightarrow \nu_e - n \geq \frac{(1-\rho^{n-1})(1-\rho^{n+1})}{\rho^{n-1}(1-\rho)^2} \\ &\Leftrightarrow \frac{R\mu}{C} \geq n + \frac{(1-\rho^{n-1})(1-\rho^{n+1})}{\rho^{n-1}(1-\rho)^2}.\end{aligned}$$

It can be seen that the function $h(n)$ with

$$h(n) = n + \frac{(1-\rho^{n-1})(1-\rho^{n+1})}{\rho^{n-1}(1-\rho)^2} \qquad [4.5]$$

is increasing in n, so a unique number n_{prof} exists such that $h(n) \leq \frac{R\mu}{C}$, for $n \leq n_{prof}$, whereas $h(n) > \frac{R\mu}{C}$, for $n > n_{prof}$. Therefore, the function $S^{(obs)}_{prof}(n)$ is unimodal with the maximum at n_{prof}. Therefore, we have the following result:

THEOREM 4.5.– For the monopolist's admission problem in the observable M/M/1 queue, the n_{prof}-threshold strategy with

$$n_{prof} = \max\left\{n : h(n) \leq \frac{R\mu}{C}\right\},$$

where the function $h(n)$ is given by [4.5], is the revenue-maximizing strategy. The threshold is induced by the administrator of the system by imposing an admission fee $p_{prof} = R - \frac{Cn_{prof}}{\mu}$.

Using some quite involved algebraic manipulations, it can be seen that the inequality $n_{prof} \leq n_{soc} \leq n_e$ holds. This is known as *Naor's inequality* and is valid in a number of situations. Recently, (Hassin and Snitkovsky 2018) provided general conditions for the validity of this inequality in a general framework. If a state-dependent admission fee is permitted, then due to information symmetry a monopolist can extract all customer surplus and hence the monopolist's optimal threshold strategy and the socially optimal threshold strategy coincide, as in the unobservable model (for details, see Chen and Frank (2001)).

4.5. Comparison of the unobservable and the observable models

Hassin (1986) compared the observable and the unobservable versions of the M/M/1 queue with strategic customers regarding their join-or-balk dilemma, by focusing on the social welfare and a monopolist's revenue under a revenue-maximizing admission fee. Let λ, μ, R and C be the parameters of a model, defined as in section 4.3. Hassin showed that if $R\mu \leq 2C$, then the revenue under a revenue-maximizing admission fee is larger for the observable model, for all $\lambda > 0$. Hence, a monopolist prefers to reveal the queue length to the customers. If, however, $R\mu > 2C$, then a unique potential arrival rate λ^Z exists such that the revenue under a revenue-maximizing admission fee is larger for the observable model, if and only if $\lambda \geq \lambda^Z$. Thus, in this case, a monopolist prefers to reveal the queue length only when $\lambda \geq \lambda^Z$. Thus, there is a range of the parameters ($R\mu > 2C$ and $\lambda < \lambda^Z$), where the provision of more information to the customers hurts the service provider. The same properties hold also for the social welfare under a revenue-maximizing fee, but with a different critical value λ^S in place of λ^Z. Therefore, in this case, for $\lambda < \lambda^S$, it is socially preferable for the monopolist to be unable to inform customers of the queue length. Thus, there is a range of the parameters where the provision of more information hurts the society as a whole. Note also that $\lambda^S < \lambda^Z$, so for arrival rates λ with $\lambda^S < \lambda < \lambda^Z$, the profit maximizer prefers to conceal the queue length, although it is socially preferable to induce him to reveal it. However, when the profit maximizer prefers to reveal the queue length (i.e. when $\lambda \geq \lambda^Z$), this is certainly socially preferable as well.

Chen and Frank (2004) compared the observable and the unobservable versions of the M/M/1 queue by focusing on the equilibrium effective arrival rate (which is the same as the throughput since there are no abandonments), under an arbitrary fixed admission fee. For a given potential arrival rate λ, let $\lambda_e^{(obs)}(\lambda)$ and $\lambda_e^{(un)}(\lambda)$ denote the corresponding equilibrium effective arrival rates in the observable and unobservable versions, respectively. Chen and Frank proved that $\lambda_e^{(obs)}(\lambda) - \lambda_e^{(un)}(\lambda)$ monotonically increases in λ and there exists a critical value λ^* such that $\lambda_e^{(obs)}(\lambda^*) - \lambda_e^{(un)}(\lambda^*) = 0$. Therefore, to attract more customers to the system, it is advisable to conceal the queue length for potential arrival rates λ with $\lambda < \lambda^*$, and to reveal it when $\lambda > \lambda^*$. This is intuitively plausible. Indeed, in an unobservable M/M/1 queue with low arrival rates all customers join, whereas in the corresponding observable queue there will be always some customers (those that arrive during high congestion periods) who balk. Hence, for sufficiently low arrival rates the equilibrium effective arrival rate is higher for the unobservable model. For high arrival rates, we have the opposite situation, i.e. all customers balk in the unobservable model, whereas a positive fraction of customers who find low congestion do enter the observable model.

Shone *et al.* (2013) considered the same problem of the comparison of the equilibrium throughputs $\lambda_e^{(obs)}$ and $\lambda_e^{(un)}$, between the observable and unobservable versions of the M/M/1 queue. They provided necessary and sufficient conditions on the system parameters under which the equilibrium throughputs are equal in the two versions. Moreover, they investigated the behavior of the equilibrium throughputs in the two informational cases as functions of the normalized service value $\frac{R\mu}{C}$. In particular, they showed that the number of distinct values of the normalized service value for which $\lambda_e^{(obs)} = \lambda_e^{(un)}$ is monotonically increasing with respect to the utilization rate $\rho = \frac{\lambda}{\mu}$ and tends to infinity as $\rho \to 1$.

Guo and Zipkin (2007) compared the observable, the unobservable and the workload-observable versions of the M/M/1 queue, under a general reward–cost structure that generalizes the standard Naor's linear reward–cost structure. Under this framework, the service value is R, but a customer's waiting cost is $\theta E[c(W)]$, where W stands for the steady-state waiting time, $c(w)$ is a common basic cost function for all customers and θ is a customer-specific parameter that represents the sensitivity to delay. In other words, a customer with delay sensitivity θ has expected utility $R - \theta E[c(W)]$, if he/she decides to join. The authors showed that the maximum equilibrium throughput of the system may correspond to different information levels according to the values of the underlying parameters. The main conclusion is that the primary factor that determines whether information is good or bad for the service provider and the customers is the distribution function of the customer delay sensitivity and not the common basic cost function.

The aforementioned papers show that neither the observable nor the unobservable versions of the M/M/1 queue are preferable for the whole range of the underlying operational and economic parameters. Therefore, a number of authors studied the M/M/1 queue with strategic customers under information structures that lie between the observable and unobservable versions. To the best of our knowledge, there are three main ideas that have appeared in the literature that bridge the observable and unobservable versions of the M/M/1 queue: partially observable models, heterogeneously observable models and observable-with-delay models. We present these ideas in sections 4.6–4.8.

4.6. Partially observable models

In partially observable models, the state-space of the queue length of a given service system is partitioned into subsets and the arriving customers are not informed about the exact queue length, but rather about the subset it belongs to. If the state-space is partitioned into subsets of consecutive integers, the waiting space can be considered to be “compartmented” and the customers are informed only about the compartment in which they are going to be placed. Economou and Kanta (2008) considered the case of regular compartmentalization (all compartments being of the same size) in the M/M/1 queue and studied the customer strategic behavior and the associated social optimization and revenue maximization administrator’s problems. In what follows, we present the corresponding main results.

The model of interest is an M/M/1 queue with the same operational and economic parameters λ, μ, R and C that were introduced in section 4.3. However, we assume that the space of the system is partitioned in compartments of fixed capacity of a customers and we consider two information cases for the customers:

– N: Known compartment number: Customers observe the number of the compartment in which they are going to enter but not the position within it. More specifically, if there exist n customers in the system just before the arrival of a tagged customer, his/her information will be the compartment number $i = \lfloor n/a \rfloor + 1$ in which he/she enters if he/she decides to join the system.

– P: Known compartment position: Customers observe the position of the compartment in which they are going to enter but not the number of the compartment. The information of an arriving customer is the position $i = (n \mod a) + 1$ in which he/she enters if he/she decides to join the system.

The decisions of the customers are irrevocable, i.e. retrials of balking customers and reneging of entering customers are not allowed.

We now limit our exposition within the framework of the known compartment number (N) case. Then, a pure strategy is specified by a set $A \subseteq \{1, 2, \ldots\}$, which shows the “favorable” compartment numbers for a customer, i.e. a customer decides to join the system if he/she knows that he/she will enter a compartment with a number belonging to A. We consider a tagged customer and assume that all other customers follow a strategy A. Then the Markov chain describing the number of customers in the system will be eventually absorbed in the set $\{0, 1, 2, \ldots, i^*a\}$, where i^* is the maximum integer such that $\{0, 1, 2, \ldots, i^*\} \subseteq A$. Indeed, a moment of reflection shows that under strategy A all other states become transient and the system behaves as an M/M/1/i^*a queue. Consider now the tagged customer who is to decide whether to join the system or not, given the information that he/she can enter the compartment number i. So he/she knows that the number of customers at his/her arrival is $n \in$

$\{(i-1)a,(i-1)a+1,\ldots,ia-1\}$. If he/she decides to enter, then his/her expected net individual benefit is

$$S_{ind}^{(po-N)}(i)=R-\frac{C(E_A[Q^-|Q^-\in\{(i-1)a,\ldots,ia-1\}]+1)}{\mu}, \qquad [4.6]$$

where Q^- is a random variable with steady-state distribution of the number of customers at an arrival instant in the M/M/1/i^*a queue and the subscript "A" in the expectation signifies that the population of the customers follow the pure strategy A. Because of the PASTA property, the distribution of Q^- coincides with the distribution of the number of customers in continuous-time, Q. Using standard formulas for the M/M/1/n queue (see, e.g. section 7.3.2 in Kulkarni (2010)), we obtain that for $\rho\neq 1$

$$S_{ind}^{(po-N)}(i)=R-\frac{C}{\mu}\left(ia-\frac{a}{1-\rho^a}+\frac{1}{1-\rho}\right). \qquad [4.7]$$

For $\rho=1$, we can see that $S_{ind}^{(po-N)}(i)=R-\frac{C}{\mu}\{ia-\frac{a-1}{2}\}$, which is the limiting case of [4.7] for $\rho\to 1$. Therefore, in the rest of this section, we present the formulas under the assumption that $\rho\neq 1$, with the understanding that the results are also valid for $\rho=1$ with the appropriate limits in the formulas (as $\rho\to 1$).

The tagged customer will decide to enter if and only if $S_{ind}^{(po-N)}(i)\geq 0$, that is when $i\leq\lfloor\frac{R\mu}{aC}+\frac{1}{1-\rho^a}-\frac{1}{a(1-\rho)}\rfloor$. Therefore, we conclude with the following result:

THEOREM 4.6.– For the join-or-balk customer dilemma in the partially N-observable a-compartmented M/M/1 queue, the i_e-threshold strategy with

$$i_e=\lfloor x_e\rfloor, \qquad [4.8]$$

$$x_e=\frac{R\mu}{aC}+\frac{1}{1-\rho^a}-\frac{1}{a(1-\rho)}, \qquad [4.9]$$

that prescribes a customer to join the system as long as the compartment number that will be assigned is at most i_e is the dominant strategy (and therefore the equilibrium strategy).

