Production and Operations Management

Increasing social welfare with delays: strategic customers in the M/G/1 orbit queue

Opher Baron*
Rotman School of Management, University of Toronto,
105 St George street, Toronto, ON, Canada, M5S 3E6,
Phone: +1-416-9784164, Email: opher.baron@rotman.utoronto.ca

Antonis Economou, Athanasia Manou
Department of Mathematics, National and Kapodistrian University of Athens,
Panepistemiopolis Zografou, Athens, Greece, 15784,
Phone: +30-2107276351, +30-2107276405, Email: aeconom@math.uoa.gr, amanou@math.uoa.gr

Strategic customers typically patronize service systems at a higher rate than the socially optimal one. Much literature has focused on inducing customers to join such systems at this latter rate. This entire literature considers non-idling policies that are the focus of queueing theory. We demonstrate that strategically imposing delays into service systems can improve the social welfare using the M/G/1 queue with orbit. This versatile queueing model has been extensively studied from a performance evaluation perspective. In this system, customers who arrive and find the server idle begin immediately their service. However, strategic customers who find the server busy decide whether to balk or join a virtual queue, i.e., an orbit. Then, each time the server finishes a service, he begins to retrieve a customer from the orbit and the corresponding retrieving time is not negligible. These retrieving times function as extra delays that are imposed on customers that find a busy server. We show that when customers are strategic there are certain ranges of the parameters where delaying the orbit customers can increase the welfare of a system or even maximize it. To this end, we characterize and compute the equilibrium strategies for the customers' joining/balking dilemma. We consider both the unobservable and observable versions of the system, and provide some insight on the optimal delay and level of information in such systems. We further show that the welfare for this system is higher than the corresponding standard M/G/1 queue with the same delay.

Key words: strategic delays; social welfare optimization; customer behavior; queue with orbit; equilibrium strategies

History: Submitted: January 2021. Accepted: March 2022, by Pinedo, Michael, after 2 revisions.

1 Introduction

In this day and age with intense competition and improved access to information on lead times in many industries, it is of paramount importance for firms and organizations to consider customers' strategic behavior. Indeed, since the pioneering work of Naor (1969), there is a growing literature discussing the decisions of strategic customers in queues. A significant conclusion in this research

* corresponding author

thread is that in most cases the equilibrium behavior of utility maximizing customers results in a usage rate that is higher than the welfare-maximizing usage rate. Therefore, much literature has been dedicated to finding queueing control policies that could coordinate the system, i.e., induce the customers to join at a rate that maximizes the social welfare. As with most studies in Queueing Theory, the literature studied this issue in the context of work-conserving policies. For example, Haviv and Oz (2018) include this as a natural requirement of welfare maximizing regulation schemes. However, recent literature demonstrates that idling servers may have positive impact on the operation of some queueing systems. For example, the pioneering work of Afeche (2013) on strategic delays shows that delaying patient customers can benefit a revenue-maximizing firm. Moreover, Baron et al. (2013) and Baron et al. (2017) demonstrate that strategically idling servers in a queueing network can be and is used by firms to improve their perceived service level. These facts bring up the question of "could a non-work-conserving service discipline improve the social welfare of a system with strategic customers (or even coordinate it)?".

In this paper we consider this question and demonstrate that, under some circumstances, the answer is positive. Because in practice it is socially optimal for customers who face an idle server to join, delaying only waiting customers is an intuitively appealing method to introduce delay that would maximize the social welfare. Therefore, we focus our study on the M/G/1 queue with orbit. This queueing model with a non-work-conserving discipline imposes additional delays (retrieving times) on queued customers. We characterize customer equilibrium behavior in this system, and demonstrate how the delays can improve the welfare generated by the system.

Queues with orbit have been used to model situations where long waiting lines occur and customers' cost of virtual wait in an orbit is lower than the cost of wait in a physical queue. Examples include newly launched hi-tech products such as smart-phones and game consoles. Then, manufacturers sign up customers and allow them to make online orders. Also, queues with orbit can be used to model situations where physical waiting is impossible. Examples range from popular restaurants to day-care centers that often have waiting lists for several months in advance. Models for queues with orbit assume the following: Customers submit their petitions for service or products remotely and the server begins to serve them immediately, if he is idle. Otherwise, their petitions are queued in a virtual queue that is referred to as the *orbit.* Upon finishing a service, the server starts to retrieve a customer from the orbit to begin a new service. This retrieving time is usually non-negligible, because of processes that should be carried out before the beginning of a service. Moreover, in case of the arrival of a new customer the retrieving process is stopped and the server begins to serve the new arrival right away.

Apart from its practical relevance, the study of the model with orbit seems significant from a theoretical viewpoint. Indeed, as we will see later in the paper, the model which allows interruptions of the retrieving process outperforms in terms of equilibrium social welfare the corresponding model where interruptions are not allowed. The possibility of interruptions of the retrieving times by newly arriving customers makes the customers more reluctant to join when the server is busy (so it decreases the throughput). This leads customers to adopt an equilibrium behavior that is closer to the socially desirable one than the corresponding model without interruptions. In other words, this operation mode makes the customers to 'internalize' to some degree the negative externalities they induce on future arrivals.

We study this M/G/1 queue with orbit and general retrieving times in the presence of strategic customers and make three important contributions: First, we provide a comprehensive theoretical performance and game-theoretic analysis of this queue in both the observable and unobservable cases. Second, based upon numerical results, we provide insights on the optimal management of such queues. These results motivate the third and main contribution of the paper. Namely, we investigate the optimal, from a social welfare point of view, server idling duration for the M/M/1 queue, where idling is implemented as retrieving time from the orbit. We discuss these three contributions next:

1. *Comprehensive theoretical analysis of the M/G/1 queue with orbit, facing strategic customers.*

Although queueing models with orbit have been extensively studied in the literature, there are only a few studies for customer strategic behavior in them and resulting performance (see the literature review in section 2). In the present paper, we first study the customer strategic behavior regarding the joining/balking dilemma in a versatile framework that assumes Poisson arrivals and allows generally distributed service and retrieving times. Regarding the information that is provided to arriving customers before their decisions are made, we consider two versions: an unobservable and an observable. In the unobservable version, customers base their decisions only on the knowledge of the operational and economic parameters of the model, whereas in the observable version they are also informed about the number of customers in the orbit. Naturally, in both cases, customers are informed about the state of the server upon arrival, i.e., whether he is idle or busy.

We derive various important performance measures of the M/G/1 queue with orbit under an arbitrary customer joining strategy. The non-work-conserving, overtaking-allowing discipline of this system poses challenging problems in the quantification of customer-related performance measures. The derivations are probabilistic, shed light on the dynamics of this system, and require a radically different analysis than in the standard M/G/1 queue (see subsection 4.1 for the unobservable version and the e-companion (section EC.1) for the observable one).

Based on these measures, we characterize the equilibrium strategies for the joining/balking dilemma of arriving customers at an M/G/1 queue with orbit and general service and retrieving times. We consider both the observable and unobservable cases and develop easily implementable numerical procedures for their computation (see subsection 4.2 for the unobservable version and the e-companion (section EC. 6) for the observable one).

2. *Insights on the management of the M/G/1 queue with orbit facing strategic customers.*

Implementing the aforementioned theoretical results, we present a detailed numerical study of this model. More specifically, we study the effect of the level of information on the strategic behavior of customers and derive qualitative results comparing both models (see section 6). First, we find that a system that has long (resp. short) service times, benefits from not revealing (resp. revealing) the number of customers in the orbit, in terms of social welfare. Second, regarding the influence of retrieving times, revealing the number of customers in orbit is preferable when their mean is small whereas concealing it seems better when their mean is medium.

We further study the effect of various model parameters on customers' strategic behavior. An interesting counter-intuitive finding is that increasing variability of the service times has, in general, a non-monotonous influence on equilibrium social welfare in the observable model (see subsection 6.1), i.e., among systems with the same mean characteristics (arrival rates, mean service and retrieving times), a system with higher service time variance may induce a higher equilibrium social welfare than a system with a lower variance.

The most striking observation we make is that the social welfare is not monotone decreasing in the mean retrieving time. Another important observation is that increasing the variability of retrieving times has always a detrimental effect on the social welfare (see subsection 6.2). Together, these observations motivate that policy makers may intentionally delay the service (even when there are waiting customers) in order to improve the social welfare and that deterministic idling would be optimal.

3. *Study the optimal idling, from a social welfare point of view, in an M/M/1 queue with strategic customers: Adding deterministic retrieving times.*

We use the theory on queues with orbit in the context where a social planner strategically introduces deterministic retrieving times, d, to delay queued (orbit) customers and eventually increase social welfare. During these strategic retrieving times the server is idle even though there are queued customers. Therefore, the policy we suggest is not work-conserving.

For the unobservable M/M/1 queue with orbit, we study the equilibrium social welfare of the system as a function of d and show the rather unexpected result that it is in general non-monotonous. More concretely, it turns out that it is a quasi-convex function: It has a unique minimum at some value $d_{\min}$, it is non-increasing for $d < d_{\min}$, and non-decreasing for $d > d_{\min}$. This finding is interesting from a theoretical viewpoint and useful from a practical viewpoint.

The first implication is that the maximum equilibrium social welfare is achieved either when the delay is 0 or when it is high enough so that no customers join the orbit. This fact seems a bit trivial, but the importance of the quasi-convexity becomes evident when the maximization of the social welfare is done subject to side constraints. For example, it is reasonable that a social planner aims to maximize the social welfare taking into account that the throughput should exceed a given minimum level or that some quality-of-service measure is kept within acceptable limits. Similar considerations are relevant for a for-profit firm. In general the optimization of queueing performance measures subject to constraints is a recurrent theme in the literature, see e.g., Legros (2018), or Ewaisha and Tepedelenlioglu (2013). Such constraints imply that the maximization is done for delays d taking values in intervals other than over $[0,\infty)$. Then, our results provide the exact value of d that would maximize the social welfare. In such cases, the highest social welfare subject to the constraints is usually achieved for positive strategic delays ($d>0$), where a fraction of the customers do join the orbit.

When the delay is not a decision variable but a natural parameter of the system, its domain may be bounded below. In this case, the result can be applied to obtain an easily computable estimate for the worst-case scenario and in general for the possible values of the social welfare.

Our results enable the complete study and full characterization of the optimal delay d. In particular, we characterize the range of parameters where the equilibrium social welfare can be increased or even maximized (i.e., the system can be coordinated) using an appropriate d (see Theorem 2). Specifically, we show that the critical parameters for the coordination of the system using strategic idling are the normalized service value ν and the congestion intensity ρ. Then, the positive quadrant of the $\nu-\rho$ plane is partitioned into 5 regions: In region A (*very low service value*), the system with no delay is coordinated; in region B (*low service value*), the system can be coordinated by imposing appropriate delays; in region C (*medium service value & somewhat congested system, i.e.,* $\rho<1$, or *a very congested system, i.e.,* $\rho>1$), the social welfare can be increased by imposing appropriate delays, but coordination is not possible; in region D (*high service value & somewhat congested system*), the social welfare cannot be increased by imposing appropriate delays and coordination is not possible; and in region E (*very high service value and somewhat congested system*), the system is coordinated without imposing delays. These regions are depicted in Figure 2b.

