THE UNIVERSITY
of EDINBURGH

Essays on Strategic Queueing

Vasco F. Alves

Doctor of Philosophy
The University of Edinburgh
2016

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Vasco F. Alves*)

To Holly, the love of my life, for her patience during the writing of this thesis.
And to my parents, to whom I owe the opportunity to write it.

Abstract

This thesis includes three essays exploring some economic implications of queueing. A preliminary chapter introducing useful results from the literature which help contextualize the original research in the thesis is presented first. This introductory chapter starts by surveying queueing results from probability theory and operations research. Then it covers a few seminal papers on strategic queueing, mostly but not exclusively from the economics literature. These cover issues of individual and social welfare in the context of First Come First Served (FCFS) and Equitable Processor Sharing (EPS) queues, with one or multiple servers, as well as a discussion of strategic interactions surrounding queue cutting. Then an overview of some important papers on the impact of queueing on competitive behaviour, mostly Industrial Organization economists, is presented.

The first original chapter presents a model for the endogenous determination of the number of queues in an M/M/2 system. Customers arriving at a system where two customers are being served play a game, choosing between two parallel queues or one single queue. Subgame perfect equilibria are obtained, varying with customer characteristics and game specifications. With risk neutrality and when jockeying is not permitted, a single queue is an equilibrium, as is two queues. With risk neutrality and jockeying allowed, there is a unique two queue equilibrium. With risk aversion and no jockeying, there is a unique single queue equilibrium, and with risk aversion and jockeying, the equilibrium depends on the magnitude of risk aversion.

The second chapter analyses the individual decisions taken by consumers when deciding whether to join an M/M/1 queue where a subset of customers who interact repeatedly can both cut the queue and be overtaken once they join, by-passing occasional users. This is shown to be an equilibrium in repeated games for sufficiently patient customers. The expected sojourn time for customers under this discipline is described as a solution of a system of difference equations, and this is then used to obtain a threshold joining strategy for arrivals, which is independent of the number of regular customers in the queue, as regulars form a sub-queue under the LCFS discipline. Numerical methods are then employed to contrast sojourn times and thresholds with the equilibrium for a strict First Come First Served queueing discipline, and with the socially optimal joining rule.

Finally, the third chapter describes a duopoly market for healthcare where one of

the two providers is publicly owned and charges a price of zero, while the other sets a price so as to maximize its profit. Both providers are subject to congestion in the form of an M/M/1 queue, and they serve patient-customers with randomly distributed unit costs of time. Consumer demand (as market share) for both providers is obtained and described with its full complement of comparative statics. The private provider's pricing decision is explored, and equilibrium existence is proven. Social welfare functions are described and the welfare maximizing condition obtained. Numerical simulations with uniform and Kumaraswamy distributions are performed for several parameter values, showcasing the pricing provider's decision and its relationship with social welfare.

Contents

Part I

Queueing Theory: An Introduction and Literature Overview

Chapter 1

Introduction

Queueing has been a fruitful field of economic research since Naor's seminal paper, Naor (1969). The peculiarities inherent in the subject, combined with the fact that there is significant influence from Operations Research (OR), where most of the work on the subject has been done, means that there are many concepts required to fully grasp the research which are not part of the common economic lexicon.

Before Economics, queueing had been the bailiwick of Operations Research (OR) and Management Science. Their concerns often centred around improving efficiency in the many cases where queues arise in businesses and other organizations. It is important to keep in mind that while the most basic form of the queue is a physical line of people waiting for some kind of serviced, queueing models can capture a wider variety of realities. Waiting lists where there is no physical queue can still be modelled as an unobservable queue—these arise in health care, make-to-order transactions, call centres, and many other contexts. Queueing models have even been applied to congestion in internet or phone interchanges, and processes in a CPU, situations where there is no direct human element. It is not surprising that many modelling advances in this area came from telecommunication engineers and computer scientists. The Section 2 of this introduction presents the basic OR results essential to understanding the subsequent discussion.

There are reasons why economists should care about queueing, and contributions that Economics is uniquely positioned to make to the queueing literature, having a comparative advantage over other disciplines. The research program in OR was mostly focused on measurement and improvements in efficiency. Yet clearly be considered from an economic standpoint. The clearest route into this consideration is thinking of queueing as a rationing mechanism. It most often functions alongside price, though in some occasions in can be the only rationing mechanism, such as when the good is being given away for "free". The cost of the time spent in the queue acts in the same way as a monetary price. While the OR literature has traditionally taken supply and demand, queue discipline and other factors as either given or alterable by management, economic contributions have tended to focus on how these factors arise in

the first place. As such, starting with the landmark paper Naor (1969), a literature strand known as Strategic Queueing has developed, where economic methods, chiefly Game Theory, have been leveraged to examine interactions between several customers, customers and service providers, and different service providers, and how these interactions give rise endogenously to, *inter alia* queue discipline, provider behaviour, and demand and service rates. It is important to note, however, that Operations Research scholars have taken up the lessons from economic theory and contributed heavily to the study of strategic queueing since then, as a quick look at the bibliography of Hassin and Haviv (2003), a literature survey up to the date of publication, will show, while economists working in the field have also refined the sophistication of their models. The research presented in this thesis is performed in full view of the fruitful results of this inter-disciplinary dialogue.

Chapters 3 and 4 of this Introduction will review some key results in Strategic Queueing. As alluded to in the foregoing, the research agenda of strategic queueing, is to consider how system characteristics hitherto regarded as parameters can be endogenized when some of the actors in the system behave strategically and make rational decisions, normally modelled using game theory. Examples include capacity limits, decisions of which queue to join, customer priority and reordering, and the queueing discipline. Other considerations are the welfare implications of sojourn times, the role of providers—are they profit maximizers, do they have market power, can they vary service rates—and how queueing can impact broader economic concerns, such as competition between firms, search, firm production (when the firms must wait for intermediate goods), health care provision, etc. In line with the interests of this thesis, the topics to be covered focus on the issues surrounding the number of queues and queuing discipline. This is to prepare the ground for the original essays in the thesis, in chapters 7 and 8.

Chapter 7, entitled "Endogenous Queue Number Determination in M/M/2 Systems," uses game theory tools to investigate how in the presence of a multiplicity of service points, the number of queues can be determined endogenously by customers rather than being a parameter of the system. The multiplicity of service points naturally offers wide opportunities for research into strategic behaviour in other context: Hlynka et al. (1994) examined customer choice of queues in the presence of uncertainty about differing service rates, for instance. The research discussed below, on competition on sojourn time, is indeed at its core an extension of this type of model where multiple servers are in competition with one another. It is easy to envisage this being extended problems in labour economics: if service rates depend on server effort, for instance, how may firms promote this through contractual mechanisms, and how would servers and customers respond? How might this affect firm behaviour in deciding whether to invest in more service points or in improving the service rate? All these are microeconomic problems with direct relevance for the real world, where queueing is an almost ubiquitous feature of service provision.

Chapter 8 of this thesis, concerned with queueing discipline, is entitled "Cutting

Queues: Customer and System Behaviour in a Repeated Game." This paper is not shy about acknowledging its dependence on existing literature. It is at the crossroads of Yu et al. (2014) and Allon and Hanany (2012). The former treats an alternative discipline to the FCFS standard, replicating Naor's results for that setting, while the latter uses game theory to endogenously derive a queue reordering where customers with higher waiting costs overtake those with lower waiting costs in a repeated games setting. This somewhat echoes Gershkov and Schweinzer (2010), although the game theoretical apparatus used to solve the problem there is much different. The chapter extends Allon and Hanany (2012) to a setting where customer costs are identical, and then characterizes the equilibrium joining threshold using the tools employed in Yu et al. (2014). The literature about priority in queues is extensive, and applications are numerous. While this model focuses on endogenous overtaking, it is possible for service providers to effect the reordering of a queue as well—indeed, that is what most of the research on this area assumes. The use of payments to improve customer priority as a price discrimination device seem to be a fruitful avenue for further research, both theoretical and empirical.

Chapter 5 of this Introduction will move to the uses of queuing in Industrial Organization. This falls into the heading of strategic interactions among providers mentioned above. It discusses some key results in Industrial Organization which have seen queues applied to more general economic topics. In particular, these papers focus on how the presence of queueing in the delivery of a good. This discussion intends to prepare the ground for Chapter 9 in this thesis, entitled "Pricing and Waiting Time Decisions in a Health Care Market with Private and Public Provision," will apply this literature to the Health Economics context. This is a much more applied paper than the other two, following on the heels of Luski's IO research, and Goddard et al. (1995) and its progeny's applying of queueing models to Health Care. The issue under consideration is how a private sector health care provider operating in the same market as a public provider which is constrained to charge a price of zero sets its price, given that it is also subject to congestion and possesses market power. This is a pertinent extension, considering how prominent issues of waiting time are in health systems.

Chapter 2

Operations Research Results

This section attempts to give a short overview OR concepts and results required for understanding the rest of the material, based on a popular textbook, Gross et al. (2008). The starting point for queueing models is always the description of the system. Queueing systems can be classified according to six basic characteristics:

(1) Distribution of customer arrivals—usually, the process of customer arrival is stochastic, so a description of the system requires the probability distribution of inter-arrival times. Another element of this component is whether customers can arrive only one at a time, or if bulk arrivals are also possible, and if so, what is the probability distribution of the batch size. These probability distributions may be stationary, whereupon they are time-independent, or they may be non-stationary, i.e., changing with time. Upon arrival, customers may join the queue, or they may balk, i.e., decide not to join and leave—or they may be deterministically required to join. It may also possible for a customer to opt to join the queue, but then give up before being served, in which case he is said to have reneged. If there is more than one queue in the system, it may be possible for customers to change queues before being served—this behaviour is called jockeying.

(2) Service distribution—similarly, this aspect of the system is described by the probability distribution of customer service time, which may also be single or batch (if a server may serve more than one customer at once). The probability distributions may also be stationary or non-stationary, and may be state dependent or state independent, i.e., variant or invariant with the state of the system (usually the number of queueing customers). The probability distribution for service times is generally assumed to be independent of that for customer arrival.

(3) Queueing discipline—this parameter refers to the way customers are selected for service. By far the most common discipline is First-Come First-Served (FCFS), where, as the name indicates, customers are served in order of arrival. However, it is also possible to use other disciplines such as Last-Come First-Served, where newly arrived customers are the first to be served, Egalitarian Processor Sharing

(EPS), where servers may serve more than one customer at the same time (usually with a performance penalty), so service is performed for all customers who are present in the system at a given time (i.e., batch service, as discussed in the point above), random selection (RSS), or various priority schemes where customers are assigned priorities on arrival, and customers with higher priorities are served before those with lower ones, regardless of arrival order.

(4) System capacity—some systems have physical limitations that require customers to be turned away once the queue reaches a certain length, with no new customers being allowed to join the queue until a service is completed. This is equivalent to forcing customers to balk even if they would prefer to join the queue—though of course, in models with impatient customers, it is possible that due to the cost-benefit structure, the physical limit never needs to be reached before balking occurs endogenously.

(5) Number of servers—this generally refers to the number of servers servicing a single queue. While there can be multiple servers with a queue for each, this is usually considered to be comprised of several systems independent of each other.

(6) Number of service stages—it is possible that customers queue for a service composed of smaller sub-services, each of which may need to be queued for.

(7) Information available to customers—whether the queue is observable or not (an example of a non-observable queue is a call centre), whether the customer knows the service and arrival rates.

A specialized notation has evolved to describe queueing systems, which is now used throughout the literature. It takes the form $A/B/X/Y/Z$, where A corresponds to the customer arrival distribution, B to service time distribution, X to the number of parallel servers, Y to restrictions on system capacity, and Z to the queueing discipline. For example, the notation $M/G/2/\infty/FCFS$ represents a system with a Poisson arrival process (the M stands for 'Markovian'[1]), no specific service time distribution (the G stands for 'General'), 2 servers, no capacity restrictions, and using the FCFS discipline. It is common practice to omit the last two symbols—service-capacity and queueing discipline—if there are no restrictions on capacity, and if FCFS is used, respectively.

In the common case where inter-arrival and service times are exponentially distributed, the arrival rates is usually denoted by λ, and the service rate by μ, a convention to be followed here. Traffic congestion can then be measured by $\rho = \frac{\lambda}{s\mu}$, where s is the number of servers. If $\rho > 1 \Leftrightarrow \lambda > s\mu$, then the average number of arrivals exceeds the average service rate, and queue size will increase continually without converging to a steady state—assuming, that is, that there are no exogenous capacity limits and

[1]Referring to the Markovian property of the exponential distribution. This is used instead of E as that notation could be easily confused with E_k, the notation for the type-k Erlang distribution.

customers have infinite patience (or to put it in economic terms, waiting is costless). Under these conditions, it is necessary that $\rho < 1$ for a steady state to emerge.[2]

It is often desirable to obtain the probability distribution $N(t)$ for the total number of customers present in the system at a given time t, i.e., those customers waiting in the queue ($N_q(t)$), and those being served ($N_s(t)$). Let $p_i(t) = \Pr\{N(t) = i\}$ (so that the steady state is defined as $p_i = \Pr\{N = i\}$): then the mean number of customers in the system (L, or queue length) is given by:

$$L = E[N] = \sum_{i=0}^{\infty} i p_i, \tag{2.1}$$

with the mean number of only the waiting customers L_q (i.e., excepting those being served) being:

$$L_q = E[N_q] = \sum_{i=s+1}^{\infty} (i - s) p_i. \tag{2.2}$$

These steady state mean system sizes can be related to average customer sojourn times through Little's Formulas, or Little's Laws, from Little (1961). Sojourn time for a customer is given by $T = S + T_q$, where S is service time and T_q is waiting time, all of which are random variables. Let $W = E[T]$ and $W_q = E[T_q]$; then Little's Formulas are:

$$L = \lambda W, \text{ and} \tag{2.3}$$

$$L_q = \lambda W_q. \tag{2.4}$$

This result is important because, given μ, it is only necessary to know one of these values to obtain the other three, as $E[T] = E[S] + E[T_q]$, or $W = W_q + \frac{1}{\mu}$ (for exponential service times).

Another result which follows from Little's Laws is:

$$L - L_q = \lambda(W - W_q) = \lambda \frac{1}{\mu} = \frac{\lambda}{\mu}. \tag{2.5}$$

It is also the case that:

$$L - L_q = E[N] - E[N_q] = E[N - N_q] = E[N_s]. \tag{2.6}$$

Then combine (2.5) and (2.6), and let $r \equiv \lambda/\mu$, whence it follows that:

$$E[N_s] = r. \tag{2.7}$$

[2]The inequality is strict as, unless arrivals and services are deterministic, then randomness prevents servers ever catching up and leads to an ever-growing queue as well.

Further, by definition:

$$L - L_q = \sum_{i=0}^{\infty} ip_i - \sum_{i=1}^{\infty}(i-1)p_i = 1 - p_0. \tag{2.8}$$

Then,

$$p_b = \rho, \tag{2.9}$$

where p_b is the probability that a given server is busy in a multi-server system in the steady state, as the expected number of customers being served is r. Given servers are symmetric, the expected number of customers present at one server is given by:

$$\frac{r}{s} = \rho = 0(1 - p_b) + 1p_b. \tag{2.10}$$

2.1 *M/M/1* queues with an exogenous capacity limit

The first case to be covered, which in some ways is the benchmark for all that follows, is that of an *M/M/1/k* queue: as usual the *M* refers to the Markovian character of the distributions for customer arrival and service time, the *1* to the number of servers, and the k to the capacity constraint, with k being the maximum number of customers the system is able to hold. If the number of customers in the system is k, arriving customers are turned away until a service occurs to reduce the number of customers to $k - 1$.

Let $i = 0, 1, 2, .., k$ indicate the number of customers present in the system. This includes the customer being served as well as those waiting to be served. This can range from 0, where the server is idle, so k, which is the maximum system capacity. The probability of a customer being served is μ, the service rate (given by the rate parameter of the exponential distribution for service time). When a customer is served, the system loses one customer, and transitions from the state i to $i - 1$.

On the other hand, λ is the probability of a new customer arriving at the system (given by the rate parameter of the distribution of inter-arrival time). When a customer arrives, the system size is increased by one customer, transitioning from the state i to $i + 1$.

For any given state i, the system may transition to another state by either adding one customer (with probability λ), or losing one customer (with probability μ); therefore the total probability of changing states given a state i is equal to $\lambda + \mu$. The two states at the tail end of possible values present an exception, as they can only change in one direction: when the system is at state $i = 0$, it can only change state by adding one customer, with probability λ. Likewise, when at state $i = k$, the system can only change state by losing one customer, with probability μ, because new arrivals are turned away and do not affect the system state.

In the same way, the probability of *arriving* at any given state i given the system

is either at state $i+1$ or $i-1$ is also given by $\lambda+\mu$: if the original state is $i+1$, the system transitions to i with probability μ (one customer is served and leaves the system), and if the original state is $i-1$, the system transitions to i with probability λ (one new customer arrives and joins the system). Once again, the two tail end states present an exception. Since there is no state $i=-1$, it's only possible to arrive at $i=0$ from $i=1$, with probability μ. Likewise, as the system cannot be in the state $i=k+1$, it's only possibly to arrive at $i=k$ from $i=k-1$, with probability λ.

In the steady state, it has to be the case that the average rate at which the system enters into state i must equal the average rate at which the system exits out of that same state. If p_i is the probability of the system being in state i, then for all possible values of i:

$$\begin{aligned}
&i=0 \Rightarrow \lambda p_0 = \mu p_1 \\
&i=1 \Rightarrow \lambda p_1 + \mu p_1 = \lambda p_0 + \mu p_2 \Leftrightarrow (\lambda+\mu)p_1 = \lambda p_0 + \mu p_2 \\
&i=2 \Rightarrow \lambda p_2 + \mu p_2 = \lambda p_1 + \mu p_3 \Leftrightarrow (\lambda+\mu)p_2 = \lambda p_1 + \mu p_3 \\
&i=3 \Rightarrow \lambda p_3 + \mu p_3 = \lambda p_2 + \mu p_4 \Leftrightarrow (\lambda+\mu)p_3 = \lambda p_2 + \mu p_4 \\
&\dots \\
&i=k \Rightarrow \mu p_k = \lambda p_{k-1}.
\end{aligned} \tag{2.11}$$

The foregoing can be re-arranged as:

$$\begin{aligned}
&\lambda p_0 = \mu p_1 \Leftrightarrow p_1 = \rho p_0 \\
&(\lambda+\mu)p_1 = \lambda p_0 + \mu p_2 \Leftrightarrow p_2 = \rho^2 p_0 \\
&(\lambda+\mu)p_2 = \lambda p_1 + \mu p_3 \Leftrightarrow p_3 = \rho^3 p_0 \\
&\dots \\
&\mu p_k = \lambda p_{k-1} \Leftrightarrow p_k = \rho^k p_0.
\end{aligned} \tag{2.12}$$

By definition, the sum of the probabilities of the system being in all possible states i must equal 1, i.e.:

$$\sum_{i=0}^{k} p_i = p_0 + p_1 + p_2 + p_3 + ... + p_k = 1 \tag{2.13}$$

must hold.

Combining (2.12) and (2.13), it follows that:

$$p_0 + \rho p_0 + \rho^2 p_0 + \rho^3 p_0 + ... + \rho^k p_0 = 1, \tag{2.14}$$

in other words,

$$p_0 \left[1 + \rho + \rho^2 + \rho^3 + ... + \rho^k\right] = 1 \Leftrightarrow$$
$$p_0 = \frac{1}{1 + \rho + \rho^2 + \rho^3 + ... + \rho^k},$$

or more simply,

$$p_0 = \left(\sum_{i=0}^{k} \rho^i\right)^{-1} = \frac{1-\rho}{1-\rho^{k+1}}, \tag{2.15}$$

as $\sum_{i=0}^{k} \rho^i$ is a finite sum. Hence, p_i can be obtained from the knowledge of the model's parameters: λ, μ and k, as follows:

$$p_i = \rho^i p_0 = \rho^i \frac{(1-\rho)}{1-\rho^{k+1}}. \tag{2.16}$$

The foregoing implies that:

- $p_i > 0$ if and only if $p_0 > 0$;
- $p_0 > 0 \,\forall\, \lambda > 0$, $\mu > 0$, as $\sum i = 0^k \rho^i$ is a finite sum.

Then the expected value of i (that is, the expected steady state queue length L) can be obtained from (2.16):

$$L = E[i] = \sum_{i=0}^{k} i\rho^i \frac{(1-\rho)}{1-\rho^{k+1}} = \frac{\rho[1-(k+1)\rho^k + k\rho^{k+1}]}{(1-\rho)(1-\rho^{k+1})} = \frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}. \tag{2.17}$$

If customers arrive at a queue of length k, they will be diverted away from the system. The fraction of arrivals meeting this fate, ζ, is obtained by multiplying the arrival rate by the probability of the system being in state k:

$$\zeta = \lambda p_k = \frac{\lambda \rho^k (1-\rho)}{1-\rho^{k+1}}. \tag{2.18}$$

The expected number of customers *joining* the queue in unit time (also known as the effective arrival rate), is then simply the remainder of the arrival rate after the subtraction of ζ:

$$\lambda - \zeta = \lambda(1-p_k) = \lambda\left[1 - \frac{\rho^k(1-\rho)}{1-\rho^{k+1}}\right] = \lambda\frac{1-\rho^k}{1-\rho^{k+1}}. \tag{2.19}$$

The stations' "busy fraction", b, indicates the share of the time the server is not idle, i.e., when i is at least one:

$$b = \sum_{i=1}^{k} p_i = 1 - p_0 = \frac{\rho(1-\rho^k)}{1-\rho^{k+1}}. \tag{2.20}$$

On the other hand, the mean number of customers waiting L_q (that is, those not being served, so $i \geq 2$), can also be obtained:

$$L_q = \sum_{i=2}^{k}(i-1)p_i = \frac{\rho}{1-\rho} - \frac{\rho(k\rho^k + 1)}{1-\rho^{k+1}}. \tag{2.21}$$

The expected number of customers completing the service, per unit time, is:

$$\mu b = \mu(1-p_0) = \mu\left[1 - \frac{1-\rho}{1-\rho^{k+1}}\right]. \tag{2.22}$$

In the steady state, the number of customers joining the queue in unit time must be equal to those leaving, and it's easy to verify that:

$$\rho = \frac{1-p_0}{1-p_k}. \tag{2.23}$$

Finally, the expressions for L and L_q in (2.17) and (2.21) can be combined with Little's Laws in (2.3) and (2.4) to obtain the expected sojourn time and queueing time:

$$W = \frac{L}{\lambda} = \frac{(1-\rho^{k-1})\left(\frac{1}{\mu-\lambda} - \frac{\lambda^k(k+1)}{\mu^{k+1}-\lambda^{k+1}}\right)}{1-\rho^k}, \text{ and} \tag{2.24}$$

$$W_q = \frac{L_q}{\lambda} = \frac{(\mu\rho^k - \lambda)[\lambda^k(k-1) - k\lambda^{k-1}\mu + \mu^k)}{(1-\rho)(\mu-\lambda)(\lambda^{k+1} - \mu^{k+1}]}. \tag{2.25}$$

2.2 Boundless *M/M/1* queue (infinite capacity)

The results in the previous section can be extended to a boundless *M/M/1* queue fairly easily. The most straightforward way of doing this is to model the birth-death process in the same manner as (2.11), obtaining (2.15) in the same way. Then simply make the capacity limit k tend towards infinity, so that p_0 becomes:

$$p_0 = \left(\sum_{i=0}^{\infty}\rho^i\right)^{-1} = 1-\rho. \tag{2.26}$$

From (2.26), all results can be obtained in the same manner as for the capacity constrained queue. Starting with p_i:

$$p_i = \rho^i p_0 = \rho^i(1-\rho). \tag{2.27}$$

This implies:

- $p_i > 0$ if and only if $p_0 > 0$;
- $p_0 > 0\,\forall\lambda > 0$, $\mu > 0$ *and* $\mu > \lambda$, as it's necessary that $\rho < 1$ in order for $\sum_{i=0}^{\infty}\rho^i$ to converge.

The expected value of i/expected steady state queue length L can be obtained from (2.27):

$$L = E[i] = \sum_{i=0}^{\infty} i\rho^i(1-\rho) = \frac{\rho}{1-\rho}. \tag{2.28}$$

As the queue has no capacity constraints, no customers will be turned away, and so $\zeta = 0$.[3] The expected number of customers *joining* the queue in unit time, then, is equal to the arrival rate λ.

The stations' "busy fraction," b, is likewise given by:

$$b = \sum_{i=1}^{\infty} p_i = 1 - p_0 = 1 - (1-\rho) = \rho. \tag{2.29}$$

The mean number of customers waiting, L_q, also follows:

$$L_q = \sum_{i=2}^{\infty} (i-1)p_i = \frac{\rho^2}{1-\rho}. \tag{2.30}$$

The expected number of customers completing the service, per unit time, is:

$$\mu b = \mu(1-p_0) = \mu\rho. \tag{2.31}$$

The steady state condition that the number of customers joining the queue in unit time must be equal to those leaving is easily verified:

$$\mu b = \lambda \Leftrightarrow \mu\rho = \lambda \Leftrightarrow \mu\left(\frac{\lambda}{\mu}\right) = \lambda. \tag{2.32}$$

Finally, the expressions for L and L_q in (2.28) and (2.30) can be combined with Little's Laws in (2.3) and (2.4) to obtain the expected sojourn time and queueing time:

$$W = \frac{L}{\lambda} = \frac{\frac{1}{\mu}}{1-\rho} = \frac{1}{\mu-\lambda}, \tag{2.33}$$

and

$$W_q = \frac{L_q}{\lambda} = \frac{\frac{\lambda}{\mu}}{\mu-\lambda} = \frac{\rho}{\mu-\lambda}. \tag{2.34}$$

2.3 *M/M/s* queues with an exogenous capacity limit

It is possible to extend the *M/M/1/k* framework to an arbitrary number of servers s. Waiting times are i.d.d. across servers, following an exponential distribution with rate

[3] Of course, this assumes customers will join the queue regardless of its length and their expected sojourn time, but the present section only take into account the probabilistic results, not customer behaviour.

μ. Let $\rho = \lambda/s\mu$, and define the following discrete functions:

$$d_i = \begin{cases} \frac{(\rho s)^i}{i!} & \text{if } 0 \leq i \leq s-1, \\ \frac{(\rho s)^s}{s!}\rho^{i-s} & \text{if } i \geq s, \end{cases} \tag{2.35}$$

and

$$D_k = \sum_{i=0}^{k} d_i,\ n \geq 0. \tag{2.36}$$

The steady state probability of having i customers in the system is given by:

$$p_i = \frac{d_i}{D_k},\ 0 \leq i \leq k, \tag{2.37}$$

while expected queue length L in the steady state is:

$$L = E[i] = \sum_{i=0}^{k} i p_i. \tag{2.38}$$

Expected mean steady state sojourn time can be obtained from L and Little's Law.

The fraction of "lost" customers ζ is given by:

$$\zeta = \frac{d_k}{D_k} \tag{2.39}$$

Chapter 3

Strategic Queueing: Joining a Queue

3.1 Joining a Queue: Individual and Social Issues

This chapter will provide an overview of the economic issues surrounding the decision to join a queue, and some of the ramifications of the number of queues in the presence of multiple servers. It was this issue which introduced economic considerations to the analysis of queueing, in Naor (1969). The problem under analysis by Naor was when should a customer arriving at a system decide to join a queue, and when to leave. This can be seen as an endogenizing of the capacity parameter for queues mentioned in the previous chapter. If customers behave like economic agents with a value for time, there is only a certain amount of time they are willing to wait to receive a service. As this time is a function of queue length, it is possible to determine a 'joining threshold,' the maximum queue size for which a customer will join that queue. It is customer 'impatience,' represented as a cost of time, which allows for these strategic considerations and the subsequent decision modelling.

Consider an FCFS *M/M/1* system described above in section 2.1, consumers to be identical risk neutral expected utility maximizers, who arrive at the system and observe queue length i, after which they decide whether to join the queue or leave. As mentioned above, customer strategy can be defined by a threshold value: a strategy n_t, equal to the smallest integer i for which the customer balks. Customer utility of joining as a function of queue length is:

$$U_i = R - (i+1)c\frac{1}{\mu}, \tag{3.1}$$

where R is the net value of the good being provided, and c is the unit cost of time. These values are constant across customers. Further, let $v_t = \frac{R\mu}{c}$. In meaningful models, at least some customers must decide to use the queue, so it is assumed that $v_t \geq 1$. The outside option has a value of zero. It is worth noting that the queueing literature has

tended to eschew formulations of time cost in discounting terms. This can probably be traced back to the formulation in the seminal Naor (1969), but whatever its origins, it is the standard practice in the literature.

The strategy n_t which maximizes the customer's utility must satisfy the following two conditions:

$$R - n_t c\frac{1}{\mu} \geq 0, \text{ and} \tag{3.2}$$

$$R - (n_t + 1)c\frac{1}{\mu} < 0. \tag{3.3}$$

Eq. (3.2) refers to the case where i falls short of n_t by one, where the customer joins. On the other hand, eq. (3.3) refers to the case where queue size at least equals the threshold value, so that the customer leaves. The two inequalities can be combined into:

$$n_t \leq \frac{R\mu}{C} = v_t < n_t + 1, , \tag{3.4}$$

which can be stated as

$$n_t = \lfloor v_t \rfloor, \tag{3.5}$$

where $\lfloor \rfloor$ is the floor function, i.e., n_t is the largest integer not exceeding v_t. Note that n_t depends on μ, R and c, but is independent of λ. This value will act as the endogenous capacity limit.

3.1.1 Social Optimization

While the foregoing discussion summarized individual considerations, it is often the case the individually optimal outcomes do not lead to the maximization of social welfare. That is true in the situation in view: as a general rule, the individually optimal threshold will be larger than socially optimal. Intuitively, this is because each customer joining the queue imposes a negative externality on all future arrivals, increasing their expected sojourn time.

Formally, social welfare can be defined as the sum of net gains accruing to customers in unit time; let W represent this, under a given threshold n:

$$W = (\lambda - \zeta)R - cE[i] = \lambda R(1 - p_n) - cL = \lambda R\frac{1-\rho^n}{1-\rho^{n+1}} - c\left[\frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}\right]. \tag{3.6}$$

It can be shown that W is discreetly unimodal on n, so that a local maximum is a global maximum. The welfare maximizing strategy n_o is defined by two inequalities:

$$\lambda R\left[\frac{\rho^{n_o}(1-\rho)}{1-\rho^{n_o+1}} - \frac{\rho^{n_o+1}(1-\rho)}{1-\rho^{n_o+2}}\right] - c\left[\frac{(n_o+1)\rho^{n_o+1}}{1-\rho^{n_o+1}} - \frac{(n_o+2)\rho^{n_o+2}}{1-\rho^{n_o+2}}\right] < 0, \text{ and} \tag{3.7}$$

$$\lambda R\left[\frac{\rho^{n_o-1}(1-\rho)}{1-\rho^{n_o}} - \frac{\rho^{n_o}(1-\rho)}{1-\rho^{n_o+1}}\right] - c\left[\frac{n_o\rho^{n_o}}{1-\rho^{n_o}} - \frac{(n_o+1)\rho^{n_o+1}}{1-\rho^{n_o+1}}\right] \geq 0. \tag{3.8}$$

After some manipulations these can be combined into:

$$\frac{n_o(1-\rho)-\rho(1-\rho^{n_o})}{(1-\rho)^2} \leq \frac{R\mu}{c} < \frac{(n_o+1)(1-\rho)-\rho(1-\rho^{n_o+1})}{(1-\rho)^2}. \tag{3.9}$$

In order to obtain n_o, consider the following function of two independent variables ρ (> 0) and v_o (≤ 1):

$$v_t = [v_o(1-\rho)-\rho(1-\rho^{v_o})](1-\rho)^{-2}. \tag{3.10}$$

Where ρ is an arbitrary positive constant, v_t is a boundlessly increasing function of v_o. Therefore the integers between which v_t lies (viewed as a function of v_s and ρ), will obey the conditions at eq. (3.9), so that

$$n_o = \lfloor v_o \rfloor. \tag{3.11}$$

Further:

$$v_o \leq v_t, \tag{3.12}$$

where the equality only holds if $v_t = 1$.

The inequality at eq. (3.12), which will normally be strict, states the conclusion anticipated above, that the system will generally be over-congested. It would be desirable to reduce the individually optimal threshold to the socially optimal. This can be done either by imposing an administrative rule such that n_o is the maximum system capacity, and further arrivals are turned away by management, or by imposing a toll on joining customers such that their expected net gain is reduced in a way that n_o is the individually optimal threshold value, which is the solution proposed by Naor.

This result was extremely significant, but is reliant on several assumptions: customer risk neutrality and homogeneity, linearity of the cost function, arrivals and services being exponentially distributed, the First Come First Served discipline, and there being only one server. Naturally, the subsequent research programme consisted of relaxing these assumptions one by one. The rest of this chapter covers some of these extensions. The scope of this survey precludes the inclusion of all relevant literature; the results surveyed are those with direct relevance for the original research presented in Part II. Issues which have not been surveyed but warrant a mention are, *inter alia*, those involving customer and/or server heterogeneity, and other distribution functions for service and inter-arrival times.

3.2 Joining *M/M/s* Systems

As it turned out, research by Knudsen (1972) showed that Naor's results held even for an arbitrary number of (identical) servers, and any unspecified time cost function.