The problem of social optimization can be also solved along similar lines with the observable M/M/1 queue. Indeed, using an appropriate admission fee, the administrator of the system can induce any desired i-threshold strategy. Then, the corresponding total social welfare per time unit is $S_{soc}^{(po-N)}(i)=\lambda^{(po-N)}(i)R-CE_i[Q]$, where $\lambda^{(po-N)}(i)$ is the throughput and $E_i[Q]$ the expected steady-state number of customers in the system, given that the customers follow the i-threshold strategy. In this case, the system behaves as an M/M/1/ia queue and using the corresponding steady-state distribution yields

$$S_{soc}^{(po-N)}(i) = \lambda R \frac{1-\rho^{ia}}{1-\rho^{ia+1}} - C\left(\frac{\rho}{1-\rho} - \frac{(ia+1)\rho^{ia+1}}{1-\rho^{ia+1}}\right). \qquad [4.10]$$

The unimodality of $S_{soc}^{(po-N)}(i)$ is a result of the unimodality of $S_{soc}^{(obs)}(n)$ (since $S_{soc}^{(po-N)}(i) = S_{soc}^{(obs)}(ai)$) and we obtain the next result that characterizes the socially optimal policy:

THEOREM 4.7.– For the social planner's admission problem in the partially N-observable a-compartmented M/M/1 queue, the i_{soc}-threshold strategy with

$$i_{soc} = \lfloor x_{soc} \rfloor \qquad [4.11]$$

with x_{soc} being the unique solution of the equation $g(x) = i_e$ in $[1, \infty]$ with

$$g(x) = \frac{(xa+1)(1-\rho^a) - a(1-\rho^{xa+1})}{a(1-\rho)(1-\rho^a)} + \frac{1}{1-\rho^a} - \frac{1}{a(1-\rho)}, \qquad [4.12]$$

is the socially optimal strategy.

The monopolist's revenue-maximization problem is also solved similarly to the corresponding problem in the framework of the observable M/M/1 queue. Indeed, in light of the equations [4.8]–[4.9] the maximum entrance fee that can be imposed by the administrator of the system in order to force the customers to adopt a given threshold i is

$$p(i) = R - \frac{aC}{\mu}\left(i - \frac{1}{1-\rho^a} + \frac{1}{a(1-\rho)}\right). \qquad [4.13]$$

Then, his expected revenue per time unit is $S_{prof}^{(po-N)}(i) = \lambda^{(po-N)}(i)p(i)$ where $\lambda^{(po-N)}(i)$ is the throughput given that the customers follow the threshold i. The system behaves as an M/M/1/ia queue and therefore using the corresponding steady-state distribution (see, e.g. section 7.3.2 in Kulkarni (2010)), we obtain after some algebra that

$$S_{prof}^{(po-N)}(i) = \frac{\lambda a C(1-\rho^{ia})}{\mu(1-\rho^{ia+1})}(x_e - i), \qquad [4.14]$$

where x_e is given by [4.9]. This function is unimodal and its maximum point is characterized in the following result (for more detail, see Economou and Kanta (2008)):

THEOREM 4.8.– For the monopolist's admission problem in the partially N-observable a-compartmented M/M/1 queue, the i_{prof}-threshold strategy with

$$i_{prof} = \lfloor x_{prof} \rfloor \tag{4.15}$$

with x_{prof} being the unique solution of the equation $h(x) = x_e$ in $[1, \infty]$ with

$$h(x) = x + \frac{(1-\rho^{xa-a})(1-\rho^{xa+1})}{\rho^{xa-a}(1-\rho)(1-\rho^a)}, \tag{4.16}$$

is the revenue-maximizing strategy.

Using these theoretical results, a number of numerical experiments were carried out by (Economou and Kanta 2008) with various interesting findings. Moreover, a number of secondary theoretical results were proved. The P case has been also studied in detail by (Economou and Kanta 2008) and the equilibrium, socially optimal and revenue-maximizing strategies have been characterized in almost-explicit forms. We do not report the results in detail here, since this case is more difficult to implement. We now summarize the main findings of this study:

In both information cases, equilibrium threshold balking strategies exist. The corresponding equilibrium, social and revenue-maximizing thresholds can be explicitly computed.

In the N case, the equilibrium threshold strategy is the dominant strategy. This is quite exceptional behavior since dominant strategies are reported in the literature only for fully observable models. In addition, a kind of Naor's inequality holds: $i_{prof} \leq i_{soc} \leq i_e$, i.e. the individual optimization leads to longer queues that it is socially desirable (while the revenue maximization induces even shorter queues). The equilibrium threshold i_e is a decreasing function of λ. This is in contrast to (Naor 1969) model where the equilibrium threshold n_e does not depend on the arrival rate λ.

In the N case, the maximum number of customers in the system is $i_e a$, when the customers follow an equilibrium strategy. When keeping all parameters fixed and letting a varies, this number is minimized for compartment sizes a that are near to the integer divisors of Naor's equilibrium threshold, n_e, of the observable model. This suggests that it is desirable for the designer of the system to construct the compartments according to one of these a, if he wants to minimize the waiting space required for the proper function of the system.

And finally, perhaps the most interesting finding in the N case is that the administrator's revenue is a unimodal function of the compartment size a, when keeping the other parameters fixed. Therefore, there exists an optimal a that maximizes the administrator's revenue. This suggests that apart from other mechanisms suggested in the literature (information pricing, state-dependent pricing, service discipline, etc.), the compartmentalization and an adequate selection of the compartment size can be used for increasing the administrator's revenue.

In the P case, the equilibrium threshold is a decreasing function of λ while the social and profit maximizing thresholds exhibit unimodal behavior. This is in agreement with the corresponding findings for (Edelson and Hildebrand 1975) model. However, a crucial difference between the P case and the unobservable model is the fact that the objectives of a revenue-maximizer monopolist and a social planner do not coincide, except for the case $a = 1$. This happens because of the partial information of the customers that enable them to secure a positive surplus.