Comparing numerically the M/M/1 queue with orbit to the corresponding system, where interruptions of retrieving times by newly arriving customers are not allowed, we found that for any fixed delay, d, the equilibrium social welfare for the M/M/1 queue with orbit exceeds the equilibrium social welfare for the corresponding system without interruptions. We also consider the case where d is not fixed, but is maximizing the equilibrium social welfare under the constraint that the equilibrium throughput exceeds a minimum level. Again, in this case, the equilibrium social welfare under the optimal delay in the M/M/1 queue with orbit exceeds the equilibrium social welfare under the optimal delay in the corresponding system without interruptions.

For the observable M/M/1 queue with orbit where d is a control variable, we study the monotonicity of the equilibrium social welfare and prove structural results that enable the efficient computation of the corresponding optimal d (see the e-companion (section EC.8)). As in the unobservable counterpart, a non-zero d may be better in terms of the equilibrium social welfare. In particular, similar regions exist for the observable model as well, e.g., the social welfare of the system can be improved and in some regions the optimal choice of retrieving time can coordinate the system.

The rest of the paper is structured as follows: In section 2 we discuss previous literature. In section 3 we present the model of the single-server queue with orbit, the associated reward-cost structure for the customers and the decision framework. In section 4 we consider the unobservable version of the model, derive the necessary performance evaluation measures and proceed to the characterization and computation of the equilibrium customer strategies regarding joining/balking. In section 5, we consider the unobservable M/M/1 queue with orbit and deterministic retrieving times and study the equilibrium throughput and the equilibrium social welfare as functions of the duration of the retrieving time. In particular we discuss the optimization of the equilibrium social welfare and the coordination of the system. In section 6 we implement the theoretical results and provide an extensive numerical study of the effect of the service and retrieving times, the arrival rate and the information levels on customer behavior. In section 7 we conclude with several qualitative results and take-away messages from the study, and various directions for future research. The performance analysis and the study of the equilibrium strategies for the observable model, as well as the technical proofs of the various results are included in the e-companion.

2 Literature review

The study of strategic customer behavior in queueing systems was pioneered by Naor (1969) who studied the equilibrium, socially-optimal and profit-maximizing joining strategies for the observable M/M/1 queue, i.e., when the customers observe the queue length upon arrival, before making their joining/balking decisions. Edelson and Hildebrand (1975) considered the unobservable counterpart. Since then, there is a growing literature on strategic customer behavior for Markovian queues, usually as extensions of the M/M/1 queue, see e.g., Hassin and Haviv (1997) (M/M/1 queue with priorities), Burnetas and Economou (2007) (M/M/1 queue with setup times), Guo and Zipkin (2007) (M/M/1 queue with non-linear cost structure), Guo and Hassin (2011) (M/M/1 queue with vacations and the N-policy) etc. The book of Hassin and Haviv (2003) is the classical introduction in this subfield of Queueing Theory. Stidham (2009) is a relevant reference that summarizes various results on the optimal design of queueing systems, taking into account the strategic nature of the customers. Hassin (2016) presents a detailed and accurate overview of queueing papers with strategic considerations for the last 15 years.

The role of the level of information that is provided to customers on their strategic behavior and on the corresponding equilibrium throughput and social welfare is a central issue in this branch of queueing. Some key studies in this direction are the papers by Chen and Frank (2004), Hassin (1986), and Hassin and Roet-Green (2017). Chen and Frank (2004) compared the unobservable and the observable versions of an M/M/1 queue and reported several interesting findings, e.g., they showed that for low (high) arrival rates it is better to conceal (reveal) information from (to) the customers to increase throughput. Hassin (1986) considered the same models and compared the social welfare and the profit when a profit-maximizing strategy is imposed to the customers. He showed that a profit-maximizer prefers to conceal (reveal) the queue length for low (high) values of the arrival rate. Hassin and Roet-Green (2017) studied an M/M/1 queue where customers join, balk or pay for inspecting the queue length and then decide whether to join or balk. They proved the existence of an equilibrium and studied the effect of pricing on social welfare. Baron, Chen, and Li (2022) investigated the impact of different information available for customers who patronize service with an observable, walk in, and a non observable, online channels. They showed that online ordering inadvertently reduces customers' individual utility and social welfare when both channels are used in equilibrium. A thorough overview of the contributions in this area can be found in Hassin (2016) Chapter 3, Ibrahim (2018), and Economou (2021).

One of the main issues in the study of strategic customer behavior is the comparison of the join/balk decisions that are aligned with (i) customers maximizing their own utility, (ii) a central planner maximizing social welfare, and (iii) a monopolistic service operator that maximizes his revenue or profit. It has been shown that, typically, the resulting arrival rates are such that the monopolist's problem results in the lowest one, followed by the socially optimal one that is in the middle, and followed by the utility maximizing arrival rate that is the highest. This fact is also known as Naor's inequality. Broadly speaking, the intuition behind this inequality is that customers ignore their marginal cost on the society, as they only consider the cost they are facing, and a monopolist charges a higher price that restrict access to a lower than socially-optimal segment of the population. Hassin and Snitkovsky (2020) have recently proved sufficient conditions for Naor's inequality in a general framework.

In view of these results, there has been a lot of research dedicated to the search of control policies that induce utility-maximizing customers to act in accordance to the socially optimal policy. The standard method, introduced by Naor (1969), is to charge customers for joining the queue. Nevertheless, as highlighted by Haviv and Oz (2016) and Haviv and Oz (2018), pricing mechanisms require the establishment of additional infrastructure to collect fees, and are not robust to different system parameters, such as the arrival and service rates, the waiting cost, or the benefit to customers from service completion. Therefore, coordination mechanisms that do not involve prices are of interest both from theoretical and practical viewpoints. These papers, as well as the whole relevant literature, focus only on work-conserving policies, to the best of our knowledge. With this in mind, an important question comes up: Is it possible to increase the social welfare in the presence of strategic customers in a queue with a non-work-conserving discipline? The M/G/1 queue with orbit seems a natural framework for answering this question as the model is simple and widely studied in the literature and has the main characteristics that we want to consider, i.e., it operates under a non-work-conserving discipline that allows overtaking.

There is an extensive literature regarding the performance evaluation of queues with orbits, i.e., retrial queues (see e.g., the books of Falin and Templeton (1997), Artalejo and Gomez-Corral (2008) and the references therein). However, there are only a few papers that deal with strategic customer behavior in such systems, see e.g., the papers by Elcan (1994), Kulkarni (1983), Hassin and Haviv (1996), Wang and Zhang (2013), and Cui, Su, and Veeraraghavan (2014). Most of the literature in retrial queues is devoted to the classical retrial policy, according to which each retrying customer conducts attempts independently of other customers. But, there are many situations where it is more reasonable to assume that only the customer at the 'head' of the orbit retries. This case is equivalent to a policy where the server imposes additional delays before picking up a customer from the (head of) the queue. This is the focus of the present paper, where the server's delay is equivalent to assuming that the server spends some time to retrieve a customer from the orbit after every service completion. This retrial policy is known as the constant retrial policy.

The performance evaluation of queueing systems with orbit is challenging because they do not satisfy two of the most common assumptions in the queueing literature. First, as stated above, these queues are not work-conserving. Second, these queues are not overtake-free. That is, as in queueing systems with priorities, some, late arriving, customers may overtake other, early arriving, ones. These two differences imply that standard queueing tools are either not valid (i.e., conservation laws) or should be very carefully used in the analysis of queues with orbit. The first analysis of constant retrial queues in Fayolle (1986) considered a Markovian framework. The study of customer strategic behavior in queues with constant retrials was initiated by Economou and Kanta (2011) who considered the M/M/1 constant retrial queue and derived the equilibrium, socially-optimal and profit-maximizing joining strategies for the customers and several associated results. Various authors considered more involved Markovian extensions of this model, see e.g., Zhang, Wang, and Zhang (2014), Wang, Zhang, and Huang (2017), Do, Do, and Melikov (2020), and Li and Wang (2021). A related study concerns a model with orbit and strategic customers by Engel and Hassin (2017). In that model, there are two waiting lines, a regular queue (system queue) and an orbit (virtual queue). The arriving customers who find a busy server decide which queue to join. Customers in the system queue have non-preemptive priority over those in the virtual queue, but waiting in the service queue is more costly. The framework is Markovian as well.

However, there are no studies on the customer strategic behavior in non-Markovian queues with orbit. Therefore, an additional key objective of this paper is to fill this gap in the literature, as queueing systems with orbit, generally distributed service times and non-negligible generally distributed retrieving times appear in various contexts as we have described above. The study of the customer strategic behavior for the model under consideration requires the performance evaluation of its state-dependent version which has been recently carried out by Baron, Economou, and Manou (2018). In general, the study of strategic behavior in non-Markovian queueing systems with generally distributed service times is a subtle issue and the first studies appeared only recently. Altman and Hassin (2002) showed that, for certain service time distributions, threshold strategies cannot be equilibrium strategies for the joining/balking dilemma in the observable M/G/1 queue. Another counter-intuitive strategic behavior was described by Haviv and Kerner (2007) for a partially observable M/G/1 queue. However, the complete characterization and computation of the equilibrium strategies for the fundamental model of the observable M/G/1 queue was carried out only recently by Kerner (2011), based on the performance evaluation of its state-dependent version which was developed in Kerner (2008), using the supplementary variable method. This methodology was extended by Abouee-Mehrizi and Baron (2006). Economou and Manou (2015) and Manou, Economou, and Karaesmen (2014) used a probabilistic approach for the performance evaluation of state-dependent non-Markovian systems and the study of strategic customer behavior in them. Recently, Oz, Adan, and Haviv (2020) developed an alternative powerful methodology for the same problem based on a rate balance principle.

3 Queueing dynamics and decision framework

We consider the M/G/1 queue with orbit described as follows: Customers arrive according to a Poisson process at rate λ. There is a single server who serves them, one at a time, with no waiting space. Customers who find the server free immediately enter service, whereas customers that arrive when the server is busy can join an orbit. When the server completes a service and there are customers in the orbit, the next service is not started immediately, but there is a delay. This delay can be thought of as the time for retrieving the next customer from the orbit.

Customers in the orbit are retrieved according to the FCFS discipline. If the retrieving time (delay) is completed before the next arrival, the server begins to serve the oldest customer from the orbit. Otherwise, i.e., when a new customer arrives during the retrieving time, the retrieving process is interrupted and the new customer enters service. The retrieving process begins anew the next time the server is free.

The arrival process, the service times, and the retrieving times are assumed independent. Service times are identically distributed and we denote a typical service time by B, its expectation by $E[B]$, its distribution function by $B(x)$, its Laplace-Stieltjes transform (LST) by $\tilde{B}(s)=\int_0^\infty e^{-sx}dB(x)$, and its probability density function (in case that B has an (absolutely) continuous distribution) by $b(x)$. Similarly, retrieving times are identically distributed and we denote a generic retrieving time by A. Its related quantities are denoted similarly to the service times.

The state of the system at a time t, is described by the random variables $C(t)$, $Q(t)$, and $S(t)$ that record, respectively, the state of the server (0 or 1), the number of customers in the orbit, and the remaining time till the next service/retrieving time completion. More specifically, $C(t)=1$ when the server is busy (serving) at time t, whereas $C(t)=0$ when the server is idle or retrieving at time t, and $S(t)$ records the remaining service time at time t when $C(t)=1$ or the remaining retrieving time at time t when $C(t)=0$. For times t such that $C(t)=Q(t)=0$, the random variable $S(t)$ has no sense and is not defined.

The customers face the dilemma of whether to join or balk upon arrival at the system. We assume that they observe the state of the server before making their decisions, and consider two versions of the model regarding the information they receive about the number of customers in the orbit: In the unobservable version, they base their decisions only on their knowledge of the parameters of the system, whereas in the observable version they are also informed about the number of customers in orbit. Their objective is to maximize their (individual) net utility, which is specified as the expected reward from service minus the expected waiting cost. We assume that customers who decide to join the system do not abandon later, i.e., reneging of entering customers is not allowed.

We note that it is quite natural to assume that customers can observe the state of the server upon their arrival (busy or idle). For example this is done when the customers call and receive an idle or busy signal. Moreover, this is assumed in several models in the literature (see e.g., Haviv and Kerner (2007), Economou and Kanta (2011), and Hassin and Koshman (2017)) and perhaps it is the most natural partial information case.

We assume that customers are homogeneous in their valuations. Each customer receives service reward R for completing the service in the system. Moreover, she accumulates waiting costs at rate K per time unit as long as she stays in the system (either in orbit or in service). Suppose that $R<KE[B]$. Then, the mean waiting cost from the service time of a customer exceeds her reward from service, so the customer is not willing to join, whatever the other customers do. Therefore in this case, the ‘always balk’ strategy is the unique equilibrium joining strategy. Suppose that

$R = KE[B]$. In this case, we have that the mean waiting cost from the service time of a customer equals her reward from service. Therefore, in this case, only customers that find an idle (retrieving) server may join, since they are indifferent between joining and balking. All other customers prefer to balk. Therefore, in this case, the equilibrium joining probabilities are of the form 'join with any probability when finding an idle server and balk otherwise'.

The above reasoning reveals that the only interesting case occurs when

$$R > KE[B]. \tag{1}$$

In this case, a customer who finds an idle server prefers to enter since her expected net benefit is $R - KE[B] > 0$. In the unobservable case, the equilibrium joining strategies will be of the form 'join with probability 1 when finding an idle server and join with probability q when finding a busy server'. Similarly, in the observable case, the equilibrium joining strategies can be only of the form 'join with probability 1 when finding an idle server and join with probability q_j when finding a busy server and j customers in orbit'. From now on, we assume that (1) holds. More specifically, we will focus on symmetric equilibrium strategies, i.e., we seek for equilibrium strategy profiles where all customers use the same strategy. Such symmetric equilibrium is sensible as customers are homogeneous. Since a customer is assumed to join the system if she finds an idle (retrieving) server, in the unobservable case, a customer's joining strategy should be specified by the joining probability q, when the server is found busy upon arrival. On the other hand, in the observable version of the model, a customer's joining strategy is specified by an infinite vector $\mathbf{q} = (q_j : j = 0, 1, 2, \ldots)$, where q_j denotes the joining probability, when the system is found at state $(C(t), Q(t)) = (1, j)$ upon arrival (more precisely just before the arrival).

A moment of reflection shows that when the customers adopt a given joining strategy, the effective arrival process becomes state-dependent. Let $\lambda_{i,j}$ denote the effective arrival rate, when $C(t) = i$ and $Q(t) = j$. Then, in the unobservable case, when a joining strategy q is employed, we have

$$\lambda_{i,j} = \begin{cases} \lambda & \text{if } i = 0, j \geq 0, \\ \lambda q & \text{if } i = 1, j \geq 0. \end{cases} \tag{2}$$

Similarly, in the observable case, when a joining strategy $\mathbf{q} = (q_j : j = 0, 1, 2, \ldots)$ is adopted by the customers, the arrival rates become

$$\lambda_{i,j} = \begin{cases} \lambda & \text{if } i = 0, j \geq 0, \\ \lambda q_j & \text{if } i = 1, j \geq 0. \end{cases} \tag{3}$$

Therefore, to study the strategic behavior of the customers in both information cases, we consider the stochastic process $\{(C(t), Q(t), S(t)) : t \geq 0\}$ that corresponds to the original model with general state-dependent arrival rates $\lambda_{i,j}$. This is a continuous time Markov process with state space

$\{(0,0)\} \cup \{(0,j,x) : j = 1,2,\ldots; x \geq 0\} \cup \{(1,j,x) : j = 0,1,\ldots; x \geq 0\}$. We denote by $S_{i,j}$ a random variable that represents the limiting conditional distribution of $S(t)$ given that $(C(t),Q(t)) = (i,j)$, for $(i,j) \neq (0,0)$. In other words, $S_{1,j}$ represents the conditional remaining service time given that there are j customers in orbit, whereas $S_{0,j}$ represents the conditional remaining retrieving time given that there are j customers in orbit. Note that because of the conditional Poisson Arrivals See Time Averages (conditional PASTA) property (see van Doorn and Regterschot (1988)), the steady-state distribution of each $S_{i,j}$ is the same for arrival instants and for arbitrary instants in continuous time. For the unobservable case, we need the limiting conditional distribution of $S(t)$ given that $C(t) = i$, which we denote by S_i, for $i = 0,1$.

Using the results of Baron, Economou, and Manou (2018), we can compute the LSTs and the expected values of the random variables $S_{i,j}$, the steady-state probabilities $p_{i,j} = \lim_{t\to\infty} \Pr[C(t) = i, Q(t) = j]$, for $(i,j) \in \{0,1\} \times \{0,1,2,\ldots\}$ and the steady-state rates $\mu_{i,j}$ of service/retrieving time completions per time unit, that initiated from state $(C(t),Q(t)) = (i,j)$, for any given strategy $\mathbf{q}$ of the customers. For easy reference to the corresponding numerical schemes, we provide them in the e-companion (section EC.1). A moment of reflection on these schemes shows that $\tilde{S}_{0,j}(s)$, for $j \geq 1$, does not depend on the strategy $\mathbf{q}$. Moreover, $\tilde{S}_{1,j}(s)$, for $j \geq 0$, depends on strategy $\mathbf{q}$ only through $\mathbf{q}_j = (q_0, q_1, \ldots, q_j)$. In addition, we stress that the computations of the LSTs $\tilde{S}_{i,j}(s)$ do not need the computation of $p_{0,0}$, which requires the use of the normalization equation. It is exactly these facts that enable us to compute the equilibrium customer strategies recursively in the observable case. For the unobservable case, we can also exploit these facts to compute the equilibrium customer strategies, but a more direct probabilistic approach is possible using mean value analysis.

4 Equilibrium strategies in the unobservable model

4.1 Expected net benefit of a customer that finds a busy server

In this section we consider the unobservable model where an arriving customer observes the server state (idle or busy), but does not get any information about the number of customers in the orbit. We assume that condition (1) holds. Under this condition, customers who find the server idle prefer to join, so the only meaningful joining strategies in the unobservable case will be of the form 'join with probability 1 when finding an idle server and join with probability q when finding a busy server'. We denote by $S^{(u)}(q)$ the expected net benefit of a tagged customer that finds a busy server upon arrival and decides to join, when the others follow the above strategy. Then, we have

$$S^{(u)}(q) = R - K(E[S_1] + E[U_0]), \tag{4}$$

where S_1 is the remaining service time of the customer currently in service at the arrival instant of the tagged customer and U_0 denotes the time until the service completion of the tagged customer, after the beginning of the next retrieving time (equivalently after the end of the current service time). Denoting by Q^- the number of customers in the orbit just before the arrival instant of the tagged customer and by C^- the state of the server, we have

$$E[U_0] = (E[Q^- | C^- = 1] + 1)E[U_{0,1}], \tag{5}$$

where $U_{0,1}$ is the time until the service completion of the first (oldest) customer in orbit given that the server is idle and a retrieving time has started. Indeed, a customer that arrives when the server is busy ($C^- = 1$) has to wait for the service in process, and then $Q^- + 1$ periods distributed as $U_{0,1}$, for moving from the $Q^- + 1$ position of the orbit to the outside of the system. In general, if we denote by $U_{0,j}$ the time until the service completion of the j-th customer in the orbit given that the server is idle and a retrieving time has just started, then $U_{0,j}$ is the sum of j independent random variables distributed as $U_{0,1}$.

In light of (4) and (5), to compute $S(q)$ we need to compute $E[S_1]$, $E[U_{0,1}]$ and $E[Q^- | C^- = 1]$. The computations of $E[S_1]$, $E[U_{0,1}]$ are straightforward, using simple probabilistic arguments:

Lemma 1. *The expected remaining service time of the customer currently in service at the arrival instant of a tagged customer is given by*

$$E[S_1] = \frac{E[B^2]}{2E[B]}. \tag{6}$$

The expected time for the service completion of the j-th customer in orbit given that the server has just started a retrieving time is given by

$$E[U_{0,j}] = j\frac{1}{\tilde{A}(\lambda)}\left(\frac{1-\tilde{A}(\lambda)}{\lambda} + E[B]\right), \tag{7}$$

where $\tilde{A}(\lambda) = \int_0^\infty e^{-\lambda x} dA(x)$.

The result in (6) states that a random tagged customer arriving while the server is busy observes the expected equilibrium residual service time. The proof follows by gluing the busy periods of the server and standard results. In fact the entire distribution of the remaining service time observed by a tagged customer is the equilibrium residual service time distribution. The result in (7) follows by first observing that $E[U_{0,j}] = jE[U_{0,1}]$ (with the same reasoning as in (5)). Then, the expected time for the first such customer, $E[U_{0,1}]$, is computed by conditioning on the first event to occur after the start of a retrieving time, being an arrival or the end of the retrieving time. In particular, we use that the mean time to an arrival or the end of a retrieving time, given that the retrieving time has just started is

$$\int_0^\infty (1 - A(x))e^{-\lambda x} dx = \frac{1-\tilde{A}(\lambda)}{\lambda} \tag{8}$$

and the probability that the retrieving time ends before the arrival is

$$\int_0^\infty e^{-\lambda x} dA(x) = \tilde{A}(\lambda).$$

For details see the proof in the e-companion (section EC.2).