Where multiple servers servicing a single queue are present, the distribution of time

spent in the queue is different than that of service time. That is because the queue will lose one customer whenever any server finishes a service, so the rate is increased by the number of servers. Expected sojourn time when the system is in state i (i.e., the number of customers in the system is i) is then given by:

$$T_i = X + Y_i, \tag{3.13}$$

where X, measuring the service time, and Y_i, measuring the waiting time, are mutually independent. X follows the distribution:

$$f(t) = \mu e^{-\mu t},\ t > 0, \tag{3.14}$$

while $Y_i = 0$ when $i \leq s - 1$, and is distributed as follows when $i \geq s$:

$$g_i(t) = \frac{(s\mu)^{i-s+1}}{(i-s)!} t^{i-s} e^{-s\mu t},\ t > 0. \tag{3.15}$$

The density function of T_i is then:

$$f(t)_i = \begin{cases} f(t) & \text{if } 0 \leq i \leq s-1, \\ f(t) \times g_i(t) = \int_0^t f(t-u) g_i(u)\,\mathrm{d}u & \text{if } s \leq i. \end{cases} \tag{3.16}$$

Further, let customer utility be the function:

$$U_i = R - \gamma_i,\ i \geq 0, \tag{3.17}$$

where R is the net value of the good sought, and γ_i is expected waiting cost given the system is in state i. Let $h(t)$ be the waiting cost to a customer spending t units of time in the system. Then γ_i is given by:

$$\gamma_i = E[h(T_i)] = \int_0^\infty h(t) f_i(t)\,\mathrm{d}t, \tag{3.18}$$

where $f_i(t)$ is the density function of T_i.

Then it can be shown that:

$$\gamma_0 = \gamma_1 = \cdots = \gamma_{s-1} < \gamma_s < \gamma_{s+1} < \ldots, \tag{3.19}$$

whence it follows that:

$$U_0 = U_1 = \cdots = U_{s-1} > U_s > U_{s+1} > \ldots. \tag{3.20}$$

Intuitively, this means expected utility is the same for all customers who arrive at a system with at least one idle server, and then it is strictly decreasing in system size i when all servers are busy.

Customers decide on whether to join the queue or balk at arrival, after observing

system size i. They set a threshold strategy n_t, the smallest queue size at which they will balk: i.e., the smallest integer for which $U_i < 0$:

$$U_{n_t-1} \geq 0 > U_{n_t}, \tag{3.21}$$

where to avoid triviality it is assumed that such an n_t exists and and that $U_{s-1} \geq 0$, so that $n_t > s$.

3.2.1 Social Optimization

Naor's results for the relationship of the social and individual thresholds also hold under the relaxed conditions. Social welfare W is given by:

$$W = \lambda \sum_{i=0}^{n-1} p_i(n) U_i. \tag{3.22}$$

It can be shown that a threshold value n_o which maximizes W exists, as W is discretely unimodal in i. This is not necessarily equal to n_s, but rather $n_o \leq n_s$, where the strict inequality will be the rule rather than the exception. The same policy options presented in Naor (1969) to attain the welfare maximizing threshold can be used in this case.

3.3 Welfare Implications of Combining Queues

The previous discussion treated the number of servers as a given parameter without taking into consideration the implications of this. However, multiple servers performing the same job are present, the issue of how many queues to have emerges. It is possible to have one queue for each server, one queue for all, or a number of intermediate combinations, depending on the number of servers. Chapter 7 of this thesis addresses the strategic interactions through which customers can endogenously determine this aspect of the queueing system. However, as with most parameters which can emerge endogenously, this can also be imposed administratively by management. Either way, the outcome will affect social welfare. It is intuitively appealing to think that one single queue for multiple servers is socially optimal and reduces aggregate waiting time. This result seems to have been more or less assumed to be true without a rigorous proof, but this eventually emerged, under certain conditions, in Smith and Whitt (1981).

The proof runs along the following lines. Let $T(s, \lambda, \mu)$, be the mean steady state waiting time function for a customer in an $M/M/s$, FCFS system, such that:

$$T(s, \lambda, \mu) = \frac{C(s, \rho)}{s\mu - \lambda}, \text{ where} \tag{3.23}$$

$$C(s, \rho) = \frac{\rho^s/(s-1)!(s-\rho)}{\sum_{k=0}^{s-1}(\rho^k/k!) + \rho^s/(s-1)!(s-\rho)}, \tag{3.24}$$

which implies $T(s, \lambda, \mu)$ is a subadditive function of s and λ for a given μ.

Meanwhile, for multiple systems, average waiting time is:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} D(s_1, \lambda_1, \mu) + \frac{\lambda_2}{\lambda_1 + \lambda_2} D(s_2, \lambda_2, \mu). \tag{3.25}$$

This decomposition can be performed as many times as required to aggregate the total number of servers in the system. Then:

$$D(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \leq \frac{\lambda_1}{\lambda_1 + \lambda_2} D(s_1, \lambda_1, \mu) + \frac{\lambda_2}{\lambda_1 + \lambda_2} D(s_2, \lambda_2, \mu), \tag{3.26}$$

whence the more general result follows:

$$H\left(\sum_{i=1}^{n} s_i, \sum_{i=1}^{n} \rho_i\right) \leq \sum_{i=1}^{n} H(s_i, \rho_i), \, \forall \, n \in \mathbb{N}, \tag{3.27}$$

where

$$H(s, \rho) = \rho \frac{C(s, \rho)}{s - \rho}. \tag{3.28}$$

This follows from Little's Law:

$$L = \lambda T = \lambda T(s, \mu, \lambda) + \frac{1}{\mu}. \tag{3.29}$$

If L_s is the steady state expected queue length in the sth system, then eq. (3.26) is equivalent to:

$$L \leq L_1 + L_2. \tag{3.30}$$

These results can be readily translated into economic concepts. The intuition behind them is that where each server has its own queue, then it's possible for some servers to be idle while customers are queueing for another server, which leads to a waste of resources. If customers are like those described in Naor (1969) and Knudsen (1972), and the service station makes zero economic profits, then it is possible to obtain total welfare by aggregating it across customers. Given aggregate social welfare is decreasing on L and T, then as combining queues reduces aggregate expected waiting time and queue length, it achieves a welfare improvement. Note, however, that relaxing some of the underlying assumptions may lead to combining queues actually not providing a benefit (see Rothkopf and Rech (1987)). In particular, this assumes both servers and customers are identical. In particular, it assumes that length of service is independent of customer characteristics. This limits the application of this result to, say, supermarkets, where the expected service length is at least dependent on the number of items of shopping. In this case, having separate queues according to customer characteristics might be beneficial. As alluded to in the foregoing, Chapter 7 in this thesis discusses how the number of queues for multiple servers might be determined endogenously, and the conditions required for the socially optimal outcome of having one single queue to emerge.

3.4 Joining an EPS Queue

While FCFS is by far the predominant discipline, it is not the only one. Another queueing discipline which has received some study is Equitable Processor Sharing (EPS). Under this discipline, customers receive an identical share of the server's effort. so that if there is only one customer in the queue, they will receive all the effort, if there are two, each receives half, etc. This behaviour approximates that of, say, an internet node, or processes sharing time on a CPU. Furthermore, this discipline is formally equivalent to that where the next customer to be served is randomly chosen as a draw from a uniform distribution. It is then interesting to extend Naor's results to this discipline, which nevertheless, remains endogenously determined. Making the order of service at least partially endogenous shall be addressed in the next chapter.

Yu et al. (2014) considered this problem for an EPS queue, where as in Naor (1969), customer arrivals are a Poisson process with rate λ, and there is one server with service times distributed exponentially with rate μ: if there are n customers in the system, each will receive service at rate μ/n. Customers receive service value R at completion, and experience time cost c, so that their expected utility function is the familiar:

$$U = R - cT. \tag{3.31}$$

The usual conditions set to avoid triviality must be satisfied: a customer will desire to queue at least for an idle server, $\rho < 1$, and all stochastic processes are independent of each other.

The first step in determining the threshold joining value is to find an expression for conditional expected sojourn time T_n, where n is the number of customers observed plus 1, i.e., that which the customer would experience if they joined the queue. Unlike in FCFS, this is not a trivial problem, since the sojourn time is a function not only of the number of customers in the system at joining time, but also of the behaviour of future arrivals.

The authors find that conditional expected sojourn times can be represented by the following system of linear difference equations:

$$(n+1)E[T_{n+2}] - \left(1 + \frac{1}{\rho}\right)(n+1)E[T_{n+1}] + \frac{1}{\rho}nE[T_n] = -\frac{1}{\lambda}, \tag{3.32}$$

$$E[T_2] - \left(1 + \frac{1}{\rho}\right)E[T_1] = -\frac{1}{\lambda}. \tag{3.33}$$

It is not clear, however, whether this is an adequate representation of the system's dynamics when there is a capacity constraint, whether exogenous, or as will be assumed later in the paper, endogenous. This is because the difference equations represent a system which can take any size from 0 to infinity, which is not true if customers are following a threshold strategy. Nevertheless, the model will be described as is.

Employing a generating function and a left-multiplication transformation, the sys-

tem is solved and found to have the following solution:

$$E[T_n] = \frac{n+1}{\mu(2-\rho)}, \tag{3.34}$$

so that for any given n, expected utility is:

$$U = R - c\left(\frac{n+1}{\mu(2-\rho)}\right). \tag{3.35}$$

The optimal threshold is then found through a similar process to that in Naor (1969): find an integer n^* satisfying $R - c\left(\frac{n^*+1}{\mu(2-\rho)}\right) \geq 0$, where the condition is not satisfied for $n^* + 1$. Then the customer will join queues of length up to $n^* - 1$, so that the threshold value n_t is:

$$n_t = n^* - 1 = \left\lfloor \frac{R}{C}\mu(2-\rho) - 1 \right\rfloor - 1. \tag{3.36}$$

Note how, unlike for the FCFS case in Naor (1969), this threshold is a function of λ.

The socially optimal joining threshold outlined in Naor (1969) is valid for all *M/M/1* queues, regardless of discipline. Resorting to numerical methods, it can be shown that for the EPS case, the individual threshold is also not equal to the socially optimal threshold in the general case. However, unlike for FCFS, the socially optimal rule may yield larger queues than the individually optimal one, meaning systems can be under-congested. This would make policies to maximize social welfare harder to formulate.

Chapter 4

Strategic Queueing: Queue Reordering

This chapter considers queue reordering. The first section covers results from the mechanism design literature where a mechanism is sought through which customers with different costs of time may reorder the queue to take account of these priorities through inter-customer payments, subject to several restrictions. The second section described a repeated games result which allows for reordering without any explicit payments, only the expectation of benefiting from the mechanism in the future rounds.

The models presented are only a selection from a wider strand of this literature. Notably, they only cover interaction between customers without taking into consideration the server nor management. Another strand considers management receiving payments from customers to alter priorities. This has even been extended into political economy as "optimal bribing" models.

All the models considered assume consumer heterogeneity in respect of time costs, and seek to use reordering to improve social welfare. Chapter 8 takes a slightly different tack, assuming homogeneous consumers who still seek to change the queue order to their benefit. Unlike the models surveyed below, this will tend to have a negative impact on social welfare, highlighting that the desirability of such mechanisms depends on consumer characteristics.

4.1 Paying to Reorder

This section covers a strand of literature at the crossroads of strategic queueing and mechanism design. It considers possibilities for customers to rearrange their order in a queue, through inter-customer trade. Gershkov and Schweinzer (2010) presents such a model, where an individual rationality and a balanced budget requirement are imposed. It is found that in *any* fully deterministic discipline such as FCFS, there is no possibility of performing this rearrangement, as the individual rationality condition is not met. However, the rearrangement is found to be possible at least for a fully random

discipline.

A finite set of $n > 1$ customers has the utility function for a given customer i:

$$U_i = V - k\theta_i - p, \tag{4.1}$$

where V is the value of the good, k is sojourn time, θ_i is the unit cost of time (which varies across customers), and p is a payment made by customer $i.\theta_i \in \Theta_i = [0, 1]$ is private information, and independently distributed with density f and distribution F. Service time is assumed to be equal for all customers and normalized to 1, without loss of generality. Denote by $\Theta = [0,\, 1]^n$ the type space, and by θ any element of it.

A mechanism M species any payments customers should make and the possible stochastic order of service. This payment and order may depend on the initial allocation specified by $\sigma_i \in I_i \equiv [0,\, 1]^n$, with $\sum_j \sigma_{ij} = \sum_i \sigma_{ij} = 1$, where σ_{ij} denotes the probability agent i is served at the jth period; σ denotes the full vector of $\langle \sigma_i \rangle_{i=1}^n$, and $I = [0,\, 1]^{n\times n}$ the space of all initial allocations. A direct revelation mechanism is a vector of payments $p^M = \langle p_i^M \rangle_{i=1}^n$ and the order $\sigma^M = \langle \sigma_{ij}^M \rangle_{i,\,j=1}^n$, where $p_i^M \colon \Theta \times I \to \mathbb{R}$, and for $1 \le i,\, j \le n$, $\sigma_{ij}^M \colon \Theta \times I \to [0,\, 1]$, so that $\sum_i \sigma_{ij}^M(\theta,\, \sigma) = 1$ for each j and $\sum_j \sigma_{ij}^M(\theta,\, \sigma) = 1$ for each i. When all players report their type truthfully, the expected utility of player i with type θ_i is:

$$U_i(\theta_i,\, \sigma) = V - E\left[\sum_{k=1}^{n} \sigma_{ik}^M(\theta,\, \sigma) k\theta_i + p_i^M(\theta,\, \sigma) \middle| \theta_{-i}\right], \tag{4.2}$$

where $\theta = (\theta_i,\, \theta_{-i})$. Further,

$$P_i^M(\theta_i, \sigma) = E\left[p_i^M(\theta,\, \sigma) \middle| \theta_{-i}\right],$$
$$W_i^M(\theta_i, \sigma) = \sum_{k=1}^{n} kE\left[\sigma_{ik}^M(\theta,\, \sigma) \middle| \theta_{-i}\right],$$

that it, expected sojourn time and expected payment by player i, respectively.

Individual rationality is defined with relation to some initial allocation Z as the requirement for the target mechanism M to give at least the same expected utility, i.e., for any customer i and $\theta_i \in \Theta_i$:

$$V - W_i^M(\theta_i,\, \sigma)\theta_i - P_i^M(\theta_i,\, \sigma) \ge V - W_i^Z(\theta_i,\, \sigma)\theta_i. \tag{4.3}$$

M is incentive compatible if, for any i and any $\theta_i,\, \hat{\theta}_i \in \Theta_i$:

$$-W_i^M(\theta_i,\, \sigma)\theta_i - P_i^M(\theta_i,\, \sigma) \ge -W_i^M(\hat{\theta}_i,\, \sigma)\theta_i - P_i^M(\hat{\theta}_i,\, \sigma). \tag{4.4}$$

Define further three service schedules/disciplines:

1. *Random order*: each customer has equal probability of being at any position:

$$\sigma_{ik}^{RSS} = \frac{1}{n}, \forall\, i,\, k,\, \theta. \tag{4.5}$$

2. *FCFS*: Players are served according to their order of arrival or some other deterministic schedule, independent of waiting cost:

$$\sigma_{ik}^{FCFS} = \begin{cases} 1 \text{ if } l = k \\ 0 \text{ otherwise.} \end{cases} \tag{4.6}$$

3. *Efficient Order*: Players are served according to decreasing waiting cost, so that in $M = ef$:

$$\sigma_{ik}^{ef}(\theta,\, \sigma) = \begin{cases} 1 \text{ if } |\{j\colon \theta_j > \theta_i\}| = k-1 \text{ and } |\{j \neq i\colon \theta_j = \theta_i\}| = 0 \\ \dfrac{1}{m} \text{ if } |\{j\colon \theta_j > \theta_i\}| = l \text{ and} \\ \qquad |\{j \neq i\colon \theta_j = \theta_i\}| = m \neq 0,\, l+m \geq k > l \\ 0 \text{ otherwise.} \end{cases} \tag{4.7}$$

where $|S|$ is the number of elements of set S.

The goal is to attain to the efficient order, starting from some other order, of which FCFS and the random order are two extreme cases.

A set of Lemmas governing customer behaviour can then be set out. These can be intuitively summarized as follows:

- Players prefer to adopt any mechanism if it provides them with an increase in expected utility over the original mechanism.

- There is a "worst-off" player $\theta^*(Z)$ relative to status-quo Z—the player who gains least from moving to the efficient mechanism. As long as this player benefits from the change, all other players will as well, and it's possible to implement the efficient mechanism.

- In the efficient mechanism, type θ_i's expected sojourn time is:

$$W_i^{ef}(\theta_i,\, \sigma) = n + (1-n)F(\theta_i). \tag{4.8}$$

Assume without loss of generality that player i is served in position i under FCFS. His sojourn time is then just i, and the worst-off type $\theta^*(FCFS)$ is given by

$$\begin{aligned} &n + (1-n)F(\theta_i^*(FCFS)) = i, \text{ or} \\ &F(\theta_i^*(FCFS)) = \frac{i-n}{1-n}, \text{ and } \theta_i^*(FCFS) = F^{-1}\left(\frac{n-i}{n-1}\right). \end{aligned}$$

Sojourn time in the random queue is:

$$\frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2},$$

so that the worst-off type $\theta^*(RSS)$ is given by:

$$n + (1-n)F(\theta^*(RSS)) = \frac{n+1}{2} \Leftrightarrow F^{-1}\left(\frac{1}{2}\right).$$

- Under an incentive compatible reordering, the worst-off type cannot pay a positive transfer while reporting his type truthfully.
- The budget-balancing condition, i.e., $\sum_i p_i^M(\theta, \sigma) = 0$, satisfying both incentive compatibility and individual rationality, is as follows, for $\theta^*(Z)$ as described above:

$$\sum_{i=1}^{n}\left[\int_0^{\theta_i^*(Z)} sF(s)\mathrm{d}W_i^M(s, \sigma) - \int_{\theta_i^*(Z)}^{1} s(1-F(s))\mathrm{d}W_i^M(s, \sigma)\right] \geq 0. \tag{4.9}$$

With the foregoing results in place, it can be stated that the efficient queueing order is implementable if and only if there exists a mechanism $\langle p^M, \sigma^{ef}\rangle$ which is incentive compatible, individually rational with regard to Z and budget balanced, i.e.:

Theorem 1. Efficient scheduling is implementable with regard to schedule $Z \in \{RSS, FCFS\}$ iff:

$$\sum_{i=1}^{n}\left[\int_0^{\theta_i^*(Z)} sF(s)f(s)\mathrm{d}s - \int_{\theta_i^*(Z)}^{1} s(1-F(s))f(s)\mathrm{d}s\right] \leq 0, \tag{4.10}$$

where

$$\theta_k^*(Z) = \begin{cases} F^{-1}\left(\frac{n-k}{n-1}\right) \text{ if } Z = FCFS \\ F^{-1}\left(\frac{1}{2}\right) \text{ if } Z = RSS. \end{cases}$$

and k is the position of player i in the FCFS schedule.

So the worst possible type $\theta_k^*(FCFS)$ depends on the initial position of the customer in the FCFS queue, implying the individual rationality constraint must be checked for every one of the n slots, leading to n separate conditions for a deterministic discipline. The following two key propositions are then derived:

Proposition 1. For any distribution of types F, the efficient scheduling is implementable with regard to the random discipline.

Proposition 2. For any distribution of types F, the efficient scheduling is not implementable with regard to the FCFS discipline.

The key difference is that in the FCFS discipline, customers have certainty about their order of service, while in the random discipline, they only have a probabilistic

ticket. This makes it impossible to efficiently reschedule the FCFS queue as it will never be rational for the first customer to sell his assumed first place to a marginally higher type behind him for a merely marginal payment (note that the balanced budget constraint prevents the other customers from offering more than the value of the service minus service time).

However, when randomness is inserted into the queueing discipline, as in the random queue considered above, efficient rescheduling becomes possible. Consider a lottery which results with probability p in the random queue and with probability $(1-p)$ the FCFS queue, and let this lottery be executed if not all layers agree to participate in the efficient reordering. Then it can be shown that:

Corollary 1. Since the worst-off type in the lottery is continuous in p, for p sufficiently high, there exists an equilibrium in which the efficient allocation is implemented.

Note that this result is not extensible to non-linear cost functions, as it is impossible to successfully generalize over linear costs, balancing the budget and require an efficient allocation. The result is, however, robust to extending common service valuations to private valuations, and to relaxing the balanced budget condition to allowing for a budget surplus.

Finally, an efficient indirect mechanism implementing the efficient schedule is presented in the form of an auction game. This is equivalent to the foregoing game.

4.2 Reordering without Payment

The foregoing sections surveyed mechanisms where customers made monetary payments to each other in order to effect queue reordering. However, when customers interact repeatedly, it is possible for reordering to be effected without monetary payment, but solely by expecting future benefits.

Allon and Hanany (2012) developed a repeated game where customers allow others with higher time costs to overtake, in the expectation that if they have a high time cost in the future, they may be allowed to overtake others. This model allows for ongoing arrivals, in contrast to most of the foregoing, but is restricted to two types of customers, one with a high cost and one with a low, although this tractability assumption does not look to restrictive.

The model attempts to capture situations like a queue for airport security controls, where some customers have time to spare and others might have a flight departing very soon. The latter category may try convince the other customers to let them cut ahead. However, other customers have no way of knowing whether the customer seeking to cut ahead is telling the truth about his higher need or not, even if they were minded to let customers with genuine needs cut ahead, though they may catch liars by, say, later seeing them shopping in the duty free area. If customers interact repeatedly, these monitoring opportunities allow the development of devise grim trigger strategies

to encourage truth telling.

Let there be M customers, each with a stream of service requests modelled by a Poisson process with rate λ/M; they queue to obtain a good with value R. Customers have different time costs, which change randomly across each iteration of the game. For tractability, only two different types are considered, denoted by $t \in \{H, L\}$. Request types follow a non-degenerate Bernoulli distribution, with probability α for being of type H and $1-\alpha$ of type L. Denote by c_H, c_L, and μ_H, μ_L the costs and service rates of each type, respectively. Without loss of generality, H has a higher expected service cost than L: $c_H\mu_H > c_L\mu_L$.

Upon arrival at the queue, customers can choose to join the end of the queue (action J), or ask to cut the line (P). The excuse they give to cut the line can only be verified *ex post*. When faced with such a request, customers in the queue can accede to the request (A) or reject it (R). The only consequence of rejection is having to join the end of the line.

4.2.1 Single Stage Game

In the single stage game, the only equilibrium is the one where all cutting requests are rejected. When cutting requests happen, they begin with the customer at the end of the queue, and continue down the queue with the intention of reaching the top. Queue cutting attempts end after the first rejection. For a cutting attempt to be beneficial, at least one incumbent must accept it. The full strategy of customer $i \in \{1, ..., M\}$ is given by $(E^i, I^i) \equiv (E^i_H E^i_L, I^i_H I^i_L)$ where $E^i_t \in \{J, P\}$ and $I^i_t = \{R, A\}$ are the actions chosen when a customer is of type $t \in \{H, L\}$.

Theorem 2. In the single shot game, all customers choose $I^i = R\,R$.

This result is quite intuitive. In the absence of repeated interactions, any acceptance of a cutting request will cause sojourn times to increase without any future compensation. Accepting requests is therefore not a feasible strategy for incumbents. Knowing this, arrivals will be indifferent between cutting or joining. For the avoidance of doubt, the convention is that they join.

4.2.2 Repeated Game With Perfect Public Monitoring

Consider instead a setting where the M customers require this service repeatedly. Assume that M is so large that each customer is not likely to have two concurrent requests, and further assume that periods are defined such that there is a clear separation between the time to complete the service and the interval between service requests. This guarantees that a customer's choice in a single period affects the current payoff in a way that is separable from future period payoffs (e.g., a service is required every day, but the expected sojourn time is 20 minutes). Future period payoffs and waiting costs are discounted by a per period discounting factor $\delta \in (0, 1)$.

In the repeated game setting, for sufficiently patient customers, there is an equilibrium where all type L customers join the end of the queue without attempting to cut, while type H customers attempt to cut and are allowed to do so up to the first H-type incumbent. This occurs despite the fact that their type cannot be verified until the end of the service.

This equilibrium employs a grim trigger strategy where customers will punish deviations by switching to the $FCFS$ inducing strategy which is the equilibrium of the single shot game: reject all requests. Denote by T_t^P and T_t^J the expected sojourn time experienced by a type t customer when cutting or joining the end of the line, respectively, and when all other customers follow the cooperative strategy (PJ, RA) (i.e. type H customers jump the queue until meeting a type H incumbent and reject cutting attempts when they are incumbents; type L customers s join the end of the queue and accept cutting requests). Also let $V^{c\mu}$ be the long term expected discounted utility when all customers follow the cooperative strategy in all periods, and V^{FCFS} the long term expected discounted payoff when all customers follow the FCFS inducing strategy in each period, i.e., $I^i = RR$.

Theorem 3. (i) The strategy in which each customer cooperates by choosing (PJ, RA) if this strategy has been chosen by all customers in all previous periods, and punishes by choosing $I^i = RR$ otherwise, is an equilibrium if and only if:

$$\frac{\delta}{1-\delta} \geq \frac{c_L(T_L^J - T_L^P)}{\alpha c_H(D^{FCFS} + 1/\mu_H - T_H^P) + (1-\alpha)c_L(D^{FCFS} + 1/\mu_L - T_L^J)}, \quad (4.11)$$

where

$$D^{FCFS} = \frac{\alpha\lambda/\mu_H^2 + (1-\alpha)\lambda/\mu_L^2}{1 - \alpha\lambda/\mu_H - (1-\alpha)\lambda/\mu_L}$$

is the delay under the FCFS discipline.

(ii) If $\mu_H = \mu_L$, (i) simplifies to

$$\frac{\delta}{1-\delta} \geq \frac{c_L}{c_H - c_L}\frac{1}{\alpha(1-\alpha)}. \quad (4.12)$$

The grim trigger strategy is credible, as its punishment is the equilibrium of the single shot game, and no customers have any incentive to deviate from it once it is in place. In the equilibrium path, high-type customers have no incentive to deviate, as they can only lose by joining the queue at the end, or by accepting a cutting request. Low-type customers have an incentive to improve their position in the queue by lying about their true type. However, doing that foregoes future utility earned if they draw a high cost in the future. Full deviations are always preferable to partial ones, and a full deviation will be profitable only if its long-term discounted payoff is larger than that of maintaining the cooperative strategy. This only holds if customers are very impatient, or in other words, do not meet the condition set in eq. (4.11).

When expected service rates are identical ($\mu_H = \mu_L \equiv \mu$), both types have identical

expected sojourn time when taking the same action under $c\mu$. Let the expected sojourn time of joining the queue at the end be W^J and that of cutting W^P. Both priority rules provide the same ex ante expected sojourn time, therefore $\alpha W^P + (1-\alpha)W^J = D^{FCFS} + 1/\mu$, and eq. (4.11) simplifies to (4.12).

This is equivalent to there being two separate FCFS queues, one for high priority customers and one for low priority customers, where the low priority queue only has access to the server when the high priority queue is empty.

4.2.3 Queue Length Dependent Strategies

In the foregoing, only strategies that are independent of queue length were considered. however, it can also be shown that the cooperative strategy can be sustained when queue-length dependent strategies are taken into account.

Regimes that depend on randomization between absolute priority schemes are not sustainable in equilibrium because under such regimes it would be very hard to distinguish between a customer that misrepresented their type and a customer that randomized. The following considers queue-length-dependent priority schemes where customers decide to push or join the line upon arrival depending on observed queue length. Denote by l_t the number of type t customers observed upon a customer's arrival to the system, and let $l = (l_H, l_L)$. Note that despite the assumption that types are not ascertainable until customers leave the system, customer behaviour in equilibrium reveals their type. Thus an arriving H type customer can know l by observing their place in the queue following a cutting attempt.

A priority scheme is generated by a queue-length-dependent threshold strategy if a customer's strategy gives priority to one class over the other only when queue length is below some prescribed threshold $\bar{l}$, i.e., if for each customer i, $E^{i,\bar{l}} = PJ$ and $I^{i,\bar{l}} = RA$ when $l \equiv (l_H, l_L) \leq (\bar{l}_H, \bar{l}_L)$, and $I^{i,\bar{l}} = RR$ otherwise. The arising priority scheme is denoted by $c\mu\bar{l}$. Denote by $T_L^{P,c\mu\bar{l}}$ and $T_L^{J,c\mu\bar{l}}$ the expected sojourn time experienced by a type L customer when pushing and joining the line at the end, respectively, and when all other customers follow the $c\mu\bar{l}$ strategy. Further let $V^{c\mu\bar{l}}$ denote the long term expected discounted utility when all customers follow the $c\mu\bar{l}$ inducing strategy.

Theorem 4. Any priority scheme generated by queue-length dependent threshold strategies is sustainable in equilibrium for sufficiently patient customers (sufficiently large δ).

Proof. The above result is obtained very simply. The critical intuition behind it is that for any given threshold $\bar{l}$, it is always possible to find a δ such that the sustainability condition:

$$c_L[T_L^{J,c\mu\bar{l}}(q) - T_L^{P,c\mu\bar{l}}(l)] \leq \delta[V^{c\mu\bar{l}} - V^{FCFS}], \tag{4.13}$$

is satisfied for all $l \leq \bar{l}$. □

The foregoing discussion paves the way for Chapter 8, which examines a similar situation in a repeated game, and characterizes that system's steady state properties.

Chapter 5

Queueing in Industrial Organization

5.1 Introduction

Queueing models have also considered interaction between different suppliers. This is part of a strand of literature which applies queueing to other economic topics, using the unique insights provided by these models to illuminate some other facet of economic behaviour. This part will survey some key results stemming from the application of queueing to Industrial Organization. In particular, it will be analysed how the presence of queueing in the delivery of a good affects strategic behaviour by the good's provider when it possesses market power. This prepares the ground for Chapter 9, which applies these models to the Health Care market.

This part will first present the monopoly problem, and then present two different duopoly models, one where consumers are heterogeneous in service rate, and one where they are identical. The striking feature of the latter model is that it yields a separating equilibrium, where the each firm specializes in serving consumers with a high cost of waiting, and the other those with a low cost, despite the fact that the firms are identical. These models assume consumers take into account *ex-ante* expected waiting times. A possible extension would allow consumers to observe queue size before making a decision.

Further, there is a vast scope for extending this research to other applied fields. Transport economics could be the next step, as some consumers are obviously willing to pay a premium to get somewhere faster and avoid queues in the air travel market. Other possible applications suggesting themselves are the unobserved queues of bureaucracies, such as waiting times for permit applications, and telecommunications, almost coming full circle to the field of telephone engineering which originated so much of the earlier queueing literature!

5.2 Monopoly with Queueing

Chen and Wan (2003) and its predecessor, Chen and Wan (2003)Chen and Frank (2004), presents a model of monopoly price setting in the presence of queueing, which is then extended to a duopoly. This section will describe the former. Consider a firm with an $M/M/1$ queue for delivery of the good, a service rate μ, and potential Poisson arrival rate Λ.

Consumers have the option of buying the monopolist's good (which they value uniformly at R) for price P, incurring a cost of c per unit of sojourn time, or forego the good and receive the outside option which they value uniformly at ν.

Consumers have full knowledge of R, P, μ, c, ν, and Λ, but not queue length, so that their decision is based on *ex ante* expected queue length. Let λ be the firm's demand, i.e., Λ minus the share of consumers seeking the outside option. Then expected sojourn time $T(\lambda)$ is given by:

$$T(\lambda) = \frac{1}{(\mu - \lambda)^+}, \tag{5.1}$$

where $x^+ = \max\{x, 0\} \, \forall \, x \in \mathbb{R}$.

Consumer utility then takes the form:

$$U = R - P - cT(\lambda), \tag{5.2}$$

when consumers join the queue, and

$$U = \nu, \tag{5.3}$$

when they take the outside option.

It will be assumed throughout that a consumer will buy the good when the queue is empty, so that $R - \frac{c}{\mu} > \nu$. In equilibrium, $R - P - cT(\lambda) = \nu$, and necessarily $\lambda < \mu$. It follows that

$$\lambda = \mu - \frac{c}{R - P - \nu}, \tag{5.4}$$

and as $\lambda \leq \Lambda$,

$$\lambda = \min\left\{\Lambda, \mu - \frac{c}{R - P - \nu}\right\}. \tag{5.5}$$

The firm lacks any production or service costs, and seeks to maximize its instantaneous profit $\pi = P\lambda$. Therefore, its problem is

$$\max_P \pi = \max_P P\lambda \text{ s.t. } 0 \leq P \leq R - \nu - \frac{c}{\mu} \tag{5.6}$$

The optimal price is found to be $\max\{P_m, P_\Lambda\}$, where P_m is the *first-order price*:

$$P_m = R - \nu - \sqrt{\frac{c(R - \nu)}{\mu}}, \tag{5.7}$$

and P_Λ is the *market capture price*:

$$P_\Lambda = R - \nu - \frac{c}{\mu - \Lambda}. \tag{5.8}$$

The first-order price P_m is the optimal price if:

$$\Lambda \geq \mu - \sqrt{\frac{c\mu}{R-\nu}}, \tag{5.9}$$

and the market capture price P_Λ is the optimal price otherwise. Therefore, the firm's demand is

$$\lambda = \min\left\{\Lambda, \mu - \sqrt{\frac{c\mu}{R-\nu}}\right\}. \tag{5.10}$$

5.3 Duopoly with Identical Consumers

This duopoly model, presented in Chen and Wan (2003), extends the monopoly model with homogeneous consumers described in section 5.2 above. There are two firms, sharing the stream of potential consumers Λ; each is an $M/M/1$ system. Both homogeneous and heterogeneous firms can be considered. Each firm $i = \{1, 2\}$ has its own price P_i, service rate μ_i, good value R_i, and *waiting cost* c_i. The outside option ν is, however, of identical value for both firms.

Firms select their own price P_i knowing the other firm's reaction function, as well as consumers' reactions. All consumers are charged the same price. Consumers have full knowledge of P_i, μ_i, R_i , and c_i, as well as of ν and Λ, though they do not know the exact length of either firm's queue. Consumers have three options: join firm 1's queue, join firm 2's queue, or take the outside option ν. Consumer utility when choosing these options is, respectively:

$$U_1 = R_1 - P_1 - c_1 T(\lambda_1), \tag{5.11}$$

$$U_2 = R_2 - P_2 - c_2 T(\lambda_2), \tag{5.12}$$

$$U_o = \nu, \tag{5.13}$$

where U_o is the utility of taking the outside option, and λ_i is each firm's demand. Consumers choose the option yielding the largest expected utility.

The sum of monetary price and the expected cost of sojourn time is a firm's *full price*. In equilibrium, full prices will be identical across firms, preventing any firm switching at the aggregate level. However, at the "microscopic" level, individual consumers choose firms by a rate-based proportional randomized strategy, where each consumer selects a firm with a probability proportional to each firm's equilibrium arrival rate, so that the arrival processes across the two firms are independent Poisson processes.

It is assumed firms can attract at least one consumer if they charge a 0^+ price and

queue length is zero, i.e:

$$R_i - \frac{c_i}{\mu_i} > \nu. \tag{5.14}$$

Also, for a given μ_i, it must be the case that $\lambda_i < \mu_i$. Expected sojourn time is then:

$$T_i(\lambda_i) = \frac{1}{\mu_i - \lambda_i}. \tag{5.15}$$

Firms' service rate μ_i is exogenous, and they have no production or service costs. They seek to maximize instantaneous profit $\pi_i = P_i\lambda_i$, where price P_i is the decision variable and the competitor's price is taken as given. Let $\pi_i(P_1, P_2)$ be firm i's expected profit rate if it chooses a price P_i given firm j's price P_j, $i \neq j$, $i, j = \{1, 2\}$. A price pair (P_1^*, P_2^*) is a (pure Nash) equilibrium if it satisfies the following conditions:

$$\pi_1(P_1^*, P_2^*) \geq \pi_1(P_1, P_2^*), \forall P_1 \geq 0,$$
$$\pi_2(P_1^*, P_2^*) \geq \pi_2(P_1^*, P_2), \forall P_2 \geq 0.$$

Equilibrium prices are obtained from each firm's reaction function. Let $P_i = f_i(P_j)$ be firm i's optimal price for a given value of P_j. An equilibrium is then a pair of prices (P_1, P_2) such that $P_1 = f_1(P_2)$ and $P_2 = f_2(P_1)$, i.e., the intersection of the two reaction functions. f_2 is found by taking P_1 as fixed and solving firm 2 problem:

$$\max_{P_2>0} \pi_2 = P_2\lambda_2, \tag{5.16}$$

subject to

$$R_2 - P_2 - \frac{c_2}{\mu_2 - \lambda_2} \geq \nu, \tag{5.17}$$
$$R_1 - P_1 - \frac{c_1}{\mu_1 - \lambda_1} = R_2 - P_2 - \frac{c_2}{\mu_2 - \lambda_2}, \tag{5.18}$$
$$\lambda_1 + \lambda_2 \leq \Lambda, \tag{5.19}$$
$$0 \leq \lambda_2 < \mu_2. \tag{5.20}$$

Constraints (5.17) and (5.18) come from the consumer problem: (5.17) states that utility gained from queueing for firm 2 must be higher than that gained from taking the outside option—λ_2 would decrease until the constraint held, or it became zero; $U_2 - U_o$ is the consumer surplus, which can be either positive or zero in equilibrium. Constraint (5.18) recognizes that in equilibrium, both firms give consumers the same expected utility.

5.3.1 Homogeneous Firms

First consider the special case where the two firms are identical, that is, $R_1 = R_2 \equiv R$, $\mu_1 = \mu_2 \equiv \mu$, and $c_1 = c_2 \equiv c$. The Nash equilibrium will take three forms, depending

on the magnitude of Λ relative to two threshold levels, $\underline{\Lambda}$ and $\bar{\Lambda}$, which are defined as

$$\underline{\Lambda} = 2\mu + \frac{c}{R-\nu} - \sqrt{\frac{8c\mu}{R-\nu} + \frac{c^2}{(R-\nu)^2}}, \tag{5.21}$$

$$\bar{\Lambda} = 2\left(\mu - \sqrt{\frac{c\mu}{R-\nu}}\right). \tag{5.22}$$

Given the condition $R - \frac{c}{\mu} > \nu$, it's easy to check that $\underline{\Lambda} < \bar{\Lambda}$.

The equilibrium's three forms correspond to three possible situations:

1. *ample demand*: $\Lambda \geq \bar{\Lambda}$;

2. *moderate demand*: $\underline{\Lambda} < \Lambda < \bar{\Lambda}$;

3. *scarce demand*: $\Lambda \leq \underline{\Lambda}$.

Ample Demand: Non-Competitive

In this case, demand is greater than capacity, and some consumers will always take the outside option. Therefore, each of the firms will charge their monopoly price.

Theorem 5. Suppose that $\Lambda \geq \bar{\Lambda}$. There exists a unique equilibrium such that each firm charges its own monopoly price:

$$P_1 = P_2 = R - \nu - \sqrt{\frac{c(R-\nu)}{\mu}}, \tag{5.23}$$

and corresponding demands are:

$$\lambda_1 = \lambda_2 = \mu - \sqrt{\frac{c\mu}{R-\nu}}. \tag{5.24}$$

Price P_i in (5.23) is the first-order optimal price for a monopolist firm, as shown in (5.7). Therefore, both firms operate independently of each other, as monopolists. Here, consumer surplus is zero: the expected net benefit of joining either firm's queue is equal to the outside opportunity. A positive fraction of potential consumers is not served by either firm, except at the border $\Lambda = \bar{\Lambda}$.

Corollary 2. When $\Lambda \geq \bar{\Lambda}$, the comparative statics of the equilibrium are as follows:

$$\begin{aligned}
\frac{\partial P_i}{\partial c} &= -\frac{1}{2}\sqrt{\frac{R-\nu}{c\mu}} < 0, \\
\frac{\partial P_i}{\partial R} &= 1 - \frac{1}{2}\sqrt{\frac{c}{(R-\nu)\mu}} > 0, \\
\frac{\partial P_i}{\partial \Lambda} &= 0, \\
\frac{\partial P_i}{\partial \nu} &= \frac{1}{2}\sqrt{\frac{c}{(R-\nu)\mu}} - 1 < 0, \\
\frac{\partial P_i}{\partial \mu} &= \frac{1}{2\mu}\sqrt{\frac{c(R-\nu)}{\mu}} > 0.
\end{aligned}$$

This means firm i would raise P_i with the consumer valuation of their good R, and its service speed μ, and cut it in response to an increase in c and ν. A small change in the potential arrival rate Λ (except at the boundary threshold $\Lambda = \bar{\Lambda}$ would result in no price change.

Moderate Demand: Moderate Competition

Intuition would indicate that, holding μ, c and R constant, a decrease in Λ would lead firms to lose their monopoly position, and have to engage in competition: this is formalized in the following theorem, when $\underline{\Lambda} < \Lambda < \bar{\Lambda}$.

Theorem 6. Suppose that $\underline{\Lambda} < \Lambda < \bar{\Lambda}$ (the moderate demand case). Then any equilibrium (P_1, P_2) must satisfy:

$$2\mu - \frac{c}{R-\nu-P_1} - \frac{c}{R-\nu-P_2} = \Lambda, \tag{5.25}$$

and corresponding actual demands are:

$$\lambda_i = \mu - \frac{c}{R-\nu-P_i}. \tag{5.26}$$

In particular,

$$P_1 = P_2 = R - \nu - \frac{c}{\mu - \frac{\Lambda}{2}}, \tag{5.27}$$

is an equilibrium with the corresponding demand $\lambda_1 = \lambda_2 = \frac{\Lambda}{2}$.

Theorem 6 asserts the existence of a symmetric equilibrium, where the firms charge a higher price than in the ample demand case, where they enjoy monopoly power! As Λ increases, firms cut prices to compensate for the increased cost of waiting, and as in the ample demand case, they raise prices as a response to increases in μ and decreases in c. Theorem 6 also gives a necessary condition for the equilibrium, strongly suggesting the existence of a continuity of equilibria; these turn out to exist, at least as set out in the proposition below. Note that $\Lambda \leq \mu$ if and only if $R - \nu \leq c/\mu$, and $\underline{\Lambda} < \bar{\Lambda}$ is

equivalent to $R - \nu > c/\mu$.

Proposition 3. There exists a continuum of equilibria if $\Lambda \leq \mu$ and $\underline{\Lambda} < \Lambda < \bar{\Lambda}$.

In an asymmetric equilibrium, the firm charging the higher price takes a smaller market share. The consumer surplus is zero in either case, and no consumers take the outside option.

Scarce Demand: Highly Competitive

Where demand is scarce relative to firm capacity, the market exhibits a high degree of competition.

Theorem 7. Suppose that $\Lambda \leq \underline{\Lambda}$. Then there exists a unique equilibrium given by

$$P_1 = P_2 = \frac{4c\Lambda}{(2\mu - \Lambda)^2}, \tag{5.28}$$

and the corresponding demands rates are $\lambda_1 = \lambda_2 = \frac{\Lambda}{2}$.

No consumer takes the outside option, and consumer surplus is positive:

$$R - P_i - \frac{c}{\mu - \lambda_i} > \nu. \tag{5.29}$$

The comparative statics follow:

Corollary 3. Suppose that $\Lambda \leq \underline{\Lambda}$. The comparative statics of the equilibrium are as follows:

$$\begin{aligned}
\frac{\partial P}{\partial c} &= \frac{4\Lambda}{(2\mu - \Lambda)^2} > 0, \\
\frac{\partial P}{\partial R} &= 0, \\
\frac{\partial P}{\partial \nu} &= 0, \\
\frac{\partial P}{\partial \mu} &= -\frac{16c\Lambda}{(2\mu - \Lambda)^3} < 0, \\
\frac{\partial P}{\partial \Lambda} &= \frac{4c(2\mu + \Lambda)}{(2\mu - \Lambda)^3} > 0.
\end{aligned}$$

Marginal changes in R and ν do not change the equilibrium. Price increases with Λ, as that reduces the degree of competition; it also leads to an increase in expected sojourn time. Consumers are made strictly worse off by this increase. This differs from the symmetric equilibrium in the moderately competitive case, where an increase in potential arrivals is compensated by a reduction in prices.

Other comparative statics are less intuitive. An increase in μ or a decrease in c both lead firms to cut prices and receive a lower expected revenue; both these changes reduce the expected waiting cost for consumers, so one would expect the firms to be able to

charge a higher price and earn greater profit. But the market is highly competitive due to excess capacity; these improvements make it even more so, leading to lower prices and profits.

5.3.2 Heterogeneous Firms

The model can be extended to heterogeneous firms. In this case, five different types of equilibria may arise:

1. Dominated market: unique equilibrium where one firm takes the whole market.
2. Highly competitive market: unique equilibrium where both firms enter the market and consumer surplus is positive.
3. No equilibrium market: there is no (pure) Nash equilibrium.
4. Moderately competitive market: there is a continuum of equilibria.
5. Non-competitive market: unique equilibrium with both firms charging their respective monopoly prices.

Numerical experiments were used to determine that the cases tend to appear in the order of 1-5 as Λ increases. Cases 2, 4 and 5 correspond to the scarce, moderate and ample demand cases of the homogeneous firms setting.

Equilibrium Existence

Whether heterogeneous or not, firms will not exhibit competitive behaviour where demand is large. The following theorem generalizes theorem 5 to the heterogeneous firms case.

Theorem 8. Suppose that:

$$\Lambda \geq \bar{\Lambda}_H \colon = \mu_1 + \mu_2 - \sqrt{\frac{c_1\mu_1}{R_1 - \nu}} - \sqrt{\frac{c_2\mu_2}{R_2 - \nu}}. \tag{5.30}$$

There exists a unique equilibrium such that each firm charges its own monopoly price:

$$P_i = R_i - \nu - \sqrt{\frac{c_i(R_i - \nu)}{\mu_i}}, \tag{5.31}$$

and the corresponding demands are:

$$\lambda_i = \mu_i - \sqrt{\frac{c_i\mu_i}{R_i - \nu}}. \tag{5.32}$$

When considering heterogeneous firms, it is assumed that $\Lambda < \bar{\Lambda}_H$, i.e., the two firms will need to compete. First, consider the extreme case where one firm takes the whole market in equilibrium: a *dominated market*.

Theorem 9. Fix $i, j = \{1, 2\}$ and $i \neq j$. Suppose that:

$$\mu_i > \Lambda\,,\ \mu_j \geq \Lambda \tag{5.33}$$

and

$$R_i - \frac{c_i \mu_i}{(\mu_i - \Lambda)^2} \geq R_j - \frac{c_j(\mu_j - \Lambda)}{\mu_j^2}. \tag{5.34}$$

Then:

$$P_i = R_i - R_j - \frac{c_i}{\mu_i - \Lambda} + \frac{c_j}{\mu_j} \text{ and } P_j = 0, \tag{5.35}$$

is an equilibrium, with corresponding $\lambda_i = \Lambda$ and $\lambda_j = 0$.

Consumer surplus is positive in the dominated market equilibrium, since although firm i takes the whole market, it is not able to charge its monopoly price (compare P_i obtained above with that in theorem 8). The threat of entry from firm j makes the monopolist behave in a manner consistent with contestable markets results.

As Λ increases, it may no longer be possible for a single firm to service the entire market. Total demand may, however, still be insufficient to sustain non-competitive behaviour, which will lead to the emergence of a *highly competitive market*, with positive consumer surplus.

Theorem 10. Suppose $\Lambda \leq \bar{\Lambda}_H$. There exists at most one solution $(P_1, P_2, \lambda_1, \lambda_2)$ satisfying the following:

$$\begin{aligned}
\lambda_1 + \lambda_2 &= \Lambda, \\
R_1 - P_1 - \frac{c_1}{\mu_1 - \lambda_1} &= R_2 - P_2 - \frac{c_2}{\mu_2 - \lambda_2} > \nu, \\
\frac{c_1 \lambda_2}{(\mu_1 - \lambda_1)^2} + \frac{c_2 \lambda_2}{(\mu_2 - \lambda_2)^2} &= P_2, \\
\frac{c_2 \lambda_1}{(\mu_2 - \lambda_2)^2} + \frac{c_1 \lambda_1}{(\mu_1 - \lambda_1)^2} &= P_1, \\
0 \leq \lambda_i &\leq \mu_i.
\end{aligned}$$

If $(P_1, P_2, \lambda_1, \lambda_2)$ is such a set of solutions also satisfying:

$$\frac{c_2 \mu_2}{(\mu_2 - \lambda_2)^3} + \frac{c_1(\mu_1 - \Lambda)}{(\mu_1 - \lambda_1)^3} > 0, \tag{5.36}$$

$$\frac{c_1 \mu_1}{(\mu_1 - \lambda_1)^3} + \frac{c_2(\mu_2 - \Lambda)}{(\mu_2 - \lambda_2)^3} > 0, \tag{5.37}$$

then (P_1, P_2) is an equilibrium with λ_1 and λ_2 being the corresponding demands.

Theorem 10 is a generalization of theorem 7, and reduces to it when the firms are homogeneous, matching the scarce demand case. If the set of equations is such that the solution yields one of $\lambda_i < 0$, then one firm takes the whole market in equilibrium, falling back to the situation described by theorem 9.

For homogeneous firms in competition, the equilibrium changes from scarce to moderate demand as Λ increases, causing the respective consumer surplus to drop to zero, and the unique equilibrium to change to a continuum of equilibria. Heterogeneous firms present the same results. The following lemma presents a necessary condition.

Lemma 1. Suppose $\Lambda \leq \bar{\Lambda}_H$. If (P_1, P_2) is an equilibrium with zero consumer surplus, then it must satisfy the following set of conditions:

$$\lambda_i = \mu_i - \frac{c_i}{R_i - \nu - P_i} \geq 0, \tag{5.38}$$

$$\mu_1 + \mu_2 - \frac{c_1}{R_1 - \nu - P_1} - \frac{c_2}{R_2 - \nu - P_2} = \Lambda, \tag{5.39}$$

$$\frac{c_1\lambda_2}{(\mu_1 - \lambda_1)^2} + \frac{c_2\lambda_2}{(\mu_2 - \lambda_2)^2} \geq P_2 \geq \frac{c_2\lambda_2}{(\mu_2 - \lambda_2)^2}, \tag{5.40}$$

$$\frac{c_2\lambda_1}{(\mu_2 - \lambda_2)^2} + \frac{c_1\lambda_1}{(\mu_1 - \lambda_1)^2} \geq P_1 \geq \frac{c_1\lambda_1}{(\mu_1 - \lambda_1)^2}. \tag{5.41}$$

With the conditions in eqs. (5.38)-(5.41) the *moderately competitive market* case, with a continuum of equilibria, follows:

Theorem 11. Suppose $\Lambda \leq \bar{\Lambda}_H$. Let (P_1, P_2) be a feasible solution to (5.38)-(5.41). Then (P_1, P_2) is a Nash equilibrium (with the corresponding set (λ_1, λ_2) being the demands) if it satisfies the conditions in either of the two following cases:

1. $\mu_1 \leq \Lambda$ and $\mu_2 \leq \Lambda$: $(P_1, P_2, \lambda_1, \lambda_2)$ satisfies either:

$$P_i + \frac{c_i}{\mu_i - \lambda_i} - \frac{c_j}{\mu_j - \lambda_j} \leq \frac{[(c_1(\Lambda - \mu_1))^{\frac{1}{3}} + (c_2\mu_2)^{\frac{1}{3}}]^3}{(\mu_1 + \mu_2 - \Lambda)^2}, \text{ or} \tag{5.42}$$

$$P_i\lambda_i \geq \left[\frac{c_j}{(\mu_j - \lambda_j^L)^2} + \frac{c_i}{(\mu_i - \lambda_i^L)^2}\right](\lambda_i^L)^2, \tag{5.43}$$

   where $\lambda_i^L + \lambda_j^L = \Lambda$, and λ_i^L is the largest root of:

$$f(\lambda_i) = R_i - R_j + P_j - \frac{c_i\mu_i}{(\mu_i - \lambda_i)^2} + \frac{c_j(\mu_j - \Lambda)}{(\mu_j - \lambda_j)^2}, \tag{5.44}$$

   in the range $(\Lambda - \mu_j, \mu_i)$.

2. For $\mu_i < \Lambda$ and $\mu_j > \Lambda$: $(P_i, P_j, \lambda_i, \lambda_j)$ satisfies:

$$R_i - R_j + P_j - \frac{c_i\mu_i}{(\mu_i - \lambda_i)^2} + \frac{c_j(\mu_j - \Lambda)}{(\lambda_i + \mu_j - \Lambda)^2} \leq 0, \tag{5.45}$$

   and either

$$R_j - R_i + P_i - \frac{c_j\mu_j}{(\mu_j - \Lambda)^2} - \frac{c_i(\Lambda - \mu_i)}{\mu_i^2} \leq 0, \text{ or} \tag{5.46}$$

$$\begin{cases} R_j - R_i + P_i - \frac{c_j\mu_j}{(\mu_j - \Lambda)^2} - \frac{c_i(\Lambda - \mu_i)}{\mu_i^2} > 0, \text{ and} \\ \left(R_j - R_i + P_i - \frac{c_j}{\mu_j - \Lambda} + \frac{c_i}{\mu_i}\right)\Lambda \leq P_j\lambda_j. \end{cases} \tag{5.47}$$

Theorem 11 is a generalization of theorem 6, allowing for heterogeneity between firms. It shows that when $\underline{\Lambda} < \Lambda < \bar{\Lambda}$, there is a continuum of equilibria. The existence of these equilibria has been confirmed by numerical investigations, as it is straightforward to confirm whether a candidate equilibrium meets the conditions.

When the market does not fall into any of the four foregoing cases, the authors conjecture no equilibrium in pure strategies exists, and give a numerical example where the necessary condition in lemma 1 is not met.

Equilibrium Price and Comparative Statics

In this section, the effects of μ_i, R_i and c_i on equilibrium will be considered.

Larger capacity confers advantage on a firm: a firm with higher capacity is able to charge a higher price and capture a larger market share.

Proposition 4. All other things being identical, the firm with larger capacity, higher value of service, or lower cost of waiting can charge a higher price and capture a larger market share in the dominated market, the non-competitive market, and the highly competitive market.

This advantage is less clear in the moderately competitive market where there is a continuum of equilibria. Numerical investigations show that it is possible that the larger capacity firm will charge a lower price, capture a smaller market share, or even earn a lower profit.

Comparative statics are only well-defined when an equilibrium exists and is unique, so only the cases of a dominated, non-competitive and highly competitive markets will be considered. Take first the dominated market. Let firm 1 be the dominant firm. It follows from theorem 9 that:

$$P_1 = R_1 - R_2 - \frac{c_1}{\mu_1 - \Lambda} + \frac{c_2}{\mu_2}. \tag{5.48}$$

This clearly implies that P_1 increases with R_1, μ_1 and c_2, and decreases with c_1, R_2 and μ_2. Therefore, a dominating firm is able to raise its price when its competitive edge increases, and must cut it when its competitor competitive edge improves. As the firm takes the whole market, an increased price implies an increased profit, and vice versa.

In the non-competitive market, where each firm services at least some consumers and charges its monopoly price, any marginal change in one firm does not affect the other. Firm i is able to raise its price and increase its market share and revenue when μ_i or R_i increase, or c_i decreases.

Finally, in the highly competitive market, no explicit solution for the equilibrium was obtained. Therefore, the authors resorted to numerical investigations to analyse its comparative statics. In summary, firm i must cut (or keep unchanged) its price,

to take a smaller (or keep unchanged its) market share, and to experience a reduction of (or no change in) its profit, as either μ_j or R_j increase, or c_j decreases. As own capacity μ_i increases, a firm can usually raise its price, except when demand is scarce and there is competition, but neither firm can dominate: price will fall and market share will increase, though the effect on profit is ambiguous. As c_i increases, a firm usually cuts its price, except when demand is scarce, competition fierce, and neither firm can dominate; in that case, the firm loses market share but may raise or cut its price, and its profit will decrease.

5.3.3 Social Welfare

The monopoly model for queueing markets has been shown to be socially efficient (see Edelson and Hildebrand (1975) and Chen and Frank (2004)). In the duopoly case, that is only true in some special cases.

With demand λ_i and price P_i, consumer surplus per unit of time for firm i is $CS_i = \lambda_i(R - P_i - c_i T_i(\lambda_i) - \nu)$, and producer surplus is identical to firm profit, $PS = \lambda_1 p_1 + \lambda_2 p_2$. Social welfare is their sum:

$$SW = CS + PS = \left(R_1 - \nu - \frac{c_1}{\mu_1 - \lambda_1}\right)\lambda_1 + \left(R_2 - \nu - \frac{c_2}{\mu_2 - \lambda_2}\right)\lambda_2. \tag{5.49}$$

The social planner ignores price as an internal wealth transfer. To maximize SW, $\lambda_i < \mu_i$ must hold, otherwise $T_i(\lambda_i) = \infty$ and $SW = -\infty$. The planner's problem will take the following form:

$$\max_{\lambda_1,\lambda_2} \left(R_1 - \nu - \frac{c_1}{\mu_1 - \lambda_1}\right)\lambda_1 + \left(R_2 - \nu - \frac{c_2}{\mu_2 - \lambda_2}\right)\lambda_2, \tag{5.50}$$

subject to

$$\begin{aligned} &\lambda_1 + \lambda_2 \leq \Lambda, \\ &0 \leq \lambda_1 < \mu_1, \text{ and} \\ &0 \leq \lambda_2 < \mu_2. \end{aligned}$$

Theorem 12. Let $\bar{\Lambda}_H$ be defined as in theorem 8. If $\Lambda \geq \bar{\Lambda}_H$, then the social optimum is given by:

$$\lambda_i = \mu_i - \sqrt{\frac{c_i \mu_i}{R_i - \nu}}; \tag{5.51}$$

otherwise, it is given by:

$$\lambda_i = \mu_i - \sqrt{\frac{c_i \mu_i}{R_i - \nu - \eta}}, \tag{5.52}$$

where η is the unique solution to

$$\mu_1 + \mu_2 - \sqrt{\frac{c_1 \mu_1}{R_1 - \nu - \eta}} - \sqrt{\frac{c_2 \mu_2}{R_2 - \nu - \eta}} = \Lambda, \tag{5.53}$$

in the interval $(0, \min\{R_1 - \nu, R_2 - \nu\})$.

Proof. The first order conditions for the planner's problem are:

$$R_i - \nu - \frac{c_i \mu_i}{(\mu_i - \lambda_i)^2} - \eta = 0, \tag{5.54}$$

$$\eta(\lambda_1 + \lambda_2 - \Lambda) = 0, \tag{5.55}$$

where η is the Lagrange multiplier. The proof considers two cases: $\lambda_1 + \lambda_2 < \Lambda$ and $\lambda_1 + \lambda_2 = \Lambda$.

The equation:

$$\mu_1 + \mu_2 - \sqrt{\frac{c_1 \mu_1}{R_1 - \nu - \eta}} - \sqrt{\frac{c_2 \mu_2}{R_2 - \nu - \eta}} = \Lambda, \tag{5.56}$$

has a unique solution in the interval $(0, \min\{R_1 - \nu, R_2 - \nu\})$, as the left-hand-side is monotonically decreasing in η, and has a value greater than Λ at $\eta = 0$, approaching $-\infty$ as η increases to $\min\{R_1 - \nu, R_2 - \nu\}$. □

Comparing with theorem 8, the Nash equilibrium is found to be socially optimal in the ample demand case ($\Lambda \geq \bar{\Lambda}$). In the homogeneous firms case, it is easy to verify that it is also optimal in the scarce demand case ($\Lambda \leq \underline{\Lambda}$) (compare with theorem 7). However, in the more general case where $\Lambda < \bar{\Lambda}_H$, the Nash equilibrium is usually not socially optimal.

5.4 Monopoly with Heterogeneous Consumers

This section presents a model of a single firm servicing heterogeneous consumers. While it is not very interesting in its own right, it is useful to contrast it to duopoly model for heterogeneous consumers below. This model is based on the duopoly model presented in Luski (1976).

Consumer decisions are based on ex-ante expected benefit. Consumer arrivals follow a Poisson distribution with mean rate λ. As a simplifying assumption, $\lambda = 1$; this can be done without loss of generality as the choice of time units is arbitrary. The service rate is exogenous and constant: service times follow an exponential distribution with exogenous parameter μ. Mean length of service per consumer is therefore $\frac{1}{\mu}$.

Let λ_M be the share of arriving consumers seeking service from the monopolist M, so that $\lambda - \lambda_M = 1 - \lambda_M$ is the share of arriving consumers who balk and do not seek service.

Let T be the ex-ante expected sojourn time at the monopolist's queue. As the firm is an *M/M/1* system:

$$T = \frac{1}{\mu - \lambda_M}. \tag{5.57}$$

Consumers are identical in their valuation of service R. Their utility is further dependent on price P, which is set by the monopolist, on T, and on time cost c, which varies across consumers according to a known distribution function $f(c)$ (whose cumulative distribution function is denoted by $F(c)$). This combines to yield the following utility function:

$$U = R - cT - P. \tag{5.58}$$

A consumer will seek service from the monopolist if $U > 0$:

$$R - cT - P > 0 \Rightarrow c \leq \frac{R-P}{T}. \tag{5.59}$$

Conversely, a consumer will balk if:

$$c > \frac{R-P}{T}. \tag{5.60}$$

The monopolist's demand can be obtained from (5.59) and $F(c)$ in a straightforward manner:

$$\lambda_M = F\left(\frac{R-P}{T}\right) \tag{5.61}$$

where $F\left(\frac{R-P}{T}\right)$ is the share of consumers whose time cost c is low enough that they join the queue, and $1 - F\left(\frac{R-P}{T}\right)$ is the share of consumers whose time cost is so high that they do not.

Expected sojourn time T, a function of μ and λ_M, is then:

$$T = \frac{1}{\mu - \lambda_M}. \tag{5.62}$$

Given the price P, the parameter μ, and the distribution function $f(c)$, demand and sojourn time for can be obtained from solving the system of two equations in two unknowns made up of eqs. (5.61)-(5.62).

The monopolist has no production or service costs, and aims to maximize expected profit per unit of time, π:

$$\pi = P\lambda_M. \tag{5.63}$$

The monopolist can vary prices directly, but can only vary sojourn times through prices, as μ is exogenous. Price changes will then have two conflicting effects: decreasing price will make the firm more attractive to consumers, which will increase demand and therefore sojourn time, which in turn counteracts the benefits for consumers of a lower price. Will the monopolist serve the entire market? Only under certain conditions.

The monopolist solves the maximization problem subject to restrictions (5.61)-(5.62). Let P^* be the equilibrium price—it is such that:

$$\frac{\partial \pi}{\partial P} = 0 \, \forall P = P^*. \tag{5.64}$$

Let $\beta = \frac{R-P}{T}$. The monopolist's profit function can be given as:

$$\pi = PF(\beta), \tag{5.65}$$

whence the partial derivative follows:

$$\frac{\partial \pi}{\partial P} = F(\beta) - Pf(\beta)\frac{1}{T}\left[1 + \beta\frac{\partial T}{\partial P}\right]. \tag{5.66}$$

The partial derivative $\frac{\partial T}{\partial P}$ can be obtained in terms of the parameters by differentiating (5.62) in respect to P, and substituting that results in (5.66), yielding:

$$\frac{\partial \pi}{\partial P} = F(\beta) - Pf(\beta)\left[\frac{1}{T} + \frac{\beta f(\beta)}{1 + Tf(\beta)}\right]. \tag{5.67}$$

Then P^* is given by:

$$P^* = \frac{F(\beta)T}{f(\beta)}\left[\frac{1}{T} + \frac{\beta f(\beta)}{1 + Tf(\beta)}\right]^{-1}. \tag{5.68}$$

5.5 Duopoly with Heterogeneous Consumers

The duopoly model is arguably more interesting with heterogeneous consumers. This was developed in Luski (1976) and Levhari and Luski (1978). Two identical firms compete on price to provide an identical good to consumers with different time costs. Despite the firms being identical, consumer heterogeneity allows for differentiation through charging different prices. Indeed, a separating equilibrium is found in which one firm charges a high price, so that sojourn times are low and the firm is preferred by highly impatient customers, whereas the other firm charges a low price, so that sojourn times are high and the firm is preferred by more patient customers.

Consumer decisions are based on ex-ante expected benefit. Consumer arrivals follow a Poisson distribution with mean rate λ. As a simplifying assumption, $\lambda = 1$; this can be done without loss of generality as the choice of time units is arbitrary. The service rate is identical in both firms, and not capable of modification: service times follow an exponential distribution with exogenous parameter μ. Mean length of service per consumer is therefore $\frac{1}{\mu}$.

Two firms $i = \{1, 2\}$ provide an homogeneous good. Let λ_i be the share of arriving consumers preferring the firm i, so that $\lambda - \lambda_1 - \lambda_2 = 1 - \lambda_1 - \lambda_2$ is the share of arriving consumers who balk and do not join any queue.

Let T_i be the ex-ante expected sojourn time at firm i. As each firm is an *M/M/1* system:

$$T_i = \frac{1}{\mu - \lambda_i}. \tag{5.69}$$

Consumers are identical in their valuation of service R. Their utility is further

dependent on price P_i, which is set by firm i,[1] on T_i, and on time cost c, which varies across consumers according to a known distribution function $f(c)$ (whose cumulative distribution function is denoted by $F(c)$). This combines to yield the following utility function:

$$U_i = R - cT_i - P_i. \tag{5.70}$$

A consumer will seek service from firm 1 if $U_1 > U_2$ and $U_1 > 0$:

$$R - cT_1 - P_1 > R - cT_2 - P_2 \Rightarrow c > \frac{P_1 - P_2}{T_2 - T_1}, \text{ and} \tag{5.71}$$

$$R - cT_1 - P_1 > 0 \Rightarrow c \leq \frac{R - P_1}{T_1}. \tag{5.72}$$

On the other hand, a consumer will seek service from firm 2 if:

$$0 < c \leq \frac{P_1 - P_2}{T_2 - T_1}, \text{ and} \tag{5.73}$$

$$c \leq \frac{R - P_2}{T_2}. \tag{5.74}$$

Finally, a consumer will balk if:

$$c > \frac{R - P_1}{T_1}, \text{ and} \tag{5.75}$$

$$c > \frac{R - P_2}{T_2}. \tag{5.76}$$

As long as at least one consumer is served by each firm, it must be the case that $T_1 \leq T_2$: as $P_1 \geq P_2$, that is the only way for (5.71) to hold. It follows that

$$\frac{P_1 - P_2}{T_2 - T_1} < \frac{R - P_1}{T_1}. \tag{5.77}$$

This allows the conditions in (5.71)-(5.76) to be reduced to three. A consumer prefers firm 2 if:

$$c \leq \frac{P_1 - P_2}{T_2 - T_1}. \tag{5.78}$$

The consumer will prefer firm 1 when:

$$\frac{P_1 - P_2}{T_2 - T_1} < c \leq \frac{R - P_1}{T_1}, \tag{5.79}$$

and will balk when:

$$c > \frac{R - P_1}{T_1}. \tag{5.80}$$

Demand functions faced by each firm can be obtained from the simplified conditions

[1] Without loss of generality, it is assumed that $P_1 \geq P_2$.

at (5.78)-(5.80) and $F(c)$:

$$\lambda_1 = F\left(\frac{R-P_1}{T_1}\right) - F\left(\frac{P_1-P_2}{T_2-T_1}\right), \text{ and} \tag{5.81}$$

$$\lambda_2 = F\left(\frac{P_1-P_2}{T_2-T_1}\right), \tag{5.82}$$

where $F\left(\frac{R-P_1}{T_1}\right)$ is the share of consumers whose time cost c is low enough that they join a queue; $F\left(\frac{P_1-P_2}{T_2-T_1}\right)$ is the share with lower values of c who will seek the service from the firm with low prices and long queues. Finally, $1 - F\left(\frac{R-P_1}{T_1}\right)$ is the share of consumers whose time cost is so high that they join neither queue.

