Operations Research Letters 43 (2015) 343–348

Contents lists available at ScienceDirect

Operations Research Letters

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

Equilibrium strategies for placing duplicate orders in a single server queue

Pengfei Guo *, Refael Hassin

Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Hong Kong
Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel

ARTICLE INFO

Article history:
Received 12 June 2014
Received in revised form
19 January 2015
Accepted 8 April 2015
Available online 17 April 2015

Keywords:
Duplicate order
Equilibrium analysis
Strategic behavior
Queueing

ABSTRACT

We consider strategic behavior for customers to place duplicate orders in a single-server, random order service system with the intention of speeding up their service. We observe that follow-the-crowd (FTC) behavior may lead to two pure equilibria and one mixed equilibrium in the order size.

1. Introduction

Duplicate orders sometimes exist in queueing systems with one server. [4] models a scenario in which a command center obtains reports from many sources that may report on the same event, causing the replication of messages. The reward from a serviced message decreases with waiting time. The paper assumes that sending a message is costly and each source chooses the rate of its messages, hoping that one of them will be serviced, attempting to maximize its expected net utility. To prevent congestion in the resulting equilibrium, it is suggested in [4] that the command center regulates the cost/reward structure. Such behavior is also widely recognized in multiple-retailer–one-supplier supply chains. When a supplier is short of capacity, it sometimes satisfies its retailers proportionally according to their order size. This type of policy sometimes results in a shortage game between the retailers with each exaggerating their order size and canceling the extra orders once their true demand is satisfied [14].

We are interested in the strategic behavior of placing duplicate orders in this scenario. Understanding such behavior and its effect on systems allows managers to more accurately forecast true demand and better design and control service systems. We consider a single-server queue with a random-order service discipline in which the chance that a customer will be picked up for service is proportional to the number of duplicate orders. We assume homogeneous customers and focus on the study of a symmetric equilibrium. Particularly, we are interested in whether a customer's tendency to place duplicate orders increases with the same tendency in other customers, a phenomena known as *follow-the-crowd* (FTC) behavior [9,10]. In contrast, *avoid-the-crowd* (ATC) behavior represents a situation in which a customer's strategy is inversely affected by exposure to the same tendency in other customers. It is known that multiple equilibria can exist in an FTC situation, whereas the ATC property assures the uniqueness of an equilibrium.

We show that FTC exists; that is, if other customers place duplicate orders, a tagged customer will do so to avoid a longer wait. Multiple equilibria exist when order sizes are integers and, in this situation, a maximum of two pure equilibria represented by two consecutive integers and one mixed equilibrium exist. An important conclusion is that as system utilization increases, the equilibrium order sizes increase. Therefore, the more scarce the capacity, the more the size of the demand is exaggerated. Such behavior may, in turn, mislead the service provider to overestimate the market and overinvest in capacity.

Queues with FTC behavior have attracted intensive study in recent years. In many situations, the FTC behavior arises due to a service-rate control policy. [3] studies a system with an exponential setup time for each busy period and show that both ATC and FTC customer

* Corresponding author at: Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Hong Kong.
E-mail address: lgtpguo@polyu.edu.hk (P. Guo).

http://dx.doi.org/10.1016/j.orl.2015.04.005

behavior can occur. [11] considers an M/G/1 queue with an observable server's status and note that when the queue is empty, both ATC and FTC behavior is possible, depending on the service distribution. [7,5,8] all demonstrate the existence of FTC behavior when high congestion triggers faster service. [7] notes that the server can be turned on or off, [5] assumes that the service rate is switched between low and high states and [8] studies multiple servers, some of which can be turned on or off. FTC behavior also arises from quality issues, as in [16].

