Strategic Customer Behaviors in Observable Multi-server Batch Service Queueing Systems with Shared Fee and Server Maintenance Cost

Ayane Nakamura[1(✉)] and Tuan Phung-Duc[2]

[1] Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan
s2230122@u.tsukuba.jp

[2] Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
tuan@sk.tsukuba.ac.jp

Abstract. We consider an observable multi-server batch service queueing model where a shared server admission fee and maintenance cost for the monopolist are considered. Specifically, customers observe the number of full batches and waiting customers for an uncompleted batch, and then afterward decide to join or balk depending on their utility function. We analyze strategic customer behavior and find the equilibrium strategy. We also show numerical results for social welfare and monopolist's revenue in various settings depending on fee, batch size, and number of servers. Based on our results, we discuss the existence of optimal fee, batch size, and number of servers.

Keywords: Queueing · Strategic customers · Batch service · Shared fee · Server cost · Social welfare · Ride-sharing

1 Introduction

There has been a recent rise in share services. The most common example is public transportation. However, shared mobility services, such as ridesharing [1] have recently received attention with the spread of electronic cars and automatic vehicles. On ride-sharing platforms, customers with the same departure and arrival points form a group and travel together. A key feature of this system is that a customer's behavior is influenced by others, and vice versa. For instance, if there is less demand for ridesharing, a customer might abandon it as the waiting time for matching with other customers may be long.

In the field of queueing theory, many studies on the analysis of batch service queueing models with single, multi, and infinite servers (see e.g., [10,13,15]) have been conducted over the past half-century. Although these studies have derived the joint steady-state probabilities on the number of busy servers and batches,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
E. Hyytiä and V. Kavitha (Eds.): VALUETOOLS 2022, LNICST 482, pp. 3–13, 2023.
https://doi.org/10.1007/978-3-031-31234-2_1

most prior research has not considered customers' strategic behavior depending on system states. One useful approach to analyze such customer behaviors is the economic analysis of queueing models from a game-theoretic perspective with customer choice following [14].

However, there are few studies dealing with batch service (i.e., group service) queueing systems with customer strategic behavior. Some research on clearing systems considering customer strategic behavior has been conducted (see e.g., [7–9,12]). The strategic behavior of customers in queueing systems with catastrophes has also been studied [2,3]. In a recent study, the routing decisions of strategic passengers in a transportation station were analyzed [11]. In addition, M/M/1 queues with batch services [4–6] have been studied for unobservable, observable, and partially observable cases.

As an extension of this research on single server batch service systems, we consider an observable M/M/n queueing model with strategic customers assuming that a monopolist imposes an admission fee. Moreover, we analyze the model considering the cost of maintaining servers for the monopolist. In the context of a ride-sharing system, multiple servers in this model correspond to the vehicles on the platform. Considering such shared-mobility systems, the scenario where a monopolist (i.e., operator) prepares many vehicles is more believable. The server maintenance cost corresponds to the vehicles' gasoline and parking costs. Here, determining the optimal number of servers from the perspective of both monopolist and society, given the trade-off between customer income and server maintenance costs, is significant. We can assume that customers decide to use ridesharing depending on the expected waiting time. In addition, we assume that server service time corresponds to the time from when the car leaves the station to when it returns.

Studies on the queueing analysis of strategic customers in multi-server systems are few even though they have many practical applications. For example, studies on strategic joining in a M/M/K queue with asynchronous and synchronous multiple vacations and M/M/c/Setup queue were recently conducted [16,17]. However, to the best of our knowledge, there is no research related to multi-server batch service queueing systems with strategic customers. As our model includes elements of batch service and multi-servers, we believe the analysis of this comprehensive model will be meaningful not only from a practical perspective but also theoretical.

The rest of this study is organized as follows. In Sect. 2, we describe the model setting and find the equilibrium strategy for customers. Using this, we show numerical results in Sect. 3 and discuss the optimal fee, batch size, and number of servers. Finally, Sect. 4 presents the conclusion and directions for future research.