Expected sojourn times T_i, a function of μ and λ_i, are then:

$$T_1 = \frac{1}{\mu - \lambda_1}, \tag{5.83}$$

$$T_2 = \frac{1}{\mu - \lambda_2}. \tag{5.84}$$

Given the two prices P_1 and P_2, the parameter μ, and the distribution function $f(c)$, demand and sojourn time for both firms can be obtained from solving the system of four equations in four unknowns made up of eqs. (5.81)-(5.84). Note though that when $P_1 = P_2$, $\frac{P_1-P_2}{T_2-T_1}$ is not well defined, as the firms are identical. In this special case:

$$\lambda_1 = \lambda_2 = \frac{1}{2}F\left(\frac{R-P_1}{T_1}\right). \tag{5.85}$$

Firms have no production or service costs, and aim to maximize expected profit per unit of time, π_i:

$$\pi_i = P_i\lambda_i. \tag{5.86}$$

Each firm takes the other's price as a given and maximizes profits with respect to its own price, yielding a Cournot-type behaviour where reaction curves can be obtained.

Under some conditions, firms will sell at different prices. Firms can vary prices directly, but can only vary sojourn times through prices, as μ is exogenous. Price changes will then have two conflicting effects: decreasing price will make the firm more attractive to consumers, which will increase demand and therefore sojourn time, which in turn counteracts the benefits for consumers of a lower price.

Firms solve their profit maximization problem subject to restrictions (5.81)-(5.84), with the other firm's price assumed constant. This yields an optimal price for all of the other firm's possible prices, and the equilibrium will be the intersection of these reaction curves, at which point neither firm wants to change its price. Let P_1^* and P_2^* be these equilibrium prices. The question then is can $P_1^* = P_2^*$ be an equilibrium? This can be answered by obtaining the partial derivatives $\partial\pi_1/\partial P_1$ and $\partial\pi_2/\partial P_2$, and

calculating their values when $P_1 = P_2$. Then $P_1^* = P_2^*$ will be equilibrium prices if:

$$\begin{aligned} &\frac{\partial \pi_1}{\partial P_1} \leq 0 \,\forall\, P_1 \geq P_1^*, \text{ and} \\ &\frac{\partial \pi_2}{\partial P_2} \geq 0 \,\forall\, P_2 \leq P_2^*. \end{aligned} \tag{5.87}$$

These conditions must hold with strict equality where the demand functions are continuous and kink-free. However, it can be shown that demand is not continuous at this point, so inequality restrictions like (5.87) are appropriate. If $\partial \pi_1 / \partial P_1 > \partial \pi_2 / \partial P_2$ when $P_1 = P_2$, then the conditions at (5.87) do not hold. Let

$$\alpha = \frac{P_1 - P_2}{T_2 - T_1}, \text{ and } \beta = \frac{R - P_1}{T_1}.$$

Starting the analysis with firm 2, its profit function can be given as:

$$\pi_2 = P_2 F(\alpha), \tag{5.88}$$

whence the partial derivative follows:

$$\frac{\partial \pi_2}{\partial P_2} = F(\alpha) - \frac{P_2 f(\alpha)}{T_2 - T_1} \left[1 + \alpha \left(\frac{\partial T_2}{\partial P_2} - \frac{\partial T_1}{\partial P_2} \right) \right]. \tag{5.89}$$

The factor $\frac{\partial T_2}{\partial P_2} - \frac{\partial T_1}{\partial P_2}$ can be obtained in terms of the parameters by differentiating (5.83) and (5.84) in respect to P_2, and substituting those results in (5.89), yielding:

$$\frac{\partial \pi_2}{\partial P_2} = F(\alpha) - P_2 \left[\frac{T_2 - T_1}{f(\alpha)} + T_2^2 \alpha + \frac{T_1^2 \alpha}{\gamma} \right]^{-1}, \tag{5.90}$$

where $\gamma = 1 + T_1 f(\beta) \beta$. When $P_2 = 0$, $\partial \pi_2 / \partial P_2 = F(\alpha) > 0$.

When $P_2 \leq P_1$, as $P_2 \to P_1$, $T_2 \to T_1$. Then

$$\lim_{P_2 \to P_1} \frac{\partial \pi_2}{\partial P_2} = F(\alpha) - \frac{P_2}{T^2} \left(\frac{1}{\alpha + \alpha / \gamma} \right), \tag{5.91}$$

where at the limit, $T = T_1 = T_2$.

Similar proceedings for firm 1 yield:

$$\lim_{P_1 \to P_2} \frac{\partial \pi_1}{\partial P_1} = F(\alpha) - \frac{P_1}{T^2} \frac{[f(\beta) T \alpha + 1] / \gamma}{\alpha + \alpha / \gamma}. \tag{5.92}$$

Using γ, the derivative for firm 1 can be written as:

$$\lim_{P_1 \to P_2} \frac{\partial \pi_1}{\partial P_1} = F(\alpha) - \frac{P_1}{T^2} \left(\frac{1}{\alpha + \alpha / \gamma} \right) + \frac{P_1}{T^2} \frac{f(\beta) T (\beta - \alpha)}{\gamma \alpha + \alpha}. \tag{5.93}$$

Using (5.91), it follows that:

$$\lim_{P_1 \to P_2} \frac{\partial \pi_1}{\partial P_1} = \frac{\partial \pi_2}{\partial P_2} + \frac{P_1}{T} \frac{f(\beta)(\beta - \alpha)}{\alpha(\gamma + 1)}. \tag{5.94}$$

Given $\beta > \alpha$, the right hand side of the foregoing is positive:

$$\frac{\partial \pi_1}{\partial P_1} \geq \frac{\partial \pi_2}{\partial P_2}, \tag{5.95}$$

and the equality holds only if $f(\beta) = 0$.

The second order conditions are satisfied when $P_1^* \neq P_2^*$ and $\partial \pi_i / \partial P_i = 0$. When $P_1^* = P_2^*$, the firms are in equilibrium if the conditions in (5.87) are satisfied. When $f(\beta) > 0$, (5.95) is an inequality for all P_i pairs where $P_1 = P_2$ and (5.87) are not satisfied. $f(\beta)$ is the probability that consumers with high waiting costs do not join the queue. When $f(\beta) = 0$ all consumers get served; if $f(\beta) > 0$, some consumers balk, and the price cannot be identical across firms. But β depends on the price, so it cannot be said whether or not prices will be equal.

If the system diverges ($\lambda/2 > \mu$), and no consumer balks, sojourn time will reach infinity and $f(\beta)$ will be positive. Therefore, where $\lambda/2 > \mu$, firms will sell at different prices. Meanwhile, if $\lambda/2$ is small compared to μ, then sojourn time will be short, even if everyone joins a queue. Then, $F(\beta) = 1$, so it is reasonable that $f(\beta) = 0$ and price will be the same in both firms. The following lemma summarizes these results.

Lemma 2. Suppose that the set on which $f(x) > 0$ is a single interval. Then, a necessary condition for equal price equilibrium P is:

$$R \geq T^2 \lambda F^{-1}(0.5) + T F^{-1}(1), \tag{5.96}$$

where T is expected sojourn time in each firm such that each firm serves one-half of the stream, and no consumer balks.

Proof. If $P_1 = P_2 = P$, then $f(\beta) = 0$. The first order condition for firm 1 becomes

$$\frac{\partial \pi_1}{\partial P_1} = F(\alpha) - \frac{P}{T^2} \left(\frac{1}{2\alpha} \right) = 0. \tag{5.97}$$

As $F(\alpha) = \lambda/2$ and $\lambda = 1$, it follows that:

$$P = \lambda T^2 \alpha = \lambda T^2 F^{-1}(0.5). \tag{5.98}$$

$f(\beta)$ is zero if

$$\frac{R - P}{T} \leq F^{-1}(1). \tag{5.99}$$

Substituting (5.98) into (5.99) yields the necessary condition. □

Therefore, it is a condition for equal prices that all consumers are served. However,

the condition is not sufficient. A further requirement is that $f(\beta) = 0$.

The equilibria are further characterized in Levhari and Luski (1978), although due to complexity of the conditions in the general case, this requires simplifying assumptions and numerical results.

First, let $f(c)$ be a uniform distribution over the range $[0, A]$, so that $f(c) = \frac{1}{A}$, and $F(c) = \frac{c}{A}$. Demand for each firm is then:

$$\lambda_1 = \frac{R - P_1}{T_1}\frac{1}{A} - \frac{P_1 - P_2}{T_2 - T_1}\frac{1}{A}, \text{ and} \tag{5.100}$$

$$\lambda_2 = \frac{P_1 - P_2}{T_2 - T_1}\frac{1}{A}, \tag{5.101}$$

assuming $(R - P_1)/T_1 < A$.

Then, for a given (P_1, P_2), the system composed of eqs. (5.81)-(5.84) yields two quadratic equations. Demand for firm 1 is:

$$\lambda_1 = \frac{b - (b^2 - 4ac)^{\frac{1}{2}}}{2a}, \tag{5.102}$$

where

$$a = \left[3\frac{R - P_1}{A} + 2 + \left(\frac{R - P_1}{A}\right)^2 + \frac{P_1 - P_2}{A} + \frac{R - P_1}{A}.\frac{P_1 - P_2}{A}\right],$$

$$b = \left[3\frac{R - P_1}{A} + 2\left(\frac{R - P_1}{A}\right)^2 + 2\frac{R - P_1}{A}.\frac{P_1 - P_2}{A}\right]\mu,$$

$$c = \left[\left(\frac{R - P_1}{A}\right)^2 + \frac{R - P_1}{A}.\frac{P_1 - P_2}{A} - \frac{P_1 - P_2}{A}\right]\mu^2.$$

In the same manner, demand for firm 2 is:

$$\lambda_2 = \frac{b' - (b'^2 - 4a'c')^{\frac{1}{2}}}{2a'}, \tag{5.103}$$

where

$$a' = 2 + \frac{R - P_1}{A} + \frac{P_1 - P_2}{A},$$

$$b' = \frac{R - P_1}{A}.\mu,$$

$$c' = \frac{P_1 - P_2}{A}.\mu^2.$$

As eqs. (5.102)-(5.103) present demand as the sole function of (P_1, P_2), firm i can maximize the following problem, taking P_j, $j \neq i$ as given:

$$\underset{P_i}{\text{Max}}\, \pi_i(P_1, P_2) = \underset{P_i}{\text{Max}}\, \lambda_i(P_1, P_2)P_i. \tag{5.104}$$

The first order condition is then:

$$\frac{\partial \pi_i}{\partial P_i} = \lambda_i + P_i \frac{\partial \lambda_i(P_1, P_2)}{\partial P_i} = 0, \tag{5.105}$$

which for firm 1 is:

$$\frac{\partial \lambda_1}{\partial P_1} = -\frac{\lambda_1^2(\partial a/\partial P_1) - \lambda_1(\partial b/\partial P_1) + (\partial c/\partial P_1)}{2\lambda_1 a - b}, \tag{5.106}$$

while for firm 2 it is:

$$\frac{\partial \lambda_2}{\partial P_2} = -\frac{\lambda_2^2(\partial a'/\partial P_2) - \lambda_2(\partial b'/\partial P_2) - (\partial c'/\partial P_2)}{2\lambda_2 a' - b'}. \tag{5.107}$$

Substituting (5.106) and (5.107) into (5.104) yields each firm's reaction function. In line with Luski (1976), there are two kinds of reaction functions. The first kind occurs when all consumers are served: the curves intersect on the 45 degree line and yield a stable equilibrium where both firms charge the same price and divide demand equally. In the other case, some consumers do balk, and the two firms have different reaction functions with a discontinuity on the 45 degree line. There is then some range where two local profit maximizing prices obtain for a firm given the other firm's price. One of those prices is below the 45 degree line, and the other above it, and there is no stable Nash equilibrium on the 45 degree line, so the two firms will never charge the same price: one firm specializes in consumers with a high cost of time, and the other firm will serve the remainder.

If firms only care about local profit maximization, they will make sequential 'local' price changes and find one of the two stable local equilibrium prices. The two are symmetrical in regard to the 45 degree line, giving the prices for both without identifying which firm will charge a high price. If firms search for the optimal price across the full range of prices to obtain the global maximum, the reaction curves have a discontinuity outwith the 45 degree line and do not intersect at all. Therefore, there is no stable Cournout-Nash equilibrium and a non-convergent oscillation occurs over and under the 45 degree line. This non-existence result is not general, but is due to the particular shape of the time cost distribution.

With a different type of distribution for c, reaction curves may possess two intersection points, which the authors show using a Pareto distribution for c. In that case, the numerical analysis shows the emergence of two stable equilibria, where the first firm (with the higher price) earns smaller profits than the second (with the lower price). Which firm takes each role is impossible to predict, and price-war type behaviour may develop. It is also found that charging identical prices is a local minimum.

Detailed results about social welfare with heterogeneous consumers are absent from the discussion, possibly because the introduction of the stochastic element into consumer welfare makes the problem exponentially harder. Chapter 9 below advances the discussion of this issue in the Health Care context.

Chapter 6

Other Developments

There are many other applications of queueing theory to economic topics, but scope precludes a full treatment here. Nevertheless, short mention will be made of a small sample.

A very interesting development is the consideration of a setting where there are multiple servers with different service rates, and customers do not have full information about these. In this case, a "smart" customer may opt to wait and observe other customers being served before deciding which queue to join, considering that the gains in information may outweigh the lost time. This question has been taken up in Hlynka et al. (1994). However, this research seems hampered by tractability problems. In particular, an approach to the situation where all customers are "smart" seems absent from the literature.

On the other hand, experimental research has provided welcome verification of theoretical work. A couple of interesting example are Milgram et al. (1986), where a field experiment was conducted where researchers attempted to ask members of the public in queues to overtake them. This was extended in Schmitt et al. (1992), where it was shown that customers are more likely to allow intrusions which are perceived by them as legitimate. Helweg-Larsen and LoMonaco (2008) studied reactions of fans of the band U2 who are queueing for a concern, showing that fairness concerns influence queueing behaviour. A significant drawback of this literature is its dominance by field experiments, possibly tainting the results due to biased sampling. It would be interesting to move the experiments to a laboratory setting, where quality sample selection might yield better results.

Part II

Three Essays on Strategic Queueing

Chapter 7

Endogenous Queue Number Determination in M/M/2 Systems

7.1 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 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 treatment of queueing in microeconomic theory goes back to P. Naor's seminal paper Naor (1969). However, long before economics started considering queueing phenomena, they had been extensively described by the Operations Research (OR) literature. OR developed a specialized terminology to describe queueing systems, for an overview of which, see chapter 2 of the Introduction, or any OR textbook, such as the popular Gross et al. (2008).[1] See further chapter 3 for more detailed discussion of the various works mentioned below.

The present chapter takes place in the context of $M/M/1$ and $M/M/2$ systems, under a First Come First Served (FCFS) discipline: M/M denotes that inter-arrival and service times, in that order, are independent and exponentially distributed, and the digit at the end indicates the number of servers servicing the queue(s). The FCFS discipline indicates customers are served in the order of their arrival.

FCFS $M/M/1$ queues featured in Naor (1969), whose innovation was considering the cost to customers of time spent in the queue. In Naor (1969), risk neutral, utility maximizing customers, with a linear utility function, choose the maximum length at which they will join the queue, which is the largest size for which the expected cost of waiting is weakly smaller than the good's net value. Once this happens, customers will turn away without the need for an exogenous capacity limit: this behaviour is known as balking. Naor also posits a condition, to be followed in this article, that it's not possible for customers to leave the queue once they join it.[2] Crucially, Naor formulated the expected benefit for the customers who do queue, which shall be used in the present chapter. 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 an $M/M/j$ system, where j is 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. Knudsen

[1] See also, for an extensive review, up to the date of publication, of the strategic queueing literature spawned by Naor (1969), Hassin and Haviv (2003).

[2] Leaving a queue after joining is termed reneging in the literature.

worked from the assumption that where there is more than one server, a single queue will feed all the servers.

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 present work, where this may not hold[3]). 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 on the other queue.

Where multiple queues are present despite their inefficiency, it has been shown that under certain conditions (an $M/M/j$ system, 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 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 despite its inefficiency, 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 work 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? Does it depend on whether jockeying (changing queues after joining one) is possible?

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) present some suggestions as to why this might be the case, but even if their arguments are valid, they certainly don't explain the emergence of multiple queues where there is no managerial intervention, such as at self-service points.

The present chapter's contribution is to answer these questions by analysing the strategic interactions between customers which determine the number of queues in a system. This analysis will employ a game theoretical model of queue formation, where manager preferences are not imposed on customers, so that the number of queues is determined endogenously. This model will be developed for a system with two servers, covering in turn risk neutral and risk averse customers, and starting from a baseline where jockeying (switching queues) is not allowed, to a less restricted case where in some circumstances it is possible for customers to jockey between queues.

[3]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.

The game starts when a customer arriving at the system encounters two busy servers, but no queue (the first two 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, as that reduces both expected waiting time and the variance thereof (which is relevant when the customers are risk averse). 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 second arrival does not always have the same benefits from that single queue: if customers are risk neutral, customer 2 is indifferent to the number of queues. In the case of risk neutral customers, it will be shown how arrivals alternate between strictly preferring one queue and being indifferent to the number of queues, according to whether the index of their arrival order is odd or even. This will lead to a proof that having a single queue is an equilibrium outcome of this game. This equilibrium is not unique, however, and it will be shown that if customers can jockey, a single queue will no longer be an equilibrium outcome when customers are risk neutral.

In order to investigate circumstances where the single queue equilibrium might be more robust to jockeying, section 7.3 focuses on risk-averse customers. In this case, it's found that the preference for a smaller variance, such as the single queue state offers, will make that equilibrium more robust: even when jockeying is allowed, the equilibrium exists and is unique as long as customers are sufficiently risk averse.

Steady state properties will not be considered, as the situation being modelled takes place when the queue is starting to form, before the steady state has had a chance to emerge. Therefore joining customers will not face the steady state expected waiting time, but an individual expected waiting time which varies with their arrival order and with the system state. The strategic interactions at play are how customers deal with newcomers to the system, who might disrupt the system order by trying to change the number of queues.[4]

7.2 Queue Number Determination with Risk Neutral Customers

Let there be a stream of customers seeking a service with a value (net of price) of R monetary units; their arrivals at the service station are a Poisson process in continuous time with parameter λ. This service is provided by two identical servers $j = \{a, b\}$. Obtaining the good from these servers takes time, distributed according to an expo-

[4]While addressing a different problem, that of whether customers let others cut ahead on the queue in an $M/M/1$ system, the recent paper 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.

nential distribution with rate μ. As there are two servers, only two customers can be served simultaneously. Others will wait until 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 two parallel $M/M/1$ queues, where each server services a separate FCFS queue, or one single $M/M/2$ queue serviced by both servers. The number of queues is endogenously determined through customer choices, being the game's equilibrium outcome.

Waiting imposes a cost on customers, who in this first section are assumed to be risk neutral. They experience cost c per unit of time t, yielding the following linear net utility (U_i) function, for customer i:

$$U_i(t_i) = R - ct_i, \tag{7.1}$$

where i, $i = \{-1, 0, 1, 2, ..., \infty\}$[5] is the customer's order of arrival into the system. For the sake of notational simplicity, the assumption is made that no customer leaves the system in the period under analysis. This can be done without loss of generality, as will be shown later. From the linear form of the utility function, it is clear that the risk neutral customers' objective in the game is equivalent to minimizing waiting time t.

The game starts when customers -1 and 0 are being served, but no other customers are waiting to be served. The system can take two possible states, denoted by $Q = \{1, 2\}$, according to the following definitions:

Definition 1. A single queue state ($Q = 1$) occurs when customer 1 stands roughly halfway between -1 and 0, and takes the place of whichever of these is served first (and for notational convenience, also when there is no queue).

Definition 2. A two queue state ($Q = 2$) occurs when customers 1 and 2 queue behind customers -1 and 0, respectively, and have to wait for the customer directly ahead of them to be served before taking their place. Jockeying is not allowed in the baseline model.

Each arrival i at the system observes the system state. Let this be denoted by a state variable $\gamma_i(Q, k)$, where Q is as defined above and k is the number of queueing customers when $Q = 1$, or the number of queueing customers on the shortest/either queue when $Q = 2$.[6]

Upon arrival at the system, customers choose from one of two actions, comprising the action set $A = \{D, S\}$:[7]

1. Action S: queue for both servers;

[5] Customers -1 and 0, who are being served, do not take part in the game, but are required for its setting.

[6] E.g., if there are four customers in the system, and $Q = 2$, $k = 2$; if $Q = 1$, $k = 4$.

[7] As a simplifying assumption, the possibility of balking (i.e., leaving without joining the queue) will not be considered. It is not the focus of the chapter, 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 D: queue for whichever server has the shortest queue, or randomize evenly if the queues are of identical size (cf. Winston (1977)).

Customer arrivals trigger a new round of the game, which is played sequentially—let this first game be denoted by Game NJ. Formally, the game stages, which are common knowledge, are:

1. Nature assigns customers an arrival number $i \in \{1, ..., \infty\}$.

2. Customer i arrives at the system, and chooses from action set $A = \{D, S\}$. This choice can be discerned by any incumbent customers with perfect accuracy. The chosen action is not performed until stage 4, however. Choosing S will not allow the customer closer access to the servers than incumbents who chose D.[8] If there are more than two customers waiting and the system is in a two queue state, customers must choose D and the round terminates.[9]

3. This stage only occurs if customer $i > 1$ encounters a single queue state and chooses action D in stage 2. In that case, incumbent customers split the single queue into two 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. Customer i acts upon his choice in stage 2. He cannot change his decision to react to incumbents' moves in stage 3, if these occurred.

5. Stages 2-4 are repeated for customer $i + 1$, with i now being an incumbent, and so on for all future arrivals.

Customers' strategy space is then composed of set A. Let $\Sigma_i = \{\alpha\}$ be the strategy for any customer i, where $\alpha \in A$. Customers' waiting time is uncertain, as the queues are stochastic processes and strategic interactions with newly arrived customers may alter the system state, which affects expected waiting time. Let $t_i(\Sigma, \gamma)$ be the waiting time for customer i, as a function of i's strategy and the system state i is facing.

It will subsequently be shown that in the equilibrium path, customers arriving after customer i will not alter the number of queues, i.e., stage 3 of the game is never triggered. In a subgame perfect equilibrium (see Hassin and Haviv (2002)), customers make their decision with full knowledge, gained through backwards induction, of the strategy of future arrivals. While these assertions will all be rigorously shown below, they are noted here to explain why the notation and computation of waiting times do

[8] So this choice is never taken in the equilibrium path, as demonstrated in lemma 3.

[9] If a further arrival chose action S in the presence of an established two queue state, he would still be constrained by the downstream two queue structure; he would, in fact, be simply postponing the decision to join one queue or the other! This echoes Hlynka et al. (1994) on "smart" customers who hold off on choosing a queue, but while interesting, the issue is not the present model's concern.

not *explicitly* incorporate possible changes to waiting time caused by changes in the number of queues.

7.2.1 Waiting Times[10]

Customers' expected waiting times are a function of system state and the customer's position in the queue. For a more detailed explanation of results below, see chapters 2 and 3. Upon arrival to the system, a customer observes queue length k (were the customer to join the queue, this would increase to $k+1$). Expected waiting time t, for any customer, is then a function of two discrete components:

$$E[t(k)] = E[X] + E[Y(k)], \tag{7.2}$$

where X is the service time and $Y(k)$ is the time spent on the queue for a customer waiting behind k customers (i.e., at the $k+1$th place). Service time X is exponentially distributed with rate μ, so that its density function is:

$$f(t) = \mu \exp(-\mu t),\ t > 0. \tag{7.3}$$

On the other hand, $Y(k)$ varies according to whether customers queue for one or both servers.[11] When j is the number of servers servicing the queue, $Y(k)$ follows a Gamma distribution with density:[12]

$$g(t,\, k,\, j) = \frac{(j\mu)^{k-j+1}}{(k-j)!} t^{k-j} \exp(-j\mu t),\ t > 0, \tag{7.4}$$

so that generally, waiting time has a density function:

$$z(t,\, k) = \begin{cases} f(t) & \text{if } 0 \leq k < j \\ \int_0^t f(t-u) g(u,\, k,\, j)\, \mathrm{d}u & \text{if } k \geq j, \end{cases} \tag{7.5}$$

where in the second case, the customer queues for time u obtained from eq. (7.4), and then reaches a server and is served for an expected time derived from eq. (7.3); in the first case, a server is idle on arrival, and only service time is taken into account.

The density functions for the two cases under consideration can be easily obtained from eq. (7.5). For the case where there are two queues ($k \geq j = 1$), the density function for waiting time in each queue is:

$$z(t,\, k) = \frac{\mu^{k+1}}{k!} \exp(-\mu t) t^k, \tag{7.6}$$

[10] This subsection is largely an exposition of work from Naor (1969) and Knudsen (1972).

[11] $Y(k) = 0$ if $k \leq j-1$, i.e., if $k \leq 1$ for $j = 2$, (there is an idle server upon arrival), there is no queueing time.

[12] See Knudsen (1972) for derivation, or section 3 in the Introduction.

while for a single queue with two servers ($k \geq j = 2$), it is:

$$z_k(t,\, k) = \frac{2^{k-1}\mu^k}{(k-2)!} \exp(-\mu t) \int_0^t u^{k-2} \exp(-\mu u)\, \mathrm{d}u, k \geq 2. \tag{7.7}$$

Using the density functions in (7.6)-(7.7), the expected waiting times can be easily derived. For a system with two parallel queues expected waiting time is given by:[13]

$$E[t(\gamma(2,k))] = \frac{1}{\mu}(k+1), \tag{7.8}$$

whereas for a system where one queue feeds two servers it is:

$$E[t(\gamma(1,k))] = \frac{1}{2\mu}(k+1), \tag{7.9}$$

where the intuition behind eqs. (7.8)-(7.9) is that having one queue feed two services doubles the processing rate.

In determining customers' preferred decisions, it is helpful to be able to compare expected waiting times directly across the two possible system states, for the same number of customers in the system. This can be done by stating expected waiting time as a function of i, the index of the customer's order of arrival, rather than k, i being preferred as it is invariant to the number of queues. It is for this reason that it is assumed that no customer leaves the system in the period under analysis. This can be done without loss of generality, for the departure of one customer simply shifts all remaining customers a step ahead in the queue. Customers still take into account the possibility of services finishing, but there is no need to incorporate that into the notation.

The expected waiting times given in (7.8)-(7.9) as a function of k can be easily transformed into functions of i. The form of the transformation differs according to whether the system is in a one or two queue state, and for the latter case, whether i is odd (i_o) or even (i_e). This difference arises because in a two queue state, customers with an odd i face two queues of equal length, while those with an even i face two queues of different length, in which case the expected waiting time is given for the shortest one. When expected waiting time is transformed into a function of i rather than k, k can, for convenience, be omitted from the state variable, presented then as $\gamma(Q)$.

In a system in a one queue state, $k = i + 1$,[14] it follows from (7.9):

$$E[t_i(\gamma(1))] = \frac{1}{2\mu}[(i+1)+1] = \frac{1}{2\mu}(i+2). \tag{7.10}$$

For customers arriving at a system in a two queue state, with an odd i, $k =$

[13] 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.

[14] E.g., $i = 1$ has two customers ahead of him, so his $k = 1 + 1 = 2$.

$i/2 + 1/2$,[15] it follows from (7.8):

$$E[t_{i_o}(\gamma(2))] = \frac{1}{\mu}\left[\left(\frac{i}{2} + \frac{1}{2}\right) + 1\right] = \frac{1}{\mu}\left(\frac{i+3}{2}\right). \tag{7.11}$$

Finally, for customers arriving at a system in a two queue state, with an even i, $k = i/2$,[16] it also follows from (7.8):

$$E[t_{i_e}(\gamma(2))] = \frac{1}{\mu}\left(\frac{i}{2} + 1\right) = \frac{1}{\mu}\left(\frac{i+2}{2}\right). \tag{7.12}$$

7.2.2 Customer Behaviour

A lemma governing customer behaviour can now be given:

Lemma 3. If customer 1 chooses action D, then customer 2 will also choose action D.

Proof. Say customer 1 chose action D, queueing for server a. If 2 also chooses D, he will queue for server b, in which case $k = 1$.

Let $P[\eta]$ be the probability of server a finishing a service before server b. Assume no customer arrives before the next service.[17] If η occurs, expected waiting time ($E[t_2(S, \gamma(2))|\eta]$) is the sum[18] of the expected service time for a to finish serving customer -1,[19] with the expected waiting time for action S (from (7.9), where $k = 2$). This is because, when a finishes serving -1, he starts serving 1, and 2 is in the same position as 1 would have been in had 1 chosen action S:

$$E[t_2(S, \gamma(2))|\eta] = \left(\frac{1}{\mu} + \frac{3}{2\mu}\right) = \frac{5}{2\mu}. \tag{7.13}$$

On the other hand, if server b finishes serving 0 before a serves -1 ($\bar{\eta}$ such that $P(\bar{\eta}) = 1 - \eta$), then the expected waiting time ($E[t_2(S, \gamma(2))|\bar{\eta}]$) is equal to that customer 2 would have experienced had he chosen D anyway, obtained from (7.8) when $k = 1$:

$$E[t_2(S, \gamma(2))|\bar{\eta}] = \frac{2}{\mu}. \tag{7.14}$$

Then the expected waiting time for customer 2 of choosing action S when customer 1 has chosen D is the mean of $E[t_2(S, \gamma(2))|\eta]$ and $E[t_2(S, \gamma(2))|\bar{\eta}]$, weighted by the

[15] E.g., $i = 3$ has two customers ahead of him regardless of which queue he chooses. Hence his $k = 3/2 + 1/2 = 2$.

[16] E.g., $i = 2$ faces one queue with two customers and another with one. He chooses the latter, so that there is one customer ahead of him, and $k = 2/2 = 1$.

[17] This can be done without loss of generality, as it occurred when the second customer chose S, the third would move to the shorter queue and the second would be forced to change his choice to D to preempt the third.

[18] Because the utility and cost functions are linear on waiting time, and customers are risk neutral, the waiting times from before and after the service can be added.

[19] The memoryless property of the exponential distribution implies that this is independent of elapsed time.

probability $P[\eta]$:

$$E[t_2(S,\gamma(2))] = P[\eta]E[t_2(S,\gamma(2))|\eta] + P[\bar{\eta}]E[t_2(S,\gamma(2))|\bar{\eta}]. \tag{7.15}$$

As stated above, $E[t_2(S,\gamma(2))|\bar{\eta}]$ is the same as the expected waiting time of choosing action D ($E[t_2(D,\gamma(2,1)]$):

$$E[t_2(D,\gamma(2,1)] = E[t_2(S,\gamma(2))|\bar{\eta}] = \frac{2}{\mu}. \tag{7.16}$$

Meanwhile, $E[t_2(S,\gamma(2))|\eta]$ is greater than the expected waiting time of choosing action D ($E[t_2(D,\gamma(2,1)]$):

$$E[t_2(S,\gamma(2))|\eta] > E[t_2(D,\gamma(2,1))] \Leftrightarrow \frac{5}{2\mu} > \frac{2}{\mu}. \tag{7.17}$$

Taken together, (7.16) and (7.17) imply that for customer 2, the expected waiting time of choosing S when 1 has chosen D is greater than that of choosing D:

$$E[t_2(S,\gamma(2))] > E[t_2(D,\gamma(2,1))]. \tag{7.18}$$

□

As stated in Definition 2, as long as customers 1 and 2 are each queueing behind one customer being served, then the system is in a two queue state. Note further that while customers arriving at a system in a single queue state can change it to two queues by choosing action D and triggering stage 3 of the game, the reverse is not possible: there is no mechanism for changing the system state from two queues to one, short of the queue length falling to 1. This implies that regardless of whether the system is at state $Q = 1$ or $Q = 2$, arrivals will always get the same expected waiting time from choosing D, as if they do so on a system in a single queue state, the system will change to a two queue state before they can overtake the incumbents: $E[t_i(D,\gamma(1,k))] = E[t_i(D,\gamma(2,k))]$, $\forall\, k \geq 2$.

7.2.3 Customers' Actions and Equilibria

Customers' preferred strategy will be comprised of the actions yielding the shorter expected waiting time; if waiting times are equal, the convention is adopted that ties will be broken in favour of action S, the single queue.

In the following discussion, comparisons between expected waiting times will be performed using the values obtained from equations (7.10)-(7.12), using the i index. The number used to perform the transformations in equations (7.10)-(7.12) is the i in t_i; i_o customers are covered first, as the first customer has an odd index.

Proposition 5. All customers i_o choose action S.

Proof. For any customer i_o, expected waiting time is strictly smaller when choosing action S than it is when choosing action D:

$$E[t_{i_o}(D, \gamma(2))] = \frac{1}{\mu}\left(\frac{i+3}{2}\right) > E[t_{i_o}(S, \gamma(1))] = \frac{1}{2\mu}(i+2). \tag{7.19}$$

□

Customers make their decision with full knowledge (obtained through backwards induction) of what subsequent arrivals will decide. Proposition 6 below makes clear that i_e customers choose action S in stage 2 of the game, if they encounter a system in a one queue state. Since stage 3 of the game is not triggered, expected waiting time calculations can be performed without *explicitly* incorporating future arrivals' actions.

