European Journal of Operational Research 227 (2013) 133–141

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

# Comparisons between observable and unobservable $M/M/1$ queues with respect to optimal customer behavior

Rob Shone *, Vincent A. Knight, Janet E. Williams

*School of Mathematics, Cardiff University, Senghenydd Road, Cardiff, UK*

ARTICLE INFO

*Article history:*
Received 15 February 2012
Accepted 14 December 2012
Available online 22 December 2012

*Keywords:*
Queueing
Customer information
Equilibrium strategies
Optimal strategies

ABSTRACT

We consider an $M/M/1$ queueing system in which the queue length may or may not be observable by a customer upon entering the system. The "observable" and "unobservable" models are compared with respect to system properties and performance measures under two different types of optimal customer behavior, which we refer to as "selfishly optimal" and "socially optimal". We consider average customer throughput rates and show that, under both types of optimal customer behavior, the equality of effective queue-joining rates between the observable and unobservable systems results in differences with respect to other performance measures such as mean busy periods and waiting times. We also show that the equality of selfishly optimal queue-joining rates between the two types of system precludes the equality of socially optimal joining rates, and vice versa.

## 1. Introduction

The field of behavioral queueing theory is currently a very active research area and has evolved in many different directions in recent years. Indeed, such is the scope of this field that the problems addressed by researchers are often both highly original and genuinely insightful, with real implications for best practice. Moreover, the possibilities for strategic behavior on the part of customers and service providers are almost limitless. The strategic setting of parameters (e.g. service rates, tolls, etc.) is one method by which a service provider may exercise control over a queueing system, but another recent trend in the literature has been an interest in problems where the amount of information (or its completeness or reliability) disclosed to customers is at the provider's discretion. It is easy to conceive situations where the queue length may or may not be made known to customers; typical scenarios include telephone call centers and hospital waiting lists. In this paper we employ classical results for strategic customer behavior in $M/M/1$ queues in order to provide insights which show that, even in the standard Markovian single-server model, the question of whether or not a service provider should grant complete information to customers is far from trivial.

Recent work in this area has been very active. Allon et al. [1] consider an $M/M/1$ system in which a firm can influence its customers' behavior by using "delay announcements". It is found that some level of "intentional vagueness" on the part of the firm may be beneficial in certain circumstances. Guo and Zipkin [15] (see also [16,17]) also study a single-server Markovian system and find that providing "more accurate delay information" may improve system performance, but this is dependent upon other factors. Hassin [19] considers a number of sub-models of $M/M/1$ queues and, in each case, examines the question of whether or not the service provider is motivated to reveal information. The fact that all of these publications adopt the classic $M/M/1$ model is interesting, and may be seen as an indication that the theme of restricting customer information in queueing systems is a young and emerging one. Armony et al. [2], Burnetas and Economou [7], Economou et al. [8], Guo et al. [14] and Jouini et al. [21] also consider varying levels of information in queueing systems. Earlier contributions on this theme were made by Hassin [18], who studied the efficacy of revealing the queue length under the separate objectives of social optimization and profit maximization, and Glazer and Hassin [11], who established (in a slightly different economic setting) that a firm selling subscriptions may sometimes find it profitable to withhold information about its product from consumers.

One of the most persistent themes in the literature on behavioral queueing theory is the sub-optimality of equilibrium solutions in the context of overall social welfare. Recently, this theme has been explored in applications including queues with server breakdowns and delayed repairs [29], vacation queues with heterogeneous customers [13], queues with compartmented waiting space [9] and queues with catastrophes causing system abandonment [6]. In this paper we consider the classical model presented by Naor [25], which is an $M/M/1$ system with linear waiting costs and a fixed service value. The principle that social welfare is generally not optimized by selfish individuals is established in Naor's

* Corresponding author. Tel.: +44 (0)29 20874811.
*E-mail address:* shonerw@cardiff.ac.uk (R. Shone).

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.12.016

work, and has also been observed in various extensions and adaptations of Naor's model (see, e.g., [3,10,22–24,28,30,31]). An important property of Naor's model is that, upon entering the system, a customer is able to observe the length of the queue before deciding on any action to take. As such, he can calculate his expected waiting time as a function of the state of the system upon his arrival. Edelson and Hildebrand [10] consider a model which is similar to Naor's, but without the assumption that a customer is able to observe the queue length before deciding whether or not to join. We refer to Naor's model as *observable*, and the adapted model considered by Edelson and Hildebrand as *unobservable*. Extensions of both models have been made by several authors; see [20] for a comprehensive survey of the literature.

While the respective properties of observable and unobservable $M/M/1$ queueing systems have been analyzed extensively, comparisons between the two types of system are not abundant in the literature. Indeed, one might observe that both system types constitute their own sub-discipline of the field. This is logical to some extent, as the mathematical tools that one employs will depend on whether or not the state of the system can be observed exactly, especially when one considers more elaborate models. For example, contemporary analysis of unobservable queues often takes place in a game theoretic setting involving flow control (see, for example, [26] for a discussion of routing games), while a more natural framework for the modeling of an observable queue is a continuous-time Markov Decision Process (see, for example, [27]). While the comparison of observable and unobservable queues might involve attempting to bridge two very different methodological areas, it is a worthy endeavor because of the potential insights that can be gained into problems involving optimal control of information.

