International Journal of Production Economics 193 (2017) 726–736

Contents lists available at ScienceDirect

International Journal of Production Economics

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

On the uniqueness and stability of equilibrium in quality-speed competition with boundedly-rational customers: The case with general reward function and multiple servers

Xin Li[a], Qingying Li[b,*], Pengfei Guo[c], Zhaotong Lian[d]

[a] Department of Decision Sciences, School of Business, Macau University of Science and Technology, Macau
[b] Glorious Sun School of Business and Management, Donghua University, Shanghai, China
[c] Faculty of Business, The Hong Kong Polytechnic University, Hong Kong
[d] Faculty of Business Administration, University of Macau, Macau

ARTICLE INFO

Keywords:
Customer-sensitive service
Bounded rationality
Strategic queueing
Matrix

ABSTRACT

In this paper, we extend the service-speed competition game with boundedly rational customers considered in Li et al. (2016) to the case with general reward function and multiple servers. The N servers make strategic decisions on their service rates sequentially and repeatedly. Since the competing servers' payoff functions can only be expressed by an implicit-function set, we propose a matrix method to derive the uniqueness of the equilibrium service rates, and we establish the stability of the equilibrium through a tatônnement scheme. By conducting a sensitivity analysis regarding the number of competing servers and the demand density, we find that the server competition benefits the customers by improving their utilities as well as getting more customers to be served. Furthermore, for a fixed demand density, the equilibrium service rate increases in the market size and converges to a certain level when the market size is large enough.

1. Introduction

Customer-intensive service is the service requiring high level of contact with customers. Such service is usually provided by experts, such as doctors (for health care treatment), professors (for education), and consultants (for legal or financial consulting). A number of customer-intensive service examples have been provided in Anand et al. (2011). For such service, customers usually regard the service in higher quality for the longer processing time. According to their estimation on each expert's service quality and the expected sojourn time, customers usually have the opportunity to choose one from a number of experts. However, as many empirical studies (see, e.g., Nisbett and Ross (1980), Kalai et al. (1992)) suggest, customers may be affected by noisy terms such that they cannot formulate accurate expectations on the service qualities. In this case, customers are referred as "boundedly rational." To formulate boundedly rational customers' choice behavior, logit customer choice models are usually adopted; see, e.g., Huang et al. (2013), Huang and Chen (2015), Huang and Liu (2015) and Song and Zhao (2016).

Combing the two above-mentioned features, Li et al. (2016) investigate a customer-intensive queueing model with two service providers, where customers are assumed to be boundedly rational, and a customer's reward from receiving the service is linear in the service rate. They derive the two servers' equilibrium decision on the service rates, establish its uniqueness and the stability, and also investigate the pricing issues.

As customer-intensive service is often rendered by experts whose expertise is usually obtained by going through a lengthy and tough training process, the number of service providers is often limited for such service. For example, the number of doctors in Hong Kong is quite stable and there exists a high entrance barrier in such a system. A research question then arises: what is the impact of the number of the service providers on the equilibrium outcome of the market (in our case, the equilibrium service rate)? To answer this question, we shall extend Li et al. (2016)'s work from the two-server case to a general N-server case, where $N \geq 2$. The answer to this question can help the policy makers to better regulate the number of service providers in such a system. To make our conclusions more general, we also extend the linear reward function in Li et al. (2016) to a general one. We consider a continuous and dynamic game among the N service providers. Specifically, the N service providers play the game continuously in multiple periods, and in each period they update their strategies sequentially to perceive a positive

* Corresponding author.
E-mail addresses: xli@must.edu.mo (X. Li), liqingying@dhu.edu.cn (Q. Li), pengfei.guo@polyu.edu.hk (P. Guo), lianzt@umac.mo (Z. Lian).

http://dx.doi.org/10.1016/j.ijpe.2017.08.026
Received 24 August 2016; Received in revised form 26 August 2017; Accepted 28 August 2017
Available online 6 September 2017
0925-5273/

marginal payoff. In this circumstance, we will investigate the existence and the stability of the servers' symmetric equilibrium, and we will also investigate the sensitivity of the equilibrium and some system performance measures, including customer utility and the total market share, regarding the number of service providers N.

Regarding the equilibrium studies in the game between the product/service provider(s) and customer(s), many scholars establish the existence, uniqueness, and stability analysis of the equilibrium; see, e.g., Hong et al. (2012) and Nagurney et al. (2014). It is common that in the service competition setting, researchers usually consider explicit payoff functions. The challenge for our generalized model is that each server's payoff function cannot be shown explicitly. Actually, the payoff function can only be expressed as the solution of an implicit-function set of N servers' strategies. This causes great difficulty to investigate a tagged server's best response decision. To tackle this challenge, we propose a matrix method to derive servers' best responses and the equilibrium decisions. To the best of our knowledge, there is no existing method in the literature that can determine the equilibrium in the service competition where players' payoff functions are given by a set of implicit functions.

In the following, we will introduce the model setting in Section 2. We will analyze the best response and derive the symmetric equilibrium in Section 3, and then establish the asymptotical stability of the equilibrium through a tatônnement process in Section 4. In Section 5, we will investigate the value of server competition through sensitivity analysis. Section 6 concludes the paper. All the proofs are relegated to the Appendix.

2. The model

We extend the quality-speed competition queueing model in Li et al. (2016) to the case with general customer-intensive reward function and N service providers, where $N \geq 2$. Specifically, we consider an oligopoly market with N customer-intensive-service providers competing for boundedly rational customers. The N servers and the customers behave according to a two-stage game. In stage one, the servers set their service rates; and in stage two, customers observe the service rates, and then choose one server or leave.