Proposition 6. All customers i_e are indifferent between actions S and D.

Proof. For any customer i_e, expected waiting times are the same when taking either action S or D:

$$E[t_{i_e}(D, \gamma(2))] = \frac{1}{\mu}\left(\frac{i+2}{2}\right) = E[t_{i_e}(S, \gamma(1))] = \frac{1}{2\mu}(i+2) \tag{7.20}$$

□

As ties are broken by choosing action S, i_e customers will do so, cooperating with the i_o customers in taking the system into a single queue state.

As no customer has an incentive to deviate and take the system away from the single queue state by choosing D, the single queue state is a subgame perfect equilibrium. The game effectively leaves the decision of queue number to the first customer, 1, who strictly prefers a single queue, and gets to implement it before any of the even numbered customers, who are indifferent, choose their action. Once this single queue state exists, there is no incentive for any arrivals to deviate from it.

Note however that as i_e customers are indifferent between the two states, were one of them to choose D instead of S by breaking ties the other way, then stage 3 of the game would be triggered, and incumbent customers would change the system to a two queue state. From lemma 3, it's clear this would also be a stable equilibrium. The single queue equilibrium is therefore not unique. This fragility may be a reason for the emergence of multiple queues.

7.2.4 Relaxing the No-Jockeying Condition

As shown in Smith and Whitt (1981), multiple queues are inefficient because, when customers cannot change queues after committing to them (i.e., jockeying is not allowed), servers can be idle while there are customers waiting to be served. Then, if customers are allowed to switch queues, their expected waiting times under the multi-

ple queue state are shorter. Accordingly, if the no-jockeying condition is relaxed, one would expect the equilibrium in the foregoing game to change.

This subsection looks at the results of relaxing the no-jockeying condition. Allowing unfettered jockeying would make the problem intractable. Therefore, customers will be allowed to jockey if and only if the system is in a two queue state, and the server in the other queue is idle, that is, the other queue's length becomes zero (0). Let this occurrence be denoted by event B, the ex-ante probability of this event happening to a customer joining queue be $P[B]$, and that of it not happening ($\bar{B}$) be $P[\bar{B}] = 1 - P[B]$. Further, let n be number of customers ahead of a customer $n+1$ in the queue when event B occurs, with n_o and n_e denoting an odd or even value of n, respectively.

The new game (game J), then, follows the stages:

1. Nature assigns customers an arrival number $i \in \{1, ..., \infty\}$.

2. Customer i arrives at the system, and chooses from action set $A = \{S, D\}$. This choice can be discerned by any incumbent customers with perfect accuracy. If he chooses $D \in A$, he also chooses from action set $\Omega = \{\omega, \bar{\omega}\}$, where action ω is switching queues if and when that becomes possible (i.e., when a server becomes idle), and $\bar{\omega}$ is to stay in the same queue. This choice is made for two possible circumstances: when n is odd, and when it is even; there is no requirement that the choice be the same in the two scenarios. The chosen action is not performed until stage 4. Choosing S will not allow the customer closer access to the servers than the incumbents who chose D. If there are more than two customers waiting and the system is in a two queue state, customers must choose D and the round terminates.

3. This stage only occurs if customer $i > 1$ encounters a single queue state and chooses action D. In that case, incumbent customers split the single queue into two 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. Customer i acts upon his choice in stage 2. He cannot change his decision to react to incumbents' moves in stage 3, if these occurred.

   (a) If at any point when the system is in a two queue state a server becomes idle, customers act on their choice from Ω, according to whether their n is odd or even; customers closer to the server move first.

5. Stages 2 to 4 (including 4.a) are repeated for customer $i+1$, with i now being an incumbent, and so on for all future arrivals.

Therefore, in this version of the game the customer's strategy is expanded to include a choice from set Ω: $\Sigma_i = \{\alpha, o\}$, where $o \in \Omega$.

Lemma 4. If, when the system is in a two queue state, one of the servers (say b) becomes idle while the other one (a) is busy (i.e., stage 4.a of the game is triggered), customers with an odd n in a's queue will switch to b's queue (see Figure 7.1 for an example of this process[20]).

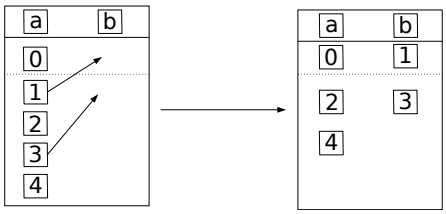

Figure 7.1: Server b becomes free, customers with an odd n change queues.

Proof. Let l be the number of customers in the queue the customer can switch to, and m the number of customers in the current queue.[21] In a similar manner to n, let l_o and l_e denote an odd or even value of l, respectively, and m_o and m_i the same for m. Further, let $E[\tau(\{D,\omega\},\gamma(2),l]$ and $E[\tau(\{D,\bar{\omega}\},\gamma(2),m)]$ be the expected waiting times τ in the queue the customer can switch to, and in the current queue, respectively, *at the time the server becomes idle.*

At stage 4.a, customers decide whether to change queues when a server becomes idle, but it's only when that happens that they can know their expected waiting times. These can be derived from (7.8), as a function of customers' choices from Ω, simply by replacing k with l when the customer switches queues, and replacing k with m when the customer does not switch.

For n_o customers, $l_o = (n-1)/2$ and $m_o = (n+1)/2$;[22] for n_e customers, $l_e = n/2$ and $m_e = n/2$.[23] Then, for n_o customers:

$$E[\tau(\{D,\omega\},\gamma(2),n_o)] = \frac{1}{\mu}(l_o+1) < E[\tau(\{D,\bar{\omega}\},\gamma(2),n_o)] = \frac{1}{\mu}(m_o+1), \qquad (7.21)$$

as $l_o < m_o$. Therefore all n_o customers are better off choosing ω and switching queues. On the other hand, for n_e customers:

$$E[\tau(\{D,\omega\},\gamma(2),n_e)] = \frac{1}{\mu}(l_e+1) = E[\tau(\{D,\bar{\omega}\},\gamma(2),n_e)] = \frac{1}{\mu}(m_e+1), \qquad (7.22)$$

as $l_e = m_e$. Therefore n_e customers have no reason to switch queues: they face the same number of customers in the queue they are at and in the queue they can switch to (because some i_o customer will always have switched just before), so they choose $\bar{\omega}$.

[20] Recall that customers can only change queues when one server is *idle*.

[21] E.g., for a customer queueing behind 3 others ($n = 3$), $n = 0$ is being served, $n = 1$ switches queues, and $n = 2$ stays in place. Then $l = 1$ and $m = 2$ for $n = 3$.

[22] E.g., for $n = 3$, $n = 0$ is being served, $n = 1$ switches queues, and $n = 2$ stays in place. Then $l = (3-1)/2 = 1$ and $m = (3+1)/2 = 2$ for $n = 3$.

[23] E.g., for $n = 4$, $n = 0$ is being served, $n = 1$ and $n = 3$ switch queues, and the third $n = 2$ stays in place. Then $l = 4/2 = 2$ and $m = 4/2 = 2$ for $n = 4$.

However, their expected waiting time is still reduced compared to game NJ, as they end up queueing behind half the customers they were previously waiting behind. □

The foregoing is clearly illustrated in Figure 7.1: customers 1 and 3 switch queues to the idle server's side, while customers 2 and 4 stay in their original queue. The reduction in the latter's waiting time is caused only by the reduction in their queue length, as there is no incentive for 2 or 4 to switch queues.

Given Lemma 4, the choice from Ω will be dropped from the notation in what follows.

When jockeying is possible, the expected waiting time for a customer in a system in a two queue state depends on whether a server becomes idle. If that does not happen, it is exactly identical to that in game NJ. However, if a server does become idle, the expected waiting time is shorter. Expected waiting time will thus depend on the probability of a server becoming idle. Let $E[t_i(D,\gamma(2))|B]$ be the expected waiting time in a system in a two queue state if a switch occurs, and $E[t_i(D,\gamma(2))|\bar{B}]$ that when a switch does not occur. The expected waiting time for a customer i in a system in a two queue state is then given by:

$$E[t_i(D,\gamma(2))] = P(\bar{B})E[t_i(D,\gamma(2))|\bar{B}] + P(B)E[t_i(D,\gamma(2))|B]. \tag{7.23}$$

The plurality of equilibria in game NJ game was due to i_e customers being indifferent between the two system states. Relaxing the no-jockeying condition eliminates this indifference, as the small probability of jockeying occurring and thereby reducing waiting time is enough to reduce the expected waiting time of the two queue state, which becomes the unique equilibrium, being strictly preferred by i_e customers.

Proposition 7. All customers i_e strictly prefer action D to S at stage 2.

Proof. Proposition 7 can be proven without obtaining a solution for the complete expected waiting time: it is only if a server becomes idle that game J differs from game NJ. Expected waiting time conditional on no server becoming idle is then the same as if no jockeying was allowed, i.e., $E[t_{i_e}(D,\gamma(2))|\bar{B}]$ from (7.23) (for an i_e customer) is identical to $E[t_{i_e}(D,\gamma(2))]$ from (7.12):

$$E[t_{i_e}(D,\gamma(2))|\bar{B}] = \frac{1}{\mu}\left(\frac{i+2}{2}\right). \tag{7.24}$$

On the other hand, if a server does become idle and the queue splits, the expected waiting time is:

$$E[t_{i_e}(D,\gamma(2))|B] = E[t_{i_e}(D,\gamma(2))|b] + E[\tau_i(o,\gamma(2))], \tag{7.25}$$

where $E[t_{i_e}(D,\gamma(2))|b]$ is the expected waiting time between joining the queue and the other server becoming idle, and $E[\tau_i(o,\gamma(2))]$ the expected waiting time between the

other server becoming idle (and customers switching queues) and the customer being served.

Crucially, $E[t_{i_e}(D,\gamma(2))|b]$ is identical across $E[t_{i_e}(D,\gamma(2))|\bar{B}]$ and $E[t_{i_e}(D,\gamma(2))|B]$, so the only difference is the expected waiting time after the other server becomes idle. It has been noted in lemma 4 that $E[\tau_i(o,\gamma(2))]$ is smaller than it would be if no switch was allowed, which implies the total expected waiting time is also smaller:

$$E[t_{i_e}(D,\gamma(2))|B] < E[t_{i_e}(D,\gamma(2))|\bar{B}]. \tag{7.26}$$

When switching was not allowed, expected waiting time for i_e customers was equal regardless of choosing action S or D ($E[t_{i_e}(S,\gamma(1))] = E[t_{i_e}(D,\gamma(2))]$). If switching is allowed, taking into account (7.23), expected waiting times are still identical if no server becomes idle ($E[t_{i_e}(D,\gamma(2))|\bar{B}] = E[t_{i_e}(S,\gamma(1))]$), but if a server does become idle they are smaller ($E[t_{i_e}(D,\gamma(2))|B] < E[t_{i_e}(S,\gamma(1))]$); then for any positive value of $P(B)$, the expected waiting time when choosing D is smaller than it would be if jockeying was not allowed, and hence smaller than the expected waiting time when choosing S: $E[t_{i_e}(S,\gamma(1))] > E[t_{i_e}(D,\gamma(2))]$. Therefore, i_e customers strictly prefer action D to action S. □

Proposition 8. If jockeying is allowed, then the two queue state is the unique subgame perfect equilibrium of the game.

Proof. In stage 2 of the game, i_o customers know by backwards induction the result from Proposition 7 that i_e customers strictly prefer action D, even given the incumbent response in stage 3 of splitting the queue. Therefore, they choose action D at stage 2, as they know the single queue state is unsustainable.[24] □

Therefore, even allowing the limited form of jockeying described above destroys the single queue equilibrium when customers are risk neutral. The two queue state is the game's unique subgame perfect equilibrium.

7.3 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, and therefore take the variance of waiting time into account, the single queue state will be strictly preferred by all customers, and thus be the unique subgame

[24] Because the assumption is being made that 2 arrives before anyone is served, customer 1 can take this decision at stage 2. Even if this simplifying assumption were not made, however, the only way the system could be in a one queue state was if there was only one customer queueing. Once the second customer arrived, 1 would always preempt him by changing the state to two queues, and therefore the assumption makes no difference for steady state equilibrium outcomes as long as the expected steady state queue length is greater than 1.

perfect equilibrium. The intuition behind this result is that the variance of waiting time is larger for the two queue state than for a single queue, so risk averse customers naturally prefer the latter.

The analysis will again start with a strict no-jockeying condition, which will be relaxed in section 7.3.4, i.e., games NJ and J are considered in turn. Subgame perfection is still the relevant equilibrium concept.

Customer risk aversion is reflected through an exponential utility function, which replaces the linear utility given at (7.1):

$$U_i(t_i) = 1 - \exp(-v(R - t)),\ v \in (0,\ \mu), \tag{7.27}$$

where v is a positive constant representing the degree of risk aversion, R is the net monetary value of the service, and t is the waiting time. The unit cost of waiting time has been normalized to 1, without loss of generality. Expected utility $U_i(t_i)$ is not defined for $v \geq \mu$, because $\lim_{v\to\mu} U_i = -\infty$, so that if $v \geq \mu$, customers would not join the queue for any values of k and R. As this applies equally to both system states, $v \in (0,\ \mu)$ will be assumed throughout the rest of the section. The exponential form was chosen for the utility function as it is widely employed in many applications and can readily capture varying dimensions of risk aversion through the parameter v.

7.3.1 Expected Utility

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 utilities must be compared instead:

$$E[U_i(t_i)] = \int_0^\infty (1 - \exp(-v(R - t)))\, z(t,\ k)\, \mathrm{d}t, \tag{7.28}$$

where $z(t,\ k)$ is the probability distribution function of waiting time.

Then for a system in a two queue state, with the queue number taken as given, the expected utility can be obtained by replacing $z(t,\ k)$ with (7.6):

$$E[U_i(t_i(\gamma(2, k)))] = 1 - \exp(-vR)\left(\frac{\mu}{\mu - v}\right)^{k+1}. \tag{7.29}$$

Likewise, for a system in a single queue state, the expected utility is obtained by replacing $z(t,\ k)$ with (7.7):

$$E[U_i(t_i(\gamma(1, k))] = 1 - \frac{\exp(-vR)\left(1 - \frac{v}{2\mu}\right)^{-k}(v - 2\mu)}{2(v - \mu)}. \tag{7.30}$$

As for the risk neutral case, it is convenient to have expected utility as a function

of arrival order i rather than k, enabling direct comparisons between system states. This is achieved with the same transformations as outlined before, only now they are applied to the utility functions given in (7.29) and (7.30). Therefore, for customers in a system in a single queue state, expected utility as a function of i is:

$$E[U_i(t_i(\gamma(1)))] = 1 - \frac{\exp(-vR)\left(1 - \frac{v}{2\mu}\right)^{-(i+1)}(v - 2\mu)}{2(v - \mu)}. \tag{7.31}$$

For i_o customers in a system in a two queue state, it is:

$$E[U_{i_o}(t_{i_o}(\gamma(2)))] = 1 - \exp(-vR)\left(\frac{\mu}{\mu - v}\right)^{\frac{i+3}{2}}. \tag{7.32}$$

Finally, for i_e customers in a system in a two queue state:

$$E[U_{i_e}(t_{i_e}(\gamma(2)))] = 1 - \exp(-vR)\left(\frac{\mu}{\mu - v}\right)^{\frac{i+2}{2}}. \tag{7.33}$$

7.3.2 Customer Behaviour

The next lemma describes the behaviour of risk averse customers, mirroring that presented for risk neutral customers in lemma 3.

Lemma 5. If customer 1 chooses action D, then customer 2 will also choose action D.

Proof. Say customer 1 chose action D, queueing for server a. If 2 also chooses D, he will queue for server b, in which case $k = 1$.

Recall $P[\eta]$ has been defined to be the probability of server a finishing a service before server b. As in the proof of lemma 3, the expected utility is the mean of the expected utility if server a finishes first (η), and the expected utility if server b is the one to finish first ($\bar{\eta}$), weighted by the probability $P[\eta]$. However, unlike the waiting times in lemma 3, expected utilities are not necessarily additive, and so need to be derived from the distribution functions of waiting time.

If η occurs, expected utility ($E[U_2(t_2(S, \gamma(2)))|\eta]$) is the expected utility from the expected time for a to finish serving -1, and from that point on the expected utility for action S (from (7.30), where $k = 2$). This is because, when a finishes serving -1 and starts serving 1, 2 is in the same position as 1 would have been had he chosen action S. Thus let $\zeta(T, k)$ be the distribution for total waiting time T if η occurs:

$$\zeta(T, k) = \int_0^T z_k(T - t)\mu \exp(-\mu t)\, \mathrm{d}t, \tag{7.34}$$

where $\zeta(T, k)$ comes from (7.5), the distribution for waiting time when the system is in a one queue state, i.e., for action S, and $\mu \exp(-\mu t)$ is the probability distribution function for service time.

The relevant form of $z(t, k)$ is $z(t, 2)$ for a system in a single queue state, i.e. from

(7.7), so that the relevant distribution of waiting time is:

$$\zeta(T,\,2) = \int_0^T z(T-t,\,2)\mu\exp(-\mu t)\,\mathrm{d}t = 2\mu((\mu T-1)\exp(-\mu T)+\exp(-2\mu T)). \quad (7.35)$$

With $\zeta_2(T,\,2)$ in hand, the expected utility when η occurs can then be derived from (7.28) (note that this is smaller than the expected utility of choosing action D ($E[U_2(t_2(D,\gamma(2,1)))]$):

$$\begin{aligned} &E[U_2(t_2(S,\gamma(2)))|\eta] = \int_0^\infty (1-\exp(-v(R-T)))\zeta(T,\,2)\,\mathrm{d}T = 1+\frac{2\exp(-vR)\mu^3}{(v-2\mu)(v-\mu)^2} \\ &E[U_2(t_2(S,\gamma(2)))|\eta] < E[U_2(t_2(D,\gamma(2,1)))]. \end{aligned} \quad (7.36)$$

On the other hand, if server b finishes serving 0 before a serves -1 ($\bar{\eta}$), then the expected utility ($E[U_2(t_2(S,\gamma(2)))|\bar{\eta}]$) is simply equal to the expected utility customer 2 would have experienced had he chosen D anyway (obtained from (7.29) when $k=1$):

$$E[U_2(t_2(S,\gamma(2)))|\bar{\eta}] = E[U_2(t_2(D,\gamma(2,1)))] = 1-\exp(-vR)\left(\frac{\mu}{\mu-v}\right)^2. \quad (7.37)$$

The expected utility for customer 2 of choosing action S when customer 1 has chosen D is then:

$$E[U_2(t_2(S,\gamma(2)))] = P[\eta]E[U_2(t_2(S,\gamma(2)))|\eta] + P[\bar{\eta}]E[U_2(t_2(S,\gamma(2)))|\bar{\eta}].$$

As $E[U_2(t_2(S,\gamma(2)))|\bar{\eta}]$ is the same as the expected utility of choosing action D (eq. (7.37)), and $E[U_2(t_2(S,\gamma(2)))|\eta]$ is lower than the expected utility of choosing action D (eq. (7.36)), the expected utility for 2 of choosing S when 1 has chosen D is lower than that of choosing D:

$$E[U_2(t_2(D,\gamma(2,1)))] > E[U_2(t_2(S,\gamma(2)))]. \quad (7.38)$$

□

As in the risk neutral scenario, if the first two customers settle on a two queue state, future arrivals choose D and the system stays in that state (see definition 2 and the formal description of the game steps), at least until the queue is next cleared of waiting customers.

7.3.3 Customers' Actions and Equilibria

As before, it is assumed that indifferent customers will break ties in favour of action S.

Proposition 9. All customers i_o choose action S.

Proof. For any customer i_o to choose S over D, it must be the case that:

$$\begin{aligned} E[U_{i_o}(t_{i_o}(S,\gamma(1)))] &= 1 - \frac{\exp(-aR)\left(1-\frac{a}{2\mu}\right)^{-(i+1)}(a-2\mu)}{2(a-\mu)} \geq \\ E[U_{i_o}(t_{i_o}(D,\gamma(2)))] &= 1 - \exp(-aR)\left(\frac{\mu}{\mu-a}\right)^{\frac{i+3}{2}}, \end{aligned} \tag{7.39}$$

where the left hand side is obtained from (7.31) and the right hand side from (7.32).

The inequality at (7.39) can be reduced to

$$\theta(v,\mu) \equiv \left(1-\frac{v}{2\mu}\right)^{i}\left(\frac{\mu}{\mu-v}\right)^{\frac{i+1}{2}} \geq 1. \tag{7.40}$$

$\theta(v,\mu)$ is increasing in v:

$$\frac{\partial\theta(v,\mu)}{\partial v} = \frac{\left(1-\frac{v}{2\mu}\right)^{i}\left(\frac{\mu}{\mu-v}\right)^{\frac{i+1}{2}}(v(i-1)+2\mu)}{2(v-2\mu)(v-\mu)} > 0, \tag{7.41}$$

as under the specified conditions, all the terms in the numerator are positive, and two terms in the denominator are negative; further, $\theta(0,\mu) = 1$, hence (7.40) holds for all positive values of v. □

As before, customer decisions are made with full knowledge of future arrivals' decisions, as outlined in Propositions 9 and 10. Since these imply that stage 3 of the game is not triggered, customers can calculate their expected utility, incorporating future arrivals' decisions.

Proposition 10. All customers i_e choose action S.

Proof. For any customer i_e to choose S over D, it must be the case that:

$$\begin{aligned} E[U_{i_e}(t_{i_e}(S,\gamma(1)))] &= 1 - \frac{\exp(-vR)\left(1-\frac{v}{2\mu}\right)^{-(i+1)}(v-2\mu)}{2(v-\mu)} \geq \\ E[U_{i_e}(t_{i_e}(D,\gamma(2)))] &= 1 - \exp(-vR)\left(\frac{\mu}{\mu-v}\right)^{\frac{i+2}{2}}, \end{aligned} \tag{7.42}$$

where the left hand side is obtained from (7.31) and the right hand side from (7.33).

The inequality at (7.42) can be reduced to

$$\nu(v,\mu) \equiv \left(1-\frac{v}{2\mu}\right)^{i}\left(\frac{\mu}{\mu-v}\right)^{\frac{i}{2}} \geq 1. \tag{7.43}$$

$\nu(v,\mu)$ is increasing in v:

$$\frac{\partial\nu(v,\mu)}{\partial v} = \frac{vi\left(1-\frac{v}{2\mu}\right)^{i}\left(\frac{\mu}{\mu-v}\right)^{i/2}}{2(v-2\mu)(v-\mu)} > 0, \tag{7.44}$$

as under the specified conditions, all the terms in the numerator are positive, and two terms in the denominator are negative; further, $\nu(0,\mu) = 1$, hence (7.43) holds and the proposition is proven. □

As there is no incentive for any customers to deviate and take the system away from the single queue state by choosing D, the single queue state is the unique subgame perfect equilibrium of the game for this type of risk averse customers. Uniqueness follows from all customers *strictly* preferring a single queue state/choosing action S.

7.3.4 Relaxing the No-Jockeying Condition

As with the risk neutral case, it is important to investigate whether there are any changes to the equilibrium when the no-jockeying condition is relaxed. This is done by considering the Game J from subsection 7.2.4, but with risk averse customers.

Lemma 6. If, when the system is in a two queue state, one of the servers (say b) becomes idle while the other one (a) is busy (i.e., stage 4.a of the game is triggered), customers with an odd n in a's queue will switch to b's queue.

Proof. This proof uses the notation introduced in the proof of lemma 4. The decision whether to switch queues at stage 4.a, if that stage is triggered, is taken by comparing expected utilities at the point in time a server becomes idle, as a function of expected waiting times at that point, τ, which can be obtained from (7.29) by replacing k with l and m.

As for the risk neutral case, all n_o customers are better off choosing ω and switching queues:

$$\begin{aligned} E[U(\tau,(\{D,\omega\},\gamma(2),n_o))] &= 1-\exp(-vR)\left(\frac{\mu}{\mu-v}\right)^{l_o+1} > \\ E[U(\tau,(\{D,\bar{\omega}\},\gamma(2),n_o))] &= 1-\exp(-vR)\left(\frac{\mu}{\mu-v}\right)^{m_o+1}. \end{aligned} \tag{7.45}$$

Similarly, n_e customers still have no reason to switch queues, even though their expected utility is higher than prior to the switch-event:

$$\begin{aligned} E[U(\tau,(\{D,\omega\},\gamma(2),n_e))] &= 1-\exp(-vR)\left(\frac{\mu}{\mu-v}\right)^{l_e+1} = \\ E[U(\tau,(\{D,\bar{\omega}\},\gamma(2),n_e))] &= 1-\exp(-vR)\left(\frac{\mu}{\mu-v}\right)^{m_e+1}. \end{aligned} \tag{7.46}$$

□

In what follows, the choice from Ω will be dropped from the notation, except where directly relevant; observance of Lemma 6 will be assumed.

Let $E[U_i(t_i(D,\gamma(2,k)))|\bar{B}]$ be the expected utility in a system in a two queue state if no switch occurs, and $E[U_i(t_i(D,\gamma(2,k)))|B]$ that when a switch *does* occur. The

expected utility for a customer i in a system in a two queue state, which depends on $P(B)$, is then given by:

$$E[U_i(t_i(D,\gamma(2,k)))] = P(\bar{B}))E[U(t_i(D,\gamma(2,k)))|\bar{B}] + P(B)E[U(t_i(D,\gamma(2,k)))|B]. \tag{7.47}$$

As the general case is extremely complex, only the limit case, where the jockeying opportunity happens immediately after the customer has joined the queue, will be considered below. In this special case, expected utility at joining is:

$$E[U_i(t_i(D,\gamma(2,k)))|B] = E[U(t_i(D,\gamma(2,k)))|b] + E[U(\tau_i(\{o,D\},\gamma(2,k)))], \tag{7.48}$$

where $E[U_i(t_i(D,\gamma(2,k)))|b] = 0$. Therefore $E[U_i(t_i(D,\gamma(2,k)))|B] = E[U_i(\tau_i(\{o,D\},\gamma(2,k)))]$.

The customer at the front, 1, has the most to gain from switching, as he will get served immediately. His utility in the event of switching is:

$$E[U_1(t_1(D,\gamma(2,0)))|B] = 1 - \exp(-aR)\left(\frac{\mu}{\mu - a}\right)^1. \tag{7.49}$$

Then, for the limit case of customer 1, it is possible to obtain the values of the parameters for which the customer would still choose action S, which follow from the condition below:

$$\begin{aligned} E[U_1(t_1(S,\gamma(1)))] > & P(\bar{B})E[U(t_1(D,\gamma(2,1)))|\bar{B}] + \\ & P(B)E[U_1(t_1(D,\gamma(2,0)))|B]. \end{aligned} \tag{7.50}$$

This yields the following parameter conditions:

- $P(B) < \frac{\mu}{2\mu - v}$;
- $v > \frac{\mu(2P(B)-1)}{P(B)}$;

This limit case indicates that for risk averse customers who can jockey, a single queue state remains the subgame perfect equilibrium for high levels of risk aversion, while for low levels of risk aversion, a two queue state becomes the equilibrium instead. However, as the foregoing analysis was dependent on there being no time between the customer's arrival and the switching event, this cannot be stated in more general terms—doing so does not seem easily tractable, and must remain a conjecture for the present.

7.4 Discussion and Conclusion

This chapter has shown that risk neutral customers derive a small benefit from combining queues. Without jockeying, half of them benefit compared to a system with two

queues, whereas the other half is indifferent between the two situations. This causes the single queue state to be a non-unique equilibrium, and if jockeying is possible the two queue state is the only equilibrium outcome.

As seen from Smith and Whitt (1981), waiting time inefficiencies in multiple queue systems are caused by servers being idle when they could be serving another customer. Allowing jockeying, even in the very limited form used in the present model, eliminates this issue, as a server does not sit idle if his queue becomes empty. It is then not surprising that when customers are risk neutral, this situation leads to a two queue state.

On the other hand, when customers are risk averse, an added source of disutility arises, the increased variance in waiting time. A two queue state requires customers to bet on which queue is going to move faster. Obviously, since the two service distributions are i.i.d., this introduces risk. 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 both servers eliminates the risk inherent in having to choose a queue. Even introducing jockeying into the game only has a limited effect: the two queue state is only an equilibrium for weakly risk averse customers. Examining in more detail the circumstances in which the single queue equilibrium breaks down under risk aversion and jockeying is an inviting topic for further research.

These results have implications for service station management. The welfare effects of non-monetary costs such as waiting time sometimes escape notice, but they negatively impact customer utility as much as price—though unlike price, they benefit no-one. There is then great scope for improving social welfare by reducing these costs.

The present chapter has shown that risk averse customers, have the most to lose from a plurality of queues, and will, in equilibrium, form a single queue when presented with two servers. It seems a reasonable assumption that customers are at least somewhat risk averse, yet combining queues is often frowned upon by managers. This work provides a counterpoint to the views expressed in Rothkopf and Rech (1987).

This does leave open the question of why it is often observed that customers form two queues even where there is no pressure from management to do so. Further research should 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 one queue equilibrium state (i.e., the observance of lemmas 4 and 6). 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, characterize more rigorously the switching conditions for

game J with risk averse customers, generalizing the problem to any number of servers, 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 Kohei Kawamura and Tim Worrall for the helpful discussions and comments, and for their reading of various drafts. 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 this chapter was written.

Chapter 8

Cutting Queues: Customer and System Behaviour in a Repeated Game

8.1 Introduction

The present paper is placed at the intersection of two strands of literature.[1] It seeks to determine whether there are any circumstances where homogeneous customers facing an *M/M/1* queue system with what at first looks to be a First Come First Served (FCFS) service discipline, allow arriving customer possible to overtake incumbents. This will be shown to be possible for a subset of customers in a setting with repeated interactions, involving the creation of a queue-within-a-queue employing the Last Come First Served (LCFS) discipline. This makes customers' sojourn times into a function of more than on queue length at arrival, and service rate: it is also affected by the rate of future arrivals. The joining threshold threshold problem is then solved for these customers.

An individual threshold strategy for FCFS queues was determined in Naor (1969), which also obtained the social welfare maximizing threshold; the former was found to be larger than the latter, so that customers joining a queue impose a negative externality on all others. The recommended prescription was charging a 'toll' to joining customers, which could reduce the Nash equilibrium threshold to the socially optimal one. These findings were generalized in Knudsen (1972) to general waiting and inter-arrival time distributions, and an arbitrary number of servers.

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 monograph by Hassin and Haviv (2003). Since then, many more papers than can be individually mentioned have been published on this subject. A few notable examples are Burnetas and Economou (2007), Boudali and Economou (2012), Sun et al. (2009) and Sun and Li (2014). Ventures beyond the FCFS discipline were rarer, as other disciplines tend to be more mathematically intractable. However, a couple of important exceptions include Hassin and Haviv (1997) and Erlichman and Hassin (2009). These cover an FCFS M/M/1 queue where, similarly to the present paper, a customer can overtake others. However, unlike the present model, this overtaking does not emerge from interactions among customers in a repeated games setting: rather customers pay the server to be allowed to cut ahead; they also do not examine customer sojourn time as the present work does. The present analysis rather draws on that performed by Yu et al. (2014) for the EPS discipline, who compute expected sojourn time for customers joining an EPS queue, using a method which is followed, *mutatis mutandis*, in the present paper, to determine the joining threshold.

As mentioned in the foregoing paragraph, one of the distinctions between the present work and Hassin and Haviv (1997) and its progeny is that the possibility of overtaking

[1]Thanks to Tim Worrall for the many helpful discussions, and to Jonathan Thomas and Paul Schweinzer, my *viva* examiners, for the helpful comments and suggestions. Any remaining errors are, of course, my own. I also thank the University of Edinburgh for funding my PhD, of which the present paper forms a part.

emerges from repeated interactions between customers rather than being part of the set of rules established by the service station manager. This leads into the discussion of the second strain of literature the present paper draws from. Queues can be classified among what Parsons (1955) described 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 by Okuno-Fujiwara and Postlewaite (1995) (and see Mailath and Samuelson (2006) 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. A few noteworthy studies can be mentioned: following on Schwartz (1975), which laid out a sociological analysis of waiting for service and customers' perceptions of the fairness of queueing disciplines, Milgram et al. (1986) described an experiment where people tried to cut ahead of several queues. While these overtaking attempts were sometimes successful, they generally met with failure. On this vein, Larson (1987) characterized deviations from FCFS as unjust, which would explain why reactions to overtaking attempts were so overwhelmingly negative. These findings were reinforced by a study of reactions to deviations from FCFS in an overnight queue for a U2 concert, in Helweg-Larsen and LoMonaco (2008), which found negative reactions even when there was little impact on outcomes. However, Oberholzer-Gee (2006) describes an experiment where customers in queues were offered payment in exchange for letting a stranger cut in line; it was found that the likelihood of the offer being accepted increased with the payment, which would favour the conclusion that self interest is the dominant factor governing behaviour in queues. This lead to Allon and Hanany (2012), which presents a model of a queue where customers with different priorities (which are reassigned after each round of the game) interact repeatedly, and concluded that sufficiently patient customers will allow those with high priority to cut ahead, as customers take into account the possibility of being allowed to cut ahead in the future, when they might have high priority.