When $a = 1$, the N case reduces to (Naor 1969) model, while the P case reduces to (Edelson and Hildebrand 1975) model. For $a \to \infty$, we have the opposite situation, i.e. the N case behaves as (Edelson and Hildebrand 1975) model and the P case as (Naor 1969) model. In general, we can say that for small values of a, the N case is more informative because then the knowledge of the compartment number determines the exact position of the customer with error of at most a. For large values of a, we know that almost all the customers enter the first compartment so the N case is less informative. Then, knowing the position in the compartment is much more important for the customer to assess his/her gains.

There are many other papers dealing with partially observable variants of the M/M/1 queue. Guo and Zipkin (2009) considered the general case of compartments with possibly different sizes and proved several interesting results about the comparison of two partitions of the state space, one a refinement of the other. More recently, (Simhon *et al.* 2016) considered the M/M/1 queue with strategic customers that face the dilemma of joining/balking, when the administrator informs the customers about the current queue length only when it is short, i.e. when it does not exceed a certain threshold D. This corresponds to the partition of the state-space to the subsets $\{0\}, \{1\}, \{2\}, \dots, \{D\}$ and $\{D+1, D+2, \dots\}$. The authors proved that the equilibrium throughput is a monotone function of D and therefore if the administrator's goal is to maximize throughput, then the optimal policy is one of the extremes, either the observable or the unobservable queue. Kim and Kim (2017) considered the generalization of the last model by assuming that the customers are informed about the current queue length only when it belongs to a subset O. This corresponds to the partition of the state-space comprising the singletons of the elements of O and the complement of O (which contains the unobservable states). The authors proved the counterintuitive result that the optimal partition for the maximization of the throughput of the system corresponds to a set O that contains all the states above a threshold, i.e. it is preferable to allow the customers to observe the queue length only when it is large. Finally, (Hassin and Koshman 2017) considered a model where the arriving customers are informed whether the queue length is less than an exogenously given threshold N or not. They focused on the monopolist's problem for the dynamic pricing version of this model (i.e. different prices are offered to the customers according to whether the queue length is below N or not) and proved the interesting result that the choice of $N = 1$ guarantees at least half of the maximum value that can be generated by the system.

4.7. Heterogeneously observable models

In heterogeneously observable models, the population of customers is divided into observing and uninformed (non-observing) customers. A simple model that encompasses this characteristic has been studied by (Economou and Grigoriou 2015) and (Hu *et al.* 2018). In what follows, we present some of the reported results for this model.

The system of interest is an M/M/1 queue with the same operational and economic parameters with the main model that we introduced in section 4.3. Every arriving customer is observing or uninformed, with probabilities p_o and $p_u = 1 - p_o$, respectively, independently of the other customers. Observing customers are informed about the number of customers in the system before making their decisions whether to join or balk, whereas uninformed customers do not. The decisions of the customers are irrevocable (i.e. neither retrials nor reneging are permitted). All the parameters of the model, including the proportion of observing customers, p_o, are assumed common knowledge for the customers.

An observing customer faces a situation like the one that we described for the observable M/M/1 queue in section 4.4. Thus, his/her dominant strategy is to join according to Naor's threshold n_e given from [4.2]. On the other hand, a general mixed strategy for an uninformed customer is specified by a single number, the join probability q.

We are now focusing on the best response of a tagged uninformed customer. To this end, suppose that the observing customers follow the n_e-threshold strategy, i.e. they join according to Naor's threshold, and that the uninformed customers follow a mixed strategy q. Then, the tagged customer bases his/her decision on the sign of his/her expected net benefit if he/she decides to join. This is given as

$$S_{ind}^{(ho)}(q) = R - \frac{C(E_q[Q^-] + 1)}{\mu}, \qquad [4.17]$$

where $E_q[Q^-]$ is the mean number of customers in system found by an arriving uninformed customer, given that the others follow the strategy (n_e, q) (i.e. the observing customers follow the n_e-threshold strategy and the uninformed customers follow the q-mixed strategy). However, the uninformed customers arrive according to a Poisson process with rate λp_u and therefore, because of the PASTA property, $E_q[Q^-]$ coincides with the mean number of customers in the system $E_q[Q]$, when the strategy (n_e, q) is employed. Therefore, for the study of strategic customer behavior, we need to compute $E_q[Q]$, for any strategy (n_e, q). Under such a strategy the number of customers in the system is a continuous-time Markov chain of birth–death type with transition rates

$$q_{i,j} = \begin{cases} \lambda_1 \text{ if } 0 \leq i \leq n_e - 1,\ j = i+1 \\ \lambda_2 \text{ if } i \geq n_e,\ j = i+1 \\ \mu \ \text{ if } i \geq 1,\ j = i-1 \\ 0 \ \text{ otherwise,} \end{cases} \qquad [4.18]$$

where

$$\lambda_1 = \lambda p_o + \lambda(1 - p_o)q,\ \lambda_2 = \lambda(1 - p_o)q. \qquad [4.19]$$

Using the well-known formula for the steady-state distribution of birth–death processes and standard summation techniques, we can easily derive the steady-state distribution in closed-form and its mean. More concretely, we have the following result:

PROPOSITION 4.1.– The steady-state distribution of the number of customers in the heterogeneously observable M/M/1 queue, when the customers follow an (n_e, q) strategy, with $n_e, q > 0$, is given by

$$\pi_n = \begin{cases} B\rho_1^n & \text{if } 0 \leq n \leq n_e - 1, \\ B\rho_1^{n_e}\rho_2^{n-n_e} & \text{if } n \geq n_e, \end{cases} \qquad [4.20]$$

where

$$\rho_1 = \frac{\lambda_1}{\mu},\ \rho_2 = \frac{\lambda_2}{\mu} \qquad [4.21]$$

and

$$B = \frac{(1-\rho_1)(1-\rho_2)}{1 - \rho_2 - \rho_1^{n_e+1} + \rho_1^{n_e}\rho_2}. \qquad [4.22]$$