2 Modelling

We consider a multi-server (n server) queueing model with batch service. We assume that the batch size is K and fee of service per batch is p, i.e., the fee for

a customer is p/K. We also assume that the service time follows an exponential distribution with parameter μ. Customers arrive at the system according to a Poisson process with parameter λ and decide whether to join or balk the system after observing the state of the system. Specifically, we assume that a customer's threshold strategy is specified by a sequence $\mathbf{q} = (q_{0,0}, q_{0,1}, q_{0,2}, \dots)$, where $q_{m,j} \in \{0,1\}, m = 0,1,2,\dots, j = 0,1,\dots,K-1$ when an arriving customer finds m full batches and j customers for an uncompleted batch. The total number of customers in the system thus becomes $mK + j$.

Suppose that a tagged customer follows strategy $\mathbf{q}' = (q'_{0,0}, q'_{0,1}, q'_{0,2}, \dots)$ while the population of other customers follows $\mathbf{q} = (q_{0,0}, q_{0,1}, q_{0,2}, \dots)$. If the tagged customer sees the state (m, j) upon his arrival to the system, his utility is expressed as follows:

$$U(\mathbf{q}', \mathbf{q}; (m, j)) = R - \frac{p}{K} - C_W \times S(m, j),$$

where R, C_W, and $S(m, j)$ denote the reward of receiving a service, waiting cost per unit time, and expected waiting time when he observes the state (m, j). Here, $R > p/K$ must always hold to avoid the obvious case where no one joins the queue. Thus, we can proceed with the analysis of multi-server model following a similar method as in the single server model of [4].

Lemma 1. *There exists a unique equilibrium strategy* $(q^e_{m,j}; m \geqq 0, 0 \leqq j \leqq K-1)$ *as*

$$q^e_{m,j} = \begin{cases} 1 \text{ if } m \leqq m^e_j, \\ 0 \text{ if } m > m^e_j, \end{cases}$$

where $m^e_0 \leqq m^e_1 \leqq \cdots \leqq m^e_{K-1}$.

Proof. The proof method mainly follows [4]. First, we focus on the case where an arriving customer observes $(m, K-1)$. Once he joins the system, his batch is completed. In this case, his waiting time becomes 0 if $m < n$, and the sum of $(m - n + 1)$ exponential service times with parameter $n\mu$ if $m \geqq n$. Therefore, the following equilibrium threshold holds:

$$q^e_{m,K-1} = \begin{cases} 1 \text{ if } m \leqq m^e_{K-1}, \\ 0 \text{ if } m > m^e_{K-1}, \end{cases}$$

where $m^e_{K-1} = \left\lfloor \frac{n\mu(R-p/K)}{C_W} + n - 1 \right\rfloor$. This strategy is dominant. Since his batch is completed once he joins the system, his decision will not be affected by other customers.

Next, we consider the case where an arriving customer sees the state $(m, K-2)$. As discussed in [4], we have $q^e_{m,K-2} = 0$ if $q^e_{m,K-1} = 0$. Additionally, $S(m, K-2)$ is an increasing function of m that tends to ∞ as $m \to \infty$, which means there exists a unique m^e_{K-2} such that $S(m^e_{K-2}, K-2) \leqq (R - p/K)/C_W < S(m^e_{K-2} + 1, K-2)$. Thus, the following holds:

$$q^e_{m,K-2} = \begin{cases} 1 \text{ if } m \leqq m^e_{K-2}, \\ 0 \text{ if } m > m^e_{K-2}, \end{cases}$$

Thus, $m^e_{K-2} \leqq m^e_{K-1}$ as $q^e_{m,K-2} = 0$ if $q^e_{m,K-1} = 0$ as described earlier. Repeating this process for $j = K-3, K-4, \ldots$, we easily obtain this lemma. □

Next, we prepare the following lemma for $S(m,j)$ to derive m^e_j.

Lemma 2. *The expected waiting time for a tagged customer who observes state* (m,j) *is*

$$S(m,j) = \begin{cases} \dfrac{K-j-1}{\lambda} & \text{if } m < n, \\ \dfrac{m-n+1}{n\mu} + \sum_{i=0}^{K-j-1} \left(\dfrac{K-j-i-1}{\lambda}\right) \binom{m-n+i}{i} & \\ \qquad \times \left(\dfrac{\lambda}{\lambda+n\mu}\right)^i \left(\dfrac{n\mu}{\lambda+n\mu}\right)^{m-n+1} & \text{if } m \geqq n. \end{cases}$$

