Computers & Industrial Engineering 87 (2015) 600–610

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Two pricing mechanisms for a service provider when customers' delay costs are value-related

Wenhui Zhou [a], Weixiang Huang [b,*]

[a] School of Business Administration, South China University of Technology, Guangzhou 510640, China
[b] Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 23 December 2014
Received in revised form 12 May 2015
Accepted 16 June 2015
Available online 24 June 2015

Keywords:
Queueing
Revenue management
Pricing
Priority auctions
Delay-sensitive

ABSTRACT

In this paper, we study two pricing mechanisms for a provider that serves delay-sensitive customers, one is the uniform pricing and the other is the priority auction. The expected delay cost of a customer is assumed to depend on his value for the service (i.e. the unit delay cost is a strictly increasing function of his value) and the expected sojourn time caused by the number of customers in the system (and his payment-based position in the queue if auction is adopted). Hence, each customer reacts to the service provider's pricing mechanism by deciding whether or not to enter the service system and how much he pays. This resulting problem is a Stackelberg game. When auction is adopted, by using of adverse selection, we derive a feasible scheme in which customers with higher value would like to pay more. We further compare the performance of these two pricing mechanisms. Our numerical examples show that auction performs better not only in terms of revenue making but also in terms of social welfare improvement. Interestingly, auction can also render more customer surplus in most instances, which differs from the common techniques in revenue management field.

1. Introduction

In many manufacturing and service systems, customers experience delays before receiving the product/service. For these systems, both the price of the product/service and the delay affect customers' purchasing decision. In this article, we consider such a system, and refer to it as a service system even though it can be equally likely a manufacturing system operating under a make-to-order policy. We assume that the service provider is a monopolist in the sense that it provides exclusive service to a designated region, or there is little competition for the service it provides. This service provider faces delay-sensitive heterogenous customers and aims to maximize the expected revenue by choosing a pricing strategy. Restaurants, hotels, and hospitals are examples.

Delay is common in many manufacturing and service systems. The delay in delivery of a purchase is over a month in some industrial markets such as airplane manufacturing, shipbuilding, textile mill products, steel, fabricated metals, nonelectric machinery, and electric machinery. There are at least three interesting features of market clearing in such industries. First, there is commonly more variation in delivery lags than in posted prices. This suggests that delay and queueing phenomena play a crucial role in clearing such markets. Second, the queue is often not directly observable by any potential customer who is considering whether or not to place an order. Carlton and Perloff (2004) provide a valuable review of the evidence on the importance of time and delay in market clearing of these industries. Third, because of customers' heterogeneity in delay (sensitivity) and self-interest maximization, firms should consider customers' strategic behavior when making decisions. See Selvaraju and Goswami (2013), Li, Wang, and Zhang (2013) and Yu, Tnag, and Wu (2014) for examples. In this study, we characterize these features and analyze two pricing mechanisms for a service provider.

There is a rich literature evolving around service pricing with facing delay-sensitive customers. Naor (1969) is the first work to study the interaction between price and queueing effect. He considers an observable M/M/1 queueing system with homogenous customers, and discusses the revenue-maximizing pricing decision of the service provider. In his work, customers' utility u in the absence of price has been modeled as the difference between the value of service and delay cost: $u(v,t) = v - C(t)$, where v is the service value, t is the delay and $C(t)$ is the delay cost, a nondecreasing function of t. In many studies, the delay cost $C(t)$ is considered as a linear function of t, i.e., $C(t) = c \cdot t$ where c is the unit delay cost/delay sensitivity (Chen & Frank, 2004; De Vany, 1976;

* Corresponding author. Tel.: +852 64369279.
E-mail address: wxhuang4-c@my.cityu.edu.hk (W. Huang).

http://dx.doi.org/10.1016/j.cie.2015.06.011
0360-8352/

Edelson & Hildebrand, 1975; Guo, Sun, & Wang, 2011; Hassin, 1986; Knudsen, 1972; Lippman & Stidham, 1977; Zhou, Chao, & Gong, 2014; Zhou, Lian, & Wu, 2014, etc.). In some studies, the delay cost is allowed for nonlinear (Dewan and Mendelson, 1990; Kittsteiner & Moldovanu, 2005; Van Mieghem, 2000, etc.). For a comprehensive review, the reader is referred to Hassin and Haviv (2002) and Stidham (2011). In these models, the delay cost $C(t)$ is mutually independent of a service's value v, i.e., the delay cost and value are additive in customers' utility.

However, under some circumstance, a customer's value of service is affected by the delay he experiences. For example, in financial markets, the expected profit of the investor is deflated by a delay in the execution of a trade if part or all of the anticipated price change occurs before the execution. Thus, delays lead to the value losses. Similar phenomenon can be seen in other industries with physical decay of products, technological obsolescence and so on. Aféche and Mendelson (2004) first model this delay-driven value losses by constructing a generalized delay cost structure where the delay decreases the customers' values, and correspondingly give a feasible priority auction pricing mechanism. In their work, a customer's utility in the absence of price is modeled as $u(v,t) = v(1 - d \cdot t) - c \cdot t$, where $d > 0$ and $c \geqslant 0$, and $d \cdot t \cdot v$ is the value losses caused by the delay he experiences.