The FTC behavior in our model arises from competition among customers for limited resources. Similar models of competition for limited capacity in the queueing literature are termed priority auctioning or bribery (see Section 4.5 in [10] for the review on this topic). [1] demonstrates FTC behavior in observable queues in which customers can buy priority. Both of these types of models can be viewed as dealing with expenditures to obtain faster service, but the difference is that a fee increases a customer's absolute priority in those papers and increases a customer's relative priority in our duplicate-order queue. Within the literature, processor-sharing models in which duplicate orders can be used as an attempt to reduce the expected waiting time are very close to our one-server duplicate order model. A particularly interesting example is in [17], in which customers purchase their priority index when the server allocates its capacity according to a discriminatory processor sharing (DPS) policy. Recently, [2] considers how to assign multi-class customers to a discriminatory processor sharing (DPS) system, which dynamically allocates the service capacity among customers according to their weights. The paper demonstrates that the social optimality can be achieved through pricing control. [6] studies the equilibrium for priority purchasing in a DPS system with multiple classes of customers. Each class tries to minimize its payment for the priority under the constraint of meeting the deadline of their jobs. The paper demonstrates the existence of equilibrium for the general case and fully characterizes it for the special case with two classes of customers and exponential service times. Our work differs from these two work on the following three aspects. First, the service capacity in our work is always fully allocated to a customer instead of being shared among all customers. Second, the fee for priority purchasing in the two papers is continuous while it is discrete in our model. Third, we fully characterize the existence of FTC and ATC behavior in such a game and characterize both pure and mixed equilibria. The foregoing two papers do not provide such discussions on the game property and also do not discuss the mixed equilibrium.

2. Duplicate orders in a single-server system

We consider a single-server queueing system. The service times follow an exponential distribution with rate μ. The server adopts a random order service discipline. That is, upon completing one job, the server randomly picks up another job for processing from the list of waiting jobs.

Customers arrive according to a Poisson process with rate λ. We assume that balking is not allowed so that every customer places at least one order. For stability, we assume $\lambda < \mu$. The incoming customer can place one order or duplicate orders. In the latter case, cancelation of the other orders occurs when one of the customer's orders is picked up for processing.

There is a waiting cost C_w per time unit for customers in the queue, regardless of the number of orders, and an ordering cost C_o set by the server. We analyze customers' strategic behavior in choosing to make one order or duplicate orders. Choosing to make duplicate orders not only reduces the customer's expected waiting time, but they also bear additional order costs as a result. Furthermore, the customer's expected waiting time depends on how other customers choose to order. Hence, we need to obtain the *equilibrium* strategy for customers.

We assume that the waiting cost is proportional to the queueing time, excluding the service time. This assumption is without-loss-of-generality because the difference between the expected waiting time and the expected queueing time is a constant.

We consider a fully unobservable case in which the queue and the server's status are unobservable. In Section 2.1, we show that the model is FTC and we compute the symmetric equilibria to reveal that there are, at most, two pure equilibria represented by two consecutive integers. According to [9], there is another mixed strategy in which customers randomly pick up the integers in a probability and we compute that in Section 2.2.

2.1. Pure equilibria

We consider a symmetric equilibrium strategy for customers to place duplicate orders and assume all other customers place q orders. We consider an incoming 'tagged' customer's decision on the number of orders to place and assume that the tagged customer places k orders.

Because the server randomly picks up an order from the queue for processing, the probability of a customer to be picked up is proportional to their number of orders. Specifically, consider a queue with n waiting customers upon a service completion, including the tagged customer. The probability of the tagged customer being chosen for the next service is

$$\frac{k}{k+(n-1)q}.$$

This becomes a random queue with *relative priorities*. According to [12], the expected queueing time for the tagged customer can be expressed as a function of k/q. We denote it as $W(k/q)$ and

$$W(k/q) = \frac{1+(k/q)-\rho(k/q)}{1-\rho+(k/q)}\mathcal{W}, \tag{1}$$

where $\rho = \lambda/\mu < 1$ and $\mathcal{W}$ is the expected queueing time in an M/M/1 queue with arrival rate λ and service rate μ, i.e., $\mathcal{W} = \frac{\rho}{1-\rho}\frac{1}{\mu}$. Clearly, when $k = q$, $W(k/q) = \mathcal{W}$.

The expected total cost for the tagged customer is

$$C(k, q) = C_o k + C_w W(k/q). \tag{2}$$

We have the following conclusion on the monotonicity of minimizer $K(q)$.

Proposition 1. *The minimizer $K(q)$ of the function $C(k, q)$ is increasing in q if $k > q(1-\rho)$ and decreasing in q if $k < q(1-\rho)$.*