Proof. The proof method mainly follows [4]. Given the case $j = K-1$, the waiting time becomes 0 if at least one server is available, i.e., $m < n$, since a tagged customer is the last customer for completing a batch. If there is no available server, i.e., $m \geqq n$, a tagged customer must wait until $m-n+1$ batches are served. Given the case $j \neq K-1$. If $m < n$, a tagged customer must wait until his batch is completed, i.e., $K-j-1$ customers must arrive. Otherwise, if $m \geqq n$, the expected waiting time can be expressed using the maximum value of two random variables as follows:

$$S(m,j) = \mathrm{E}[\max(Y_{m-n+1}, Z_{K-j-1})], \qquad \text{if } j \neq K-1, m \geqq n,$$

where Y_{m-n+1} and Z_{K-j-1} denote Erlang-$(m-n+1)$ and Erlang-$(K-j-1)$ random variables with rates $n\mu$ and λ, respectively. The former corresponds to the total service time until a server becomes available and the latter to the waiting time until his batch is completed. A waiting time of a tagged customer equals the maximum value of these two times.

Let $N(Y_{m-n+1})$ denote the number of newly arrived customers during time Y_{m-n+1}. We can thus obtain the following transformation:

$$\begin{aligned} \mathrm{E}[\max(Y_{m-n+1}, Z_{K-j-1})] &= \mathrm{E}[Y_{m-n+1}] + \mathrm{E}[\max(0, Z_{K-j-1} - Y_{m-n+1})] \\ &= \frac{m-n+1}{n\mu} + \sum_{i=0}^{\infty} \mathrm{P}(N(Y_{m-n+1}) = i) \times \\ &\qquad \mathrm{E}[\max(0, Z_{K-j-1} - Y_{m-n+1}) \mid N(Y_{m-n+1}) = i] \\ &= \frac{m-n+1}{n\mu} + \sum_{i=0}^{K-j-1} \mathrm{P}(N(Y_{m-n+1}) = i) \times \\ &\qquad \frac{K-j-i-1}{\lambda}. \end{aligned}$$

Here, we obtain

$$\begin{aligned} \mathrm{P}(N(Y_{m-n+1}) = i) &= \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^i}{i!} \frac{(n\mu)^{m-n+1}}{(m-n)!} x^{m-n} e^{-\mu n x} dx \\ &= \binom{m-n+i}{i} \left(\frac{\lambda}{\lambda+n\mu}\right)^i \left(\frac{n\mu}{\lambda+n\mu}\right)^{m-n+1}. \end{aligned}$$

This concludes the proof. □

Therefore, we can immediately obtain the following theorem.

Theorem 1. *The thresholds* m_j^e $(0 \leqq j \leqq K-1)$ *of Lemma 1 are*

$$m_{K-1}^e = \left\lfloor \frac{n\mu(R - p/K)}{C_W} + n - 1 \right\rfloor,$$

$$m_j^e = \max\left\{ m; 0 \leqq m \leqq m_{j+1}^e \text{ and } S(m,j) \leqq \frac{R - \dfrac{p}{K}}{C_W} \right\}, \qquad 0 \leqq j \leqq K-2.$$

We can thus obtain the steady state probability of the system $\pi_{j,m}$ numerically under the state space $S = \{(j,m); j \in \{0,1,\ldots,K\}, m \in \{0,1,\ldots,m_j^e + 1\}\}$. In addition, we obtain the blocking probability P_b, expected number of busy servers $\mathrm{E}[S_b]$, expected number of idle servers $\mathrm{E}[S_i]$, and expected number of waiting customers $\mathrm{E}[L]$, respectively, as follows (note that we use Little's law for $\mathrm{E}[S_b]$, and also PASTA for P_b):

$$P_b = \sum_{j=0}^{K-1} \pi_{j,m_j^e+1},$$

$$\mathrm{E}[S_b] = \frac{\lambda(1-P_b)}{\mu K},$$

$$\mathrm{E}[S_i] = n - \frac{\lambda(1-P_b)}{\mu K},$$

$$\mathrm{E}[L] = \sum_{(j,m)\in S} \{\max(0, m-n)K + j\}\pi_{j,m}.$$

Furthermore, we consider server maintenance cost such as energy or parking costs in a ride-sharing platform. Let C_b and C_i denote the maintenance costs of a server when it is busy and idle, respectively. Here, we simply assume that $C_b > C_i$. We can derive the expected server maintenance cost $\mathrm{E}[M]$ as follows:

$$\begin{aligned}\mathrm{E}[M] &= C_b\mathrm{E}[S_b] + C_i\mathrm{E}[S_i] \\ &= (C_b - C_i)\frac{\lambda(1-P_b)}{\mu K} + C_i n.\end{aligned}$$