The present paper begins by setting up a repeated game scenario which is heavily indebted to Allon and Hanany (2012), although it deals with homogeneous customers, a subset of which is known to each other. The single shot equilibrium is presented, and then a grim-trigger strategy where customers allow overtaking attempts from their 'acquaintances' in the repeated-game setting is set forth. It is shown that this strategy, where customers allow some attempts to overtake them, is an equilibrium for sufficiently patient customers, and further that queue length-dependent strategies will not affect this equilibrium (again, for sufficiently patient customers).

Once the repeated game equilibrium is established, it will be seen to create a queue-within-a-queue which uses the LCFS discipline. The method for deriving customer

sojourn time in this sub-queue will be discussed; it will be seen that it can be obtained using a method similar to that outlined in Yu et al. (2014) for the EPS discipline. Then the joining threshold (maximum length for which an arriving customer will join the queue) will be investigated. An overview of Naor (1969)'s findings regarding the socially optimal threshold for an M/M/1 queue follows, after which the sojourn times and joining thresholds obtained for the overtaking system presented here are contrasted with that for the FCFS discipline, and with the socially optimal value, through numerical simulations. While the form of the results prevents general conclusions from being drawn, numerical simulations seem to indicate that queue cutting reduces social welfare and should be discouraged. The conclusion, outlining avenues for further research, follows at the end.

8.2 A Repeated Game Model of Queue Cutting

This model is an adaptation of that presented in Allon and Hanany (2012), stripped of its element of customer heterogeneity and with other changes to make it address the problem at hand. In Allon and Hanany (2012), type heterogeneity was the driver for the sustainability of the queue cutting discipline: customers let others with high cost of time overtake them, in order to benefit from that possibility when their type was higher in the future. In the present model, customers are homogeneous in regards to their cost function. However, a subset of customers interacts repeatedly. Under these circumstances, the queue cutting discipline is still found to be an equilibrium in repeated games for sufficiently patient customers, showing heterogeneity is not required for this result. Nevertheless, as the numerical results in section 8.5 indicate, when customers are homogeneous this might have a negative effect on social welfare, unlike Allon and Hanany (2012), where it seems beneficial. Why is this important? The scenario may be thought to be unrealistic, but anyone who has gone to a sufficiently rowdy high school could readily attest otherwise—the lunch line problem. More generally, wherever enforcement of the FCFS discipline by management is lax, customer who are regular users of the service will have an incentive to 'collude' against occasional users through allowing other regulars to overtake them. The welfare issues alluded to suggest enforcement of FCFS by the system managers can be a welfare enhancing measure, so that this is a repeated game where it might improve social welfare to move to the single-shot equilibrium!

Consider a Poisson stream of identical, risk neutral utility maximizing customers, demanding a certain service from an M/M/1 queue; these customers have the utility function $U = \nu - cE[W]$, where ν is the service value, c the unit cost of time, and $E[W]$ the expected sojourn time. In each instance of the game, inter-arrival times are modelled by a Poisson process with rate λ. Customers are identical, and obtain value ν at the end of the service, incurring a waiting cost c per unit of sojourn time. A share of customers $\alpha \in (0, 1)$ expects to use the service repeatedly, while the remainder $1 - \alpha$ does not expect to use it again; denote the former by i_α and the latter by $i_{\bar{\alpha}}$.

Information about repeated user status is common knowledge *shared by all regular users*, but not for occasional users, who are unaware of this.

Repeated user status is assigned by nature before any instance of the game occurs. The single shot game is then comprised of the following steps, which approximate the game presented in Allon and Hanany (2012) without completely reproducing it:

1. Nature assigns customers their arrival order.

2. Customer arrives at the queue and observes the queue length (which for a customer N is $N-1$), and in the case of regular users, the positions of other regulars. Customers then decides whether to join the queue or balk. If they balk, the game ends for them.

3. If the customer joins the queue, they can make one attempt at cutting it: they will ask one "known" incumbent customer (and it is both trivial and intuitive that they will ask the regular incumbent who is closest to the server) to be allowed to cut ahead (action P).[2] Only the consent of the customer who is being overtaken is required, so that if one customer in the middle of the queue allows an arrival to cut ahead, all the customers behind the incumbent just have to 'lump it', something which is well within the purview of the high school lunch problem! This action is costless. The customer can also choose to join the end of the queue without trying to cut (action J).

4. The incumbent customer who received the cutting request can accept it (action A): in that case, if the incumbent is the nth customer in the queue, the arrival takes the nth place, and the incumbent is now the $n+1$th; customers behind them will likewise see their place in the queue increase by one. It is never optimal for a customer to accept some cutting attempts and reject others.[3] The incumbent can also reject the cutting attempt, in which case the arrival joins the queue at its end: action R. Regardless of where customers join the queue, they cannot leave it until they are served.[4]

The full strategy of customer i is then given by (E^i, I^i):

- $E^i \in \{J, P\}$,
- $I^i \in \{R, A\}$,

[2] Only the interactions of regular customers with each other are considered. It's obvious that occasional users of the system will reject all requests to cut ahead in all circumstances, and no requests by them will be accepted, so that there is no loss of generality in not considering them, and the benefits for elegance of presentation are significant.

[3] As all customers have identical cost functions, costs are linear in waiting time, and the increase in expected sojourn time from letting an arrival overtake is always the same (this takes into account the fact that future arrivals might let other customers cut ahead—see section 8.3.1). This accords with the setup in Allon and Hanany (2012).

[4] Note, however, that it is possible to set up a version of the game where customers are allowed to leave, the results of which are not too dissimilar qualitatively. The author hopes to set out this version of the game in future research.

where E^i describes the choices available to arrivals, and I^i those available to incumbents.

8.2.1 Single Shot Game

In the single stage game, only one equilibrium can be sustained: in this equilibrium, all requests to cut the line are rejected, and thus each arriving customer joins the queue at its end. The following theorem sets out this equilibrium formally; for the proof of this and subsequent theorems consult the Appendix.

Theorem 13. The strategy where customers choose $E^i = J$, $I^i = R$ is an equilibrium, which yielding a pure FCFS queue. There are no other equilibria for the single stage game.

Obviously, in the single shot game it makes no sense to consider whether customers are regular users or not, and there is no incentive for any incumbent to allow other customers to overtake them, as this will only increase their expected sojourn time. Therefore, they will refuse all cutting requests. This is true regardless of queue length and the number of requests, and will describe the behaviour of the occasional customers even in the repeated games setting, as for them, it is a single-shot game.

8.2.2 Repeated Game With Perfect Public Monitoring

Consider now the setting where some players use the service repeatedly. As in Allon and Hanany (2012), it will be assumed that each customer will not have concurrent requests, that periods are such that the length of time between service requirements is clearly separated from waiting times (e.g. say a service is needed every day, and expected sojourn time is 20 minutes). Future period payoffs and waiting costs are discounted by a factor $\delta \in (0, 1)$.

Note that the FCFS inducing strategy which is the equilibrium of the single stage game is also an equilibrium of the repeated game. Nevertheless, there are conditions where it is an equilibrium outcome for regular customers to allow others to cut the queue.

Consider the following grim trigger strategy profile, set out in Allon and Hanany (2012): incumbents who are regular customers agree to all requests for cutting ahead made by other regulars. If one refuses, the punishment strategy is triggered. This punishment is the FCFS inducing strategy which is the equilibrium of the single stage game, as is required to avoid deviations. The punishment strategy is anonymous, i.e., it doesn't target a specific customer.

Let W^A and W^R be the expected sojourn time experienced by a regular customer as an incumbent when agreeing and when refusing to let an arrival cut ahead, respectively, given that all other regular customers follow the cooperative strategy of agreeing to cutting requests from other regulars (P, A). Further let V^A be the long term expected

discounted payoff when all regular customers follow the cooperative strategy in all periods, and V^{FCFS} the long term expected discounted payoff all customers follow the punishment strategy (J, R).

Theorem 14. The strategy in which every regular customer chooses (A, P) if this choice was taken by all regular customers in previous periods, and punishes by choosing (R, J) if some regular customer deviates from this, is an equilibrium if and only if:

$$\frac{\delta}{1-\delta} \geq \frac{W^A - W^R}{D^{FCFS} + \frac{1}{\mu} - W_\iota^A}, \tag{8.1}$$

and

$$W_\iota^A < D^{FCFS} + \frac{1}{\mu}, \tag{8.2}$$

where D^{FCFS} is the *a priori* expected waiting time under the FCFS discipline (produced by the punishment strategy), and W_ι^A is the *a priori* expected waiting time conditional on the cooperative strategy being followed.

Theorem 14 shows that the cooperative strategy, where regular users, as incumbents, allow arriving regular users to cut ahead, can be sustained in equilibrium for sufficiently patient customers, as the joining customers are willing to forego the present benefit of a shorter sojourn time in exchange for the future benefits of being able to cut ahead. For this to happen, however, it must also be the case that *ex ante* expected sojourn time is lower when all regular customers follow the cooperative strategy (eq. (8.2)). For further discussion of this condition, see the Appendix at 8.7.4.

When the cooperative equilibrium is sustainable, arriving regular customers will ask the (regular) incumbent closest to the server to overtake them. This can be seen to lead to a sort of queue-within-a-queue, as set out in the following corollary:

Corollary 4. When the cooperative strategy is being implemented, regular customers will form a LCFS queue within the FCFS queue.

Essentially, new arrivals of regular customers will ask the first regular to let them overtake. When their request is accepted, they will take the first place among the regulars. However, there might be occasional customers in front of sub-queue of regulars, as well as behind them, and these will behave according to FCFS.

Queue-Length-Dependent Strategies

So far, only ex-ante decisions on whether to join the end of the queue or request to cut ahead, and whether to accept or reject these requests, have been considered. This section explores whether the equilibrium where customers allow others to cut the queue can be sustained when customers' strategies vary with queue length.

The approach to this problem is borrowed from Allon and Hanany (2012), focusing on queue-length-dependent strategies, i.e., a strategy where customers' choice of

whether or not to agree to a cutting request depends on observed queue length. Let the system size observed by customer N at arrival be $N-1$; further denote by $n^{\alpha*}$ the place occupied by the first regular customer (i.e., where the arrival would join).

Further, let $\overline{n^{\alpha*}}$ be a threshold defining the queue-length-dependent strategy, such that for $n^{\alpha*} \leq \overline{n^{\alpha*}}$, incumbent regular customers will play A, and for $n^{\alpha*} > \overline{n^{\alpha*}}$ they will play R. Then let $W^{A,\overline{n^{\alpha*}}}(n^{\alpha*})$ denote the expected sojourn time for an incumbent of letting an arrival cut ahead, and $W^{R,\overline{n^{\alpha*}}}$ that of refusing to do so, given that in both cases all other regular customers are following the queue-length-dependent strategy. Finally, let $V^{A,\overline{n^{\alpha*}}}$ be the long-term expected discounted utility when all regular customers follow the queue-length-dependent strategy for all periods.

Theorem 15. When regular customers follow queue-length-dependent strategies, for any threshold $\overline{n^{\alpha*}}$ there is a value of δ such that the threshold is never reached in equilibrium, i.e., it is never the case that $n^{\alpha*} > \overline{n^{\alpha*}}$, and so the 'long-queue' strategy is never employed: incumbent regular customers always play A.

The intuition behind Theorem 15 is that if customers employ queue length dependent strategies, where there is a maximum length above which they will reject cutting attempts, then there is always some value of delta for which customers are patient enough that this cutoff is never reached and only the short queue strategy is employed.

8.3 Sojourn Time and Joining Decision

This section considers sojourn time for customers who decide to join the queue, and subsequently considers their decision of whether or not to join. This decision is taken in stage 2 of the game, as detailed in Section 8.2. This consideration only comes into play if the conditions for the cooperative equilibrium to be sustainable, as defined in theorem 14, are met, otherwise incumbents allow no cutting attempts and the discipline is FCFS. Further, the criteria employed by regular customers who can cut ahead of other regulars when making a joining decision will differ from those of occasional service users.

On arrival, customer N observes queue size $N-1$; if she is a regular customer, she will also observe its composition, i.e., which members are also regular customers. If the customer is an occasional user, she does not have access to this information, and may only join the queue at its end, taking position N. She will do so as long as her expected utility from doing so is positive:

$$U = \nu - cE[W_N],$$

where W_N, $(N = 1, 2, ...)$ is the sojourn time for a customer arriving at a system containing $N-1$ customers.

On the other hand, regular users have the possibility of asking another regular user to overtake them. As described previously in Section 8.2, the customer from whom this will be requested is to be the regular customer who is closer to the server—although,

of course, it is possible that there are no other regulars in the queue, in which case the customer will just join the queue at the end. A share α of customers is regular, and under the cooperative equilibrium, regular customers join the queue by taking the place of the first regular customer. Therefore, conditional on their deciding to join the queue, their place within it will be determined by a bounded geometric distribution with parameter α, where $p_n(n, N)$ denotes the probability the first regular incumbent is in the nth position in the queue, $n \in \{2, ..., N\}$ and $N \geq 3$, as a function of n and N. Note that if the number of trials exceeds the number of customers in the queue, the customer will just take the last place: this reflects the situation where no other regulars are in the queue. Further, it is not possible to overtake the first customer, as her service has started, so the distribution support starts at $n = 2$. The probability distribution function is then:

$$p_n(n, N) = \begin{cases} \alpha(1-\alpha)^{n-2} & \text{if } n \in \{2, ..., N-1\} \\ (1-\alpha)^{n-2} & \text{if } n = N \end{cases}, \forall N \geq 3, \tag{8.3}$$

where obviously for $N = 2$, the customer cannot overtake anyone and so takes the place $n = 2$.

To this p.d.f. corresponds the following cumulative distribution function:

$$P_n(n, N) = \begin{cases} 1 - (1-\alpha)^{n-1} & \text{if } n \in \{2, ..., N-1\} \\ 1 & \text{if } n = N \end{cases}, \forall N \geq 3. \tag{8.4}$$

Finally, the *ex-ante* expected value of n as a function of N is:

$$\begin{aligned} E[n(N)] = \sum_{n=2}^{N} p_n n = \sum_{n=2}^{N-1} \alpha(1-\alpha)^{n-2} + (1-\alpha)^{N-2} N = \\ \frac{\alpha^2 + (1-\alpha)^n - 1}{\alpha(\alpha - 1)}, \forall N \geq 3. \end{aligned} \tag{8.5}$$

Regular customers will of course also only join the queue as long as their expected utility from doing so is positive. Unlike occasional customers, however, their place in the queue upon joining is not necessarily N, but rather $n(N)$, i.e., n as a function of N, so that utility is a function of W_n^N, where $N \in \{1, 2, \dots\}$ and $n = 1$ when $N = 1$, $n = 2$ when $N = 2$, and $n \in \{3, \dots, N\}$ when $N \geq 3$:

$$U = \nu - cE[W_n^N].$$

Nonetheless, the customer's decision of whether to join the queue or not is made with knowledge of n, i.e., of the result of the draw from the p.d.f. at (8.3). Customer N will of course join the queue if and only if:

$$U = \nu - cE[W_n^N] \geq 0, \tag{8.6}$$

which can be written as:

$$E[W_n^N] \leq \frac{\nu}{c}. \tag{8.7}$$

8.3.1 Expected Sojourn Time

Occasional customers are unaware of the cooperation between regular customers. Therefore, while their actual waiting time will not be according to the FCFS discipline, they act on the mistaken assumption that it is. Under the FCFS rules, expected sojourn time is given by:

$$E[W_N] = \frac{N}{\mu}, \tag{8.8}$$

so that occasional customers will join the queue if (see Naor (1969)):

$$\nu - c\frac{N}{\mu} \geq 0. \tag{8.9}$$

On the other hand, determining the expected sojourn time for regular customers is more challenging, because it is a function of future arrivals as well as the existing queue size: as regular customers form a LCFS queue, future arrivals of regular customers will overtake the customer and increase the expected time to service completion. On the other hand, the existing queue size is only relevant insofar as it is composed of occasional customers (who will not allow anyone to overtake them) ahead of the LCFS queue of regulars, and the customer being served (who cannot be overtaken)—at the time of joining, queue size beyond the first regular customer is irrelevant for determining the sojourn time of a regular customer.

The approach followed here to determine expected sojourn time for regular customers follows that used by Yu et al. (2014) in their solution of the joining problem for the EPS discipline, although it has been modified to take account of the specificities of the present LCFS-like problem. The essence of this approach is to describe the queueing problem through a system of linear difference equations.

Before setting out how to solve this problem, however, it is important to posit three conditions: (1) $\rho \equiv (\lambda/\mu) < 1$ and $\varrho \equiv (\Lambda/\mu) < \rho$, where $\Lambda = \alpha\lambda$ i.e., departures from the system must on average exceed arrivals—this is required to prevent the system from growing explosively, and the arrival rate for regular customers is lower than the total arrival rate; (2) all stochastic processes are independent of each other; (3) the inequality $\nu \geq \frac{c}{\mu}$ holds—this is required to avoid triviality: it must be profitable for a customer to join when the queue length is zero and the server is idle, otherwise the system would never be used.

Let W_n denote the expected sojourn time of an arbitrary regular customer, chosen to be 'tagged', joining at the nth position in the queue, where $n = n_{\bar{\alpha}}^* + n_\alpha$: $n_{\bar{\alpha}}^*$ is the number of occasional customers queueing ahead of the LCFS queue of regular customers, and n_α the customer's position on the sub-queue of regular customers. Under the cooperative strategy, arriving regular customers overtake the first regular

in the queue, so that $n_\alpha = 1$ on joining if $n^*_{\bar{\alpha}} \geq 1$, or $n_\alpha = 1$ on joining if $n^*_{\bar{\alpha}} = 0$, though future arrivals of regular customers will overtake the 'tagged' customer and cause her value of $n_\alpha = 1$ to change. Regular customers on the system at arrival, and occasional customers behind the sub-queue of regular customers, have no effect on the expected sojourn time of a joining customer, so total queue size is irrelevant for regular customers' joining decision: it will only be a function of service and arrival rates, and the number of occasional customers ahead of the sub-queue of regulars.

Further let A denote the time to the next customer arrival, and S the time to completion of the currently ongoing service. Once a customer chooses to join, and takes her place in the queue, either $n = n^*_{\bar{\alpha}} + 1$ when $n^*_{\bar{\alpha}} \geq 1$, or $n = 0 + 2$ when $n^*_{\bar{\alpha}} = 0$, there are two possibilities for the next relevant event[5]:

- The server finishes servicing the first customer, who leaves the system, before a new customer arrival: $A > S$.

- A new regular customer arrives before service completion, and cuts ahead of all the other regulars, including the 'tagged' customer: $A < S$. In this case, the 'tagged' customer's place in the LCFS sub-queue, n_α, increases by 1.

Further let:

$$I_B = \begin{cases} 1, & \text{if event } B \text{ occurs,} \\ 0, & \text{if event } B \text{ does not occur.} \end{cases}$$

Then the expected sojourn time can be decomposed, using a first-step argument, into two components corresponding to the two cases outlined above. In order to perform the decomposition, it is convenient to take the Laplace transform of expected sojourn time (see Yu et al. (2014)):[6]

$$\begin{aligned} \exp\left(-s\, W_{n_{\bar{\alpha}^*}+n_\alpha}\right) = &\exp\left(-s\, W_{n_{\bar{\alpha}^*}+n_\alpha}\, I_{\{S<A\}}\right) + \exp\left(-s\, W_{n_{\bar{\alpha}^*}+n_\alpha}\, I_{\{A<S\}}\right) \\ &, \forall\, n_{\bar{\alpha}^*} \geq 1, \text{ and } n_\alpha \geq 1 \end{aligned} \tag{8.10}$$

$$\begin{aligned} \exp\left(-s\, W_{0+n_\alpha}\right) = &\exp\left(-s\, W_{0+n_\alpha}\, I_{\{S<A\}}\right) + \exp\left(-s\, W_{0+n_\alpha}\, I_{\{A<S\}}\right) \\ &, \forall\, n_{\bar{\alpha}^*} = 0, \text{ and } n_\alpha \geq 3, \end{aligned} \tag{8.11}$$

$$\begin{aligned} \exp\left(-s\, W_{0+2}\right) = &\exp\left(-s\, W_{0+2}\, I_{\{S<A\}}\right) + \exp\left(-s\, W_{0+2}\, I_{\{A<S\}}\right) \\ &, \forall\, n_{\bar{\alpha}^*} = 0, \text{ and } n_\alpha = 2. \end{aligned} \tag{8.12}$$

After a series of lengthy considerations, for which see the Appendix, the following form for the difference equations emerges, where $\Lambda = \alpha\lambda$, i.e., the share of arrivals who are regular customers:

[5] I.e., excluding arrivals of occasional users, which do not affect waiting times of regular customers, since they all join the end of the queue.

[6] As customers are risk neutral, the expectations operator will be dropped in order to save space.

$$W_{n_{\bar{\alpha}^*}+n_\alpha} = \frac{1}{\Lambda+\mu} + \frac{\mu}{\Lambda+\mu} W_{(n_{\bar{\alpha}^*}-1)+n_\alpha} + \frac{\Lambda}{\Lambda+\mu} W_{n_{\bar{\alpha}^*}+(n_\alpha+1)}, \tag{8.13}$$

for $n_{\bar{\alpha}^*} \geq 1$, and $n_\alpha \geq 1$,

$$\begin{aligned} &W_{0+n_\alpha} = \frac{1}{\Lambda+\mu} + \frac{\mu}{\Lambda+\mu} W_{0+(n_\alpha-1)} + \frac{\Lambda}{\Lambda+\mu} W_{0+(n_\alpha+1)}, \\ &\forall\, n_{\bar{\alpha}^*} = 0, \text{ and } n_\alpha \geq 3, \end{aligned} \tag{8.14}$$

$$W_{0+2} = \frac{1}{\Lambda+\mu} + \frac{\mu}{\Lambda+\mu} W_{0+1} + \frac{\Lambda}{\Lambda+\mu} W_{0+3}, \forall\, n_{\bar{\alpha}^*} = 0, \text{ and } n_\alpha = 2, \tag{8.15}$$

and finally for the customer being served:

$$W_{0+1} = \frac{1}{\mu}, \forall\, n_{\bar{\alpha}^*} = 0, \text{ and } n_\alpha = 1. \tag{8.16}$$

The intuition behind eqs. (8.13)-(8.16) if relatively easy to grasp: the term $1/(\Lambda+\mu)$ is the expected time to the next relevant event, whether an arrival or service completion; $\mu/(\Lambda+\mu)$ is the probability of the next relevant event being a service completion, in which case the customer's place in the queue falls by one, either by the service of an occasional or a regular customer; $\Lambda/(\Lambda+\mu)$ is the probability that the next relevant event is a new arrival of a regular customer, in which case the customer's place in the queue increases by one, through the addition of a regular customer.

As eqs. (8.13)-(8.16) describe a system of linear second order difference equations with two boundary conditions, it is solvable for any values of μ, λ and α. This can be done through a guess and verify approach, where one conjectures that the solution takes the following form:

$$W_{n_{\bar{\alpha}^*}+n_\alpha} = \beta_0 + n_{\bar{\alpha}^*}\beta_1 + n_\alpha\beta_2. \tag{8.17}$$

Once this is applied to the system, the following values emerge:

$$\begin{aligned} \beta_0 &= -\frac{\varrho}{\mu-\Lambda} \\ \beta_1 &= \frac{1}{\mu-\Lambda} \\ \beta_2 &= \frac{1}{\mu-\Lambda}, \end{aligned}$$

so that expected sojourn time is:

$$W_{n_{\bar{\alpha}^*}+n_\alpha} = \frac{n_{\bar{\alpha}^*} + n_\alpha - \varrho}{\mu-\Lambda}. \tag{8.18}$$

8.3.2 Threshold Value

Occasional users by the very nature of their occasional use do not know of the cooperation between regular users until they observe it at play. Instead, they take the queue to obey FCFS rules, so that their joining threshold $\overline{N_{\bar{\alpha}}}$, the highest place in the queue they will be willing to take, is that obtained in Naor (1969):

$$\overline{N_{\bar{\alpha}}} = \left\lfloor \frac{\nu\mu}{c} \right\rfloor . \tag{8.19}$$

On the other hand, regular users know about their cooperative equilibrium and how they will form an LCFS sub-queue, so that if they decide to join, they will always join the beginning of the LCFS sub-queue, i.e. $n_\alpha = 1$ when $n_{\bar{\alpha}^*} \geq 1$, or $n_\alpha = 2$ when $n_{\bar{\alpha}^*} = 0$ (though this may of course change as other customers join). However, it follows from (8.18) that $W_{1+1} = W_{0+2}$, so that if a customer will join the queue for $n_{\bar{\alpha}^*} = 1$, she will also do so when $n_{\bar{\alpha}^*} = 0$, and so only the former case need be considered. The joining threshold is then $\overline{n_{\bar{\alpha}^*}}$, the maximum number of occasional customers ahead of the LCFS sub-queue for which an arriving regular customer will join the queue:

$$\overline{n_{\bar{\alpha}^*}} = \left\lfloor (\mu - \Lambda)\frac{\nu}{c} - (1 - \varrho) \right\rfloor . \tag{8.20}$$

It is evident from (8.20) that regular customers' joining decisions are independent of total queue length. This implies that *the queue may grow to any arbitrary length, even if customers are impatient*, although the condition that $\rho < 1$ ensures that it will never experience explosive growth.

A few comparative statics results for the threshold can be outlined:

Lemma 7. The threshold for regular customers, $\overline{n_{\bar{\alpha}^*}}$ is increasing on the value of the final service ν.[7]

Lemma 8. The threshold for regular customers, $\overline{n_{\bar{\alpha}^*}}$ is decreasing on unit cost of time c.

Lemma 9. The threshold for regular customers, $\overline{n_{\bar{\alpha}^*}}$ is increasing on the service rate μ.

Lemma 10. The threshold for regular customers, $\overline{n_{\bar{\alpha}^*}}$ is decreasing on the arrival rate λ.

Lemma 11. The threshold for regular customers, $\overline{n_{\bar{\alpha}^*}}$ is decreasing on the share of regular customers α.

[7] Though note that because $\overline{n_{\bar{\alpha}^*}}$ is a discrete variable, a small increase in ν may not be enough to change the threshold value. The same applies, *mutatis mutandis*, to lemmas 8-10.

8.4 Social Optimization

This section considers the system from the point of view of a social planner who is concerned to maximize aggregate utility accruing to the stream of customers, and can set a maximum length for the queue such that when this threshold is reached, arriving customers are turned away until a service occurs (cf. Naor (1969)).

Given the system is M/M/1, absent individual optimization its queue length distribution is the same regardless of service discipline, i.e. identical to the FCFS discipline (see Federgruen and Groenevelt (1988) and Guillemin and Boyer (2001)). Let $\pi_N(N = 0, 1, ...)$ be the stationary probability of there being N customers in the queue. Then for a boundless system:

$$\pi_N = \rho^N(1-\rho),\ N = 0, 1, 2, \ldots. \tag{8.21}$$

This means that the approach to social optimization pursued by Naor (1969) can be followed here as well, as is done by Yu et al. (2014) for the EPS discipline. This approach is outlined below.

Let $\bar{N}$ be a size limit for the system, such as can be set by the social planner; the distribution of queue lengths becomes:

$$\pi_N = \frac{\rho^N(1-\rho)}{1-\rho^{\bar{N}+1}},\ N = 0, 1, 2, \ldots, \bar{N}, \tag{8.22}$$

and expected queue size is:

$$E[N] = \frac{\rho}{1-\rho} - \frac{(\bar{N}+1)\rho^{\bar{N}+1}}{1-\rho^{\bar{N}+1}}. \tag{8.23}$$

Then let $\bar{N}_s$ be maximum system size set by the social planner. Instantaneous utility accruing to the stream of customers is given by:

$$\begin{aligned} U_s(\bar{N}_s) &= \nu\lambda(1-\pi_{\bar{N}_s}) - cE[N] \\ &= \nu\lambda\frac{1-\rho^{\bar{N}_s}}{1-\rho^{\bar{N}_s+1}} - c\left[\frac{\rho}{1-\rho} - \frac{(\bar{N}_s+1)\rho^{\bar{N}_s+1}}{1-\rho^{\bar{N}_s+1}}\right]. \end{aligned} \tag{8.24}$$

This expression was obtained in Naor (1969), where U_s was also shown to be discreetly unimodal in N_s. The first term is the value of the service multiplied by the effective arrival rate, i.e., λ multiplied by the probability of the system being at the maximum size, when arrivals are turn away; the second term is the unit cost of time times the expected queue size $E[N]$.

Discrete unimodality implies that the socially optimal threshold $\bar{N}_s^*$ may be found through iterative substitutions of successive integers into (8.24) until the maximum is achieved. This can also equivalent to finding a value for $\bar{N}_s^*$ satisfying the two following

inequalities:

$$U_s(\bar{N}_s^* - 1) \leq U_s(\bar{N}_s^*)$$
$$U_s(\bar{N}_s^*) \geq U_s(\bar{N}_s^* + 1).$$

8.5 Numerical Investigations

The present section will employ numerical simulations to compare the several disciplines with respect to sojourn times and joining thresholds. Table 8.1 presents the maximum queue length for the FCFS discipline (that is, the highest place in the queue a customer is willing to take), which is given by $n_t^{FCFS} = \lfloor \frac{\nu}{c}\mu \rfloor$; the threshold $\overline{N_{\bar{\alpha}}}$ for occasional customers, as given in (8.19); the threshold $\overline{n_{\bar{\alpha}^*}}$ for regular customers under the cooperative equilibrium, as given in (8.20); and the socially optimal threshold as described in section 8.4:

Table 8.1: Threshold comparison across disciplines for $\nu = 12$, $c = 6$, $\lambda = 2$, $\mu = 3$, and $\alpha = 0.5$.

Discipline	FCFS	$\overline{N_{\bar{\alpha}}}$	$\overline{n_{\bar{\alpha}^*}}$	Social
Threshold	6	6	3	2

Meanwhile, Figure 8.1 shows how, for a fixed α, the threshold function (prior to the application of the floor operator) varies with λ and μ, and Figure 8.2 how it varies with λ and α for a fixed μ, reflecting the behaviours described in Lemmas 7-11:

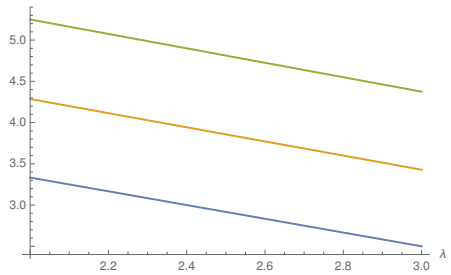

Figure 8.1: Threshold function for $\nu = 12$, $c = 6$, and $\alpha = 0.5$.

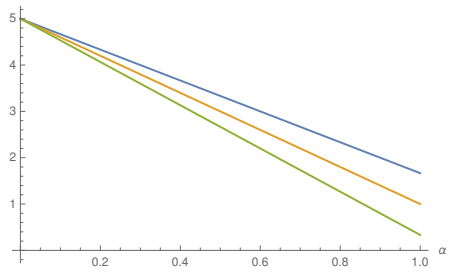

Figure 8.2: Threshold function for $\nu = 12$, $c = 6$, and $\mu = 3$.

Table 8.2 presents the sojourn times for customers at the respective threshold value (as presented in table 8.1), in a pure FCFS queue, for regular customers in the queue

considered in this paper, as well as customers in the FCFS discipline for the socially optimal threshold. The chosen parameters are, as before, $\nu = 12$, $c = 6$, $\lambda = 2$, $\mu = 3$, and $\alpha = 0.5$. It is easy to verify that these values meet the condition $\nu \geq c/\mu$, which must hold so that at least one customer uses the system.[8]

Table 8.2: Sojourn time comparisons, $\nu = 12$, $c = 6$, $\lambda = 2$, $\mu = 3$, and $\alpha = 0.5$.

Case	Time
FCFS	2
Regulars	1.8333
Social	0.6667

Meanwhile, Figure 8.3 shows how the sojourn time and the threshold varies with α:

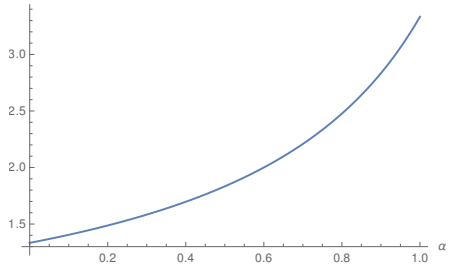

Figure 8.3: Expected sojourn time at the threshold for for $\nu = 12$, $c = 6$, $\mu = 3$, and $\lambda = 2$

According to Naor (1969), social welfare is discreetly unimodal on queue length. As queue cutting reduces the threshold for regular customers, it might naively be construed improving welfare. However, this ignores the fact that the threshold only refers to the maximum number incumbent occasional users ahead of the the LCFS sub-queue of regular customers, which can itself increase boundlessly—which will tend to lower social welfare. However, this is contingent on the relative sizes of the shares of regular and occasional customers. It can be noted, nevertheless, that for the situation where all customers are regular (i.e., $\alpha = 1$), and there is a pure LCFS queue, for the parameters chosen above, customers always join the queue and expected queue length[9] $E[N] = 4$, which is almost twice as large as that for the FCFS queue, where the threshold of 6 yields an expected queue length $E[N] = 2.01$. These results are somewhat expected, as the cutting process aggravates the worst aspect of the FCFS queue: on that discipline, joining customers impose a negative externality on all future arrivals who decide to join; here, not only is that still the case, but they also impose that externality on customers who were already in the queue.

[8]For the FCFS discipline, expected sojourn time is identical to that for occasional customers as given in (8.8).

[9]The expected steady state queue length of an M/M/1 system, conditioned on threshold N_t and ρ, is $E[N] = \frac{\rho}{1-\rho} - \frac{(N_t+1)\rho^{N_t+1}}{1-\rho^{N_t+1}}$, or $E[N] = \frac{\rho}{1-\rho}$ for unbounded queues(see Naor (1969)).