It remains to determine $E[Q^-|C^-=1]$. To this end, let C and Q be respectively the state of the server and the number of customers in orbit in steady-state (at an arbitrary instant), and S the sojourn time of an arriving customer. Let $E[Q|C=i]$ denote the mean number of customers in orbit given that the server's state is i, $\Pr[C=i]=p_i$ denote the steady-state probability that the server is at state i, and $E[S|C^-=i,\text{join}]$ denote the expected sojourn time of an arriving customer that finds the server at state i and joins the system, $i=0,1$. Note that, by the conditional PASTA property (see van Doorn and Regterschot (1988)) we have that $E[Q^-|C^-=i]=E[Q|C=i]$, $i=0,1$. Thus, one way to compute $E[Q^-|C^-=i]$ is by using the formula $E[Q^-|C^-=i]=\frac{\sum_{j=0}^\infty j p_{i,j}}{p_i}$, where the steady-state probabilities $p_{i,j}$ and $p_i=\sum_{j=0}^\infty p_{i,j}$ are computed using the recursive schemes that are mentioned in the e-companion (section EC.1) with $q_j=q$. As a byproduct of this computation, we also obtain the stability condition of the system when the joining probability q is employed: The system is stable (i.e., the process $\{(C(t),Q(t))\}$ is positive recurrent), if and only if

$$\tilde{A}(\lambda)-\lambda q E[B]>0. \tag{9}$$

This approach requires lengthy calculations, using a similar path as in the proof of Theorem 5.4. in Baron, Economou, and Manou (2018). Therefore, we choose to present here the computation using a more instructive probabilistic approach, based on mean value analysis. Mean value analysis has been also applied successfully for the study of customers' equilibrium behavior in other unobservable non-Markovian queueing systems (see e.g., Economou, Gomez-Corral, and Kanta (2011)). We stress that the mean value analysis here is significantly different from the mean analysis of the M/G/1 queue (see e.g., Adan and van der Wal (2010) Section 13.3.2). Indeed, note that the M/G/1 with orbit model is not equivalent to an M/G/1 queue where the service times are the sum of a delay and a service time and an exceptional first service time in a busy period (only service without delay). Specifically, in the latter system only arrivals when the system is empty enter service immediately, whereas in the former both arrivals when the system is empty and during the delay enter service immediately. So, arrivals may have an exceptional service time even if they arrive in the midst of a busy period. In short, an easy modification of the mean value analysis of the M/G/1 system cannot apply in the present framework because our system is not work-conserving and allows customers to overtake customers that arrived before them. We begin the mean value analysis by determining the probabilities p_0 and p_1.

Lemma 2. *The steady-state probabilities of idle/retrieving and serving server are given respectively by*

$$p_0 = \frac{1-\lambda q E[B]}{1-\lambda q E[B]+\lambda E[B]} \tag{10}$$

and

$$p_1 = \frac{\lambda E[B]}{1-\lambda q E[B]+\lambda E[B]}. \tag{11}$$

The results in Lemma 2 follow by applying Little's law for the server using the effective arrival rate, i.e., $p_1 = (\lambda p_0 + \lambda q p_1)E[B]$, and the normalization equation. We then apply Little's law to the whole system, and obtain

$$E[Q+C] = \lambda E[S], \tag{12}$$

where

$$E[Q+C] = p_0 E[Q|C=0] + p_1(E[Q|C=1]+1) \tag{13}$$

and

$$E[S] \, = p_0 E[S|C^- = 0] + p_1 q E[S|C^- = 1, \text{join}],$$

since a balking customer has a zero sojourn time. Note that $E[S]$ refers to the mean sojourn time of an arriving customer, whether she enters or not.

Consider, now a customer who arrives, finds the server busy and decides to join the retrial orbit. Using conditional PASTA, we obtain that her mean sojourn time is

$$E[S|C^- = 1, \text{join}] = E[S_1] + (E[Q|C=1]+1)\, E[U_{0,1}].$$

Thus,

$$E[S] \, = p_0 E[B] + p_1 q \left(E[S_1] + (E[Q|C=1]+1)\, E[U_{0,1}]\right). \tag{14}$$

Note that we have three equations, (12), (13), and (14), to determine four unknowns: $E[S]$, $E[Q+C]$, $E[Q|C=0]$, and $E[Q|C=1]$. To obtain an additional equation, we will relate $E[Q|C=0]$ and $E[Q|C=1]$. To this end we will prove a lemma that is of independent interest.

Lemma 3. *(i) The steady-state probability of empty orbit after a service completion, r_e, is given by*

$$r_e = 1 - \frac{\lambda E[B] q}{\tilde{A}(\lambda)}. \tag{15}$$

(ii) The mean time between two successive service completions is

$$E[D] = r_e \frac{1}{\lambda} + (1-r_e)\frac{1-\tilde{A}(\lambda)}{\lambda} + E[B].$$

(iii) The expected duration of a (server's) idle/retrieving period is

$$E[I] = \frac{1-\lambda E[B]q}{\lambda}. \tag{16}$$

(iv) The conditional expectations of the orbit length given the state of the server are related through the equation

$$E[Q|C=0] = \frac{1-\tilde{A}(\lambda)}{\tilde{A}(\lambda)}\frac{p_1}{p_0}q(E[Q|C=1]+1). \tag{17}$$

(v) The expected number of customers in the orbit given that the server is busy is

$$E[Q|C=1] = \frac{\lambda q\left(-2\tilde{A}(\lambda)E^2[B]+\tilde{A}(\lambda)E[B^2]+2E^2[B]\right)}{2E[B]\left(\tilde{A}(\lambda)-\lambda E[B]q\right)}. \tag{18}$$

Part (i) of Lemma 3 provides an intuitive justification of the stability condition (9). Indeed, the process $\{(C(t),Q(t))\}$ is positive recurrent, if and only if the system visits state $(0,0)$ infinitely often and the mean inter-visit time is finite. But, each inter-visit time to state $(0,0)$ can be decomposed to intervals between successive service completions of finite mean length with the average inter-completion time, $E[D]$, given in part (ii). The expression for $E[D]$ follows as the average service time plus the average time to the next service start (i.e., $\frac{1}{\lambda}$ with probability r_e, when the orbit is empty after the previous service completion, or the expression in (8) with probability $1-r_e$). The system is stable if and only if it remains empty after a service completion with positive probability, i.e., if $r_e>0$, which yields (9). The average inter-completion time, E[D], also includes some portion of periods during which the server is idle (no customers in the system during retrieving times), as given in part (iii), which follows the logic leading to part (ii) and some algebra. The proof of part (iv) relies on part (iii) and that, since the customers arrive and depart one by one, the distributions of what is seen by arrivals and by departures are identical (see e.g., Kulkarni (2010) Theorem 7.2). For details see the proof in the e-companion (section EC.2). Lemmas 1 and 3 provide all the required quantities to compute $S(q)$ given in (4).

Proposition 1. *The expected net benefit of an arriving customer who finds a busy server and decides to join, when the other customers follow a strategy q is given by*

$$S^{(u)}(q) = R-K\left(\frac{E[B^2]}{2E[B]}+\frac{(-2\lambda qE^2[B]+\lambda qE[B^2]+2E[B])\left(1-\tilde{A}(\lambda)+\lambda E[B]\right)}{2E[B]\left(\tilde{A}(\lambda)-\lambda E[B]q\right)\lambda}\right), \tag{19}$$

and is (strictly) decreasing in q.

The proof of Proposition 1 follows by substituting (6), (18), and (7) in (4) to get (19) and then veryfying that the derivative with respect to q is strictly negative.

4.2 Equilibrium joining probability

As we have already seen in section 3, the equilibrium joining probability for an arrival that finds an idle server is immediately determined by comparing R and $KE[B]$. We now proceed to find the equilibrium joining probability when the server is found busy. We denote this probability by $q_e^{(u)}$ to emphasize that we refer to the unobservable model. The next theorem shows that this equilibrium probability is unique and is given in closed form. The result is presented under the condition

$$\tilde{A}(\lambda) - \lambda E[B] > 0, \tag{20}$$

i.e., if the system is stable when all customers decide to join. We will see how this result can be adapted in the case where this stability condition fails to hold, in the discussion following Theorem 1.

Theorem 1. *In the unobservable M/G/1 queue with orbit, where the condition* (20) *holds, a unique equilibrium joining probability $q_e^{(u)}$ exists. Let*

$$t^{(u)} = \frac{R}{K} - \frac{E[B^2]}{2E[B]}. \tag{21}$$

Then, we have the following mutually exclusive cases:

1. 0 *is the unique equilibrium joining probability $q_e^{(u)}$, if and only if*

$$t^{(u)} \le \frac{1 - \tilde{A}(\lambda) + \lambda E[B]}{\tilde{A}(\lambda)\lambda}. \tag{22}$$

2. *A unique equilibrium joining probability $q_e^{(u)}$ exists which lies strictly between* 0 *and* 1, *if and only if*

$$\frac{1 - \tilde{A}(\lambda) + \lambda E[B]}{\tilde{A}(\lambda)\lambda} < t^{(u)} < \frac{(-2\lambda E^2[B] + \lambda E[B^2] + 2E[B])\left(1 - \tilde{A}(\lambda) + \lambda E[B]\right)}{2E[B]\left(\tilde{A}(\lambda) - \lambda E[B]\right)\lambda}. \tag{23}$$

In this case, the equilibrium joining probability, $q_e^{(u)}$, is given by

$$q_e^{(u)} = \frac{2\lambda E[B]\tilde{A}(\lambda)t^{(u)} - 2E[B](1 - \tilde{A}(\lambda) + \lambda E[B])}{2\lambda^2 E^2[B]t^{(u)} + (\lambda E[B^2] - 2\lambda E^2[B])(1 - \tilde{A}(\lambda) + \lambda E[B])}. \tag{24}$$

3. 1 *is the unique equilibrium joining probability $q_e^{(u)}$, if and only if*

$$t^{(u)} \ge \frac{(-2\lambda E^2[B] + \lambda E[B^2] + 2E[B])\left(1 - \tilde{A}(\lambda) + \lambda E[B]\right)}{2E[B]\left(\tilde{A}(\lambda) - \lambda E[B]\right)\lambda}. \tag{25}$$

Note that $t^{(u)}$ in (21) is the difference between the relative service reward with respect to the unit waiting cost (R/K) and the expected remaining service time of the customer in service at the arrival instant of a tagged customer (as given in (6), Lemma 1).

When the condition (20) does not hold, the system is unstable for q such that $\tilde{A}(\lambda) \le \lambda q E[B]$. Therefore, case 3 of Theorem 1 never holds. Then, it is easy to see that we only have cases 1 and 2, where the right-hand side inequality of (23) is omitted.

4.3 The model without interruptions

If interruptions of retrieving times of customers in the orbit by newly arriving customers are not allowed, then the unobservable model becomes a standard M/G/1 queue where a generic service time of customers who find an empty system is distributed as B, whereas a generic service time of customers who find a non-empty system is distributed as $A+B$, with A and B being independent. The analysis of this system is easier than the analysis of the M/G/1 queue with orbit which we presented in subsections 4.1 and 4.2. The reason is that overtaking is prohibited in this model. In addition, this standard M/G/1 queue seems a reasonable model since the interruptions may result in customers' frustration that can be translated to lower service reward and decrease in demand. However, as we will see in section 5, the equilibrium social welfare for the M/G/1 queue with orbit for a given delay d always exceeds the equilibrium social welfare for the corresponding standard M/G/1 queue with the same delay parameter. Also, when d is not fixed, but maximizes the equilibrium social welfare subject to a lower bound on the equilibrium throughput, the equilibrium social welfare under the optimal delay for the queue with orbit exceeds the equilibrium social welfare for the corresponding standard M/G/1 queue. Moreover, if the system is unobservable and the retrieval process does not require customer's involvement, then the frustration does not occur.

Therefore, the consideration of the M/G/1 queue with orbit for introducing delays as a means for improving the social welfare is preferable in comparison to the corresponding standard M/G/1 queue (without interruptions). However, for the sake of completeness, we report the main results for the corresponding standard M/G/1 queue, i.e., the analogues of Proposition 1 and Theorem 1, in the e-companion (section EC.3). Using these results, one can carry out the whole analysis of section 5 for the corresponding standard M/G/1 queue with delays and obtain essentially the same structural results (analogues of Propositions 2, 3 and Theorem 2). Operating the standard M/G/1 may be beneficial for systems where the retrieving process involves customers and thus interruptions are frustrating and may be unacceptable. Moreover, in section 5, we will use the results for the model without interruptions to establish that the M/G/1 queue with orbit 'beats' the corresponding standard M/G/1 queue in terms of equilibrium social welfare.

5 Social welfare maximization in the unobservable M/M/1 queue with deterministic retrieving times

5.1 Computation of equilibrium throughput and social welfare.

We focus on the case of an M/M/1 with orbit and deterministic retrieving times. We let the exponential service rate be μ and the deterministic retrieving time be $A=d\geq 0$. The retrieving times can in general be genuine pre-processing times for starting service. However, in this section, they are primarily thought of as added delays to the system with the objective of improving its equilibrium throughput and/or social welfare. We introduce the notation

$$\rho = \frac{\lambda}{\mu} \text{ and } \nu = \frac{R\mu}{K}, \tag{26}$$

i.e., ρ is the congestion intensity and ν is the normalized service reward and note that $\nu > 1$ by (1). We will use the results of the previous section to present an economic analysis of the model. The expected net benefit formula (19) and the right-hand-side quantity in (24) assume respectively the forms:

$$S^{(u)}(q) = \frac{K}{\mu}\left[\nu - 1 - \frac{1 - e^{-\lambda d} + \rho}{\rho(e^{-\lambda d} - \rho q)}\right], \tag{27}$$

$$q^{(u)}(d) = \frac{(\rho(\nu-1)+1)\, e^{-\lambda d} - (1+\rho)}{\rho^2(\nu-1)}. \tag{28}$$

Let $TH^{(u)}(q,d)$ and $SW^{(u)}(q,d)$ be the throughput and the social welfare generated by the system, respectively, given that the customers enter with probability q when finding a busy server and that the retrieving times are equal to d. Then, using (10)-(11), we have that

$$TH^{(u)}(q,d) = \lambda p_0 + \lambda(1-p_0)q, \tag{29}$$

$$SW^{(u)}(q,d) = \lambda p_0\left(R - \frac{K}{\mu}\right) + \lambda(1-p_0)qS^{(u)}(q). \tag{30}$$

We focus on the case where customers act individually, i.e., according to their equilibrium strategy, and wish to explore whether we can increase throughput and social welfare by imposing a retrieving time. Denoting by $q_e^{(u)}(d)$ the equilibrium customer strategy given in Theorem 1, we have that the equilibrium throughput and equilibrium social welfare, as functions of the retrieving time d, are given respectively by

$$TH_e^{(u)}(d) = TH^{(u)}(q_e^{(u)}(d), d) \text{ and } SW_e^{(u)}(d) = SW^{(u)}(q_e^{(u)}(d), d). \tag{31}$$

We are interested in maximizing the equilibrium throughput and equilibrium social welfare with respect to d. The maximization of the throughput is trivial:

Lemma 4. *The equilibrium throughput is non-increasing in d. Thus, the maximum equilibrium throughput is attained when there is no retrieving time.*

Indeed, we have that

$$TH_e^{(u)}(d) = \lambda p_0 + \lambda(1-p_0)q_e^{(u)}(d) = \frac{\lambda}{1 - \rho q_e^{(u)}(d) + \rho}$$

and the result follows since $q_e^{(u)}(d)$ is non-increasing in d.

A similar analysis can be done for the model without interruptions during the retrieving times, that is for the corresponding standard M/G/1 queue (described in subsection 4.3). We can then obtain the following counterparts of formulas (27) and (28):

$$S^{MG1(u)}(q) = \frac{K}{\mu}\left[\nu - 1 - \frac{\lambda d + \rho + \rho^2 q}{\rho(1 - \lambda d q - \rho q)}\right], \tag{32}$$

$$q^{MG1(u)}(d) = \frac{\rho(\nu - 2) - \lambda d}{\rho^2(\nu - 1) + \rho(\nu - 2)\lambda d - \lambda^2 d^2/2}, \tag{33}$$

Let $TH^{MG1(u)}(q,d)$ and $SW^{MG1(u)}(q,d)$ denote the throughput and the social welfare generated by this system, respectively, given that the customers enter with probability q when finding a busy server and that the retrieving time is d. Moreover, let $q_e^{MG1(u)}(d)$ be the equilibrium customer strategy and $TH_e^{MG1(u)}(d)$, $SW_e^{MG1(u)}(d)$ the associated equilibrium throughput and social welfare. Then, formulas (29), (30) and (31) remain valid for the corresponding MG1-superscript quantities. As in the model with orbit, the equilibrium throughput is maximized for $d = 0$. Therefore, we will focus on the study of the equilibrium social welfare.

An analytical comparison for any d between the equilibrium social welfare functions $SW_e^{(u)}(d)$ and $SW_e^{MG1(u)}(d)$ is not possible due to their involved dependence on d. However, when $d = 0$ or $d \to \infty$ the equilibrium social welfare functions have the same value (with and without interruptions). Indeed, for $d = 0$ both models are reduced to the same regular M/G/1 queue with service times distributed as B, and when $d \to \infty$, both models are reduced to the same M/G/1/1 queue with service times distributed as B. Moreover, extensive numerical experiments have shown that the inequality

$$SW_e^{(u)}(d) \geq SW_e^{MG1(u)}(d), \quad d \geq 0,$$

is always valid.

Numerical experiments also suggest that when we impose a delay that maximizes $SW_e(d)$ under the constraint that the equilibrium throughput, $TH_e(d)$, exceeds a minimum level TH, i.e., when we solve the problem

$$\begin{array}{rl} SW_e^*(TH) = & \max_{d \geq 0} SW_e(d) \\ & \text{s.t.} \;\; TH_e(d) \geq TH, \end{array}$$

for both systems, we have that

$$SW_e^{*(u)}(TH) \geq SW_e^{*MG1(u)}(TH).$$

Numerical results that illustrate these points can be found in the e-companion (section EC.4). Therefore, we understand that the orbit model ‘beats’ the corresponding standard queue model without interruptions. This result is counter-intuitive at first glance. However, a moment of reflection shows that the orbit model that allows retrieving time interruptions favors the increase of the social welfare in two ways: First, it makes customers that find a busy server balk more frequently, since they may be overtaken by newly arriving customers. Second, it enables serving immediately some customers who arrive during retrieving times without further waiting. Therefore, hereafter we will focus our study for the equilibrium social welfare exclusively on the orbit model.

5.2 Maximization of social welfare.

We now move to the study of the monotonicity and the optimization of the equilibrium social welfare $SW_e^{(u)}(d)$ with respect to d. To this end, we let $SW_e^{(u)*} = \max_d SW_e^{(u)}(d)$ be the maximum equilibrium social welfare that can be achieved by tuning the retrieving time, and $d_e^{(u)*} = \arg\max_d SW_e^{(u)}(d)$ the corresponding optimal retrieving time. The results of the next proposition are demonstrated in Figures 1a and 1b. The various cases depend on the values of ρ (subcritical: $\rho \in (0,1)$, critical: $\rho \in [1,\infty)$) and of ν (very low: $\nu \in (1,2]$, low: $\nu \in (2, \frac{2-\rho}{1-\rho})$, medium: $\nu \in [\frac{2-\rho}{1-\rho}, \frac{2}{1-\rho})$, high: $\nu = \frac{2}{1-\rho}$, very high: $\nu \in (\frac{2-\rho}{1-\rho}, \infty)$).

Proposition 2. *In the unobservable $M/M/1$ queue with orbit and deterministic retrieving time d, let $q^{(u)}(d)$ be given by (28) and set*

$$d_1^{(u)} = \frac{1}{\lambda} \ln\left(\frac{\rho(\nu-1)+1}{1+\rho}\right), \tag{34}$$

and

$$d_2^{(u)} = \frac{1}{\lambda} \ln\left(\frac{\rho(\nu-1)+1}{1+\rho+\rho^2(\nu-1)}\right). \tag{35}$$

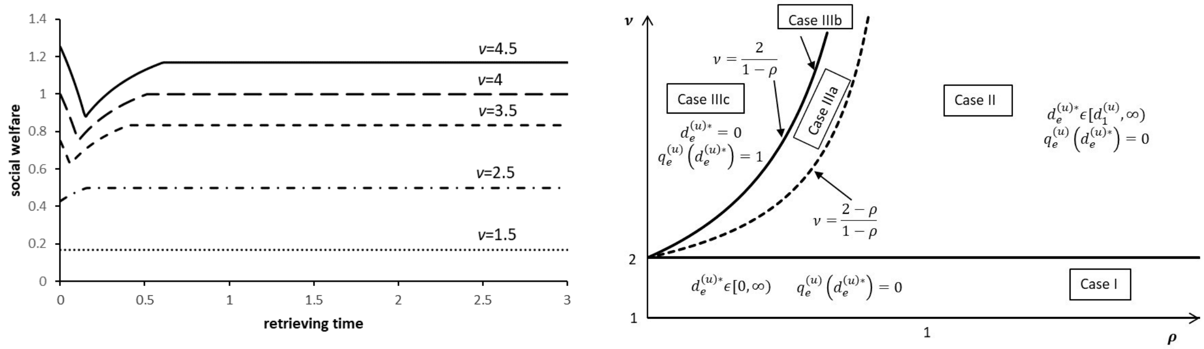
Then, we have the following mutually exclusive cases regarding the equilibrium joining probability $q_e^{(u)}(d)$ and the equilibrium social welfare $SW_e^{(u)}(d)$, as functions of d, and the values of $SW_e^{(u)*}$ and $d_e^{(u)*}$:

Case I. *(Very low service reward) Here, $\nu \in (1,2]$.*
In this case, $q_e^{(u)}(d) = 0$, for $d \in [0,\infty)$, and $SW_e^{(u)}(d)$ is constant in $d \in [0,\infty)$.
Thus, $SW_e^{(u)*} = K\frac{\rho}{1+\rho}(\nu-1) = \frac{(R\mu-K)\rho}{1+\rho}$ and $d_e^{(u)*} \in [0,\infty)$.