Using these performance measures, we obtain social welfare represented by SW as

$$\begin{aligned}SW &= \lambda\left(R - \frac{p}{K}\right)(1-P_b) - C_W\mathrm{E}[L] + \lambda(1-P_b)\frac{p}{K} \\ &= \lambda R(1-P_b) - C_W \sum_{(j,m)\in S} \{\max(0,m-n)K + j\}\pi_{j,m}.\end{aligned}$$

Then, the monopolist's revenue MR is calculated by

$$\begin{aligned} MR &= \lambda(1 - P_b)\frac{p}{K} - \mathrm{E}[M] \\ &= \left(p - \frac{C_b - C_i}{\mu}\right)\frac{\lambda(1 - P_b)}{K} - C_i n. \end{aligned}$$

3 Numerical Examples

This section shows numerical results for social welfare SW and monopolist's revenue MR. We show the results of those measures for various admission fee p, batch size K, and number of servers n. We discuss the optimal values of those parameters using our numerical results.

3.1 Fee (p)

There are optimal p such that they maximize both SW and MR. This result means that if a monopolist chooses the optimal p, this is also approximately social optimal. One reason considered is that the common function P_b in SW and MR has a large impact on the plummeting point of p. Additionally, customers never join the system when p increases beyond a certain level, resulting in a sharp decrease in both measures.

3.2 Batch Size (K)

Fig 2 shows the results for batch size K under the parameter setting $\mu = 30$, $n = 3$, $p = 20$, $C_W = 100$, $R = 30$, $C_b = 300$, and $C_i = 50$. For SW, there exists an optimal K unless λ is small, i.e., there is some customer demand for the system. As K increases, more customers can use the system by sharing resources (i.e., low fee). Therefore, people tend to obtain a reward R, increasing social welfare. However, if K becomes large, the customer waiting time and blocking probability customers will be large, reducing SW. Previous research on batch service M/M/1 queues without admission fees [4] showed that equilibrium social welfare tendency decreases as batch size increases. In this study, the effect of the sharing economy for SW will be clear considering shared fees (p/K).

On the other hand, MR decreases as K increases for any λ under the same parameter setting. This is probably due to the fee setting. We simply assume that one batch service is for p regardless of batch size; thus, fee per customer becomes p/K. By this assumption, customer fee decreases inversely as batch size increases. Although we consider server cost, this effect is large. In Sect. 4, we discuss how a more realistic and elaborate fee setting in future work may improve this tendency.

3.3 Number of Servers (n)

Finally, we show numerical examples for n. We assume that the parameter setting as $\mu = 30$, $K = 3$, $p = 50$, $C_W = 100$, $R = 30$, $C_b = 10$, and $C_i = 5$. In both graphs, there exist optimal n for large values of λ. These results show the trade-off between the increase in joining customers and server maintenance cost for n. Observing the graph in more detail, the optimal n in terms of social welfare is the same or larger than that of monopolist's revenue. Naturally, the monopolist hesitates to introduce additional servers from the perspective of server maintenance cost. However, from the social view, it seems better to introduce more additional servers to receive more customers and to reduce the waiting time and the blocking probability.