The corresponding mean steady-state number of customers is

$$E_q[Q] = \frac{(1-\rho_2)[(n_e-1)\rho_1^{n_e+1} - n_e\rho_1^{n_e} + \rho_1]}{(1-\rho_1)[1-\rho_2-\rho_1^{n_e+1}+\rho_1^{n_e}\rho_2]} + \frac{(1-\rho_1)[n_e\rho_1^{n_e} - (n_e-1)\rho_1^{n_e}\rho_2]}{(1-\rho_2)[1-\rho_2-\rho_1^{n_e+1}+\rho_1^{n_e}\rho_2]}. \qquad [4.23]$$

The cases where one or both of n_e and q are 0 can be easily derived from the above formulas by taking the appropriate limits. We do not report the formulas for brevity, but we discuss them briefly: For $n_e = 0$ and $q = 0$, the system is continuously empty. For $n_e = 0$ and $q > 0$, the system behaves as an M/M/1 queue with arrival rate $\lambda_2 = \lambda(1-p_o)q$ and service rate μ. Similarly, for $n_e > 0$ and $q = 0$, the system behaves as an M/M/1/n_e queue with arrival rate $\lambda_1 = \lambda p_o$ and service rate μ. Note also that the quantity $E_q[Q]$ is increasing in q. This is intuitively clear, but it can be also formally proven by using (Kirstein 1976) rate sufficient conditions for the strong comparability of Markov chains.

We can now return to the study of a tagged uninformed customer's best response when a strategy (n_e, q) is employed by the other customers. Let $\hat{q}_e$ be the root of $S_{ind}^{(ho)}(q)$. Then, the analysis proceeds along the same lines with the unobservable model of section 4.3. The set of best responses against (n_e, q), $BR((n_e, q))$, is (as far as $q \in [0,1]$ and $\lambda(1-p_o)q < 1$)

$$BR((n_e,q)) = \begin{cases} \{(n_e,0)\}, & \text{if } q > \hat{q}_e, \\ \{n_e\} \times [0,1]\,, & \text{if } q = \hat{q}_e, \\ \{(n_e,1)\}, & \text{if } q < \hat{q}_e. \end{cases}$$

We can now proceed to the computation of the equilibrium strategies:

The strategy $(n_e, 0)$ is equilibrium strategy, if and only if $(n_e, 0) \in BR((n_e, 0))$, i.e. $0 \geq \hat{q}_e$, which reduces to $R \leq \frac{C(E_0[Q]+1)}{\mu}$.

A strategy (n_e, q_e) with $q_e \in (0,1)$ is equilibrium strategy, if and only if $(n_e, q_e) \in BR((n_e, q_e))$, i.e. $q_e = \hat{q}_e$. This is valid as far as $\hat{q}_e \in (0,1)$, which occurs if and only if $\frac{C(E_0[Q]+1)}{\mu} < R < \frac{C(E_1[Q]+1)}{\mu}$.

Finally, the strategy $(n_e, 1)$ is the equilibrium strategy, if and only if $(n_e, 1) \in BR((n_e, 1))$, i.e. $1 \leq \hat{q}_e$, which reduces to $R \geq \frac{C(E_1[Q]+1)}{\mu}$.

In summary, we have the following result:

THEOREM 4.9.– For the join-or-balk customer dilemma in the heterogeneously observable M/M/1 queue, where a fraction of customers are observing and the rest customers are uninformed, a unique equilibrium strategy (n_e, q_e) exists. The threshold n_e for the entrance of the observing customers is Naor's threshold given from [4.2]. The join probability q_e for the uninformed customers is given by the formula

$$q_e = \begin{cases} 0, & R \leq \frac{C(E_0[Q]+1)}{\mu}, \\ \hat{q}_e, & \frac{C(E_0[Q]+1)}{\mu} < R < \frac{C(E_1[Q]+1)}{\mu}, \\ 1, & R \geq \frac{C(E_1[Q]+1)}{\mu}, \end{cases}$$

where $\hat{q}_e$ is the unique root of $S_{ind}^{(ho)}(q)$ given from [4.17] and $E_q[Q]$ is computed from [4.23].

Economou and Grigoriou (2015) determined the equilibrium strategies in the slightly more general case where the service value R and the waiting cost rate C are different for informed and uninformed customers and provided a preliminary analysis. Hu *et al.* (2018) considered the homogeneous reward–cost framework of this section and studied in depth the effect of the fraction p_o of observing customers on the equilibrium threshold and social welfare. They showed that different behaviors emerge, according to the arrival rate of the customers. If the throughput is the focal performance measure, then $p_o = 1$ maximizes throughput if the arrival rate is high enough, whereas $p_o = 0$ maximizes throughput if the arrival rate is low enough. If the arrival rate is in an intermediate range, the maximum throughput is achieved at a p_o strictly between 0 and 1. Therefore, in this range, it is optimal to have a segment of uninformed customers or reveal the queue-length only to a fraction of customers.

If social welfare is the focal performance measure, the service provider should reveal the queue length information and encourage its dissemination when the arrival rate is relatively small. In other situations, it is optimal to have a segment of uninformed customers or, equivalently, to hide the queue-length from a certain fraction of customers.

In a nutshell, the results of (Hu *et al.* 2018) showed that throughput and social welfare are in general unimodal and not monotonous in the fraction of observing customers. In other words, information heterogeneity in a population can lead to more efficient outcomes, in terms of the system throughput or social welfare, than information homogeneity. Moreover, it was shown that for an overloaded system (with utilization factor sufficiently higher than 1), social welfare always attains its maximum when some fraction of customers is uninformed.

4.8. Observable-with-delay models

In models with delayed observations, the customers decide whether to join or balk without knowing the state of the system, but later on they are informed about their current position and may renege. Burnetas *et al.* (2017) considered a simple model of this kind, which boils down to an M/M/1 queue where the administrator of the system makes periodic announcements to the customers about their current positions. The model was motivated by a situation that occurs when people submit petitions through certain web-based systems. Then, upon submission, the customers receive a confirmation message with the registration number of their petition. Later on they learn the number of pending petitions in front of them. This is done either by periodic refreshments of a web page that indicates the registration number of the currently processed petition or by periodic bulk emails that announce the status of pending petitions. In what follows, we describe the model and the corresponding findings in some detail.