Case II. *(Low/medium service reward with subcritical congestion or low to very high service reward with critical congestion) Here, $\left(\rho < 1 \text{ and } \nu \in \left(2, \frac{2-\rho}{1-\rho}\right)\right)$ or $(\rho \geq 1 \text{ and } \nu \in (2,\infty))$.*
In this case,

$$q_e^{(u)}(d) = \begin{cases} q^{(u)}(d) & d \in [0, d_1^{(u)}), \\ 0 & d \in [d_1^{(u)}, \infty), \end{cases}$$

and $SW_e^{(u)}(d)$ is increasing in $d \in [0, d_1^{(u)}]$ and constant in $d \in [d_1^{(u)}, \infty)$.
Thus, $SW_e^{(u)*} = K\frac{\rho}{1+\rho}(\nu-1)$ and $d_e^{(u)*} \in [d_1, \infty)$.



(a) $SW_e^{(u)}(d)$, for $\rho = \frac{1}{2}$ and $A = d \in [0,1]$. (b) Cases for Proposition 2.

Figure 1 Effect of retrieving time on social welfare and equilibrium probability

Case III. *(Medium to very high service reward and subcritical congestion) Here, $\rho < 1$ and $\nu \in \left[\frac{2-\rho}{1-\rho}, \infty\right)$.*

In this case,

$$q_e^{(u)}(d) = \begin{cases} 1 & d \in [0, d_2^{(u)}], \\ q^{(u)}(d) & d \in (d_2^{(u)}, d_1^{(u)}), \\ 0 & d \in [d_1^{(u)}, \infty), \end{cases}$$

and $SW_e^{(u)}(d)$ is decreasing in $d \in [0, d_2^{(u)}]$, increasing in $d \in [d_2^{(u)}, d_1^{(u)}]$ and constant in $d \in [d_1^{(u)}, \infty)$.

Subcase a: If $\nu \in \left[\frac{2-\rho}{1-\rho}, \frac{2}{1-\rho}\right)$, then $SW_e^{(u)*} = K\frac{\rho}{1+\rho}(\nu - 1)$ and $d_e^{(u)*} \in [d_1^{(u)}, \infty)$.

Subcase b: If $\nu = \frac{2}{1-\rho}$, then $SW_e^{(u)*} = K\frac{\rho}{1+\rho}(\nu - 1)$ and $d_e^{(u)*} \in \{0\} \cup [d_1^{(u)}, \infty)$.

Subcase c: If $\nu > \frac{2}{1-\rho}$, then $SW_e^{(u)*} = K\rho(\nu - 1) - K\frac{\rho^2}{1-\rho}$ and $d_e^{(u)*} = 0$. This equals to the sum of $SW_e^{(u)*}$ of all other cases plus $A := \frac{K[\nu(1-\rho)-2]\rho^2}{1-\rho^2}$, so on the range where A is positive we have a higher social welfare than in other cases.

The proof of Proposition 2 follows by considering the changes in equilibrium strategy with d and then investigating the resulting changes to the social welfare. From Proposition 2, it is obvious that as the retrieving time increases, customers that find the system busy join with lower probabilities ($q_e^{(u)}(d)$ is non-increasing in d). However, as we can see in Figure 1a, the social welfare does not always decrease with the retrieving time, i.e., an increase in retrieving time does not always harm customers. On the contrary, in most cases, the social welfare is not decreasing in d. Similar observations occur when increasing the mean of generally distributed retrieving times, see Section 6. Intuitively, with no delay strategic customers arrive at a higher than optimal rate for maximizing the social welfare (as they ignore the externalities they impose on other customers). The introduction of delays then reduces the arrival rate and thus increases the social welfare.

We also notice that in case I, when $\nu \leq 2$, i.e., when the normalized service reward is very low, customers that find the server busy will not join for any d. Thus, in this extreme case, the joining behavior of customers and consequently the social welfare do not change by imposing a retrieving time. This is depicted in Figure 1a for $\nu = 1.5$. Then, for case II, every delay that keeps customers out of the orbit, i.e., such that only customers arriving at an empty system join, maximizes the social welfare. This is depicted in Figure 1a for $\nu = 2.5$. Similar results hold, for cases IIIa and IIIb (which is a special case where a zero retrieving time also maximizes the social welfare). These are depicted in Figure 1 for $\nu = 3.5$ and $\nu = 4$, respectively. Finally, in the other extreme case, it is socially optimal not to impose a retrieving time. This is depicted in Figure 1a for $\nu = 4.5$. Note that in the two extreme cases (I and IIIc), we cannot increase the social welfare by imposing a certain retrieving time. Moreover, the maximal social welfare in all but case IIIc is given as $\frac{(R\mu-K)\rho}{1+\rho}$, and the one under case IIIc is higher.

Figure 1b summarizes the findings of Proposition 2 on the optimal retrieving time, $d_e^{(u)*}$, and the equilibrium joining strategy under $d_e^{(u)*}$, $q_e^{(u)}(d_e^{(u)*})$, with respect to the congestion rate and the normalized service reward.

So far, we have proved that there are cases where we can increase the equilibrium social welfare by imposing a certain retrieving time. In these cases, when $d = 0$, customers that act individually join ignoring the externalities they impose to others. Thus, the social welfare under the equilibrium strategy is less than the social welfare under the socially optimal strategy. Hence, if $q_s^{(u)}$ denotes the socially optimal strategy when there is no delay (i.e., $d = 0$) and $SW_s^{(u)}$ is the corresponding social welfare, we have that $SW_e^{(u)}(0) < SW_s^{(u)}$. We are now interested in finding out whether we can not only increase the social welfare by imposing a certain retrieving time, but also reach $SW_s^{(u)}$. In other words, we check if retrieving times can be used as coordination mechanisms that will force customers to behave in socially optimal fashion and achieve the maximum social welfare. To this end, we first compute the socially optimal strategy for $d = 0$ and the corresponding social welfare. The results are summarized in the following proposition and are illustrated in Figure 2a.

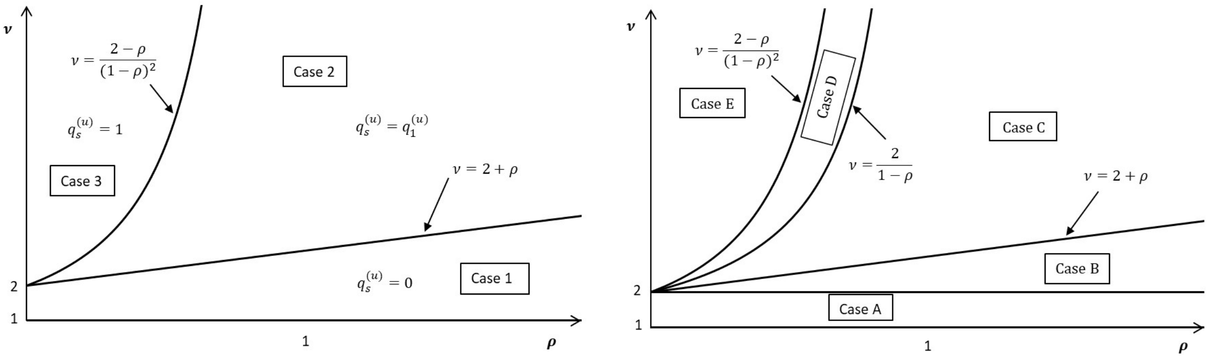
Proposition 3. *In the unobservable $M/M/1$ queue with orbit without retrieving time (d=0), the socially optimal strategy, $q_s^{(u)}$, and the corresponding social welfare, $SW_s^{(u)}$, are given below. Specifically, if*

$$q_1^{(u)} = \frac{1}{\nu\rho}\left(\nu - 1 - \sqrt{\nu\rho + 1}\right), \tag{36}$$

we have the following mutually exclusive cases:

Case 1. *(Very low and low service reward) Here, $\nu \in (1, 2+\rho]$.*
In this case, $q_s^{(u)} = 0$ and $SW_s^{(u)} = K\frac{\rho}{1+\rho}(\nu - 1)$.

Case 2. *(Medium/high service reward with subcritical congestion or medium to very high service reward with critical congestion) Here, $\left(\rho < 1 \text{ and } \nu \in \left(2+\rho, \frac{2-\rho}{(1-\rho)^2}\right)\right)$ or $(\rho \geq 1 \text{ and } \nu \in (2+\rho, \infty))$.*
In this case, $q_s^{(u)} = q_1^{(u)}$ and $SW_s^{(u)} = \frac{K\rho(\nu-1) - \lambda R\rho q_1^{(u)}}{\left[1+\rho(1-q_1^{(u)})\right]\left[1-\rho q_1^{(u)}\right]}$.



(a) Cases for Proposition 3.

(b) Cases for Theorem 2.

Figure 2 $(\nu,\rho)-$regions for Proposition 3 and Theorem 2

Case 3. *(Very high service reward and subcritical congestion) Here,* $\rho < 1$ *and* $\nu \in \left[\frac{2-\rho}{(1-\rho)^2}, \infty\right)$. *In this case,* $q_s^{(u)} = 1$ *and* $SW_s^{(u)} = K\rho(\nu - 1) - K\frac{\rho^2}{1-\rho}$.

The proof of Proposition 3 follows by comparing the resulting social welfare from different equilibrium joining strategies and finding the region where the different strategies maximize the social welfare. We can now compare $SW_s^{(u)}$ to $SW_e^{(u)*}$ to see if there are cases where the maximum social welfare can be achieved by imposing a certain retrieving time. Combining the various cases of Propositions 2 and 3 we have our main results in Theorem 2 comprising five cases that correspond to the five regions of Figure 2b.