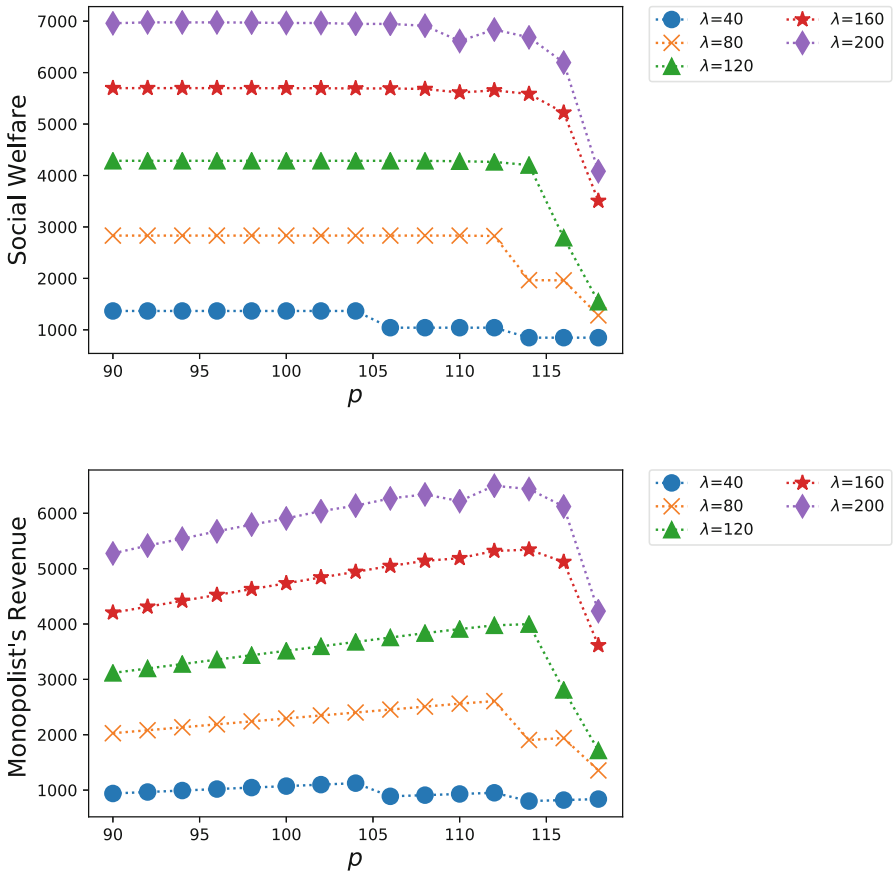

Fig. 1. Social welfare (SW) and monopolist's revenue (MR) against p.

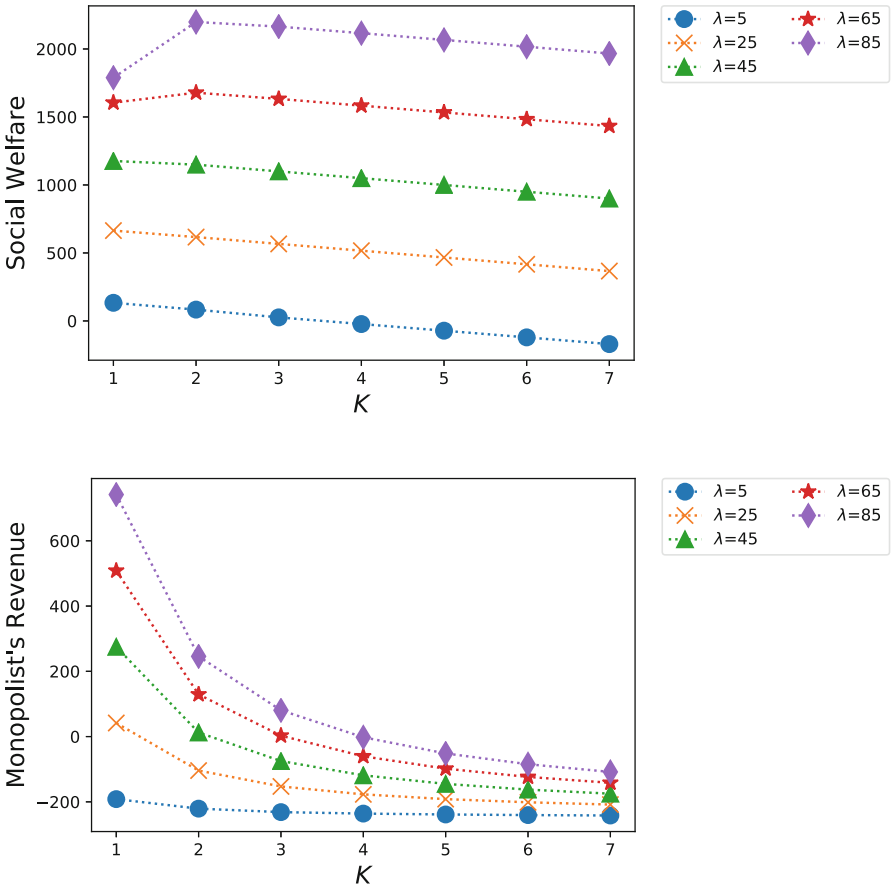

Fig. 2. Social welfare (SW) and monopolist's revenue (MR) against K.