8.6 Conclusion

The present article has built on the advances of Naor (1969) and Yu et al. (2014) to analyse a queueing system which is a hybrid between the FCFS and LCFS disciplines. The repeated games framework presented by Allon and Hanany (2012) was adapted to construct a scenario where a subset of incumbent customers in an otherwise FCFS queue, those who use it regularly, allow other arriving regular customers to overtake them.

The methods used by Yu et al. (2014) to obtain the expected waiting time in an EPS queue were adapted to produce the expected waiting time in the discipline considered here. This was then used to determine the joining threshold, which was found to refer only to the number of 'occasional' customers ahead of the sub-queue of regular customers, not the maximum queue size, which can grow boundlessly.

Numerical simulations comparing the system under analysis here with the FCFS discipline, both under the endogenous as well as the socially optimal thresholds, were performed, where it was found to differ in behaviour from both the former studied disciplines, and particularly to lead to greater potential queue lengths, over and above the socially optimal threshold. This indicates the potential for this arrangement to reduce welfare *vis-a-vis* the FCFS queue, which itself is not socially optimal. While it is important not to exaggerate this potential, it could be an operational problem, implying that system managers should discourage queue cutting.

Further research could address generally distributed service time, take into consideration sub-groups of regular users, and perform a deeper analysis of the welfare properties of the system. This would, of course, result in higher complexity. Another important avenue for further research is to characterize the steady-state properties of the cooperative equilibrium, particularly *ex-ante* expected sojourn time, and how it relates to that under FCFS.

8.7 Appendix: Proofs

8.7.1 Single and Repeated Game

Proof of Theorem 13. This proof follows that for the similar theorem in Allon and Hanany (2012). There is no incentive for incumbents to allow arrivals to cut ahead: this will only increase their expected sojourn time. Therefore, they will refuse all requests. This is true regardless of queue length. Arrivals are then indifferent between J and P. Thus, the strategy profile $E^i = J$, $I^i = R$ forms an equilibrium in weakly dominant strategies. If all cutting attempts are rejected, the queue operates according to the FCFS discipline.

Suppose incumbent customer i chose $I^i = A$. If there exists any new arrival i' that chooses $E^{i'} = P$, then i will deviate to R to avoid being overtaken. On the other

hand, if a newcomer i' has a positive probability of encountering incumbent i, they will deviate to $E^{i'} = P$ to seek to cut ahead. That strategy profile is therefore not an equilibrium. □

Proof of Theorem 14. This proof follows that for the similar theorem in Allon and Hanany (2012). Once a deviation is observed, the threat to punish by rejecting all cutting attempts is credible, as it is the equilibrium of the single shot game. Because all customers do so, a single customer has no incentive to deviate and allow cutting. Therefore, the grim trigger is sustained in all future interactions.

In the equilibrium path, arrivals have no incentive to deviate when incumbents choose A, as they get to cut ahead. Focusing on the incumbents, they have an incentive to improve their waiting time by rejecting cutting requests (R), but doing so results in losing all future benefits from being able to cut ahead. Partial deviations can be ignored as customers would always be better off with a full deviation than a partial one.

An incumbent regular customer will not deviate from A to R if and only if:

$$\nu - cW^A + \delta V^A \geq \nu - cW^R + \delta V^{FCFS}, \tag{8.25}$$

where the left hand side is the long term expected discounted payoff from allowing the arrival to cut ahead, and all customers continuing with that strategy, and the right hand side is the long term expected payoff of refusing to allow the arrival to cut ahead, which triggers the FCFS inducing punishment strategy in all future periods. This can be rewritten as:

$$\frac{\delta}{(1-\delta)} \geq \frac{c(W^A - W^R)}{(1-\delta)(V^A - V^{FCFS})}. \tag{8.26}$$

The expected sojourn time for a customer under the FCFS rule is $D^{FCFS} + \frac{1}{\mu}$, and so the difference between the two rules in terms of the per period long-term expected discounted payoff is:

$$(1-\delta)(V^A - V^{FCFS}) = [\nu - cW_\iota^A] - \left[\nu - c\left(D^{FCFS} + \frac{1}{\mu}\right)\right], \tag{8.27}$$

where W_ι^A is the a priori expected waiting time conditional on the cooperative strategy being followed.

Substituting (8.27) into (8.26) yields the following:

$$\frac{\delta}{1-\delta} \geq \frac{W^A - W^R}{D^{FCFS} + \frac{1}{\mu} - W_\iota^A},$$

proving the theorem. □

Proof of Theorem 15. This proof follows that for the similar theorem in Allon and Hanany (2012). Fix a strategy where the threshold is $\overline{n^{\alpha*}}$. When $n^{\alpha*} > \overline{n^{\alpha*}}$, no

customer will agree to a request to cut ahead, and no requests will be made. Below the threshold, there is no incentive for arrivals to deviate, as they are can improve their expected waiting time by cutting ahead. For incumbents, it must be the case, for all $n^{\alpha*} \leq \overline{n^{\alpha*}}$, that:

$$\nu - cW^{A,\overline{n^{\alpha*}}}(n^{\alpha*}) + \delta V^{A,\overline{n^{\alpha*}}} \geq \nu - cW^{R,\overline{n^{\alpha*}}}(n^{\alpha*}) + \delta V^{FCFS}, \tag{8.28}$$

which simplifies to

$$c[W^{A,\overline{n^{\alpha*}}}(n^{\alpha*}) - W^{R,\overline{n^{\alpha*}}}(n^{\alpha*})] \leq \delta[V^{A,\overline{n^{\alpha*}}} - V^{FCFS}]. \tag{8.29}$$

As waiting times are finite for $n^{\alpha*} < \overline{n^{\alpha*}}$, and so is the region, the left-hand side of (8.29) is bounded over the region; this bound is constant in δ. On the other hand, the right hand side is increasing in δ, as it is $\delta/(1-\delta)$ times the positive difference between the stage game payoff under the cooperative strategy and FCFS. Thus, there exists $\overline{\delta} \in (0,1)$ such that for all $\delta \in (\overline{\delta}, 1)$, (8.29) is satisfied for all $n^{\alpha*} \leq \overline{n^{\alpha*}}$. □

8.7.2 Individual Joining Decision

The difference equations given in (8.13)-(8.15) were obtained as follows. First, set out the various elements stemming from the decomposition in (8.10), for a queue with some occasional users ahead of the sub-queue of regulars (i.e., $n_{\bar{\alpha}^*} \geq 1$). Starting with the case when the first relevant event (i.e. excluding arrivals of occasional users, so that the relevant arrival rate is $\alpha\lambda = \Lambda$) is a service completion of an occasional customer, and applying the memoryless property of the exponential distribution:

$$\begin{aligned}
&\exp\left(-s\, W_{n_{\bar{\alpha}^*}+n_\alpha}\, I_{\{S<A\}}\right) = \\
&\int_0^\infty \exp(-sy)\,\mu\exp(-\mu y)\, Pr\{A > y\}\exp\left(-sW_{(n_{\bar{\alpha}^*}-1)+n_\alpha}\right)\,\mathrm{d}y = \\
&\int_0^\infty \mu\exp(-y(s+\Lambda+\mu))\exp\left(-sW_{(n_{\bar{\alpha}^*}-1)+n_\alpha}\right)\,\mathrm{d}y = \\
&\frac{\mu}{s+\Lambda+\mu}\exp\left(-sW_{(n_{\bar{\alpha}^*}-1)+n_\alpha}\right).
\end{aligned} \tag{8.30}$$

Then the converse case where inter-arrival time (for regular users) is shorter than service time, so that queue length is increased by 1:

$$\begin{aligned}
&\exp\left(-s\, W_{n_{\bar{\alpha}^*}+n_\alpha}\, I_{\{A<S\}}\right) = \\
&\int_0^\infty \exp(-sx)\,\Lambda\exp(-\Lambda x)\, Pr\{x < S\}\exp\left(-sW_{n_{\bar{\alpha}^*}+(n_\alpha+1)}\right)\,\mathrm{d}x = \\
&\int_0^\infty \Lambda\exp(-x(s+\Lambda+\mu))\exp\left(-sW_{n_{\bar{\alpha}^*}+(n_\alpha+1)}\right)\,\mathrm{d}x = \\
&\frac{\Lambda}{s+\Lambda+\mu}\exp\left(-sW_{n_{\bar{\alpha}^*}+(n_\alpha+1)}\right).
\end{aligned} \tag{8.31}$$

Substituting (8.30)-(8.31) into (8.10) then yields:

$$\begin{aligned}\exp\left(-s\,W_{n_{\bar{\alpha}^*}+n_\alpha}\right) =& \frac{\mu}{s+\Lambda+\mu}\exp\left(-sW_{(n_{\bar{\alpha}^*}-1)+n_\alpha}\right)+\\ & \frac{\Lambda}{s+\Lambda+\mu}\exp\left(-sW_{n_{\bar{\alpha}^*}+(n_\alpha+1)}\right),\end{aligned} \tag{8.32}$$

for $n_{\bar{\alpha}^*} \geq 1$, and $n_\alpha \geq 1$.

The next step is to reverse the Laplace transform by taking the first derivative of (8.32) with regard to s, multiplying by -1 and setting $s = 0$, which yields:

$$W_{n_{\bar{\alpha}^*}+n_\alpha} = \frac{1}{\Lambda+\mu} + \frac{\mu}{\Lambda+\mu}W_{(n_{\bar{\alpha}^*}-1)+n_\alpha} + \frac{\Lambda}{\Lambda+\mu}W_{n_{\bar{\alpha}^*}+(n_\alpha+1)},$$

for $n_{\bar{\alpha}^*} \geq 1$, and $n_\alpha \geq 1$, which corresponds to (8.13).

A similar process can be employed for (8.11)-(8.12), yielding (8.14) and (8.15), respectively.

8.7.3 Threshold Value

Let:

$$\theta(\nu,\, c,\, \mu,\, \lambda,\, \alpha) = (\mu - \alpha\lambda)\frac{\nu}{c} - \left(1 - \frac{\alpha\lambda}{\mu}\right), \tag{8.33}$$

i.e., $\theta(.)$ is the threshold $\overline{n_{\bar{\alpha}^*}}$ from (8.20) before the floor function is applied to make it into an integer.

Proof of Lemma 7. An increase in ν increases $\theta(.)$:

$$\frac{\partial\theta(.)}{\partial\alpha} = \frac{\mu-\alpha\lambda}{c} > 0,$$

as $\mu > \lambda$.

If the increase in ν is high enough, this might lift $\theta(.)$ to the next larger integer and increase the threshold. □

Proof of Lemma 8. An increase in c decreases $\theta(.)$:

$$\frac{\partial\theta(.)}{\partial c} = -(\mu-\alpha\lambda)\frac{\nu}{c^2} < 0,$$

as $\mu > \lambda$.

If the increase in c is high enough, this might lower $\theta(.)$ to the next smaller integer and decrease the threshold. □

Proof of Lemma 9. An increase in μ increases $\theta(.)$:

$$\frac{\partial\theta(.)}{\partial\mu} = \frac{\nu}{c} - \frac{\alpha\lambda}{\mu^2} > 0,$$

as $\nu > \frac{c}{\mu}$ per the triviality avoidance clause.

If the increase in μ is high enough, this might lift $\theta(.)$ to the next larger integer and increase the threshold. □

Proof of Lemma 10. An increase in λ decreases $\theta(.)$:

$$\frac{\partial \theta(.)}{\partial \lambda} = \alpha \left(-\frac{\nu}{c} + \frac{1}{\mu} \right) < 0,$$

as $\nu > \frac{c}{\mu}$ per the triviality avoidance clause.

If the increase in λ is high enough, this might lower $\theta(.)$ to the next smaller integer and decrease the threshold. □

Proof of Lemma 11. An increase in α decreases $\theta(.)$:

$$\frac{\partial \theta(.)}{\partial \lambda} = \lambda \left(-\frac{\nu}{c} + \frac{1}{\mu} \right) < 0,$$

as $\nu > \frac{c}{\mu}$ per the triviality avoidance clause.

If the increase in α is high enough, this might lower $\theta(.)$ to the next smaller integer and decrease the threshold. □

8.7.4 FCFS Ex-Ante Expected Sojourn Times

As alluded to in Theorem 14, one of the conditions for the cooperative equilibrium to hold is that *a priori* expected sojourn time for regular customers must be smaller when all regular customers choose the cooperative strategy than when they use the FCFS-inducing strategy.

Total *a priori* expected sojourn time under either discipline consists of service time μ^{-1} plus waiting time. Under the FCFS-inducing strategy, there is no differentiation between regular and occasional customers and thus expected sojourn time W^{FCFS} follows readily from Little's Laws (see Little (1961) and Federgruen and Groenevelt (1988)). Average queue length under FCFS, L^{FCFS} is a function of the threshold $\overline{N_{\bar{\alpha}}}$:

$$W^{FCFS} = \sum_{n=0}^{\overline{N_{\bar{\alpha}}}} n p_n = \sum_{n=0}^{\overline{N_{\bar{\alpha}}}} n \, \rho^n \frac{1-\rho}{1-\rho^{\overline{N_{\bar{\alpha}}}+1}} = \frac{\rho}{1-\rho} - \frac{(\overline{N_{\bar{\alpha}}}+1)\rho^{\overline{N_{\bar{\alpha}}}+1}}{1-\rho^{\overline{N_{\bar{\alpha}}}+1}}. \tag{8.34}$$

Then $W^{FCFS} = L^{FCFS}/\lambda$:

$$W^{FCFS} = \frac{1}{\mu-\lambda} - \frac{(\overline{N_{\bar{\alpha}}}+1)\rho^{\overline{N_{\bar{\alpha}}}}}{(1-\rho^{\overline{N_{\bar{\alpha}}}+1})\mu} \tag{8.35}$$

Expected waiting time D^{FCFS}, as relevant for eq. (8.2) in Theorem 14, then can

be obtained from W^{FCFS} by subtracting service time $1/\mu$:

$$D^{FCFS} = W^{FCFS} - \frac{1}{\mu} = \frac{1}{\mu - \lambda} - \frac{\overline{N_{\bar{\alpha}}}\,\rho^{\overline{N_{\bar{\alpha}}}} + 1}{(1 - \rho^{\overline{N_{\bar{\alpha}}}+1})\mu} \tag{8.36}$$

Chapter 9

Pricing and Waiting Time Decisions in a Health Care Market with Private and Public Provision

9.1 Introduction

Waiting times are a problem affecting most health care systems. This is especially true where rationing is required, although its root cause is simply that capacity is limited while treatments take time to perform. Queueing imposes a further cost on the patient-consumer, and in this way, health care fits into a wide body of literature which deals with the economic causes, behaviour and consequences of queues.

Rationing is common in public health care systems, such as the British NHS. Where systems of this type are present, a private sector often also coexists, being able to offer shorter waiting times to those willing to pay. The present work deals with the pricing decisions of a single profit maximizing private provider of healthcare, which operates in a market also served by a public provider offering free treatment. The approach to modelling the health care market follows a simplified version of that outlined in Goddard et al. (1995). However, unlike that model, where the focus was on the NHS sector and its outcomes, the present work focuses on the pricing decisions of the private sector: in particular, in Goddard et al. (1995) the private sector set a market clearing price by design, and was not subject to congestion, whereas here the sole private provider exercises market power and is subject to congestion in a manner similar to the NHS. Although as will be shown, congestion will be lower in equilibrium in the private sector compared to the public sector. The present model yields similar results to Goddard et al. (1995) regarding the responses of demand for public and private sector treatment when capacity increases, namely, that an increase in capacity reduces expected waiting time in the public sector, and therefore increases demand for that sector while reducing it for the private. However, because in the present model price is a decision variable, the price's response to an increase in capacity will be dependent on the distribution of consumer waiting time cost.

The private provider's decision problem is solved with recourse to the basic model of competition in waiting times presented in Luski (1976) and Levhari and Luski (1978), taking a special case where one provider charges price zero (i.e., the NHS). This work's chief contribution is the analysis of private sector decisions when it is subject to congestion and possesses market power. It is appropriate to use queueing theory to model the waiting time features of the Health Care market, as persuasively argued in Goddard et al. (1995). Queueing has been a subject of economic analysis since the seminal paper Naor (1969), which dealt with optimal queue sizes in $M/M/1$ FCFS queues, and sketched a framework for individual decisions taken by impatient consumers which has been almost universally followed. These results were extended to an arbitrary number of queues by Knudsen (1972). See chapter 3 in the Introduction for an overview of these results.

9.1.1 Related Literature

Luski (1976) and Levhari and Luski (1978) employed the queueing framework to model

duopolistic competition between two identical providers selling a good whose provision required queueing. The two providers' simultaneously choose price, and this decision affects demand both directly and indirectly through waiting time. Consumers are differentiated through a randomly distributed unit cost of waiting time. The key result of Levhari and Luski (1978) is to show there is a separating equilibrium where one provider specializes in serving consumers with high waiting costs at a high price, and the other provider serves consumers with low waiting costs at low prices. This framework is adapted for the present chapter, with the public sector charging a price of zero, and the outside option being constructed such that no consumer will choose it in equilibrium. See Chapter 5 in the Introduction for an overview of these and other important results from the application of queueing theory to Industrial Organization.

A model similar to that in Levhari and Luski (1978) was devised by Chen and Frank (2004) for a monopolist provider, and then extended to a duopoly in Chen and Wan (2003). Their model assumes the presence of an outside option, and allows for heterogeneous firms. Lederer and Li (1997) has some crossover application, as it deals with consumer heterogeneity in delay cost, which is important in the health care context. Finally, Li and Lee (1994), describe a model of competition in three product characteristics: price, quality and service speed, where the good's value declines with time. This value-decay assumption is often used in the health economics literature.

Moving away from industrial organization to the fields of health economics and the economics of publicly provided private goods, Barzel (1974) presents a theory of rationing "free" goods through waiting which is relevant to public health provision, but does not have queueing as an important feature. Iversen (1993) has a model of resource allocation to meet waiting times within a national health service setting. Hoel and Sæther (2003) also consider competition between private and public health providers—with an important contribution to the welfare economics of the issue. This draws on previous work by Lindsay and Feigenbaum (1984) (and see Cullis and Jones (1986)), where rationing by waiting lists is used, with value decay rather than cost of time used to represent consumer impatience. Private providers have no capacity constrains, and therefore no queueing in equilibrium. This value decay approach is followed by Iversen (1997), who however is chiefly concerned with whether rationing occurs in the public sector, and does not use queuing models. Propper (2000) continued in the same vein as Lindsay and Feigenbaum (1984) by considering no capacity constraints in the private sector, and using this to analyse empirically the demand for private care in the UK.

Finally, Farnworth (2003) is quite similar to the present chapter, including capacity constraints on the fee-charging provider. The difference between it and the present work is that while there is a fee-charging provider "competing" with a "free" one, the former has the same 'social' objectives as the latter and is therefore not profit maximizing, presenting a completely different problem. Indeed, the, author specifically describes the model as presenting "two private hospitals that are publicly funded," and the fact that one charges a price while the other does not is attributed to "policy makers," (who

determine the price charged by the former) for "equity reasons"—it is the service rate that is the hospitals' decision variable. Conversely, the sole objective of the private sector provider in the present model is profit maximization, and it operates without reference to externally set social policy. Nevertheless, he reaches a similar result in regard to waiting times' response to a price increase: a decrease (increase) in the fee-charging (free) sector, providing corroboration to one of the present model's results.

9.2 The Model

The model combines Goddard et al. (1995)'s approach to queueing in health care markets with the framework for duopoly competition in price and waiting time in Levhari and Luski (1978). Adaptations to the latter include: i) restricting the public provider to charge a price of zero; and ii) instead of there being a disutility of waiting in the queue, and positive utility coming from a good bought after being served, disutility accrues regardless of whether one chooses to queue or not (reflecting the suffering caused by disease) and the 'good' to be purchased is the treatment of disease, eliminating the suffering. This is similar to what has been developed in Lindsay and Feigenbaum (1984), Iversen (1997), Farnworth (2003), and others. However, the framework represents an ongoing disutility rather than an exponential decay of treatment value, as generally assumed. While it can be shown that the two approaches are formally equivalent (for which, see the Appendix), the one adopted here is more tractable.

9.2.1 Consumers

Consumers are risk neutral expected utility maximizers, whose utility function is linear in waiting time, and who have a expected lifespan from the emergence of their disease of τ.[1] The intuition behind the specification adopted here is that in healthcare, the value of a treatment is not so much a pure benefit, but the removal of a preexisting condition causing disutility.

Consumers suffer a disutility of c per unit of time, which can be interpreted as the severity of the consumer's illness. As illness severity varies across consumers, c is a random variable following a continuous probability distribution function:

$$f(c),\ x \in [0, \bar{c}], \tag{9.1}$$

where $\bar{c}$ is finite and $F(c)$ is the corresponding cumulative distribution function.

Disease cost c is suffered until removed by treatment at T_i, the waiting time for provider $i = \{n, p\}$, where n denotes the public and p the private providers. If the consumer does not seek treatment, they will suffer c until death, i.e., for τ units of

[1] As consumers are risk neutral, the expectations term is omitted.

time. The expected utility of seeking treatment from provider i is then:

$$U_i = -cT_i - P_i, \tag{9.2}$$

where $P_n = 0$, and $P_p = P$ is the price charged by the private provider. The expected utility of foregoing treatment, U_o, is:

$$U_o = -c\tau$$

Total potential demand for health care is given by the exogenous parameter λ. This can be thought of as representing a patient population which is fixed in the short run. The two providers will each service a share of this total, which will be the arrival rate for that provider's queue.[2] Following the literature (e.g., Levhari and Luski (1978)), this parameter is normalized such that $\lambda = 1$, without any loss of generality, because the unit of time t is arbitrary, and therefore it is always possible to find one measure of time for which the arrival rate is 1.

Unlike Goddard et al. (1995), there is no consideration of consumer income—all consumers can afford the private sector treatment. While this is a restrictive assumption, made in the interests of simplicity, the problem is treated in this way to allow for a clear focus on the private provider's pricing decisions in response to the public sector provision. Interaction with consumer income constraints and heterogeneity, while doubtlessly important, is left for further research.

9.2.2 Providers

It is assumed that neither the private nor the public sectors have costs, or in what is perhaps a preferable interpretation, those costs are sunk in the short run. It is further assumed that for both the private and public sectors, service times are identically and independently distributed along an exponential distribution with rate μ. The assumption is made that $\mu > 1$, meaning that either server is capable of, by itself, serving the entire demand stream $\lambda = 1$. The assumption is required in order to avoid explosive growth of waiting times. This rate is exogenous and common to both servers. This parameter can be taken to reflect the state of technology. For example, improvements in surgical techniques allowing for hospital stays to be reduced would lead to an increase in the service rate. In what is a simplifying departure from Goddard et al. (1995), it is assumed that only one consumer can be treated at a time. This turns the queueing process into the well known $M/M/1$ system.

Expected waiting time for each provider, as a function of arrival and service rates, can be obtained from a well known result in queueing theory (see, *inter alia*, Gross

[2] It is not required that the two shares add up to 1, as patients can opt for not seeking treatment. Nevertheless, given the way the model is constructed, the shares of the two providers will always add up to 1 in equilibrium. An explanation follows below.

et al. (2008), or chapter 2 in the Introduction, where this result is given in eq. (2.33)):

$$T_i = \frac{1}{\mu - \lambda_i},\ i \in \{n, p\}. \tag{9.3}$$

As the present work is concerned with the analysis of competition between public and private providers in health care, the assumption will be made that both the public and private providers exist and are used by at least one consumer. For this to happen, it must be the case that $U_n = -cT_n > U_o = -c\tau$, which implies that $\tau > T_n$. Since, by assumption, the public sector does not charge for treatment, regardless of demand, and $\tau > T_n$, no consumers will forego treatment, and public sector supply is equal to its demand.

Meanwhile, private sector demand, given the exogenous parameter μ, is a function of P, which is set at the value that maximizes its instantaneous profit function:

$$\max_P \pi = \max_P \lambda_p P. \tag{9.4}$$

9.3 Demand

9.3.1 Individual Choice

An individual consumer chooses to seek care from a particular provider if two conditions are satisfied: his expected utility is larger than that of joining the waiting list for the other provider, and it is larger than that of foregoing treatment.

For a consumer with a given value of c, expected utility when seeking care from the private or public provider can be obtained from eq. (9.2). They are, respectively:

$$U_p = -cT_p - P, \tag{9.5}$$

and

$$U_n = -cT_n. \tag{9.6}$$

A consumer will then seek care from the private sector when:

$$c > \frac{P}{T_n - T_p}. \tag{9.7}$$

On the other hand, the consumer will choose the public provider when (assuming indifferent consumers will opt for the public sector):[3]

$$c \leq \frac{P}{T_n - T_p}. \tag{9.8}$$

[3]Returning to the assumption that both a public and a private sector exist, and are used by at least one consumer, it must be the case that $T_p < T_n$, and therefore $T_p < \tau$. Proof by contradiction: for $U_p > U_n$, it is necessary that $c > \frac{P}{T_n - T_p}$. But if $T_p > T_n$, this requires that $c < 0$, which is not a possible value of c. Cf. a similar discussion in Luski (1976).

Let c^* be the critical value of c for which a consumer is indifferent between the two providers:

$$c^*(P, T_n, T_P) = \frac{P}{T_n - T_p}. \tag{9.9}$$

9.3.2 Market Demand

Market demand for the private and public providers is obtained from the individual optimization process described in subsection 9.3.1. Potential demand λ divides itself between the two providers, with the share of consumers with a $c \leq c^*$ choosing the public provider, and the share with $c > c^*$ the private provider, so that as c^* decreases, the private provider's market share increases. The two demand functions are then:

$$\lambda_p = \int_{c^*}^{\bar{c}} f(c)\,\mathrm{d}c \Rightarrow \lambda_p(c^*) = 1 - F(c^*), \text{ and} \tag{9.10}$$

$$\lambda_n = \int_{0}^{c^*} f(c)\,\mathrm{d}c \Rightarrow \lambda_n(c^*) = F(c^*). \tag{9.11}$$

It can be seen from eq. (9.9) that c^* is itself a function of T_n, T_p, and P. The former two values (as set out in eq. (9.3)) are themselves a function of demand and the parameter μ. Ultimately, demand is then determined by the shape of the distribution of c, and the values of μ and P, the latter being the private sector provider's decision variable, as set out in subsection 9.4.3 below.

9.4 Supply

9.4.1 Providers

Waiting times T_n and T_p depend on of P, but only indirectly through λ_n and λ_p, both of which are a function of c^*. So it is possible, through implicit differentiation, to show that c^* is increasing in P even when taking into account the indirect effects, where c^*, λ_p and λ_n are as presented in (9.9)-(9.11). The direct effect corresponds to the direct increase of c^* in response to P, while the indirect effects are those mediated by the increase in T_n and decrease in T_p caused by the shift in the respective demands in response to an increase in price. Let ${c^*}'_P \equiv \frac{\partial c^*(P, T_n(\lambda_n(c^*),\mu), T_p(\lambda_p(c^*),\mu))}{\partial P}$, the total derivative of c^* with regard to P. Then:

$${c^*}'_P = \frac{1}{T_n - T_p} + \frac{\partial c^*}{\partial T_n}\frac{\partial T_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial c^*}\frac{\partial c^*}{\partial P} + \frac{\partial c^*}{\partial T_p}\frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}\frac{\partial c^*}{\partial P} > 0.$$

Solving, it yields:

$${c^*}'_P = \left[\left(1 - \frac{\partial c^*}{\partial T_n}\frac{\partial T_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial c^*} - \frac{\partial c^*}{\partial T_p}\frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}\right)(T_n - T_p)\right]^{-1} > 0, \tag{9.12}$$

$$+$$

where

$$\underset{-}{-\frac{\partial c^*}{\partial T_n}}\underset{+}{\frac{\partial T_n}{\partial \lambda_n}}\underset{+}{\frac{\partial \lambda_n}{\partial c^*}} > 0, \quad \underset{+}{-\frac{\partial c^*}{\partial T_p}}\underset{+}{\frac{\partial T_p}{\partial \lambda_p}}\underset{-}{\frac{\partial \lambda_p}{\partial c^*}} > 0,$$

with the signs under the partial derivatives indicating whether they are negative or positive. This result leads to the following two lemmas, which establish that private sector demand is bounded between 0 and half of the market:

Lemma 12. For any continuous distribution function $f(c)$, $x \in [0, \bar{c}]$, there is a price $\bar{P}$ which at the limit makes demand for the private provider equal to zero.

Proof. As c^* is a continuous and strictly increasing function of P (${c^*}'_P > 0$), and $\lambda_p(c^*)$ is a continuous and decreasing function of c^*, as defined at (9.9), it follows that:

$$\lim_{c^* \to \bar{c}} \lambda_p = 0, \tag{9.13}$$

where $\bar{P}$ is such that $c^*(\bar{P}, T_n, T_p) = \bar{c} = F^{-1}(1)$. □

Lemma 13. For any distribution function $f(c)$, $x \in [0, \bar{c}]$ meeting the conditions in (9.1), when $P \to 0$, demand for the private sector equals that for the public sector: $\lambda_n(c^*) = \lambda_p(c^*) = \frac{1}{2}$.

Proof. This is the reverse of Lemma 12. As c^* is a continuous and strictly increasing function of P (${c^*}'_P > 0$), and $\lambda_p(c^*)$ is a continuous and decreasing function of c^*, as defined at (9.9), it follows that when $P \to 0$, the two providers are indistinguishable, as none of them charges for treatment and they offer the same service rate μ. In this case the market outcome is a Bertrand equilibrium with the two providers splitting the market equally: $\lim_{P \to 0} c^*(P, T_n, T_p) = F^{-1}\left(\frac{1}{2}\right)$. □

As $\partial \lambda_p / \partial P < 0$, this further implies that demand for private sector treatment will never exceed that for the public sector.

9.4.2 Comparative Statics

The same process can be used to ascertain the effects a shock to μ will have on c^*, if P does not change. Let ${c^*}'_\mu \equiv \frac{\partial c^*(P, T_n(\lambda_n(c^*), \mu), T_p(\lambda_p(c^*), \mu))}{\partial \mu}$, the total derivative of c^* with regard to μ. Then:

$${c^*}'_\mu = \frac{\partial c^*}{\partial T_n}\left(\frac{\partial T_n}{\partial \mu} + \frac{\partial T_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial c^*}\frac{\partial c^*}{\partial \mu}\right) + \frac{\partial c^*}{\partial T_p}\left(\frac{\partial T_p}{\partial \mu} + \frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}\frac{\partial c^*}{\partial \mu}\right).$$

Solving, it yields:

$${c^*}'_\mu = \left(\frac{\partial c^*}{\partial T_n}\frac{\partial T_n}{\partial \mu} + \frac{\partial c^*}{\partial T_p}\frac{\partial T_p}{\partial \mu}\right) \Big/ \left(1 - \frac{\partial c^*}{\partial T_n}\frac{\partial T_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial c^*} - \frac{\partial c^*}{\partial T_p}\frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}\right). \tag{9.14}$$

It follows from the definition of c^* at (9.9) that:

$$\frac{\partial c^*}{\partial T_p} = -\frac{\partial c^*}{\partial T_n} = \frac{P}{(T_n - T_p)^2} > 0.$$

It also follows from the waiting time expression at (9.3) that:

$$\frac{\partial T_i}{\partial \mu} = -\frac{1}{(\mu - \lambda_i)^2} = -T_i^2 < 0.$$

So ${c^*}'_\mu$ in eq. (9.14) can be written as:

$$c^{*\prime}{}_\mu = \underset{+}{\frac{P}{(T_n - T_p)^2}\left(T_n^2 - T_p^2\right)} \Big/ \left(1 - \underset{+}{\frac{\partial c^*}{\partial T_n}\frac{\partial T_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial c^*}} - \frac{\partial c^*}{\partial T_p}\frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}\right) > 0, \forall\ P > 0. \tag{9.15}$$

Demand for either provider, as set out in (9.10)-(9.11), is a direct function of c^* alone, it is easy to sign their comparative statics in respect of both P and μ, considering P as a parameter. Let ${\lambda_p}'_P \equiv \frac{\partial \lambda_p(c^*(P,T_n(\lambda_n,\mu),T_p(\lambda_p,\mu)))}{\partial P}$, ${\lambda_n}'_P \equiv \frac{\partial \lambda_n(c^*(P,T_n(\lambda_n,\mu),T_p(\lambda_p,\mu)))}{\partial P}$, ${\lambda_p}'_\mu \equiv \frac{\partial \lambda_p(c^*(P,T_n(\lambda_n,\mu),T_p(\lambda_p,\mu)))}{\partial \mu}$, ${\lambda_n}'_\mu \equiv \frac{\partial \lambda_n(c^*(P,T_n(\lambda_n,\mu),T_p(\lambda_p,\mu)))}{\partial \mu}$, the derivatives of λ_p and λ_n in regard to P and μ, respectively.

$${\lambda_p}'_P = \frac{\partial \lambda_p}{\partial c^*}\frac{\partial c^*}{\partial P} = -f(c^*){c^*}'_P < 0 \tag{9.16}$$

$${\lambda_n}'_P = \frac{\partial \lambda_n}{\partial c^*}\frac{\partial c^*}{\partial P} = f(c^*){c^*}'_P > 0 \tag{9.17}$$

$${\lambda_p}'_\mu = \frac{\partial \lambda_p}{\partial c^*}\frac{\partial c^*}{\partial \mu} = -f(c^*){c^*}'_\mu < 0 \tag{9.18}$$

$${\lambda_n}'_\mu = \frac{\partial \lambda_n}{\partial c^*}\frac{\partial c^*}{\partial \mu} = f(c^*){c^*}'_\mu > 0. \tag{9.19}$$

These indicate that, as expected, a price increase will cause demand for the private sector to fall and that for the public sector to increase. This effect reduces the expected waiting time for the private sector, for which the consumers with higher time costs are willing to pay the increased price. Section 9.4.3 describes the private sector's pricing decision.