Theorem 2. *In the unobservable M/M/1 queue with orbit, we have the following mutually exclusive cases:*

Case A. *(Very low service reward)* $\nu \in (1, 2]$.
Here, Case I of Proposition 2 and Case 1 of Proposition 3 hold. Thus, $q_e^{(u)}(0) = q_s^{(u)} = 0$, *i.e., when* $d = 0$ *customers that find the server busy do not join and this is also socially optimal.* ***The system is coordinated without imposing delays.***

Case B. *(Low service reward)* $\nu \in (2, 2+\rho]$.
Here, Case II of Proposition 2 and Case 1 of Proposition 3 hold. Thus, $q_e^{(u)}(0) > q_s^{(u)} = 0$, *i.e., customers do not behave socially optimally when* $d = 0$. *Nevertheless, we can achieve the maximum social welfare* $SW_s^{(u)}$ *by imposing a retrieving time* $d \geq d_1^{(u)}$. *Thus,* $SW_e^{(u)*} = SW_s^{(u)}$. ***The system can be coordinated by imposing appropriate delays.***

Case C. *(Medium service reward with subcritical congestion or medium to very high service reward with critical congestion)*
$\left(\rho < 1 \text{ and } \nu \in \left(2+\rho, \frac{2}{1-\rho}\right]\right)$ or $(\rho \geq 1 \text{ and } \nu \in (2+\rho, \infty))$.
Here, Case II, IIIa or IIIb of Proposition 2 and Case 2 of Proposition 3 hold. Thus, $q_e^{(u)}(0) > q_s^{(u)} =$

$q_1^{(u)}$*, i.e., customers do not behave socially optimally when* $d=0$*. Although we can increase the social welfare by imposing seeking time* $d \geq d_1^{(u)}$*, we cannot achieve the maximum social welfare, i.e.,* $SW_e^{(u)}(0) \leq SW_e^{(u)*} < SW_s^{(u)}$*.* ***The social welfare can be increased by imposing appropriate delays, but full coordination is not possible.***

Case D. *(High service reward and sub-critical congestion)* $\rho < 1$ and $\nu \in \left(\frac{2}{1-\rho}, \frac{2-\rho}{(1-\rho)^2}\right)$*. Here, Case IIIc of Proposition 2 and Case 2 of Proposition 3 hold. Thus,* $1 = q_e^{(u)}(0) > q_s^{(u)} = q_1^{(u)}$*, i.e., customers do not behave socially optimally when* $d=0$*. Additionally, we cannot increase social welfare by imposing any retrieving time, i.e.,* $SW_e^{(u)}(0) = SW_e^{(u)*} < SW_s^{(u)}$*.* ***The social welfare cannot be increased by imposing appropriate delays and the coordination is not possible.***

Case E. *(Very high service reward and sub-critical congestion)* $\rho < 1$ *and* $\nu \in \left[\frac{2-\rho}{(1-\rho)^2}, \infty\right)$*. Here, Case IIIc of Proposition 2 and Case 3 of Proposition 3 hold. Thus,* $q_e^{(u)}(0) = q_s^{(u)} = 1$*, i.e., customers behave socially optimally when* $d=0$*.* ***The system is coordinated without imposing delays.***

Concluding, in the two extreme cases, i.e., when the service reward is very low ($\nu \leq 2$) or the service reward is very high and the system is somewhat congested, i.e., $\rho < 1$ and $\nu \in \left[\frac{2-\rho}{(1-\rho)^2}, \infty\right)$, customers join at a socially optimal rate. Otherwise, customers join at a rate that is higher than the welfare-maximizing rate. In the latter case, we may be able to force customers join at the welfare-maximizing rate and achieve maximum social welfare by imposing seeking time $d \geq d_1^{(u)}$ (Case B), we may be able to increase social welfare but not achieve the maximum (Case C), or we may not be able to increase social welfare at all (Case D).

Theorem 2 shows that the unconstrained optimization of the social welfare using delays is quite trivial, since the optimal welfare corresponds to either $d=0$ or to imposing d high enough so that no customer joins the orbit. However, the quasi-convex form of the equilibrium social welfare with respect to d suggests a solution for the constrained optimization of the social welfare subject to side constraints, using delays in a non-trivial way, i.e., with $d>0$ and a fraction of customers who join the orbit. Such side constraints correspond to additional objectives of a social planner, for example when a minimum throughput is required. Based on Lemma 4, the requirement of a minimum equilibrium throughput induces an upper bound on the delay that can be imposed. The optimization of queueing performance measures subject to constraints is a recurrent theme in the literature, see e.g., Legros (2018) equation (43), or Ewaisha and Tepedelenlioglu (2013) equation (4).

6 The effect of the level of information and of the various parameters: Numerical results

In this section, we present the results of numerical experiments that were conducted to determine the effect of the arrival rate, the mean of the service and retrieving times on the social welfare in equilibrium. Moreover, we aim to study which level of information implies higher equilibrium social welfare. To this end, we have performed a large number of numerical experiments, which have illustrated similar results. We further investigate numerically the impact of the variances of the service and retrieving times on these. In the following subsections, we present the most significant findings in the context of representative scenarios. For example, while we do not include results on the M/M/1 with deterministic retrieving times, the qualitative results on the impact of their means is identical to the reported below. One important take away is that the deterministic retrieving times are likely to support the best social welfare.

6.1 Effect of service time on social welfare

In this part, we explore the effect of service time on equilibrium social welfare under both levels of information. Specifically, we wish to identify how the mean and the variance of the service time affect the social welfare. Also, we wish to conclude under which level of information the social welfare is higher depending on the mean service time.

In order to find the effect of the mean service time on the social welfare, we consider a numerical scenario with $\lambda = 1,\ R = 5,\ K = 1,\ A \sim \text{Erlang}(2, 3)$ and $B \sim \text{Erlang}(3, \mu),\ \mu \in [1.5, 30]$. Thus, $E[B]$ varies from 0.1 to 2. We may expect that as the service becomes slower, social welfare decreases. However, Figure 3a indicates that this is not necessarily the case. Indeed, in the observable case, the social welfare may be increasing in some intervals of values of $E[B]$. This occurs when the equilibrium strategy is of mixed-threshold type, i.e., $\mathbf{q} = (1, \ldots, 1, q, 0, \ldots)$, with $q \in (0, 1)$. In this case, as the mean service time increases, q decreases. However, customers that join with probability q have zero expected net benefit and do not contribute to the social welfare. As their joining probability decreases, their proportion decreases and the proportion of the others that join with probability 1 and have positive expected net benefit increases. Thus, although the throughput decreases, the social welfare increases.

In the unobservable case, we do not have the same situation. Again, when the equilibrium is a strategy $q \in (0, 1)$, customers that find the server busy and join do not contribute to the social welfare. Also, as $E[B]$ increases, q decreases and we have a positive effect since the proportion of customers that find the server idle increases. Although the proportion of those customers increases, their expected net benefit, that is $R - KE[B]$, decreases and the latter negative effect outweighs the former.

Service Time Distribution	Variance	Equilibrium Strategy - Obs	Social Welfare - Obs	Equilibrium Strategy - Unobs	Social Welfare - Unobs
Erlang(1,0.6)	2.7778	$(0,0,0,\ldots)$	1.2500	0	1.2500
Erlang(2,1.2)	1.3889	$(0.0545,0,0,\ldots)$	1.1857	0.0009	1.2487
Erlang(3,1.8)	0.9259	$(0.9155,0,0,\ldots)$	0.8193	0.0157	1.2294
Erlang(4,2.4)	0.6944	$(1,0,0,\ldots)$	0.8083	0.0232	1.2194
Erlang(5,3)	0.5556	$(1,0,0,\ldots)$	0.8104	0.0277	1.2133

Table 1 Effect of the variance of service times on the social welfare in the observable and the unobservable case.

Comparing the two levels of information, we find that the observable case is socially preferable when the expected service time is small.

To explore the effect of service time variance on the customer behavior and social welfare for both levels of information, we consider a numerical scenario with $\lambda = 1,\ R = 5,\ K = 1,\ A \sim \text{Erlang}(2,3)$, and $B \sim \text{Erlang}(n, 0.6n),\ n \in \{1,2,3,4,5\}$. Thus, as n increases, the variance of the service time B decreases, but its mean value remains the same. We report for each service time distribution the equilibrium strategies and resulting social welfare for the observable and unobservable cases in Table 1.

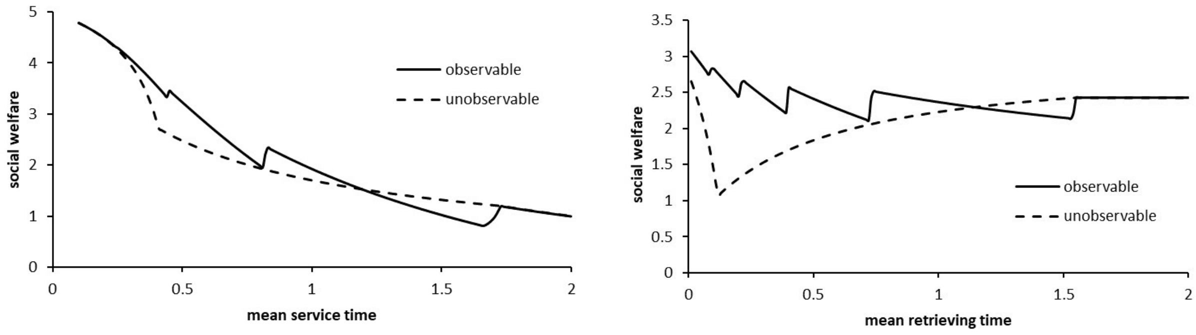
Table 1 indicates that, in the observable case, as the variance of the service time distribution decreases, customers become more willing to join. We can also see that usually smaller variance in the service time reduces the social welfare. As we explained above, this happens because, although more customers join, the fraction of customers that have low or zero expected net benefit increases and the fraction of those that have high expected net benefit, i.e., those that find the server idle, decreases. The effect is similar in the unobservable case. As variance decreases, the social welfare decreases.

Finally, in our results, the effect of the variance in the observable case is stronger than in the unobservable case.

6.2 Effect of retrieving time on social welfare

In this subsection, we present the effect of the mean retrieving time and the effect of the variance of the retrieving time on the social welfare. We also compare the social welfare for the two levels of information.

To explore the effect of mean retrieving time on the social welfare, we consider a numerical scenario with $\lambda = 1,\ R = 5,\ K = 1,\ A \sim \text{Erlang}(2,\mu),\ \mu \in [1,20]$, and $B \sim \text{Erlang}(3,4)$. Thus, $E[A]$ varies from 0.1 to 2. Figure 3b indicates that the social welfare is not always decreasing in $E[A]$. Indeed, it is constant when $E[A]$ is very large and all customers that find the server busy balk. Also, it is increasing whenever customers join according to a mixed strategy. Moreover, the observable case provides higher social welfare than the unobservable case, when the expected retrieving time is relatively short. However, it does not provide higher social welfare when the expected retrieving



(a) Social welfare vs. $E[B]$

(b) Social welfare vs. $E[A]$

Figure 3 Social welfare as function of the mean service and mean retrieving times.

Retrieving Time Distribution	Variance	Equilibrium Strategy - Obs	Social Benefit - Obs	Equilibrium Strategy - Unobs	Social Benefit - Unobs
Erlang(1,2.6)	0.1479	$(1,1,1,0,\ldots)$	2.3033	0.7127	1.6276
Erlang(2,5.2)	0.0740	$(1,1,1,0,\ldots)$	2.2250	0.6806	1.6785
Erlang(3,7.8)	0.0493	$(1,1,0.1576,\ldots)$	2.5109	0.6685	1.6972
Erlang(4,10.4)	0.0370	$(1,1,0,\ldots)$	2.5608	0.6621	1.7069
Erlang(5,13)	0.0296	$(1,1,0,\ldots)$	2.5544	0.6582	1.7129

Table 2 Effect of the variance of retrieving times on the expected social benefit in the observable and the unobservable case.

time is relatively long. These results provide further intuition on how the optimal deterministic retrieving time changes and support that in many cases this time may be large.

To explore how retrieving time variance affects the customer behavior and the social welfare for both levels of information, we consider a numerical scenario with $\lambda = 1,\ R = 5,\ K = 1,\ A \sim$ Erlang$(n, 2.6n),\ n \in \{1,2,3,4,5\}$, and $B \sim$ Erlang$(3,4)$. We report for each retrieving time distribution the equilibrium strategies and resulting social welfare for the observable and unobservable cases in Table 2.

