Endogenous Queue Number Determination in G/M/s Systems

Alves, Vasco

DOI:
10.1007/s10288-020-00437-y

License:
Unspecified

Citation for published version (Harvard):
Alves, V 2020, 'Endogenous Queue Number Determination in G/M/s Systems', *4OR*.
https://doi.org/10.1007/s10288-020-00437-y

Endogenous Queue Number Determination in G/M/s Systems

Vasco F. Alves
Department of Economics
University of Birmingham
Edgbaston
Birmingham, Warwickshire, B15 2TT
United Kingdom
v.alves@bham.ac.uk

March 20, 2020

Abstract

This paper presents a model for the endogenous determination of the number of queues in an G/M/s system. Customers arriving at a system where s customers are being served play a game, choosing between s parallel queues or one single queue. Equilibria are obtained for risk-neutral and risk-averse customers. With risk-neutral customers, both a single queue and multiple queues are equilibrium states, and there is scope for mixed strategies. When risk-averse customers are considered, there is a unique single queue equilibrium. These results are discussed and suggestions for further research put forth.

JEL: C73 C72

Keywords: Queues - Applications: strategic interactions. Queues - Multichannel: determining number. Games/group decisions: strategic queueing.

1 Introduction

Queues form naturally whenever there is some delay in service time necessary for the provision of a good, and the number of providers is smaller than the current number of customers. Queues force customers to suffer the cost of time spent in the queue, as well as the monetary cost of the good. Customers will want to minimize this cost, and increasing queueing efficiency can yield significant social benefits: witness the rise of self-service check out points at supermarkets.

The present paper takes place in the context of s parallel G/M/1 systems, and pooled G/M/s systems, where S is any finite number of servers, under a First Come First Served (FCFS) discipline where reneging is not allowed. It sits within the strategic queueing literature, where strategic interactions between customers in queues are modelled through game theory.

The seminal Naor (1969) considered the setting of an FCFS M/M/1 queue. In this model, risk-neutral, utility maximizing customers, with a linear utility function, choose a joining threshold, the largest queue length for which the expected cost of waiting is weakly smaller than the service's net value. Once this happens, customers will balk without the need for an exogenous capacity limit. Naor showed that in such a queue, average queue length grows beyond the social welfare maximizing level, and that a social planner can improve social welfare, attaining

a first-best optimum where aggregate waiting time is minimized. This is achieved by shifting the cost structure faced by arriving customers, through levying a toll on customers who join the queue, thereby adding its cost to the cost of waiting and reducing the threshold at which customers join the queue.

Naor's result was extended in Knudsen (1972) to a general cost function, and a system with a single queue served by any finite number of servers. Knudsen found Naor's result on tolling held even under these relaxed conditions, and crucially for the present purposes, extended his framework for individual optimization to the more general case.

Naor's paper was followed by a variety of further articles examining customers' strategic queueing behaviour, especially in M/M/1 FCFS queues. For a good overview of the literature up to publication, see the review monographs Hassin and Haviv (2003) and Hassin (2016). Since then, many more papers than can be individually mentioned have been published on this subject.

The second strain of literature relevant for the present paper centres on queues being considered among what is described in Parsons (1955) as social systems, in that they involve interactions between individuals according to some set of socially agreed upon norms. These sorts of interactions can be modelled as a game, which can then be investigated with standard game theoretic tools, such as the theory of repeated games, as described in Okuno-Fujiwara and Postlewaite (1995) (and see Mailath and Samuelson (2006), *inter alia*, for a thorough review of the repeated games literature). Kandori (1992) showed the applicability of this type of analysis to situations where game 'partners' change by describing a process where 'punishment' for deviating from social norms is meted out by the community rather than by the aggrieved individuals only. The extent to which queueing is governed by these social norms has been the object of research in the Psychology and Sociology literatures, following on Schwartz (1975), which laid out a sociological analysis of waiting for service and customers' perceptions of the fairness of queueing disciplines. Allon and Hanany (2012) studies a setting where, in the context of repeated interactions and changing priorities, customers allow queue cutting when their priority is low, with the expectation of being allowed to cut ahead in future rounds of the game, when their priority is high. Erlichman and Hassin (2015) looks at a similar problem, but with priorities being sold by the server. The slightly different case of an unobservable M/G/1 queue is analysed in Haviv and Ravner (2016), where an efficiency enhancement pricing mechanism is also presented.

Returning to the issue of the number of queues for multiple service points, while it seems intuitively appealing that a single queue for two servers is more socially efficient than one queue for each, this was only formally demonstrated in Smith and Whitt (1981) (but see Rothkopf and Rech (1987) for some situations, not relevant to the current paper, where this may not hold[1]). The source of this inefficiency is that if customers cannot switch queues, then one of the servers may be idle while there are customers waiting to be served in other queues. The recent work in Sunar et al. (2017), however, has shown that when customers are risk-neutral, delay sensitive, and may balk, dedicated queues may be preferable to combining them.