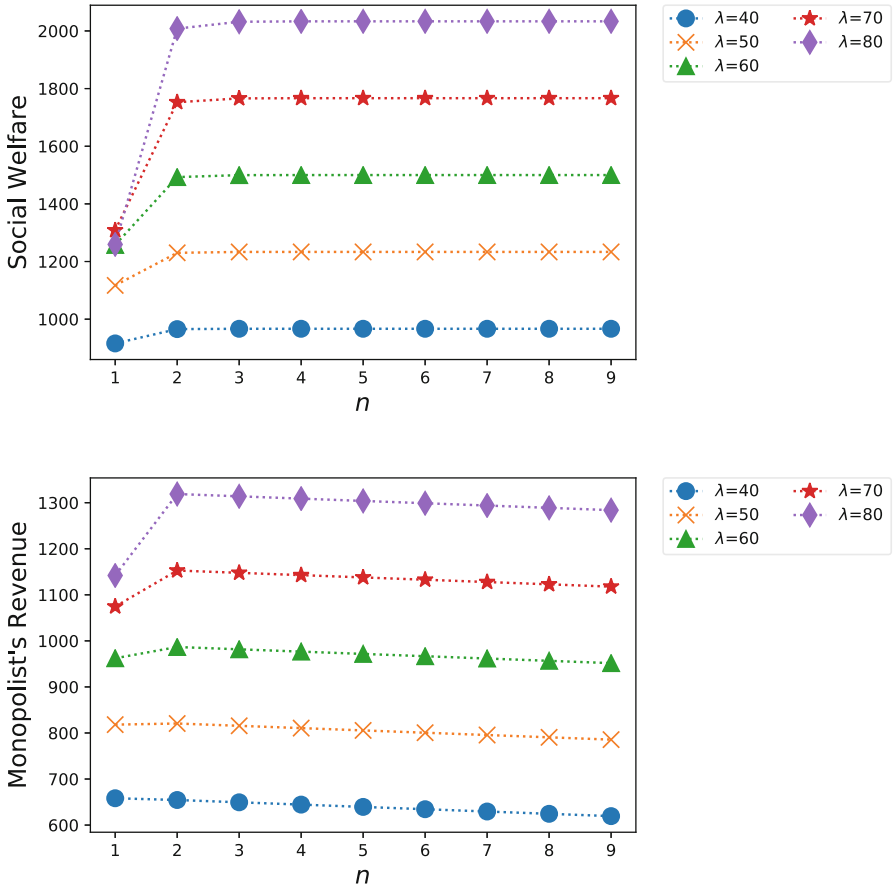

Fig. 3. Social welfare (SW) and monopolist's revenue (MR) against n.

4 Conclusion

In this study, we studied the observable batch service M/M/n queueing system mainly for ridesharing applications, considering shared customer admission fee and server maintenance cost. Using analysis and numerical experiments, we have shown various tendencies of social welfare and monopolist's revenue against fees, batch sizes, and numbers of servers. For fees and number of servers, there exist some optimal points that maximize the two measures for relatively large customer arrival rates. For batch size, we have observed the optimal value to maximize social welfare. However, the monopolist's revenue decreases as batch size increases under the admission fee setting in this study.

An interesting topic for future research would be to investigate better ways to share fees such that the monopolist's revenue and social welfare increase. Although our equal distribution assumption is simple, it may be possible to

consider a distribution method in which the longer the waiting time, the lower the fee. Moreover, a fee of K customers should not equal p. From the perspective of the monopolist, it may be more natural that a customer pays a fixed basic fee and an additional fee depending on batch size K. Finding a beneficial fee policy that improves both social welfare and monopolist's revenue in batch service systems will be challenging and significant future work.

We must also consider the effect of introducing multi-servers to society. In this study, we assumed that social welfare (SW) does not depend on server maintenance cost, since we assume that society's constituents receive the cost from the monopolist. However, excessive servers may negatively impact the environment and traffic congestion in the context of transportation systems. Deriving the optimal number of servers considering these elements is thus highly meaningful.