Table 2 shows that in both levels of information, as $Var[A]$ decreases, customers tend to join less. This occurs because, as the variance decreases, the probability of having very short retrieving times usually decreases and the probability of a new arrival before the retrieving completion $\Pr(T_\lambda \leq A) = 1 - \tilde{A}(\lambda)$ usually increases. Thus, customers that find the server busy become less willing to join, as they expect a higher number of new arrivals during the idle periods of their sojourn time. Also, in the observable model, social welfare is not monotonous with $Var[A]$. In broad terms, the social welfare decreases as $Var[A]$ decreases when customers follow a pure equilibrium strategy, whereas it increases when they follow a strictly mixed equilibrium strategy. These results support that using a deterministic retrieving time that minimizes the variance of the retrieving time is likely to be the most efficient retrieving time.

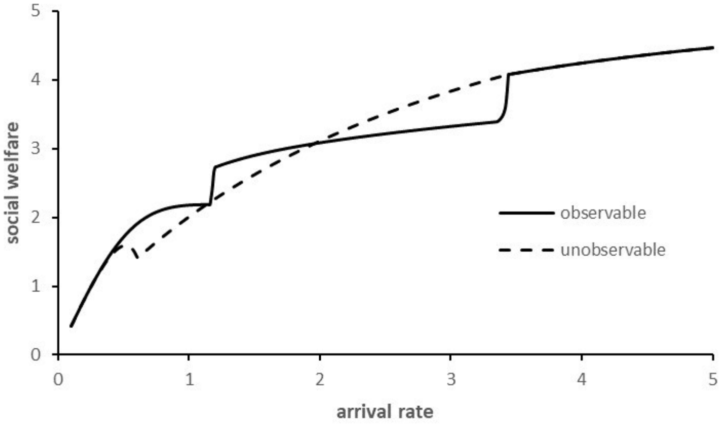


Figure 4 Expected social benefit vs. λ.

6.3 Effect of arrival rate on social welfare

In this subsection, we explore the effect of the arrival rate on the social welfare and compare the two levels of information. Figure 4 presents a numerical scenario with $\lambda \in [0.1, 5]$, $R = 5$, $K = 1$, $A \sim \text{Erlang}(2,3)$, and $B \sim \text{Erlang}(3,4)$. As the arrival rate increases, the social welfare increases both in the unobservable and the observable models. However, because then the system is busier, customers expect higher waiting times and become less willing to join. Thus, as the arrival rate increases, the slope of social welfare typically decreases.

We also compare the two levels of information for the various values of λ. The social welfare, is usually higher in the observable case when the arrival rate is low. Otherwise, the social welfare in the unobservable model is at least as high as in the observable case.

7 Conclusions - Extensions

We analyzed an M/G/1 queue with orbit and arbitrarily distributed retrieving times serving strategic customers under both the unobservable and observable cases. We provided computationally efficient methods to evaluate the performance measures of the resulting state-dependent queues and customers' equilibrium behavior. We further numerically compared the influence of the service and retrieving times, and the arrival rate on the resulting equilibrium social welfare. These comparisons provide insight about how to improve the performance of queues with orbit serving strategic customers.

Our derivations provide the foundation for the analysis of more complicate queueing models with orbit facing strategic customers. Such systems may include customer abandonment, cases where all customers can retry, the provider's improved control of their retrieving times (e.g., where a faster retrieving process is available), and methods to help a provider facing a high demand control the retrial rate of customers in orbit.

In addition, our results suggest that intentionally delaying some customers can help to improve the social welfare. The well studied queue with orbit provides a natural model to investigate the benefit of policies with idling to social welfare. Moreover, as, in applications, it is socially optimal for customers who face an idle server to join, delaying waiting customers in orbit, is an appealing method to introduce delay to coordinate the system.

There is a place to investigate alternative policies with idling and their ability to improve the social welfare. We believe that the regions we found here for both the observable and unobservable models would exist for other such policies. We leave the study of these and additional extensions for future work.

Acknowledgments

The authors gratefully thank the review team of the paper.

References

Abouee-Mehrizi, H. and Baron, O. (2016) State-dependent M/G/1 queueing systems. *Queueing Systems* **82**, 121-148.

Adan, I. and van der Wal, J. (2010) Mean value techniques. In *Boucherie, R.J and van Dijk, N.M. (eds) Queueing Networks: A Fundamental Approach, Springer (Chapter 13).*

Afeche, P. (2013) Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management* **15**, 423-443.

Altman, E. and Hassin, R. (2002) Non-threshold equilibrium for customers joining an M/G/1 queue. In *Proceedings 10th International Symposium on Dynamic Games and Applications*. Saint-Petersburg, Russia.

Artalejo, J.R. and Gomez-Corral, A. (2008) *Retrial Queueing Systems, A Computational Approach.* Springer.

Baron, O., Berman, O., Krass, D. and Wang, J. (2013) Using strategic idleness to improve customer service experience in service networks. *Operations Research* **62**, 123-140.

Baron, O., Berman, O., Krass, D. and Wang, J. (2017) Strategic idleness and dynamic scheduling in an open-shop service network: Case study and analysis. *Manufacturing & Service Operations Management* **19**, 52-71.

Baron, O., Chen, X., and Li, Y. (2022) Omnichannel Services: The False Premise and Operational Remedies. Forthcoming in *Management Science*

Baron, O., Economou, A. and Manou, A. (2018) The state-dependent M/G/1 queue with orbit. *Queueing Systems* **90**, 89-123.

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

Chen, H. and Frank, M. (2004) Monopoly pricing when customers queue. *IIE Transactions* **36**, 569-581.

Cui, S., Su, X. and Veeraraghavan, S.K. (2014) A model of rational retrial in queues. *Operations Research* **67**, 1699-1718.

Do, N.H., Do, T.V. and Melikov, A. (2020) Equilibrium customer behavior in the M/M/1 retrial queue with working vacations and a constant retrial rate. *Operational Research* **20**, 627-646.

Economou, A. (2021) The impact of information structure on strategic behavior in queueing systems. *Chapter 4 in* Anisimov, V. and Limnios, N. (eds.) *Queueing Theory 2: Advanced Trends.* ISTE Ltd and John Wiley & Sons Inc. DOI:10.1002/9781119755234.CH4

Economou, A. and Kanta, S. (2011) Equilibrium customer strategies and social-profit maximization in the single-server constant retrial queue. *Naval Research Logistics* **58**, 107-122.

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

Economou, A. and Manou, A. (2015) A probabilistic approach for the analysis of the $M_n/G/1$ queue. *Annals of Operations Research*, article in press. DOI: https://doi.org/10.1007/s10479-015-1943-0

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

Elcan, A. (1994) Optimal customer return rate for an M/M/1 queueing system with retrials. *Probability in the Engineering and Informational Sciences* **8**, 521-539.

Engel, R. and Hassin, R. (2017) Customer equilibrium in a single-server system with virtual and system queues. *Queueing Systems* **87**, 161-180.

Ewaisha, A. and Tepedelenlioglu, C. (2013) Throughput maximization in multi-channel cognitive radio systems with delay constraints. *2013 Asilomar Conference on Signals, Systems and Computers* pp. 1463-1467, doi: 10.1109/ACSSC.2013.6810538

Falin, G.I. and Templeton, J.G.C. (1997) *Retrial Queues.* Chapman and Hall, London.

Fayolle, G. (1986) A simple telephone exchange with delayed feedbacks. In Boxma, O.J., Cohen, J.W. and Tijms, H.C. (editors) *Teletraffic Analysis and Computer Performance Evaluation*, Elsevier, Amsterdam, pp. 245-253.

Guo, P. and Hassin, R. (2011) Strategic behavior and social optimization in Markovian vacation queues. *Operations Research* **59**, 986-997.

Guo, P. and Zipkin, P. (2007) Analysis and comparison of queues with different levels of delay information. *Management Science* **53**, 962-970.

Hassin, R. (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* **54**, 1185-1196.

Hassin, R. (2016) *Rational Queueing.* CRC Press, Boca Raton, FL.

Hassin, R. and Haviv, M. (1996) Optimal and equilibrium retrial rates in a busy system. *Probability in the Engineering and Informational Sciences* **10**, 223-227.

Hassin, R. and Haviv, M. (1997) Equilibrium threshold strategies: The case of queues with priorities. *Operations Research* **45**, 966-973.

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

Hassin, R. and Roet-Green, R. (2017) The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* **65**, 804-820.

Hassin, R. and Snitkovsky, R. I., (2020) Social and monopoly optimization in observable queues. *Operations Research* **68**, 1178-1198.

Hassin, R. and Koshman, A. (2017) Profit maximization in the M/M/1 queue. *Operations Research Letters* **45**, 436-441.

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

Haviv, M. and Oz, B. (2016) Regulating an observable M/M/1 queue. *Operations Research Letters* **44**, 196-198.

Haviv, M. and Oz, B. (2018) Self-regulation of an unobservable queue. *Management Science* **64**, 2380-2389.

Ibrahim, R. (2018) Sharing delay information in service systems: a literature survey. *Queueing Systems* **89**, 49-79.

Kerner, Y. (2008) The conditional distribution of the residual service time in the $M_n/G/1$ queue. *Stochastic Models* **24**, 364-375.

Kerner, Y. (2011) Equilibrium joining probabilities for an M/G/1 queue. *Games and Economic Behavior* **71**, 521-526.

Kulkarni, V.G. (1983) On queueing systems with retrials. *Journal of Applied Probability* **20**, 380-389.

Kulkarni, V.G. (2010) *Modeling and Analysis of Stochastic Systems, 2nd Edition* (CRC Press).

Legros, B. (2018) M/G/1 queue with event-dependent arrival rates. *Queueing Systems* **89**, 269-301.

Li, K. and Wang, J. (2021) Equilibrium and balking strategies in the single-server retrial queue with constant retrial rate and catastrophes. *Quality Technology and Quantitative Management* **18**(2), 156-178.

Manou, A., Economou, A. and Karaesmen, F. (2014) Strategic customers in a transportation station: When is it optimal to wait? *Operations Research* **62**, 910-925.

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

Oz, B., Adan, I. and Haviv, M. (2017) A rate balance principle and its application to queueing models. *Queueing Systems* **87**, 95-111.

Shaked, M. and Shanthikumar, G. (2007) *Stochastic Orders*. Springer.

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

van Doorn, E.A. and Regterschot, D.J.K. (1988) Conditional PASTA. *Operations Research Letters* **7**, 229-232.

Wang, J. and Zhang, F. (2013) Strategic joining in M/M/1 retrial queues. *European Journal of Operational Research* **230**, 76-87.

Wang, J., Zhang, X. and Huang, P. (2017) Strategic behavior and social optimization in a constant retrial queue with the N-policy. *European Journal of Operational Research* **256**, 841-849.

Zhang, Z., Wang, J. and Zhang, F. (2014) Equilibrium customer strategies in the single-server constant retrial queue with breakdowns and repairs. *Mathematical Problems in Engineering*. Article ID 379572. DOI: http://dx.doi.org/10.1155/2014/379572