More often, a customer's sensitivity to delay is correlated with his value for the service. Furthermore, they are often positively correlated. That is, a customer with higher value for the service is always impatient (as he really needs it), so he will bear a higher delay cost than those who have a lower value for the service even though their delays are equal, i.e., a customer with higher value has a higher unit delay cost (Zhou, Chao et al., 2014; Zhou, Lian et al., 2014). In other words, a customer with higher value has higher time values. This assumption is also applied in the marketing literature. For instance, Guo and Meng (2014) assume that consumers with higher value for product support have a higher marginal value of time. In reality, this phenomenon is commonly seen in the industries like healthcare and automotive repairing. The reader is also referred to Section 6.3.3 of Mandelbaum, Sakov, and Zeltyn (2001) for some empirical evidence on this phenomenon in a call center. This paper studies such a phenomenon by modeling a customer's utility in the absence of price as $u(v,t) = v - D(v) \cdot t$, where the unit delay cost, $D(v)$, is nonnegative and strictly increasing in customer's value of the service. Interestingly, if the unit delay cost is a linear function of values, i.e. $D(v) = d \cdot v - c$, our model degenerates into that of Aféche and Mendelson (2004). Thus, the proposed model in this paper incorporate theirs, regardless of that the issue studied in these two papers are different. More specifically, they assume that customers' values are delay sensitive, while customers' delay costs are value-related in our paper.

Apart from that customers' values and delays (delay costs) are related, another key feature of customers is that they have different values for the same type of service. In other words, customers are heterogenous in terms of service values. Note that the unit delay cost is value-related in this paper, this indicates that customers are also heterogenous in the aspect of unit delay cost/delay sensitivity. When customers are heterogenous, in the sense that they have different values for the service, different sensitivities to delay, or different time for process accomplishment, most of the existing research focus on priority pricing. That is, the schedules of the customers and pricing decisions are jointly made by the service provider. However, customers may not tell their class identification for their own sake, thus making the priority pricing infeasible. Mendelson and Whang (1990) is perhaps the first to give a incentive-compatible pricing scheme which induces the customers to reveal their true class identification. The extension of their work includes Bradford (1996), Hassin and Haviv (1997), Rao and Petersen (1998), and Van Mieghem (2000).

If there are a continuum of unobserved customer types, priority auction pricing is the most efficiency strategy to maximize the expected revenue of the monopolistic service provider (Aféche & Mendelson, 2004). Kleinrock (1967) is the first to study the allocation of priorities based on payments made by customers. In Kleinrock's model, a new arriving customer offers a nonnegative payment to the queue manager (these payments are called "bribes" in Kleinrock's work), and this customer is then assigned a position in the queue such that all those customers who made larger payment are assigned in the front of him, while all the customers who made smaller payment are assigned behind him. To the customers who paid equally, they follow first-come-first-serve (FCFS) discipline. Based on this rule, Kleinrock derives the steady-state expected waiting time for an arbitrary customer. Liu (1985) and Glazer and Hassin (1986) revise Kleinrock's model and assume that the customers make payments to minimize the total cost which is the sum of their delay cost and payment. Kittsteiner and Moldovanu (2005) show how the convexity or concavity delay cost function with respect to sojourn time determines the queue discipline, shortest-processing-time-first (SPT) or longest-processing-time-first (LPT), under the scenario that the processing time is customers' private information. In this paper, we construct a feasible priority auction mechanism to induce customers to reveal their true class identification. Unlike the mentioned works, we assume that a customer's unit delay cost is a function of his service value.

Although we use the term "bid" in the priority auction mechanism, customers do not necessarily engage in a bidding process. This is because the service provider will design a price menu under which customers' true types are revealed and any customer will choose the price equals to his bid. Thus, this mechanism is quite applicable in practise. By using of adverse selection, we derive such a feasible priority auction mechanism. For the sake of comparison, we also study the uniform pricing mechanism in this paper. Unfortunately, it turns out that we cannot obtain the closed-form solution. However, we can find some useful property in pricing. For the uniform pricing (priority auction), if the price (admission fee) is so high, then no customer will enter the system; if the price (admission fee) is sufficiently low, then all the customers will enter the system and the revenue is decreasing in the price (admission fee). Hence, the optimal strategy of the service provider lies at a closed range. We further conduct several numerical examples to get some managerial insight. Our numerical results show that priority auction performs better not only in terms of revenue making but also in terms of social welfare improvement. Interestingly, priority auction can also result in more customer surplus in most instances, which differs from common technique in revenue management field. In other words, priority auction can achieve "Pareto Improvement" in terms of the service provider, customers and the whole system over the uniform pricing in most instances.

The rest of the paper is organized as follows. In Section 2, we describe the model in detail. The equilibrium customer behavior under uniform pricing mechanism is given in Section 3 feasible priority auction mechanism is derived in Section 4. In Section 5, we compare the performance of two pricing mechanisms by numerical examples. Finally, Section 6 concludes this paper.

2. Description of model

We consider a monopolist that provides a type of service to the market. The problem is modeled as a single-server queueing system. The service time of the customers are independent and