We consider an M/M/1 queue with the same operational and economic parameters with the main model that we introduced in section 4.3. Each customer, upon arrival, decides whether to join or balk, without observing the number of customers in the system. However, the administrator of the system announces to the customers their positions in the system, at the points of a Poisson process at rate θ. The customers, after an announcement, reevaluate their expected benefit of staying in the system, and will renege if it is negative.

Because of the exponentiality assumptions, it makes sense that a customer makes decisions only at the instants of system state transitions, i.e. at his/her arrival instant and at the times of the subsequent announcements of the administrator of the system. At his/her arrival instant, the system is unobservable to the customer. Therefore, his/her join-or-balk strategy is specified by a join probability, say q.

The analysis of the reneging behavior of a customer is trivial. At the epoch of the first announcement following his/her entrance, the system becomes observable to his/her. Therefore, the customer is willing to stay if his/her position n in the system is such that $R - C\frac{n}{\mu} \geq 0$, i.e. he/she decides to stay in the system, only if $n \leq n_e$, where n_e is Naor's threshold given by [4.2]. If a customer does not renege after the first announcement, he/she will not renege later. This fact follows from the Markovian (memoryless) nature of the model.

Therefore, the behavior of the system is specified by two parameters: the join probability q of the customers and their renege threshold n_e at the time of the first announcement after their arrival. Therefore, we need to analyze the behavior of the system when the customers follow a strategy (n_e, q). Under such a strategy, the number of customers in the system is represented by a continuous-time Markov chain with transition rates

$$q_{i,j} = \begin{cases} \lambda q, & \text{if } i \geq 0, \quad j = i+1, \\ \mu, & \text{if } i \geq 1, i \neq n_e + 1, \quad j = i-1, \\ \theta, & \text{if } i \geq n_e + 2, \quad j = n_e, \\ \mu + \theta, & \text{if } i = n_e + 1, \ j = n_e, \\ 0, & \text{otherwise.} \end{cases} \qquad [4.24]$$

The steady-state distribution of the number of customers in the system, Q, and its mean can be computed in closed-form, by solving the system of the corresponding balance equations.

PROPOSITION 4.2.– The steady-state distribution of the number of customers in the observable-with-delay M/M/1 queue, when the customers follow an (n_e, q) strategy, with $n_e, q > 0$ and $\lambda q \neq \mu$, is given by [4.20] where

$$\rho_1 = \frac{\lambda q}{\mu}, \ \rho_2 = \frac{\lambda q + \mu + \theta - \sqrt{(\lambda q + \mu + \theta)^2 - 4\lambda q \mu}}{2\mu}, \qquad [4.25]$$

and B is given by [4.22]. The corresponding mean steady-state number of customers is given by [4.23].

The singular cases where $n_e = 0$ or $q = 0$ or $\lambda q = \mu$ can be easily derived from the formulas of Proposition 4.2 by taking appropriate limits.

We now define $S^{(owd)}_{ind-cond}(n|n_e)$ to be the conditional expected net benefit of a customer who joins the system, given that he/she finds n customers, when all customers (including herself/himself) use the same renege threshold n_e. To compute it, for a certain n, consider a tagged customer who arrives and decides to join, when the system has n other customers. We consider two cases according to whether $n \leq n_e - 1$ or not.

In the first case, where $n \leq n_e - 1$, the customer will certainly receive the service reward R, since he/she has no incentive to renege later. Moreover, his/her sojourn time in the system will be the sum of $n+1$ service times and we conclude that

$$S^{(owd)}_{ind-cond}(n|n_e) = R - \frac{C(n+1)}{\mu}. \qquad [4.26]$$

In the second case, where $n \geq n_e$, the net benefit of the tagged customer, S_n, has the representation

$$S_n = (R - C(Y_n + Z))1_{\{X \geq Y_n\}} - CX1_{\{X < Y_n\}}, \qquad [4.27]$$

where X, Y_n and Z are independent random variables with X being an exponential distribution with rate θ, Y_n being an Erlang distribution with $n+1-n_e$ phases and rate μ and Z being an Erlang distribution with n_e phases and rate μ. This representation is deduced by interpreting X as the time until the first announcement after the arrival of the tagged customer, Y_n the time of $n+1-n_e$ service times until the tagged customer has no incentive to renege and Z the time after Y_n until the departure of the tagged customer (if he/she does not renege). Moreover, note that the customer reneges if $X < Y_n$, in which case his/her sojourn time is X. On the other hand, when $X \geq Y_n$, the customer stays in the system for $Y_n + Z$ time units and receives the reward for service. Taking expected values in [4.27] yields

$$S^{(owd)}_{ind-cond}(n|n_e) = (R - CE[Z])\mathrm{P}[X \geq Y_n] - CE[\min(X, Y_n)]. \qquad [4.28]$$

We can easily deduce that

$$\mathrm{P}[X \geq Y_n] = \left(\frac{\mu}{\mu+\theta}\right)^{n-n_e+1}, \quad E[\min(X, Y_n)] = \frac{1}{\theta}\left(1 - \left(\frac{\mu}{\mu+\theta}\right)^{n-n_e+1}\right),$$

and $E[Z] = \frac{n_e}{\mu}$. Combining and simplifying [4.26] and [4.28] (for details, see Burnetas *et al.* (2017)) yields the following result:

LEMMA 4.1.– Consider the observable-with-delay M/M/1 queue, where the customers follow an (n_e, q) strategy. We have

$$S^{(owd)}_{ind-cond}(n|n_e) = \begin{cases} R - \frac{C(n+1)}{\mu}, & \text{if } 0 \leq n \leq n_e - 1, \\ \left(R - \frac{Cn_e}{\mu} + \frac{C}{\theta}\right)\left(\frac{\mu}{\mu+\theta}\right)^{n-n_e+1} - \frac{C}{\theta}, & \text{if } n \geq n_e. \end{cases}$$

[4.29]

We can now compute the (unconditional) expected net benefit for a customer who decides to join, when the others follow a strategy (n_e, q). We denote this quantity by $S^{(owd)}_{ind}(q)$. Then,

$$S^{(owd)}_{ind}(q) = \sum_{n=0}^{\infty} \pi_n S^{(owd)}_{ind-cond}(n|n_e),$$

where (π_n) is the steady-state distribution given in Proposition 4.2 and the quantity $S^{(owd)}_{ind-cond}(n|n_e)$ is computed in Lemma 4.1. Evaluating the corresponding geometric sums (for details, see Burnetas *et al.* (2017)), we deduce the following result:

PROPOSITION 4.3.– Consider the observable-with-delay M/M/1 queue, where the customers follow an (n_e, q) strategy, with $n_e, q > 0$ and $\lambda q \neq \mu$. The expected net benefit of a customer who decides to join is given by

$$\begin{aligned} S_{ind}^{(owd)}(q) = B\left(R - \frac{C}{\mu}\right)\frac{1-\rho_1^{n_e}}{1-\rho_1} - B\frac{C}{\mu}\frac{(n_e-1)\rho_1^{n_e+1} - n_e\rho_1^{n_e} + \rho_1}{(1-\rho_1)^2} \\ +B\left(R - \frac{Cn_e}{\mu} + \frac{C}{\theta}\right)\frac{\mu\rho_1^{n_e}}{\mu+\theta-\mu\rho_2} - B\frac{C}{\theta}\frac{\rho_1^{n_e}}{1-\rho_2},. \end{aligned} \qquad [4.30]$$

where B, ρ_1 and ρ_2 are given in proposition 4.2.

Furthermore, using coupling arguments and basic properties of stochastic orders, it has been shown that the expected net benefit $S_{ind}^{(owd)}(q)$ is a strictly decreasing function of q, which shows an avoid-the-crowd behavior in the model. Using this fact, it is easy to show the existence and uniqueness of the equilibrium strategy.

More concretely, we consider a tagged arriving customer and assume that the strategy (n_e, q) is employed by the other customers. Let $\tilde{q}_e$ be the root of $S_{ind}^{(owd)}(q)$. Then, the analysis proceeds along the same lines with the heterogeneously observable model of section 4.7. The set of best responses against (n_e, q), $BR((n_e, q))$, is (as far as $q \in [0, 1]$)

$$BR((n_e, q)) = \begin{cases} \{(n_e, 0)\}, & \text{if } q > \tilde{q}_e, \\ \{n_e\} \times [0, 1]\,, & \text{if } q = \tilde{q}_e, \\ \{(n_e, 1)\}, & \text{if } q < \tilde{q}_e. \end{cases}$$

We can now proceed to the computation of the equilibrium strategies:

The strategy $(n_e, 0)$ is equilibrium strategy, if and only if $(n_e, 0) \in BR((n_e, 0))$, i.e. $0 \geq \tilde{q}_e$, which reduces to $S_{ind}^{(owd)}(0) \leq 0$.

A strategy (n_e, q_e) with $q_e \in (0, 1)$ is equilibrium strategy, if and only if $(n_e, q_e) \in BR((n_e, q_e))$, i.e. $q_e = \tilde{q}_e$. This is valid as far as $\tilde{q}_e \in (0, 1)$, which occurs if and only if $S_{ind}^{(owd)}(1) < 0 < S_{ind}^{(owd)}(0)$.

Finally, the strategy $(n_e, 1)$ is the equilibrium strategy, if and only if $(n_e, 1) \in BR((n_e, 1))$, i.e. $1 \leq \tilde{q}_e$, which reduces to $S_{ind}^{(owd)}(1) \geq 0$.

In summary, we have the following result:

THEOREM 4.10.– For the join-or-balk and stay-or-renege customer dilemmas in the observable-with-delay M/M/1 queue, a unique equilibrium strategy (n_e, q_e) exists.

The threshold n_e for reneging at the first announcement instant following customers' arrivals is Naor's threshold given from [4.2]. The join probability q_e is given by the formula

$$q_e = \begin{cases} 0, & S_{ind}^{(owd)}(0) \leq 0, \\ \tilde{q}_e, & S_{ind}^{(owd)}(1) < 0 < S_{ind}^{(owd)}(0), \\ 1, & S_{ind}^{(owd)}(1) \geq 0, \end{cases}$$

where $\tilde{q}_e$ is the unique root of $S_{ind}^{(owd)}(q)$ given from [4.30].

For the social optimization and the profit maximization in this model, the administrator may impose two kinds of fees: an admission fee paid by all customers that decide to join the system and a service fee paid by those customers who do receive service (i.e. those who do not abandon before being served). Burnetas *et al.* (2017) studied these problems using a combination of theoretical results and numerical experiments. The more important managerial take-away messages that have been deduced from the analytical results and the numerical experimentation are discussed in the sequel.

An important finding is that the equilibrium throughput of the system is a unimodal function of the announcement rate θ. Thus, if the administrator of the system is interested in maximizing the throughput (for example if he receives an exogenous payment per served customer), then there is an ideal announcement rate to achieve this objective. The optimal announcement rate lies strictly between 0 and ∞. In other words, some delay in providing information to the customers is beneficial in terms of throughput.

In the unobservable model, (Edelson and Hildebrand 1975) showed that the socially optimal join probability is always smaller than or equal to the equilibrium join probability (because of the negative externalities – see the comment just after theorem 4.2). However, in the observable-with-delay model this is no longer valid: The socially optimal join probability may be greater than the equilibrium join probability in some cases. It is important to stress here that this not due to positive externalities, but due to n_{soc} being less than or equal to n_e.