On the other hand, when the service rate increases, expected waiting time falls for both sectors. This reduces the private sector's advantage over the public sector, so that for a constant value of P, its demand would fall—similarly, to retain the same market share, it would have to reduce P.

9.4.3 Private Provider Optimization

The profit maximization process follows the model presented in Luski (1976) and Levhari and Luski (1978), but with the public provider constrained to charge a price of zero. The private provider will set P at the level P^* which maximizes its instantaneous

profit function:

$$P^* = \arg\max_P \pi = \arg\max_P \lambda_p P. \tag{9.20}$$

Note that waiting time in both sectors is then determined by the private provider, as it is a function of P and parameter μ (via c^*). Let $\pi'_P \equiv \frac{\partial \pi(\lambda_p(c^*(P,T_n(\lambda_n,\mu),T_p(\lambda_p,\mu))),P)}{\partial P}$, the total derivative of π with regard to P. Then performing the maximization yields the following first order condition:

$$\pi'_P = \lambda_p(c^*(P, T_n(\lambda_n,\mu), T_p(\lambda_p,\mu)) + P{\lambda_p}'_P = 0, \tag{9.21}$$

where ${\lambda_p}'_P$ follows from (9.16) and is known to be negative. P^* then follows from solving the first order condition for P:

$$\begin{aligned} \int_{c^*}^{\bar{c}} f(c)\,\mathrm{d}c - Pf(c^*){c^*}'_P = 0 \Leftrightarrow \\ [1 - F(c^*)] = Pf(c^*){c^*}'_P \Leftrightarrow \\ P^* = \frac{1 - F(c^*)}{f(c^*)\,{c^*}'_P} = \frac{\lambda_p}{f(c^*)\,{c^*}'_P}. \end{aligned} \tag{9.22}$$

Theorem 16. There exists at least one value of P which maximizes the private provider's profits.

Proof. At least one value P^* exists if the first order condition (9.21) has at least one zero in the domain $P \in (0, \bar{P})$, where $\bar{P}$ is the value for which $\lambda_p = 0$, or $c^* = \bar{c}$. When $P \to 0$, (9.21) takes the form:

$$\lim_{P\to 0} \left[(1 - F(c^*)) + P{\lambda_p}'_P\right] > 0,$$

whereas when $P \to \bar{P}$, $c^* \to \bar{c}$, and (9.21) takes the form:

$$\lim_{P\to \bar{P}} \left[(1 - F(\bar{c})) + \bar{P}{\lambda_p}'_P\right] = 0 - \bar{P}f(\bar{c}){c^*}'_P < 0.$$

Since the FOC takes a positive value at one end of the domain, a negative value at the other, and is continuous across the domain, the intermediate value theorem implies there is at least one zero. □

The firm's profit in equilibrium, π^* is then given by:

$$\pi^* = P^*\lambda_p(c^*(P^*, T_n(\lambda_n,\mu), T_p(\lambda_p,\mu)) = \frac{{\lambda_p}^2}{f(c^*)\,{c^*}'_P} = \frac{(1 - F(c^*))^2}{f(c^*)\,{c^*}'_P} \tag{9.23}$$

While existence could be shown from the first order condition alone, consideration of the uniqueness of P^* requires engaging with the second order condition as well.

Lemma 14. It is a sufficient condition for P^* to be a unique local maximum in the

range $(0, \bar{P})$ for:

$$\frac{\partial f(c^*)}{\partial c^*}\left(c^{*\prime}{}_P\right)^2 + f(c^*)\frac{\partial^2 c^*}{\partial P^2} > 0.$$

Proof. If the derivative of the first order condition is negative across the specified range, P^* will be unique in that range. This follows from the proof of theorem 16: as one end of the domain is negative and the other positive, and the FOC is continuous, if its derivative is negative it will only have one zero along that domain.

This condition is given by:

$$-2f(c^*)c^{*\prime}{}_P - P\left[\frac{\partial f(c^*)}{\partial c^*}\left(c^{*\prime}{}_P\right)^2 + f(c^*)\frac{\partial^2 c^*}{\partial P^2}\right] < 0, \tag{9.24}$$

evaluated at P^*. As $c^{*\prime}{}_P > 0$, (9.24) holds under the condition presented in lemma 14. □

Lemma 14 presents only a sufficient condition, so one cannot guarantee uniqueness for distributions which do not meet the conditions presented. However, numerical simulations have failed to produce a counter-example where P^* is not an unique maximum on the relevant range, even when the sufficient conditions outlines in the lemma were not met.

9.5 Welfare

In the present context, welfare for the consumers seeking treatment from provider i can be defined as the gain in disease-free time from seeking treatment $(\tau - T_i)$, across all consumers seeking treatment from that provider, reflected by the integral term, the expected value of c for each demand stream, minus the price of seeking treatment in the case of those consumers choosing the private sector. Therefore, welfare accruing to consumers in unit time, W_i, $i \in \{n, p\}$, is given by the following expressions:

$$W_n(c^*, T_n) = \left(\int_0^{c^*} cf(c)\,\mathrm{d}c\right)(\tau - T_n) \tag{9.25}$$

$$W_p(c^*, T_p, P) = \left(\int_{c^*}^{\bar{c}} cf(c)\,\mathrm{d}c\right)(\tau - T_p) - \lambda_p P. \tag{9.26}$$

Let W be aggregate social welfare, obtained from the sum of the firm's profit and the welfare of the two demand streams. The second term in W_p cancels out the firm's profit, so that W is given by:

$$W = W_n + W_p + \pi = \left(\int_0^{c^*} cf(c)\,\mathrm{d}c\right)(\tau - T_n) + \left(\int_{c^*}^{\bar{c}} cf(c)\,\mathrm{d}c\right)(\tau - T_p)$$

$$W(c^*, T_n, T_p) = \left(\int_0^{\bar{c}} cf(c)\,\mathrm{d}c\right)\tau - \left(\int_0^{c^*} cf(c)\,\mathrm{d}c\right)T_n - \left(\int_{c^*}^{\bar{c}} cf(c)\,\mathrm{d}c\right)T_p. \tag{9.27}$$

The first term $\left(\int_0^{\bar{c}} cf(c)\,\mathrm{d}c\right)\tau = E[c]\tau$ is the expected disutility of seeking no treatment, for all consumers. It is a constant term determined exogenously, the product of the expected value of c across the consumer population and expected lifespan τ. The two subsequent terms are the product of the expected value of the share of consumers seeking demand from each provider, and the expected waiting time for that provider, so that as waiting time falls, disease free time increases and so does welfare.

Both $\int_0^{c^*} cf(c)\,\mathrm{d}c$ and $\int_{c^*}^{\bar{c}} cf(c)\,\mathrm{d}c$ can be rewritten using integration by parts:

$$\int_0^{c^*} cf(c)\,\mathrm{d}c = c^*F(c^*) - \int_0^{c^*} F(c)\,\mathrm{d}c$$
$$\int_{c^*}^{\bar{c}} cf(c)\,\mathrm{d}c = \bar{c} - c^*F(c^*) - \int_{c^*}^{\bar{c}} F(c)\,\mathrm{d}c.$$

Then W_i and W can be written as follows:

$$W_n = \left(c^*F(c^*) - \int_0^{c^*} F(c)\,\mathrm{d}c\right)(\tau - T_n) \tag{9.28}$$

$$W_p = \left(\bar{c} - c^*F(c^*) - \int_{c^*}^{\bar{c}} F(c)\,\mathrm{d}c\right)(\tau - T_p) - \lambda_p P \tag{9.29}$$

$$\begin{aligned} W = &\left(\bar{c} - \int_0^{\bar{c}} F(c)\,\mathrm{d}c\right)\tau + c^*F(c^*)(T_p - T_n) - \bar{c}T_p + \\ &\left(\int_0^{c^*} F(c)\,\mathrm{d}c\right)T_n + \left(\int_{c^*}^{\bar{c}} F(c)\,\mathrm{d}c\right)T_p. \end{aligned} \tag{9.30}$$

The final expression is especially useful to determine how welfare responds to changes in price and processing rate, as $\bar{c}$ and τ are constants, while the integrand is the cumulative distribution function. Nevertheless, while the foregoing set is more tractable, it is preferable to use eqs. (9.25)-(9.27) for intuitive reasoning about welfare, as the interpretation of eqs. (9.28)-(9.30) is not straightforward.

Let $W'_P \equiv \frac{\partial W(c^*(P,T_n(\lambda_n,\mu),T_p(\lambda_p,\mu)),T_n(\lambda_p,\mu),T_n(\lambda_p,\mu))}{\partial P}$, the total derivative of W with regard to P. Further, let P^W be the social welfare maximizing price, such that $W'_P|_{P=P^W} = 0$; substituting this value back into W yields the largest possible social welfare W^*.

$$
\begin{aligned}
&W'_P = \left(c^{*\prime}{}_P F(c^*) + c^* f(c^*) c^{*\prime}{}_P\right)(T_p - T_n) + c^* F(c^*)\left(\frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}c^{*\prime}{}_P - \frac{\partial T_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial c^*}c^{*\prime}{}_P\right) \\
&- \bar{c}\frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}c^{*\prime}{}_P + \frac{\partial}{\partial P}\left(\int_0^{c^*} F(c)\,\mathrm{d}c\right)T_n + \left(\int_0^{c^*} F(c)\,\mathrm{d}c\right)\frac{\partial T_n}{\partial \lambda_n}\frac{\partial \lambda_n}{\partial c^*}c^{*\prime}{}_P \\
&+ \frac{\partial}{\partial P}\left(\int_{c^*}^{\bar{c}} F(c)\,\mathrm{d}c\right)T_p + \left(\int_{c^*}^{\bar{c}} F(c)\,\mathrm{d}c\right)\frac{\partial T_p}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial c^*}c^{*\prime}{}_P \Leftrightarrow \\
&W'_P = c^{*\prime}{}_P f(c^*)\Bigg[c^*(T_p - T_n) - T_p^2\left(c^* F(c^*) - \bar{c} + \left(\int_{c^*}^{\bar{c}} F(c)\,\mathrm{d}c\right)\right) + \\
&T_n^2\left(\left(\int_0^{c^*} F(c)\,\mathrm{d}c\right) - c^* F(c^*)\right)\Bigg].
\end{aligned}
\tag{9.31}
$$

The expression for $\partial W/\partial P$ in (9.31) above is not easily tractable through analytical means as long as $f(c)$ is not specified. Therefore, section 9.6 below presents the results for a sample of tractable distributions.

9.6 Results for Selected Distribution Functions

In this section, the model presented above is developed for two distribution functions $f(c)$, the Uniform and Kumaraswamy distributions. This is necessary because some of the results are intractable when the distribution function is not specified. Section 9.7 below presents numerical results for both the distributions considered here, and discussion of the results will be postponed until then, so that in this section only an overview is provided.

9.6.1 Uniform Distribution

Consider first a uniform distribution such that:

$$f(c) = (\bar{c})^{-1}, \tag{9.32}$$

$$F(c) = \frac{c}{\bar{c}}. \tag{9.33}$$

Under this distribution, demand and waiting time functions take the following forms:

$$\lambda_n = \frac{c^*}{\bar{c}}\ , \qquad \lambda_p = 1 - \frac{c^*}{\bar{c}},$$

$$T_n = \frac{1}{\mu - \frac{c^*}{\bar{c}}}\ , \qquad T_p = \frac{1}{\mu - (1 - \frac{c^*}{\bar{c}})},$$

which taken together form a system of four equations in four unknowns, easily solvable analytically.

The derivative of c^* with regard to price follows easily from c^*:

$$c^{*\prime}{}_P = \left[T_n - T_p + \frac{c^*}{\bar{c}}(T_n^2 + T_p^2)\right]^{-1},$$

which allows analytical derivation of the comparative statics outlined in eqs. (9.16)-(9.17):

$$\lambda_p{}'_P = -\frac{1}{\bar{c}}c^{*\prime}{}_P = -\frac{1}{\bar{c}}\left[T_n - T_p + \frac{c^*}{\bar{c}}(T_n^2 + T_p^2)\right]^{-1}$$
$$\lambda_n{}'_P = \frac{1}{\bar{c}}c^{*\prime}{}_P = \frac{1}{\bar{c}}\left[T_n - T_p + \frac{c^*}{\bar{c}}(T_n^2 + T_p^2)\right]^{-1}.$$

The private provider's optimization problem is presented and solved below:

$$\max_P \pi = \max_P \lambda_p P = \max_P \left(1 - \frac{c^*}{\bar{c}}\right) P$$
$$\pi'_P = 0 \Leftrightarrow \lambda_p + P\lambda_p{}'_P = 0$$
$$P^* = (\bar{c} - c^*)\left(T_n - T_p + \frac{c^*}{\bar{c}}(T_n^2 + T_p^2)\right).$$

Once P^* is known, equilibrium profit follow easily:

$$\pi = P^*\lambda_p(P^*) = \frac{(\bar{c} - c^*)^2}{\bar{c}}\left(T_n - T_p + \frac{c^*}{\bar{c}}(T_n^2 + T_p^2)\right).$$

For price P^*, welfare levels W_n, W_p, and W are:

$$W_n = \frac{1}{2}\frac{(c^*)^2}{\bar{c}}(\tau - T_n)$$
$$W_p = \frac{1}{2}\left(\bar{c} - \frac{(c^*)^2}{\bar{c}}\right)(\tau - T_p) - \pi$$
$$W = \frac{1}{2}\left[\bar{c}\tau - T_n\left(\frac{(c^*)^2}{\bar{c}}\right) + T_p\left(\frac{(c^*)^2}{\bar{c}} - \bar{c}\right)\right].$$

The derivative of W with regard to P is then as follows:

$$W'_P = c^{*\prime}{}_P\left[T_p\left(\frac{c^*}{\bar{c}} + \frac{T_p}{2}\left(1 - \left(\frac{c^*}{\bar{c}}\right)^2\right)\right) - T_n\left(\frac{c^*}{\bar{c}} + \frac{T_n}{2}\left(\frac{c^*}{\bar{c}}\right)^2\right)\right].$$

This can be set equal to 0 and solved for P, yielding P^W, the social welfare maximizing price.

Consider further a situation where there are two public providers, each charging a price of 0, and sharing consumers equally, i.e., $\lambda_{N_1} = \lambda_{N_2} = 1/2$; waiting times are therefore identical as well $T_{N_1} = T_{N_2} = T = (\mu - 1/2)^{-1}$. The following lemma can be stated:

Lemma 15. When c is uniformly distributed, it increases social welfare improving for

one of the providers to charge a positive price.

Proof. When there are two identical providers charging a price of 0 and c is uniformly distributed, the derivative of social welfare with regard to price takes the following form:

$$W'_P = {c^*}'_P \left[\frac{T^2}{2} \left(1 - 2 \left(\frac{c^*}{\bar{c}} \right)^2 \right) \right].$$

When demand is shared equally between the two providers, this is equivalent to $c^* = F^{-1}\left(\frac{1}{2}\right) = \frac{1}{2}\bar{c}$, so that $\frac{c^*}{\bar{c}} = \frac{1}{2}$. Then the derivative is

$$W'_P = {c^*}'_P \frac{T^2}{4},$$

which is positive, so that charging a positive price increases social welfare from the situation where the two providers charge a price of zero. □

Note, however, that there are positive prices which yield worse outcomes, and this result does not guarantee that the profit maximizing price will improve welfare.

9.6.2 Kumaraswamy distribution

This subsection performs the same exercise as above, but for a Kumaraswamy distribution with parameters $a = 1$ and $b = 2$, and $\bar{c} = 1$. This is defined as follows:

$$f(c) = 2(1 - c) \tag{9.34}$$

$$F(c) = c(2 - c). \tag{9.35}$$

This a bounded continuous distribution which depending on the parameters can take a variety of shapes, although it is much more tractable than the more widely known Beta distribution. Under the chosen values, the cumulative distribution function is concave. This places it under the set of distributions for which uniqueness of equilibrium price cannot be generally proven. Nevertheless, numerical simulations did not present any non-unique counter-example.

Under this distribution, demand and waiting time take the following forms:

$$\lambda_n = c^*(2 - c^*) \,, \qquad \lambda_p = 1 - c^*(2 - c^*),$$

$$T_n = \frac{1}{\mu - c^*(2 - c^*)} \,, \qquad T_p = \frac{1}{\mu - (1 - c^*(2 - c^*))},$$

which taken together form a system of four equations in four unknowns, easily solvable analytically.

The derivative of c^* with regard to price follows:

$${c^*}'_P = \left[T_n - T_p + 2(1 - c^*)c^*(T_n^2 + T_p^2) \right]^{-1},$$

which allows analytical derivation of the comparative statics outlined in eqs. (9.16)-(9.17):

$$\lambda_p{}'_P = -2(1-c^*)\left[T_n - T_p + 2(1-c^*)c^*(T_n^2+T_p^2)\right]^{-1}$$
$$\lambda_n{}'_P = 2(1-c^*)\left[T_n - T_p + 2(1-c^*)c^*(T_n^2+T_p^2)\right]^{-1}.$$

The private provider's optimization problem is presented and solved below:

$$\max_P \pi = \max_P \lambda_p P = \max_P \left(1 - c^*(2-c^*)\right) P$$
$$\pi'_P = 0 \Leftrightarrow \lambda_p + P\lambda_p{}'_P = 0$$
$$P^* = (1-c^*(2-c^*))\left(\frac{T_n - T_p}{2(1-c^*)} + c^*(T_n^2+T_p^2)\right).$$

Once P^* is known, equilibrium profit follow easily:

$$\pi = P^*\lambda_p(P^*) = (1-c^*(2-c^*))^2\left(\frac{T_n - T_p}{2(1-c^*)} + c^*(T_n^2+T_p^2)\right).$$

When price is P^*, welfare levels W_n, W_p, and W are:

$$W_n = (c^*)^2\left(1 - \frac{2}{3}c^*\right)(\tau - T_n)$$
$$W_p = \left(\frac{1}{3} + (c^*)^2\left(\frac{2}{3}c^* - 1\right)\right)(\tau - T_p) - \pi$$
$$W = \frac{1}{3}\tau + T_n(c^*)^2\left(\frac{2}{3}c^* - 1\right) + T_p\left((c^*)^2\left(1 - \frac{2}{3}c^*\right) - \frac{1}{3}\right).$$

The derivative of W with regard to P is then as follows:

$$W'_P = c^*{}'_P\Big[2(1-c^*)T_p\left(c^* + T_p\left(\frac{1}{3} - (c^*)^2\left(1 - \frac{2}{3}c^*\right)\right)\right)$$
$$- 2(1-c^*)T_n\left(c^* - T_n(c^*)^2\left(\frac{2}{3}c^* - 1\right)\right)\Big].$$

This can be set equal to 0 and solved for P, yielding P^W, the social welfare maximizing price.

Consider again situation where there are two public providers, each charging a price of 0, presented at the end of section 9.6.1. The following lemma can be stated:

Lemma 16. When c follows the Kumaraswamy distribution at (9.34), it increases social welfare improving for one of the providers to charge a positive price.

Proof. When there are two identical providers charging a price of 0 and c follows the Kumaraswamy distribution at (9.34), the derivative of social welfare with regard to

price takes the following form:

$$W'_P = c^{*\prime}{}_P 2(1-c^*)T^2\left[\frac{1}{3}+(c^*)^2\left(\frac{4}{3}c^*-2\right)\right].$$

When demand is shared equally between the two providers, this is equivalent to $c^* = F^{-1}\left(\frac{1}{2}\right) = 1-\frac{\sqrt{2}}{2}$. Then the derivative is

$$W'_P = c^{*\prime}{}_P \frac{\sqrt{2}}{3}T(5+\sqrt{2}),$$

which is positive, so that charging a positive price increases social welfare from the situation where the two providers charge a price of zero. □

As before, however, there are positive prices which yield worse outcomes, and this result does not guarantee that the profit maximizing price will improve welfare.

9.7 Numerical Simulations

This section presents the results of numerical simulations for the distributions $f(c)$ outlines in section 9.6, showing the values obtained for waiting times, market share, price, profit and social welfare, including the social welfare when there are two public providers charging a price of 0 and sharing demand equally, denoted by W_{2n}. Tables 9.1 and 9.2 use a Uniform distribution, while table 9.3 uses the Kumaraswamy distribution. There follows a section discussing the results presented here and in section 9.6.

Table 9.1: Numerical Simulations for a Uniform Distribution $[0,\bar{c}]$, $\bar{c}=1$, and $\tau=8$.

μ	T_n	T_p	λ_n	λ_p	P^*	π	W_n	W_p	W	P^W	W^*	W_{2n}
1.2	2.719	0.969	0.832	0.168	1.457	0.244	1.829	0.836	2.909	0.163	3.304	3.286
1.5	1.443	0.765	0.807	0.193	0.547	0.106	2.134	1.157	3.397	0.090	3.510	3.500
2	0.831	0.557	0.796	0.204	0.218	0.045	2.272	1.318	3.635	0.045	3.672	3.667
4	0.312	0.264	0.790	0.210	0.038	0.008	2.399	1.446	3.853	0.010	3.858	3.857

Table 9.2: Numerical Simulations for a Uniform Distribution $[0,\bar{c}]$, $\bar{c}=8$, and $\tau=8$.

μ	T_n	T_p	λ_n	λ_p	P^*	π	W_n	W_p	W	P^W	W^*	W_{2n}
1.2	2.719	0.969	0.832	0.168	11.656	1.955	14.631	6.689	23.275	1.304	26.430	26.286
1.5	1.443	0.765	0.807	0.193	4.472	0.845	17.074	9.257	27.175	0.724	28.080	28.000
2	0.831	0.557	0.796	0.204	1.745	0.356	18.179	10.544	29.078	0.359	29.373	29.333
4	0.312	0.264	0.790	0.210	0.301	0.063	19.193	11.569	30.826	0.076	30.865	30.857

Table 9.3: Numerical Simulations for a Kumaraswamy Distribution $[0,1]$, and $\tau=8$.

μ	T_n	T_p	λ_n	λ_p	P^*	π	W_n	W_p	W	P^W	W^*	W_{2n}
1.2	2.897	0.948	0.855	0.145	1.206	0.175	1.148	0.589	1.912	0.127	2.208	2.190
1.5	1.490	0.753	0.829	0.171	0.432	0.074	1.363	0.825	2.262	0.071	2.343	2.333
2	0.846	0.550	0.818	0.182	0.170	0.031	1.452	0.941	2.423	0.035	2.449	2.444
4	0.314	0.262	0.811	0.189	0.029	0.006	1.532	1.032	2.569	0.008	2.572	2.571

9.7.1 Discussion

Analysis of the simulation results yields some notable findings. In the first instance, for the Uniform distribution, varying the values of $\bar{c}$ results only in proportional changes to welfare and profits, while market shares remain unaltered. Hence only one different value is presented to illustrate this. Meanwhile, it is noteworthy that, for both distributions, as μ increases, W approaches W^* even as P and π fall, though the latter two values remain strictly positive. This can be intuitively explained as the increase in μ improving the service of the public sector *vis-a-vis* the private sector, therefore forcing it to reduce its price. This is in line with the result in eq. (9.18) showing λ_p is decreasing in μ for a constant price—clearly, even when the private sector changes its price to respond to the fall in demand, the downward pressure on demand still dominates.

There is a unique value of P^*—in fact, numerical simulations beyond those displayed here were not able to produce any counter-example where multiple equilibria emerged.

Comparing the results for the two distributions, one observes similar movements in welfare, prices and market shares when μ increases. However, welfare seems to be lower for the Kumaraswamy distribution, as are private sector market share and profits. This can be explained as being due to the Kumaraswamy distribution, with the chosen parameters, concentrating consumers towards the bottom of the range, i.e., more consumers have c closer to 0 than 1, and therefore the private sector chooses to charge lower prices in equilibrium. Despite that, it still receives a lower market share than for the Uniform distribution case.

When the social welfare is compared with the situation where there are two public providers charging a price of 0, regardless of the distribution, it is the case that the private sector equilibrium price produces an inferior outcome to the case where there are two public providers instead. However, this is always inferior to the socially optimal outcome where the private provider price P^W. This suggests an argument for price regulation of the private sector. Presumably this is due to the private sector functioning as an escape valve for consumers with very high disease costs who are willing to pay a lot to be served more quickly.

9.8 Conclusion

Waiting times affect competition in healthcare in a way that is not adequately captured by traditional models. The present chapter, using methods derived from the applications of queueing theory to Industrial Organization and Health Economics, contributes to the understanding of this phenomenon, following the lead of Lindsay and Feigenbaum (1984) and Farnworth (2003), among others.

When a profit maximizing private provider acts as a competitor to a public provider of health care, and that private provider faces capacity constraints resulting in queues, in a similar manner the public sector, but can vary its demand by varying the price,

there will always exist at least one price which maximizes the private sector's profit. This equilibrium's uniqueness will depend on the distribution of consumers' cost of waiting time/illness severity.

Comparative statics can be obtained for demand functions for any given distribution of c. Some of the most noteworthy results include that an increase (decrease) in the price charged by the private sector causes an increase (decrease) in demand for the public sector, as well as their waiting time; that a positive price for the private sector is welfare enhancing when compared with free treatment in two providers, and that an increase in the service rate μ decreases waiting times for both sectors, reducing the attractiveness of the private sector and lowering its demand. This forces the private sector to reduce its price, approximating the welfare maximizing value with higher values of μ. The most promising lead for further research is relaxing the assumption that both providers have the same service rates, and allowing at least the private sector to choose its own rate.

Further, the numerical results strongly suggest that the private sector equilibrium price is too high from a social welfare point of view. Further research should explore the effects of the public sector charging a regulated small price compared to a private sector charging a profit maximizing price.

Other warranted topics for further research include, in the first instance, to incorporate consumer income constraints and distribution into the decision process. Other interesting extensions are to incorporate long run capacity decisions by providers, and allowing for a plurality of private (and possibly public) providers. Numerical simulations of model outcomes for different distributions are also warranted, as is a further examination of how intertemporal considerations and discounting might affect consumer decisions.

9.9 Appendix: Equivalence of Exponential and Linear Formulation

Many works addressing waiting times in health care have formulated consumer utility using exponential decay. The following is an example, taken from Lindsay and Feigenbaum (1984):

$$U_i = V(\bar{u}, p)e^{-gt}. \tag{9.36}$$

In (9.36), $V(\bar{u}, p)$ is the value of the good, in this case the treatment, which is a function of a vector of parameters $\bar{u}$, and of price p. This value decays at rate g per unit of time t. Applying this function to the present problem, if $\bar{u}$ is held constant and V is a linear function of p such that $V = v - p$, where v is the value of the good before its price is deducted, (9.36) becomes:

$$U_i = (v - p)e^{-gt}. \tag{9.37}$$

This can then be transformed by taking logs:

$$\tilde{U}_i = \ln U_i = \ln(v - p) - gt. \tag{9.38}$$

Both v and p are arbitrary, so they can be redefined to new values such that $\Lambda = \ln(v - p)$:

$$\tilde{U}_i = \Lambda - gt, \tag{9.39}$$

where g is equivalent to parameter c in the utility function at equation (9.2).

There is still one outstanding issue. The presence of value parameter Λ introduces some mathematical complications. Moreover, as discussed above, medical treatments, especially of chronic diseases, can hardly be said to possess intrinsic value for consumers, unless perhaps they suffer from Münchausen syndrome. Rather, they are valuable in so far as they removes an illness. Therefore, take price p as being paid to remove disutility g. It's perfectly possible to postulate a function V of this kind, say $V = -p$, i.e., there is no benefit from the treatment itself other than it causing g to stop. This yields the following utility function:

$$U_i = -pe^{-gt}. \tag{9.40}$$

Taking logs of (9.40) yields

$$\tilde{U}_i = \ln U_i = \ln(-p) - gt. \tag{9.41}$$

Then once P is defined such that $P = \ln(-p)$, the utility function from (9.2) emerges.

Acknowledgements

Grateful thanks are extended to Tim Worrall for the many helpful discussions. Any remaining errors are, of course, my own.

Chapter 10

Bibliography

Allon, G. and Hanany, E. (2012). Cutting in Line: Social Norms in Queues. Management Science *58*, 493–506.

Barzel, Y. (1974). A Theory of Rationing by Waiting. Journal of Law and Economics *17:1*, 73–95.

Boudali, O. and Economou, A. (2012). Optimal and equilibrium balking strategies in the single server Markovian queue with catastrophes. European Journal of Operational Research *218*, 708–715.

Burnetas, A. and Economou, A. (2007). Equilibrium customer strategies in a single server Markovian queue with setup times. Queueing Systems *56*, 213–228.

Chen, H. and Frank, M. (2004). Monopoly pricing when customers queue. IIE Transactions *36:6*, 569–581.

Chen, H. and Wan, Y.-W. (2003). Price Competition of Make-to-Order Firms. IIE Transactions *35:9*, 817–832.

Cullis, J. G. and Jones, P. R. (1986). Rationing by Waiting Lists: An Implication. The American Economic Review *76:1*, 250–256.

Edelson, N. M. and Hildebrand, D. K. (1975). Congestion tolls for queueing processes. Econometrica *43:6*, 81–92.

Erlichman, J. and Hassin, R. (2009). Equilibrium Solutions in the Observable M/M/1 Queue with Overtaking. In Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools VALUETOOLS '09 pp. 64:1–64:9, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels.

Farnworth, M. G. (2003). A game theoretic model of the relationship between prices and waiting times. Journal of Health Economics *22:1*, 47–60.

Federgruen, A. and Groenevelt, H. (1988). Characterization and optimization of achievable performance in general queueing systems. Operations Research *36*, 733–741.

Gershkov, A. and Schweinzer, P. (2010). When queueing is better than push and shove. International Journal of Game Theory *39*, 409–430.

Goddard, J. A., Malek, M. and Tavakoli, M. (1995). An economic model of the market for hospital treatment for non-urgent conditions. Health Economics *4:1*, 41–55.

Gross, D., Shortle, J. F., Thompson, J. M. and Harris, C. M. (2008). Fundamentals of Queueing Theory. 4th edition, John Wiley and Sons, New Jersey.

Guillemin, F. and Boyer, J. (2001). Analysis of the M/M/1 queue with processor sharing via spectral theory. Queueing Systems *39*, 377–397.

Hassin, R. and Haviv, M. (1997). Equilibrium Threshold Strategies: The Case of Queues with Priorities. Operations Research *45*, 966–973.

Hassin, R. and Haviv, M. (2002). Nash Equilibrium and Subgame Perfection in Observable Queues. Annals of Operations Research *113*, 15–26.

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

Helweg-Larsen, M. and LoMonaco, B. L. (2008). Queuing among U2 fans: Reactions to social norm violations. Journal of Applied Social Psychology *38*, 2378–2393.

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

Hoel, M. and Sæther, E. M. (2003). Public health care with waiting time: the role of supplementary private health care. Journal of Health Economics *22:4*, 599–616.

Iversen, T. (1993). A theory of hospital waiting lists. Journal of Health Economics *12:1*, 55–71.

Iversen, T. (1997). The effect of a private sector on the waiting time in a national health service. Journal of Health Economics *16:4*, 381–396.

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.

Larson, R. C. (1987). Perspectives on queues: Social justice and the psychology of queueing. Operations Research *35*, 895–905.

Lederer, P. J. and Li, L. (1997). Pricing, Production, Scheduling, and Delivery-Time Competition. Operations Research *45:3*, 407–420.

Levhari, D. and Luski, I. (1978). Duopoly pricing and waiting lines. European Economic Review *11:1*, 17–35.

Li, L. and Lee, Y. S. (1994). Pricing and Delivery-Time Performance in a Competitive Environment. Management Science *40:5*, 633–646.

Lindsay, C. M. and Feigenbaum, B. (1984). Rationing by Waiting Lists. The American Economic Review *74:3*, 404–417.

Little, J. D. C. (1961). A Proof for the Queueing Formula L=λW. Operations Research *9*, 383–387.

Luski, I. (1976). On Partial Equilibrium in a Queuing System with Two Servers. The Review of Economic Studies *43:3*, 519–525.

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

Milgram, S., Liberty, H. J., Toledo, R. and Wackenhut, J. (1986). Response to intrusion into waiting lines. Journal of Personality and Social Psychology *51*, 683–689.

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

Oberholzer-Gee, F. (2006). A market for time fairness and efficiency in waiting lines. Kyklos *59*, 427–440.

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.

Propper, C. (2000). The demand for private health care in the UK. Journal of Health Economics *19:6*, 855–876.

Rothkopf, M. H. and Rech, R. (1987). Perspective on Queues:Naor, 1969 Combining Queues Is Not Always Beneficial. Operations Research *35*, 906–909.

Schmitt, B. H., Dube, L. and Leclerc, F. (1992). Intrusions into waiting lines: Does the queue constitute a social system? Journal of Personality and Social Psychology *63*, 806–815.

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

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

Sun, W., Guo, P. and Tian, N. (2009). Equilibrium threshold strategies in observable queueing systems with setup/closedown times. Central European Journal of Operations Research *18*, 241–268.

Sun, W. and Li, S. (2014). Equilibrium and optimal behavior of customers in Markovian queues with multiple working vacations. TOP *22*, 694–715.

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

Yu, M., Tang, Y. and Wu, W. (2014). Individually and socially optimal joining rules for an egalitarian processor-sharing queue under different information scenarios. Computers and Industrial Engineering *78*, 26–32.