We assume that each server's service time has an exponential distribution. For each server i, where $i = 1, \ldots, N$, we let μ_i be the service rate, which is server i's decision variable. Potential customers arrive according to a Poisson process with rate Λ. Upon arrival, a customer decides whether to join one of the servers or balk according to the utility value. The utility value of joining a specific server depends on the reward from receiving the service, price and congestion. If a customer decides to balk, the utility is zero. If a customer decides to join server i, where $i = 1, 2, \ldots, N$, then reneging is not allowed. The reward from receiving the service is denoted as $R(\mu_i)$, which is a general function depending on the service rate μ_i. For customer-intensive service, customer service reward decreases with service rate. Hence, a large service rate reduces service facility congestion, but it also reduces the service value for customers. The following assumption about the service reward function $R(\mu_i)$ is made.

Assumption 1. *Service reward function $R(\mu_i)$ is decreasing and concave in service rate μ_i, i.e., $R'(\mu_i) < 0$ and $R''(\mu_i) \leq 0$.*

This assumption fits the characteristics of the customer-intensive service. The inequality $R'(\mu_i) < 0$ implies that a customer's service reward decreases in the service rate. Regarding $R''(\mu_i) \leq 0$, it is easy to check that under this condition, the reward function is increasing and concave in the expected service time, i.e., $1/\mu_i$. This makes sense: when the average service time is short, the marginal effect of increasing one unit service time is significant to customers; however, when the average service time is relatively long, the marginal effect of increasing one unit service time shall be less significant. Note that the linear reward function considered in Li et al. (2016) satisfies $R'(\mu_i) < 0$ and $R''(\mu_i) = 0$, which is a special case of our model.

Once joining server i, a customer needs to pay a price p for the service and experience a certain period of congestion at a unit-time waiting cost C. We assume that queues are unobservable, and thus the congestion is estimated from the long-run average performance. Given that each customer chooses to join server i with probability ϕ_i, which is also referred to as server i's market share, the effective arrival rate to server i is $\phi_i\Lambda$. Customers arriving at server i can still be regarded as an $M/M/1$ queue. Thus, the expected sojourn time (including the service time) is $1/(\mu_i - \phi_i\Lambda)$. Therefore, the expected utility for a customer joining the queue at server i is the reward from receiving the service less the price paid and the expected sojourn cost, as given as follows.

$$U_i(\mu_i) = R(\mu_i) - p - \frac{C}{\mu_i - \phi_i\Lambda}. \tag{1}$$

Note that an arriving customer has $N + 1$ options: either balking or joining server i, for $i \in \{1, \ldots, N\}$. For simplicity, we denote balking as join server $i = 0$ and we set $U_0(\cdot) = 0$. Note that ϕ_0 can be determined when the probabilities of joining N servers are determined, i.e., $\phi_0 = 1 - \sum_{j=1}^{N}\phi_j$. Next, we will focus on ϕ_i, where $i = 1, 2, \ldots, N$.

Recall that we consider boundedly rational customers. That is, customers have limited cognitive ability in assessing service quality and sojourn time. Therefore, like Huang et al. (2013) and Li et al. (2016), we use the multinomial logit choice model to formulate boundedly rational customers' choices, and we focus on the probability of a representative customer choosing server i, i.e., ϕ_i, $i = 0, 1, 2, \ldots, N$. From McKelvey and Palfrey (1995), we have

$$\phi_i = \frac{e^{U_i(\mu_i)/\beta}}{\sum_{j=0}^{N} e^{U_j(\mu_j)/\beta}} = \frac{e^{U_i(\mu_i)/\beta}}{1 + \sum_{j=1}^{N} e^{U_j(\mu_j)/\beta}}, \quad i = 1, 2, \ldots, N, \tag{2}$$

where the parameter $\beta > 0$ measures the level of bounded rationality (see Huang et al. (2013)). If $\beta \to 0$, customers are fully rational. Note that Li et al. (2016) consider a queueing system with two servers, i.e., the special case $N = 2$, and they explain in details the logic on how to derive the above probabilities from McKelvey and Palfrey (1995). Although Li et al. (2016)'s analysis is based on the two-server case, it is straightforward to extend their analysis to the N-server case. Therefore, we omit the derivation.

From McKelvey and Palfrey (1995), there is a unique solution $(\phi_1, \phi_2, \ldots, \phi_N)$ satisfying the equation set (2). There are N equations in set (2). Solving these N equations, we can obtain customers' joining probabilities ϕ_i, $i = 1, 2, \ldots, N$. To facilitate the analysis, we rewrite (2) as follows.

$$\beta \ln \frac{\phi_i}{1 - \sum_{j=1}^{N}\phi_j} = R(\mu_i) - p - \frac{C}{\mu_i - \phi_i\Lambda}, \quad i = 1, 2, \cdots, N. \tag{3}$$

3. Servers' best responses and the equilibrium

The equation set (3) gives customers's joining probabilities in implicit forms, from which we will derive servers' best response functions, and then the equilibrium strategy. For notational convenience, we denote $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$. To determine the Nash equilibrium on the service rates, we consider a tagged server's decision, say, server k, for any $k = 1, 2, \ldots, N$, and we assume other servers' decisions $\mu_1, \mu_2, \ldots, \mu_{k-1}, \mu_{k+1}, \ldots, \mu_N$ are given.

Follow the assumption in Anand et al. (2011) and Li et al. (2016), we assume that any server's capacity cost is fixed and is not affected by the service rate. Then, the tagged server k's best response μ_k maximizes the corresponding market share ϕ_k. From the implicit function (3), ϕ_k depends on $\boldsymbol{\mu}$. Since we assume that the other $N - 1$ servers' decisions are given, we denote $\phi_k(\mu_k)$ for simplicity. We first investigate the properties