Finally, the possibility of using two prices for the entrance and the service of the customers may be efficient for coordinating a given system. Even if this is not possible, this pricing mechanism can induce an almost efficient system, where the equilibrium behavior of the customers is very near to the socially desirable.

Another model with delayed observation characteristics is the so-called "armchair decision" problem introduced by (Hassin and Roet-Green 2014) (see also Roet-Green (2013)). In this model, the customers observe the queue length before reaching it, using probably some web-based application. Then, they decide whether to leave their armchair and go to the service facility or not, but when they arrive at the system they are informed about the current queue length and should make their second decision to join or balk.

4.9. Conclusions and literature review for further study

The advancement of technology and its incorporation in the contemporary service systems have provided a great flexibility to the administrators of the systems to share real-time information about congestion with the customers. Yet, given the strategic and usually selfish nature of the customers, the provision of more information is not always beneficial for the administrator or for them. In this work, we presented a number of models that lie between full information and no information for the customers and showed that the optimal level of information is strictly between these two extremes, in most cases. However, much more work remains to be done. In particular, it seems important to extend the studies for the comparison of information levels in more general models. This will enable to drive robust conclusions about the effect of information provision in service systems. Another important issue is the study of pricing for information acquisition. The pricing can be used as an additional mechanism for differentiating customers and may be particularly useful for social optimization and profit maximization.

Apart from the three basic ideas for bridging observable and unobservable models that have been presented in this chapter, a recent study of (Dimitrakopoulos *et al.* 2019) shows that alternating a system between observable and unobservable periods can be also advantageous for increasing the equilibrium throughput and/or the equilibrium social welfare.

After reading the present short introduction, readers who wish to deepen their understanding of the subject should read the papers that are mentioned in the references, in particular the recent and excellent survey of (Ibrahim 2018) on sharing delay information in service systems and the recent book of (Hassin 2016).

4.10. Acknowledgments

I would like to express my gratitude to my friend and colleague Professor Apostolos Burnetas who introduced me in the field of Strategic Queueing, during a series of seminars, some years ago. I also owe many thanks to Professors Refael Hassin and Moshe Haviv who shaped this field with their monograph (Hassin and Haviv 2003). It is certainly the book that has influenced most my academic career. In addition, various inspiring discussions with Professor Hassin sowed the seed for the present work. Finally, I would like to thank the editors of this volume, Professors

Nikolaos Limnios and Vladimir Anisimov for their kind invitation to write this chapter and an anonymous referee who helped a lot with feedback to improve the content and the presentation of the chapter.

4.11. References

Burnetas, A., Economou, A., Vasiliadis, G. (2017). Strategic customer behavior in a queueing system with delayed observations. *Queueing Syst.*, 86, 389–418.

Chen, H., Frank, M. (2001). State dependent pricing with a queue. *IIE Trans.*, 33, 847–860.

Chen, H., Frank, M. (2004). Monopoly pricing when customers queue. *IIE Trans.*, 36, 569–581.

Dimitrakopoulos, Y., Economou, A., Leonardos, S. (2019). Strategic customer behavior in a queueing system with alternating information structure. Working Paper.

Economou, A., Grigoriou, M. (2015). Strategic balking behavior in a queueing system with a mixed observation structure. *Proceedings of the 10th Conference on Stochastic Models of Manufacturing and Service Operations (SMMSO 2015).* University of Thessaly Press, Volos, 51–58.

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

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

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

Guo, P., Zipkin, P. (2009). The effects of the availability of waiting-time information on a balking queue. *Eur. J. Oper. Res.*, 198, 199–209.

Hassin, R. (1986). Consumer information in markets with random products quality: The case of queues and balking. *Econometrica*, 54, 1185–1195.

Hassin, R. (2016). *Rational Queueing*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

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

Hassin, R., Koshman, A. (2017). Profit maximization in the M/M/1 queue. *Oper. Res. Lett.*, 45, 436–441.

Hassin, R., Roet-Green, R. (2014). The armchair decision: Depart or stay home. Working Paper.

Hassin, R., Snitkovsky, R. (2018). Social and monopoly optimization in observable queues. Working Paper.

Haviv, M., Oz, B. (2018). Self-regulation of an unobservable queue. *Manage. Sci.*, 64, 2380–2389.

Hu, M., Li, Y., Wang, J. (2018). Efficient ignorance: Information heterogeneity in a queue. *Manage. Sci.*, 64, 2650–2671.

Ibrahim, R. (2018). Sharing delay information in service systems: A literature survey. *Queueing Syst.*, 89, 49–79.

Kim, B., Kim, J. (2017). Optimal disclosure policies in a strategic queueing model. *Oper. Res. Lett.*, 45, 181–186.

Kirstein, B.M. (1976). Monotonicity and comparability of time-homogeneous Markov processes with discrete state space. *Math. Operationsforsch. Statist.*, 7, 151–168.

Kulkarni, V.G. (2010). *Modeling and Analysis of Stochastic System*, 2nd ed. CRC Press, Taylor and Francis Group, Boca Raton, FL.

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

Roet-Green, R. (2013). Information in queueing systems with strategic customers. PhD Thesis, School of Mathematical Sciences, Tel-Aviv University.

Shone, R., Knight, V.A., Williams, J.E. (2013). Comparisons between observable and unobservable M/M/1 queues with respect to optimal customer behavior. *Eur. J. Oper. Res.*, 227, 133–141.

Simhon, E., Hayel, Y., Starobinski, D., Zhu, Q. (2016). Optimal information disclosure policies in strategic queueing games. *Oper. Res. Lett.*, 44, 109–113.

Stidham, S. Jr. (1985). Optimal control of admission to a queueing system. *IEEE T. Automat. Contr.*, 30, 705–713.

Stidham, S. Jr. (2009). *Optimal Design of Queueing Systems*. CRC Press, Taylor and Francis Group, Boca Raton, FL.