Furthermore, the extensions to unobservable or partially observable queues are also significant. Recently, the concept of Mobility as a Service (MaaS) has expanded widely [18]. In this framework, an operator (in this study, a monopolist) manages various transportation services and the recommended transportation modes for customers considering fee and waiting time through a web platform. Therefore, the observability of customers will be more flexible than the old transportation system such as buses or trains that depart according to fixed timetables and fees. The comparison among various policies, i.e., unobservable, partially observable, and observable is necessary for building an effective platform.

In this study, we have made simple assumptions using various parameters (e.g., K, C_W, C_b and C_i). However, it is essential to estimate realistic values using real data. This will be the focus of our next study.

References

1. Agatz, N., Erera, A., Savelsbergh, M., Wang, X.: Optimization for dynamic ride-sharing: a review. Eur. J. Oper. Res. **223**(2), 295–303 (2012)
2. Boudali, O., Economou, A.: Optimal and equilibrium balking strategies in the single server markovian queue with catastrophes. Eur. J. Oper. Res. **218**(3), 708–715 (2012)
3. Boudali, O., Economou, A.: The effect of catastrophes on the strategic customer behavior in queueing systems. Naval Res. Logistics (NRL) **60**(7), 571–587 (2013)
4. Bountali, O., Economou, A.: Equilibrium joining strategies in batch service queueing systems. Eur. J. Oper. Res. **260**(3), 1142–1151 (2017)
5. Bountali, O., Economou, A.: Equilibrium threshold joining strategies in partially observable batch service queueing systems. Ann. Oper. Res. **277**(2), 231–253 (2019)
6. Bountali, O., Economou, A.: Strategic customer behavior in a two-stage batch processing system. Queueing Syst. **93**(1), 3–29 (2019)
7. Canbolat, P.G.: Bounded rationality in clearing service systems. Eur. J. Oper. Res. **282**(2), 614–626 (2020)
8. Czerny, A.I., Guo, P., Hassin, R.: Hide or advertise: the carrier's choice of waiting time information strategies. SSRN 3282276 (2018)
9. Economou, A., Manou, A.: Equilibrium balking strategies for a clearing queueing system in alternating environment. Ann. Oper. Res. **208**(1), 489–514 (2013)

10. Ghare, P.: Multichannel queuing system with bulk service. Oper. Res. **16**(1), 189–192 (1968)
11. Logothetis, D., Economou, A.: Routing of strategic passengers in a transportation station. In: Ballarini, P., Castel, H., Dimitriou, I., Iacono, M., Phung-Duc, T., Walraevens, J. (eds.) EPEW/ASMTA -2021. LNCS, vol. 13104, pp. 308–324. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91825-5_19
12. Manou, A., Economou, A., Karaesmen, F.: Strategic customers in a transportation station: when is it optimal to wait? Oper. Res. **62**(4), 910–925 (2014)
13. Nakamura, A., Phung-Duc, T.: Stationary analysis of infinite server queue with batch service. In: Ballarini, P., Castel, H., Dimitriou, I., Iacono, M., Phung-Duc, T., Walraevens, J. (eds.) EPEW/ASMTA -2021. LNCS, vol. 13104, pp. 411–424. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91825-5_25
14. Naor, P.: The regulation of queue size by levying tolls. Econometrica J. Econometric Soc. **37**, 15–24 (1969)
15. Neuts, M.F.: A general class of bulk queues with poisson input. Ann. Math. Stat. **38**(3), 759–770 (1967)
16. Nguyen, H.Q., Phung-Duc, T.: M/m/c/setup queues: conditional mean waiting times and a loop algorithm to derive customer equilibrium threshold strategy. In: Gilly, K., Thomas, N. (eds.) Computer Performance Engineering. LNCS, vol. 13659, pp. 68–99. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25049-1_6
17. Wang, J., Zhang, Y., Zhang, Z.G.: Strategic joining in an M/M/K queue with asynchronous and synchronous multiple vacations. J. Oper. Res. Soc. **72**(1), 161–179 (2021)
18. Wong, Y.Z., Hensher, D.A., Mulley, C.: Mobility as a service (MaaS): charting a future context. Transp. Res. Part A: Policy Pract. **131**, 5–19 (2020)