[1]For instance, management may want to use separate queues as a discrimination mechanism: supermarkets often have queues for customers with less items. This, however, requires customer heterogeneity, which is not a feature of the model outlined here.

Where multiple queues are present despite their inefficiency, it has been shown that in M/M/s systems (where s is any finite number of servers) where all servers have the same service time distribution, customers should join the shortest queue, and break ties arbitrarily (Winston (1977)). Where expected waiting times vary with servers, there have been attempts to determine if customers might be better off waiting to gain information about these, such as that in Hlynka et al. (1994).

Nevertheless, in the light of its inefficiency, the persistence of multiple parallel queues presents something of a conundrum. While combining queues seems to be optimal, it often does not match the observed behaviour of customers in day to day transactions. This may be due to managers enforcing a multiple queue discipline, but in many cases managers don't seek to direct customers one way or the other. Why is it, then, that customers sometimes form multiple queues for multiple service points, and other times only one? The motivation behind the present paper is to discover whether and in what circumstances this socially optimal outcome is sustainable without management intervention—is it individually optimal? Is the incidence of this behaviour related to customers' risk aversion? Armony and Plambeck (2005) studies a related problem on unobservable queues, where customers can place duplicate orders in the presence of two service points, to protect themselves against supply shocks. Dehghanian et al. (2016) considers jockeying by strategic, risk-neutral customers, between two parallel queues (assumed as the given system structure), finding it may not be optimal to initially join the shortest queue. Likewise Ganesh et al. (2012) studies jockeying between parallel queues, showing that 'smart' jockeying does not significantly affect system-wide sojourn times compared to a 'random' strategy. Ata and Olsen (2009) studies the case of a monopolistic server faced with, *inter alia*, risk-averse customers, and prescribes asymptotically optimal pricing policies.

The literature has usually assumed that the number of queues which will form in the presence of multiple servers is the choice of the service station manager. As such, they would be the ones to blame for the formation of multiple queues. Rothkopf and Rech (1987) presents some suggestions as to why this might be the case, but even if these arguments are valid, they do not explain the emergence of multiple queues where there is no managerial intervention, such as at self-service points. Zhang et al. (2008) considers the concept of a 'blind' scheduler who makes scheduling decisions without knowledge of the system state, another setting where management intervention is limited.

The present paper attempts to answer this question by setting forth a model where strategic interactions between customers determine the number of queues in a system. Anecdotally, when they are not prompted to form a given number of queues, customers faced with busy service points but no queue most often attempt to form a single queue for all of them, and move to the first service point to become free. The problem with this strategy is that this position straddling multiple service points can be interpreted by new arrivals as permission to queue for only one of the servers, and the first customer cannot stop this as any attempt to move to block the new arrival forces the incumbent to move away from the other service points and commit to that one anyway; most readers can probably relate to this experience.

The model setting is a system with multiple servers under a no-jockeying condition, covering in turn risk-neutral and risk-averse customers. The game starts when a customer arriving at the system encounters all servers as busy, but no queue (if at least one server is idle, customers'

decision is trivial). It will be outlined how the number of queues is determined through this multi-stage game, whereby later arrivals can disrupt a single queue, and so their potential future decisions must be accounted for by earlier customers. The first arrival will be demonstrated to strictly prefer a single queue.

The intuition behind this preference for the single queue is that this customer can be served as soon as the first service occurs, rather than having to guess at which server will finish the current task first. On the other hand, the sth customer (where s is the number of servers) does not always have the same benefits from that single queue: if customers are risk-neutral, customer s is indifferent to the number of queues. In the case of risk-neutral customers, it will be shown how customers alternate between strictly preferring one queue and being indifferent to the number of queues, in blocks of s customers. This will lead to a proof that having a single queue is an equilibrium outcome for this game. This equilibrium is not unique, however, with s queues also being an equilibrium state.

In order to address the presence of multiple equilibria, section 3 focuses on risk-averse customers, arguably a more true-to-life setting. It is found that risk aversion quashes the multiple queue equilibrium, leaving the single queue state as the unique equilibrium.

Steady-state properties will not be considered, as the situation being modelled takes place when the queue is starting to form, before a steady-state has emerged. Therefore joining customers will not face the steady-state expected waiting time, but an individual expected waiting time which varies with the system state at their arrival. The strategic interactions modelled in the game relate to how incumbent customers deal with arrivals to the system, who might disrupt the present order by trying to change the number of queues.[2]

The model presented here is especially relevant for situations where there is no channel for managers to interact with customers to establish the number of queues, such as at any self-service point, or where for some reason engagement with the public is discouraged—such as when selling tickets behind bullet-proof glass windows on dangerous parts of a transport network. Further, the model advances the analysis of strategic interactions between customers.

2 Queue Number Determination with risk-neutral Customers

Consider a stream of customers seeking a service the provision of which requires a queue; their arrivals at the service station may follow any general distribution for inter-arrival time. This service is provided by s identical servers. Obtaining the good from these servers takes time, distributed according to an exponential distribution with rate μ. While the arrival process is not relevant for the game's equilibrium outcome, the results rely on the exponential distribution of service times, in particular the exponential distribution's memoryless property. Generalizing beyond this distribution is left for further research.

As there are s servers, only s customers can be served simultaneously. Others will wait until

[2]While addressing a different problem, that of whether customers let others cut ahead on the queue in an M/M/1 system, Allon and Hanany (2012) also addresses how customers deal with violations of social norms, and reaches a conclusion with a similar tenor: undirected customers can, at least in some circumstances, reach socially efficient outcomes through strategic interaction, although it's important to note that unlike the present model, Allon and Hanany (2012) is set in the context of repeated games.

a server becomes available, and are served in order according to the First Come, First Served (FCFS) discipline. It is possible for the system to be organized as s parallel G/M/1 queues, where each server services a separate FCFS queue, and customers must choose one queue to join, or as one single G/M/s queue serviced by all s servers, where the customer at the head of the queue is served by the first server to become free. The number of queues is endogenously determined through customer choices, being the game's equilibrium outcome.

While it is possible to imagine sub-groups of queues, e.g., one for servers 1-$s/2$, and another for servers $s/2+1$-s, or other, possibly assymetric combinations. However, only total pooling or separation are considered in this setting. This is for two reasons. First, for any reasonably small number of queues, this kind of 'partial pooling' is not consistent with observed patterns of endogenous customer behaviour. As such, any results would have limited application. Second, as in principle 'partial pooling' can be asymmetric (indeed must be so if the number of servers is odd), and the number of possible combinations increases with the number of servers, the mathematical complexity of the problem is greatly increased for limited benefit. Research along these lines is left for future work.

Only situations where all servers are active will be considered here, so this can be assumed and need not be explicitly stated in characterizing the system state, and the queue lengths do not include them (i.e., they number only the customers waiting). This state can be described by a matrix $\boldsymbol{\Theta_Q}$ composed of $Q \in \{1, \ldots, n\}$ column vectors $\boldsymbol{\theta_q}$, each with $I \in \{1, 2, ..., \infty\}$ elements, where Q is the number of queues in the system, q the (arbitrary) index of each queue, and I the maximum length of each queue, where each element $\theta_{i,j}$ is 0 or 1 depending on whether a customer is queueing in the place in the queue corresponding to that element. If a given element $\theta_{\underline{i},\underline{q}} = 1$, it must be the case that $\theta_{i,\underline{q}} = 1 \, \forall \, i < \underline{i}$, i.e., the queue cannot have gaps in it. Further, if queue length is $\underline{i}$, then $\theta_{i,\underline{q}} = 0 \, \forall \, i > \underline{i}$. Queue length $L_q = \sum_{i=1}^{I} \theta_{i,q}$ for a given q, and total number of customers waiting in the system $L = \sum_q L_q$. Finally, assume by convention that when a system has no waiting customers, $Q = 1$.

Waiting imposes a cost on customers. Balking will not be considered, so the efficiency issues raised in Sunar et al. (2017) are not relevant. Therefore, only the cost function is required to analyse customer behaviour. Since they will initially be taken as risk-neutral, the cost function $C_{i,q}$ of customer i, q will be linear with unit cost of time c:

$$C_{i,q}(t_{i,q}(\boldsymbol{\Theta_Q})) = c \, t_{i,q}(\boldsymbol{\Theta_Q}), \tag{1}$$

where $t_{i,q}(\boldsymbol{\Theta_Q})$ is expected waiting time for customer i, q, a function of system state. From the linear form of the cost function, it is clear that the risk-neutral customers' objective in the game is to minimize expected waiting time t.

The game starts when all servers are working, but no customers are waiting to be served. Each arrival at the system observes the system state, described by matrix $\boldsymbol{\Theta_Q}$.

There are two possible actions available to customers, comprising the action set $A = \{S, M\}$:[3]

1. Action S: queue for both servers and form a **S**ingle queue;

[3]As previously mentioned, the possibility of balking (i.e., leaving without joining the queue) will not be considered. It is not the focus of the paper, and is not relevant to the determination of the number of queues. It is safe to assume that the reward is large relative to waiting time, taking the possibility out of consideration.

2. Action M: queue for whichever server has the shortest queue, or randomize with equal probability if at least two queues are of identical size and form **M**ultiple queues (cf. Winston (1977); if this is done when the customer faces a single queue, it will force the creation of multiple queues, as explained in more detail below. In this case, the customer again joins the end of the shortest resulting queue).

However, action S is not available when $Q > 1$, i.e. when the system is in a multiple queue state. This reflects the asymmetry between the two states, as it is much more difficult to persuade customers in two separate queues to combine than to split one single queue into two. So for S to be available to an arriving customer, all incumbents must have previously chosen S—i.e., the system must be in a single queue state, $Q = 1$. Obviously, a customer arriving at a system with no waiting customers may take either action as well, which is why it's defined that $Q = 1$ in that case.

Each new customer arrival triggers a new round of the game, which is played sequentially. Formally, the game stages, which are common knowledge, are:

1. A customer i arrives at the system, observes its state, and chooses from action set A. This choice can be discerned by any incumbent customers with perfect accuracy. The chosen action is not performed until stage 3, however. If there is at least one customer waiting, and that customer has taken action M so that the system is in a multiple state, customers must choose M and the round terminates.

2. This stage only occurs if an arriving customer encounters a system $\boldsymbol{\Theta_1}$ where $L \geq 1$, i.e., a single queue with at least one customer, and chooses action M in stage 2. In that case, incumbent customers split the single queue into separate queues, changing the system state. They will choose which server to queue for, in turns, with incumbents placed closer to the server in the single queue moving first: choosing the server with the shortest queue or randomizing between queues of equal length. They do this before customer i can act on the choice made at step 2.[4]

3. Customer i acts upon his choice in stage 1.

4. The customer remains in the queue until service completion, acting as an incumbent *vis-à-vis* future arrivals.

Customers' strategy space is then composed of a choice from set A for each possible system state $\boldsymbol{\Theta}$, so that $\boldsymbol{\Sigma}$, a vector whose elements are either of the possible actions in A for each possible state $\boldsymbol{\Theta}$, denotes the strategy for any customer. Customers' waiting time is uncertain, as the queues are stochastic processes and strategic interactions with newly arrived customers may alter the system state. Let then $t_{i,q}(\alpha, \boldsymbol{\Theta_Q})$ be the *ex-post* waiting time for customer i, q, as a function of α, the action prescribed by strategy $\boldsymbol{\Sigma}$ for state $\boldsymbol{\Theta_Q}$.

2.1 Waiting Times

Given a strategy Σ, customers' expected waiting times are a function of system state, and the customer's position in the queue. Upon arrival to the system, a customer observes system

[4]This response could be endogenized, but to avoid triviality it was embedded into the game.

state Θ_Q. From this, the customer learns their place i, q for each of their possible actions. Expected sojourn time is the sum of the exponentially distributed service time (with rate μ), and waiting time which follows a Gamma (Erlang) distribution for a given Q. Therefore when $Q = s$ expected sojourn time is given by:[5]

$$E[t_{i,q}(\boldsymbol{\Theta_s})] = \frac{1}{\mu} + \frac{i}{\mu}, \tag{2}$$

whereas for a system where one queue feeds s servers it is:[6]

$$E[t_{i,q}(\boldsymbol{\Theta_1})] = \frac{1}{\mu} + \frac{i}{s\mu}, \tag{3}$$

where the intuition behind eqs. (2)-(3) is that having one queue feed s servers multiplies the processing rate by s (as long as the customer is in the queue, not during service).

In determining customers' preferred decisions, it will be helpful to be able to compare expected waiting times directly across the possible system states, for the same number of customers in the system. This can be done by considering how customers in a single queue would be redistributed to s queues if the system state changed in the way prescribed in stage 2 of the game.

Let then $i_{\mathbf{1}}$ be the customer's position on the queue when $Q = 1$, and $i_{\boldsymbol{s}}$ their position on the shorter s queue(s) if the system state changes to $\boldsymbol{\Theta_s}$.[7] Then:

$$i_{\boldsymbol{s}} = \left\lceil \frac{i_{\mathbf{1}}}{s} \right\rceil, \tag{4}$$

so that, e.g., for $s = 2$, the first and second customers in the single queue take the first places in the two queues, and so on. Then a customer arriving at a system $\boldsymbol{\Theta_1}$ will have the following waiting times depending on system state (which they might influence through their action choice), and without taking future arrivals' actions into account:

$$E[t_{i_{\mathbf{1}},q}(S, \boldsymbol{\Theta_1})] = \frac{1}{\mu} + \frac{i_{\mathbf{1}}}{s\mu}, \tag{5}$$

$$E[t_{i_{\boldsymbol{s}},q}(M, \boldsymbol{\Theta_s})] = \frac{1}{\mu} + \frac{1}{\mu}\left\lceil \frac{i_{\mathbf{1}}}{j} \right\rceil, \tag{6}$$

where of course the system state changes to $\boldsymbol{\Theta_s}$ if action M is chosen.

Note again that while customers arriving at a system in a single queue state can change it to multiple queues by choosing action M and triggering stage 3 of the game, the reverse is not possible: there is no mechanism for changing the system state from multiple queues to one, other than the queue clearing. This implies that regardless of whether $Q = 1$ or $Q = s$, arrivals will always get the same expected waiting time from choosing M, as if they do so on a system in a single queue state, the system will change to a multiple queue state before they can overtake the incumbents.

[5] See Knudsen (1972) and Naor (1969) for derivation of these results, although their intuition is simple: customers must wait i service completions to begin service.

[6] At this juncture, strategic interactions are not being considered, and the number of queues is taken as given, so t is presented as independent of customer choices.

[7] When there are multiple queues, the i refers to the queue chosen by the customer, with the q term kept implicit to simplify the notation.

2.2 Customers' Actions and Equilibria

Customers' preferred strategy will be comprised of the actions yielding the shorter expected waiting time for any given system state. As the decision of a customer faced with multiple queues is trivial under pure strategies, analysis will focus on customers arriving at a single queue. For these purposes, it will be convenient to divide customers into two sets:

- Set $\mathbb{O} = \{i \mid i \neq ns,\, n \in \mathbb{Z}^+\}$
- Set $\mathbb{E} = \{i \mid i = ns,\, n \in \mathbb{Z}^+\}$.

Set $\mathbb{O}$ comprises those customers whose *arriving* place in the queue is *not* a multiple of the number of servers, while $\mathbb{E}$ comprises those for whom it is.

Proposition 1. If a customer is in set $\mathbb{O}$, it is a dominant strategy to choose action S.

Proof. For any place in a single queue system which is not a multiple of s, expected waiting time is strictly smaller than for the corresponding place in a multiple queue system were the system to change state:

$$\begin{aligned} &E[t_{i_{\boldsymbol{s}},q}(M,\Theta_s)] > E[t_{i_{\mathbf{1}},1}(S,\Theta_1)]\,\forall\, i_{\mathbf{1}} \in \mathbb{O}\,\&\, q \in \{1,\ldots,s-1\} \Rightarrow \\ &E[t_{i_{\boldsymbol{s}},q}(M,\Theta_s)] = \frac{1}{\mu} + \frac{1}{\mu}\left\lceil \frac{i_{\mathbf{1}}}{s} \right\rceil > E[t_{i_{\mathbf{1}},1}(S,\Theta_1)] = \frac{1}{\mu} + \frac{i_{\mathbf{1}}}{s\mu}. \end{aligned} \tag{7}$$

as the change of state takes place according to (4).

Therefore, these customers strictly prefer action S when arriving at a single queue system where their place would be $i \in \mathbb{O}$, as they prefer that place to the corresponding place in a multiple queue system. Given these customers have no incentive to deviate from the strategy of always choosing S when arriving at a system in a single queue state, it is a dominant strategy to do so. □

The foregoing times are conditioned on all future arrivals choosing to preserve the single queue state. *However*, since customers can always queue ahead of new arrivals who choose to split the queue, and obtain expected waiting time $E[t_{i_{\boldsymbol{s}},q}(M,\Theta_s)]$ anyway, this does not provide them with a reason to deviate from the foregoing strategy, regardless of future arrivals' choices.

Proposition 2. If a customer is in set $\mathbb{E}$, they are indifferent in choosing between actions S and M. Therefore, any choice defines an equilibrium.

Proof. For any place in a single queue system which is a multiple of s, the expected waiting time is identical with that of the corresponding place in a multiple queue system were the system to change state:

$$\begin{aligned} &E[t_{i_{\boldsymbol{s}},s}(M,\Theta_s)] = E[t_{i_{\boldsymbol{s}},1}(S,\Theta_1)]\,\forall\, i_{\mathbf{1}} \in \mathbb{E} \Rightarrow \\ &E[t_{s,s}(M,\Theta_s)] = \frac{1}{\mu} + \frac{1}{\mu}\left\lceil \frac{i_{\mathbf{1}}}{s} \right\rceil = E[t_{i_{\mathbf{1}},1}(S,\Theta_1)] = \frac{1}{\mu} + \frac{i_{\mathbf{1}}}{s\mu}. \end{aligned} \tag{8}$$

Since these customers are indifferent between the two possible states, they are indifferent between the two possible actions S and M. □

It is therefore the case that if customers in set $\mathbb{E}$ choose action S, the single queue state will emerge, whereas if they break ties the other way and choose action M, the multiple queue state will emerge. The corollary follows:

Corollary 1. Both the single queue state and the multiple queue state are equilibria in pure strategies of this game.

It is worth noting, however, that the first customer to arrive strictly prefers a single queue, and gets to implement it before any of the indifferent customers choose their action. Once this single queue state exists, there is no incentive for any arrivals to deviate from it. This might lead one to expect single queue states would be more prevalent. However, only one arrival needs to deviate from S to M to establish the other equilibrium. This fragility of the single queue equilibrium may be a reason for the emergence of multiple queues in real world scenarios.

3 Queue Number Determination with risk-averse Customers

The results in the previous section relied on risk-neutrality: customers only took expected waiting time into account. In this section, it will be shown that if customers are risk-averse, the single queue state will be strictly preferred by all customers, and thus be the unique equilibrium of the game. The intuition behind this result is that the risk associated with the multiple queue state is higher, as in the single queue state active servers can keep the queue moving even while some are faced with a low-probability high service time; in the multiple queue state, this safety valve is not present for any individual queue, so risk-averse customers naturally prefer the former.

The analysis will mirror that presented in Section 2, with an identical game being played. Let the customer cost function $C_{i,q}(t_{i,q}(\boldsymbol{\Theta_Q}))$ be strictly convex in time, instead of the linear utility given at (1), such that it reflects risk aversion:

$$\begin{aligned} C'_{i,q}(t_{i,q}(\boldsymbol{\Theta_Q})) > 0, \\ C''_{i,q}(t_{i,q}(\boldsymbol{\Theta_Q})) > 0, \end{aligned} \tag{9}$$

where as before, $t_{i,q}(\boldsymbol{\Theta_Q})$ is the customer's waiting time conditioned on system state $\boldsymbol{\Theta_Q}$.

3.1 Expected Cost

When customers are risk-averse, comparing expected waiting times is not enough to determine their preferred action, as an action might yield a lower expected waiting time, and still be passed over because the customer considers it too risky. Expected costs must be compared instead. Since expected service time is separable from expected waiting time, and the former is equal regardless of the number of queues, only the latter is going to be considered in the following discussion, as this simplifies the distribution functions without any loss of generality. Expected cost is given by:

$$E[C_{i,q}(t_{i,q}(\boldsymbol{\Theta_Q}))] = \int_0^\infty c(t)\, z(t(\boldsymbol{\Theta_Q}))) \, \mathrm{d}t, \tag{10}$$

where $z(t(\boldsymbol{\Theta_Q})))$ is the probability distribution function of waiting time, i.e.:

$$z(t(\boldsymbol{\Theta_s})) = \frac{\mu^i}{(i-1)!}\exp(-\mu t)\, t^{i-1}, \forall\, q \in \{1,\ldots,j\}\, i \in \{1,...,\infty\} \text{ when } Q = s \tag{11}$$

for a system in a multiple queue state, and:

$$z(t(\boldsymbol{\Theta_1}))) = \frac{(s\mu)^i}{(i-1)!}\exp(-s\mu t)\, t^{i-1},\ \forall\, i \in \{1,...,\infty\},\ \text{when } Q = 1 \tag{12}$$

for a system in a single queue state. To these correspond the cumulative probability functions $Z(t(\boldsymbol{\Theta_s}))$ and $Z(t(\boldsymbol{\Theta_1}))$, respectively.

3.2 Customers' Actions and Equilibria

When customers are risk-averse, all customers will strictly prefer a place in a single queue to the corresponding place in a multiple queue state.

Proposition 3. It is a dominant strategy for customers to choose action S, regardless of their position in the queue.

Proof. In order for a customer to prefer the single queue state to the multiple queue state, it must be the case that the expected cost of the former is smaller than that of the latter, for corresponding places in the queue:

$$\int_0^\infty c(t) z(t(\boldsymbol{\Theta_s}))\,\mathrm{d}t > \int_0^\infty c(t) z(t(\boldsymbol{\Theta_1}))\,\mathrm{d}t. \tag{13}$$

Define $S(t) = \int_0^t F(t)\,\mathrm{d}t$. After some manipulation, and integration by parts, (13) becomes:

$$c'(\infty)(E[t(\boldsymbol{\Theta_s})] - E[t(\boldsymbol{\Theta_1})]) + \int_0^\infty c''(t)[S(t(\boldsymbol{\Theta_s})) - S(t(\boldsymbol{\Theta_1}))]\,\mathrm{d}t > 0. \tag{14}$$

As $c'(t) > 0$ and $c''(t) > 0$, in order for (14) to hold it is sufficient that:

$$E[t(\boldsymbol{\Theta_s})] \geq E[t(\boldsymbol{\Theta_1})], \text{ and} \tag{15}$$

$$S(t(\boldsymbol{\Theta_s})) \geq S(t(\boldsymbol{\Theta_1})), \tag{16}$$

with at least one of the inequalities being strict.

The condition at (15) is equivalent to

$$E[t_{i_{\mathbf{2}},q}(M,\Theta_s)] \geq E[t_{i_{\mathbf{1}},1}(S,\Theta_1)])\,\forall\, i_{\mathbf{1}},\ i_{\boldsymbol{s}} \in \{1,...,\infty\}, \tag{17}$$

which was shown in Propositions 1 and 2.

On the other hand, (16) corresponds to:

$$\int_0^t [Z(t_{i_{\boldsymbol{s}},q}(M,\Theta_s))]\ \mathrm{d}t > \int_0^t [Z(t_{i_{\mathbf{1}},1}(S,\Theta_1)))]\ \mathrm{d}t\,\forall\, i_{\mathbf{1}},\ i_{\boldsymbol{s}} \in \{1,...,\infty\}, \tag{18}$$

which can be shown from the results in section 4.2 of Seth and Yalonetzky (2014) for stochastic ordering of Gamma distributions, *mutatis mutandis* for the present case dealing with cost rather than utility functions.

As the customer is both cost minimizing and risk-averse ($c''(t) > 0$), and the single queue state always offers lower risk and a weakly lower expected waiting time, it is always strictly preferred to the multiple queue state, for the corresponding queue states. Therefore, customers choose action S when arriving at a single queue system. □

It can also be added that since S is always a dominant strategy, there is no scope for the use of mixed strategies when customers are risk-averse.

The corollary follows:

Corollary 2. The single queue state is the equilibrium of the game.

4 Discussion and Conclusion

This paper has shown that risk-neutral customers derive a small benefit from combining queues whereas the remainder is indifferent between the two situations. This implies that both the single and multiple queue states are equilibria in pure strategies.

On the other hand, when customers are risk-averse, risk becomes another source of disutility, as the multiple queue state shows greater dispersion in waiting times, as it requires customers to bet on which queue is going to move faster. It's then quite intuitively appealing, and rigorously confirmed above, that risk-averse customers would prefer single queues more strongly than risk-neutral ones, as having a single queue for all servers eliminates the risk inherent in having to choose a queue. This is why only the single queue is an equilibrium for risk-averse customers.

It has been shown that risk-averse customers have the most to lose from a multiplicity of queues, and will, in equilibrium, form a single queue when presented with multiple servers. It seems a reasonable assumption that customers are at least somewhat risk-averse, yet combining queues is often frowned upon by managers. This paper provides a counterpoint to the views expressed in Rothkopf and Rech (1987). These results have implications for service station management, as there is great scope for improving social welfare by reducing the cost of multiple queues, which can be done in a Pareto improving manner (assuming the conditions in Sunar et al. (2017) do not hold).

While the results hold for any queue length, it is acknowledged that they are more relevant to short queues, especially when there is only one customer waiting. This is because the more customers there are present in a single queue, the greater the social pressure to conform to it. So while the proofs were kept as general as possible, it is worth keeping in mind that the model was intended to address the context of few customers waiting.

This does leave open the question of why it is often observed that customers form multiple queues even where there is no pressure from management to do so. As pure strategies are dominant and independent of future arrivals' strategies, there is no motivation to consider mixed strategies. However, it is a plausible conjecture that jockeying plays a role here. Indeed, for the case of risk-neutral customers, it was seen that both a multiple queue state and a single queue state were equilibria in pure strategies. While this is left for future research, under different equilibrium concepts, such as a trembling-hand equilibrium, the irreversibility of the multiple queue state might explain its emergence in real world applications, even though this would be against the wishes of other customers. It is harder to see why this equilibrium would occur when customers are risk-averse, and further research along these lines is required.

On a similar vein, in contexts where balking is permitted, the results in Sunar et al. (2017) indicate that under some conditions, social welfare is improved by having separate queues. Extending the present model to allow for balking would be a fruitful avenue for further research,

as it's not clear whether the results described in the foregoing would hold. It is worth noting, however, that the results of that paper only considered risk-neutral customers, and it is not clear whether they themselves would hold if customers are risk-averse.

The possibility of jockeying is just the sort of small disturbance which might favour the multiple queue equilibrium: if jockeying were to be permitted, then in the low probability event of a server clearing a queue, or at least reducing its length significantly compared to the other, customers could switch queues and reduce their expected waiting time. And even if this is a low probability event, it's enough to reduce *ex-ante* expected waiting time and make the previously indifferent customers prefer the multiple queue equilibrium instead.

With risk-averse customers, what would happen were jockeying to be allowed is not so clear: even though the expected value of waiting time in a multiple queue state might fall below that of a single queue state for some customers, the single queue state would still be less risky. One may conjecture that the degree of risk aversion possessed by customers would affect the resultant equilibrium, with more risk-averse customers preferring the single queue equilibrium more strongly.

Examining in more detail the circumstances in which the single queue equilibrium breaks down when jockeying is possible is an inviting topic for further research, although there are significant tractability problems to consider. Further research should then investigate customers' judgement of the probability of jockeying being possible, their degree of risk aversion in this specific context, and on a slightly behavioural tack, whether they judge their fellow customers to be rational when it comes to actions which might disturb the single queue equilibrium state. While it might be quite complex mathematically, it would be interesting to explore the impact of either server or customer heterogeneity in expected service time. It might also be interesting to investigate the impact on equilibrium robustness of repeated interactions as in Allon and Hanany (2012).

Other avenues for further research include the steady-state properties of a system with risk-averse customers, and providing a full formal treatment of social welfare issues with risk-averse customers, which still seems to be absent from the literature, as is research into management incentives when dealing with these customers.

Acknowledgements

I thank my PhD supervisors Kohei Kawamura and Tim Worrall, as well as Jonathan Thomas and Paul Schweinzer, my *viva* examiners for the helpful discussions and comments, and their reading of various drafts. I also thank various anonymous referees for their comments. Any remaining errors are, of course, my own. I also thank the University of Edinburgh School of Economics for funding my PhD, during the course of which the initial drafts of this paper were written.

Ethical Statement

The author states that there are no conflicts of interest, no funding sources to report, and as the present paper is a theoretical work, there are no ethical considerations to report.

References

Allon, G. and Hanany, E. (2012). Cutting in line: Social norms in queues. Management Science *58*, 493–506.

Armony, M. and Plambeck, E. L. (2005). The impact of duplicate orders on demand estimation and capacity investment. Management science *51*, 1505–1518.

Ata, B. and Olsen, T. L. (2009). Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. Operations Research *57*, 753–768.

Dehghanian, A., Kharoufeh, J. P. and Modarres, M. (2016). Strategic dynamic jockeying between two parallel queues. Probability in the Engineering and Informational Sciences *30*, 41–60.

Erlichman, J. and Hassin, R. (2015). Strategic overtaking in a monopolistic M/M/1 queue. IEEE Transactions on Automatic Control *60*, 2189–2194.

Ganesh, A., Lilienthal, S., Manjunath, D., Proutiere, A. and Simatos, F. (2012). Load balancing via random local search in closed and open systems. Queueing systems *71*, 321–345.

Hassin, R. (2016). Rational Queueing. CRC Press, Boca Raton, Florida.

Hassin, R. and Haviv, M. (2003). To Queue or not to Queue: Equilibrium Behavior in Queuing Systems. Kluwer Academic Publishers, Norwell, Massachusetts.

Haviv, M. and Ravner, L. (2016). Strategic bidding in an accumulating priority queue: equilibrium analysis. Annals of Operations Research *244*, 505–523.

Hlynka, M., Stanford, D. A., Poon, W. H. and Wang, T. (1994). Observing Queues Before Joining. Operations Research *42*, 365–371.

Kandori, M. (1992). Social norms and community enforcement. Review of Economic Studies *59*, 63–80.

Knudsen, N. C. (1972). Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure. Econometrica *40*, 515–528.

Mailath, G. L. and Samuelson, L. (2006). Repeated Games and Reputations: Long-Run Relationships. Oxford University Press, New York.

Naor, P. (1969). The Regulation of Queue Size by Levying Tolls. Econometrica *37*, 15–24.

Okuno-Fujiwara, M. and Postlewaite, A. (1995). Social norms and random matching games. Games and Economic Behavior *9*, 79–109.

Parsons, T. (1955). The Social System. Psychology Press, London.

Rothkopf, M. H. and Rech, R. (1987). Perspective on Queues: Combining Queues Is Not Always Beneficial. Operations Research *35*, 906–909.

Schwartz, B. (1975). Queuing and Waiting. University of Chicago Press, Chicago.

Seth, S. and Yalonetzky, G. (2014). Stochastic Dominance with Parametric Distributions.

Smith, D. R. and Whitt, W. (1981). Resource Sharing Efficiency in Traffic Systems. Bell System Technical Journal *60*, 39–55.

Sunar, N., Tu, Y. and Ziya, S. (2017). Pooled or Dedicated Queues When Customers Are Delay-Sensitive.

Winston, W. (1977). Optimality of the Shortest Line Discipline. Journal of Applied Probability *14*, 181–189.

Zhang, L., Wu, F. and Huberman, B. A. (2008). Games and Queues.