Hassin and Haviv [20] observe that the field of behavioral queueing theory is "lacking continuity" and "leaves many issues uncovered". This paper contributes to the literature by providing new results based on comparisons between classical results for observable and unobservable queues, and thereby helping to unify the field. We address the problem of a revenue-maximizing server who has the opportunity to suppress information on queue length to customers entering the system. Assuming that the server's objective is simply to maximize the throughput of customers, the problem of whether or not to reveal the queue length is non-trivial. This paper does not aim to directly tackle problems of revenue maximization and, as such, admission fees (or tolls) are not explicitly included in the analyses. Instead, a parameter $R > 0$ is assumed to represent the net worth of service to a customer after any admission fees or tolls have been accounted for. The value of $R$ may be considered adjustable, which allows for the possibility of tolls being imposed (or payments offered) to customers.

In the spirit of trying to maximize customer throughput, we are primarily interested in comparisons between observable and unobservable systems with respect to the average rates at which customers join the queue for service (referred to as *joining rates*) and various other measures of system performance, including waiting times and mean busy periods. Two types of *optimal* joining rates are considered separately. *Selfishly optimal* joining rates are understood to be those which occur under (Nash) equilibrium conditions, and *socially optimal* joining rates are those which maximize the collective welfare of customers. As we have mentioned, the contrast between "selfish" and "social" optimization is an important theme in the literature and most of our results are presented in a bipartite form, with selfish and social customer behavior considered as separate cases.

The main contributions of this paper are as follows:

- Necessary and sufficient conditions for the equality of optimal queue-joining rates between the observable and unobservable systems are established under both types of customer behavior.
- A non-existence theorem is provided, showing that the equality of selfishly optimal joining rates precludes the equality of socially optimal joining rates and vice versa.
- Performance measure comparisons are presented for the observable and unobservable systems under the equality of optimal joining rates.

*Note on terminology*: The emphasis of this paper is on comparisons between observable and unobservable systems. As such, when "two types of system" are mentioned, these two types are *observable* and *unobservable*. When "two types of optimal joining rate" are mentioned, these two types are *selfishly optimal* rates, i.e. those which occur under (Nash) equilibrium conditions, and *socially optimal* rates, which maximize the collective welfare of customers. When "the equality of selfishly (or socially) optimal joining rates" is discussed, this refers to an observable and unobservable system sharing the same selfishly (or socially) optimal joining rate.

## 2. The model

Following the assumptions of Naor, we assume a stationary Poisson stream of customers with parameter $\Lambda > 0$ arriving at a single server system. Service times are independently, identically and exponentially distributed with mean $1/\mu$. A customer incurs a cost $C > 0$ per unit time while waiting in the system, but also obtains a fixed reward $R > 0$ upon successful completion of service. We assume $R \geqslant C/\mu$ in order to avoid the case where a customer would be unwilling to wait even for his own service, which would lead to trivialities. The queue discipline is first-come first-served (FCFS) and a newly-arrived customer may choose one of two alternatives: he may either join the queue, in which case he incurs the associated cost of waiting but obtains the reward $R$ upon completion of service, or "balk" from the system by exiting immediately, without incurring any cost or reward. We assume that customers are risk-neutral, and choose to join the queue if and only if the expected cost of waiting is less than or equal to $R$.

A diagrammatic representation of the system as described is given in Fig. 1.

The relative traffic intensity for the system is denoted $\rho = \Lambda/\mu$. We assume $\Lambda > 0$ and hence $\rho > 0$. Note, however, that we do not assume $\rho < 1$ (a common assumption for $M/M/1$ queues which ensures that "steady state" conditions are attained). This is because the attainment of steady state conditions merely requires the effective queue-joining rate (as opposed to the system arrival rate) to be smaller than the service rate $\mu$. As we shall see, this is always the case under both types of optimal customer behavior, regardless of whether or not the system is observable.

We use $\tilde{\lambda}^{[O]}$ and $\lambda^{*[O]}$ to denote, respectively, the selfishly and socially optimal queue-joining rates for the observable system, and $\tilde{\lambda}^{[U]}$ and $\lambda^{*[U]}$ to denote the corresponding measures for the unobservable system. Similar notation is adopted for other output and performance measures, and this is summarized in Table 1.

## 3. Summary of known results

### 3.1. Observable systems

In the case of *observable* systems, it can be shown that customers who wish to maximize their individual welfare will follow a pure threshold strategy (this is established by [25] and can be easily shown using game theoretic arguments). This means that there exists an integer $n_s$ such that newly-arrived customers will join the queue if and only if the number of other customers already present in the system is smaller than $n_s$. Specifically, we have: