

Cutting in Line: Social Norms in Queues

Gad Allon

Kellogg School of Management, 2001 Sheridan Road Evanston , IL 60208 , g-allon@kellogg.northwestern.edu

Eran Hanany

Department of Industrial Engineering, Tel Aviv University, Tel Aviv 69978, Israel , hananye@post.tau.ac.il

July 25, 2011

While the norm in many retail banks is to serve customers on a first-come-first-served basis, some customers try to cut the line, usually by providing an excuse for their urgency. In other queues, however, this behavior is considered unacceptable and is aggressively banned. In all of these cases, customer exhibit strategies that have not yet been explored in the operations literature: they choose whether or not to cut the line and must also decide whether to accept or reject such intrusions by others. This paper derives conditions for the emergence of such behavior in equilibrium among the customers themselves, i.e. when the queue manager is not involved in granting priorities and the customers have to use community enforcement to sustain such equilibria.

1. Introduction

In many service systems the manager is not involved in the way the queue is managed. For example, while the norm in many retail banks is to serve customers on a first-come-first-served basis, some customers try to cut the line, usually by providing an excuse for their urgency. This is also a common practice in airport security queues when people may ask to cut the line to avoid missing their flight. In certain places, this practice is sometimes coined ‘I just have a short question’ to describe people trying to declare to not require too much of the service provider’s time, justifying cutting the line. It may also happen that the service manager endorses queue cutting – a famous pediatrician in Haifa, Israel had a sign on his door asking customers to allow those with more urgent issues to jump the line. In Europe similar behavior is observed when in line for train tickets: one may jump to the front of the queue when almost missing their train. In other queues, however, this behavior is considered unacceptable and is aggressively banned. Examples are lines for a limited time show or a product on rationing (e.g. iPhone during its first days). In all of these cases, customers exhibit strategies that have not yet been explored in the operations literature: they choose whether or not to cut the line and must also decide whether to accept or reject such intrusions by others. This paper derives conditions for the emergence of such behavior in equilibrium among the customers themselves, i.e.

when the queue manager is not involved in granting priorities.

The literature on queueing focused primarily on the decision whether to join or balk, from the customer point of view, or on the priority offered by the firm (and sometimes bought by consumers). The sociology literature has recognized that the waiting line, with or without the involvement of the system manager, is by itself a social system. The main focus of that literature was norms and rules in such social systems, whereby the First In First Out (FIFO) priority rule is prominent while intrusions are deviations from the norm. Customers are described as experiencing tension between social norms and economic reasoning, the latter acting as a force that pushes customers to violate the norms. However, this literature tended to disregard two somewhat important considerations: (1) queues may be characterized by normative, selective disruptions that are value-enhancing for both the customers and service manager in the long run, and (2) abiding by the queue norms may go hand in hand with economic value maximization. As Schwartz (1975), Chapter 4, pointed out, orderly queues do not necessarily presuppose consensual devotion to FIFO, but may instead give more direct expression to the mutual interest of accommodation between the more and less impatient.

Our study brings these different concepts and methods together to build a model relying on rational decision making to characterize the different social norms observed in service systems, and in particular, explain several of the norms observed in practice. In viewing the queue as a social system, we follow Parsons (1955) who listed three properties of any social system that may develop: (1) two or more players come into some mode of interaction occupying different positions, (2) some organized patterns emerge, governing the relationship between the members and describing their rights and obligations to each other, and (3) some set of common norms and values are created, together with various types of shared symbols. In order to study these social norms from the perspective of rational decision making, we adopt the view that people do not follow a norm just for the sake of it, but rather because it is in their best interest and allows them to maximize their utility. Towards this end, we assume that all customers are a-priori identical, but in different periods they may have different types of requests: with either high or low cost of time (urgency) and with either high or low work content (lengthy or quick question/service resolution). The implications of such a model are: (i) customers may have legitimate reasons to cut the line, (ii) all customers may be on both

‘sides’ of the norm – either have a reason to cut or concede to those with such a reason, and (iii) the types of requests, relating to urgency and work content, are unobserved a-priori, but observable after the service was executed, in which case it becomes clear if indeed the customer had low work content or whether it was urgent.¹

We use a game theoretic model, initially of a single stage game, and then of a repeated game. In the single stage model, each customer arrives to the queue once and decides whether to join the queue at the end or cut the line; each customer confronted with a queue jumper decides whether to allow for it or not. We characterize the possible equilibria that may arise in such a game. Then we study a game where players engage in the single stage game repeatedly over time. When studying this game, our analysis echoes Okuno-Fujiwara and Postlewaite (1995), who defined social norms as having two components: standard of behavior and a transition mapping when the standard is violated. Each customer is assumed to have at most one service request per period, guaranteeing that a customer’s choice in any period affects the current period waiting cost in a way that is separable from its effect on future period discounted waiting costs. We initially study a repeated game with public monitoring and assume that customers make decisions whether to adhere to the norm before observing the state of the queue. In this game, when a customer deviates from the norm, all other customers become aware of the deviation. This basic model is then used as a benchmark to study more complex systems in which the special features of service systems are highlighted. In particular, we then study a model in which customers have queue length dependent strategies and decide whether to cut based on their position in the line. We also study systems without perfect monitoring, where only a small fraction of the entire customer base is present at any point in time. Customers have private signals of the deviation: the customer immediately behind the deviator gets a perfect signal, while others do not get any signal. Furthermore, even when a customer observes a deviation, they cannot spread the information to all other customers immediately since, again, only a small fraction of the customers are present at any point in time. This requires developing a more sophisticated punishment strategy that also serves as a means of informing other customers about the presence of deviators that require punishing.

¹ For example, if a person was allowed to jump the line stating their urgency, their type would be observed later sitting in a restaurant. Similarly, detection would occur if a person that claimed to have only a quick question required significant time with the service provider.

Our results are as follows.

1. When customers play the game only once, we show that the only possible priority rule that can emerge in equilibrium is FIFO. In particular, intrusions must be rejected in equilibrium. In that sense we show that in systems such as an overnight line, or when the frequency of visits is very low such that the interaction is best modeled using a single stage game, the only possible norm is one that results in FIFO dynamics. In this respect we are strengthening the observation made by several sociologists regarding the prominence of the FIFO priority rule in self-regulated waiting lines.

2. When players engage in a repeated game, and there is perfect monitoring of past actions, we show that the priority rule termed $c\mu$, defined by prioritizing customers with higher expected service cost rate, can be sustained in equilibrium under certain conditions. $c\mu$ priority minimizes the total expected waiting cost in a multi-class system when types are observed, yet the actual service times are unknown in advance and the system manager is restricted to work conserving priority rules (see Cox and Smith (1961)). We show that when customers are patient enough, legitimate intrusions may be used to improve the system performance by reducing the expected waiting cost based on $c\mu$ priority, and thus may be considered part of the norm. This behavior is supported by a threat to move to a socially inferior FIFO priority, which is also inferior on an individual basis under the conditions we characterize. We also show that the larger the cost rate difference between the two types of requests (high and low), the easier it is to sustain a $c\mu$ priority scheme in equilibrium.

3. When players exercise queue length dependent strategies, we show that many other priority rules are sustainable in equilibrium – all lying in the range between FIFO and $c\mu$. We study a family of priority schemes which are induced by thresholds: customers with urgent requests (or low work content) are allowed to push only if the queue length is below a threshold. We show that any priority scheme that is induced by such a threshold is sustainable in equilibrium when customers are sufficiently patient.

4. We then study a game in which perfect monitoring is impossible, and customers can only privately monitor those in the queue when a deviation occurs. We derive simple sufficient conditions for the welfare maximizing efficient equilibria to be sustained. In characterizing these equilibria we demonstrate that an efficient priority rule is sustained as a norm via a contagious punishment process, in which customers start

punishing whenever they see a deviation, whether the first one observed or being a punishment consequence of earlier deviations.

5. Our analysis contributes to the literature on waiting lines as social systems by showing that many of the social norms observed in queues can be sustained also on the basis of rational decision making. We contribute to the queueing literature by what appears to be the first paper studying jumping and cutting by rational customers. We also contribute to the queueing literature with strategic players by showing that welfare can be maximized by virtue of community enforcement alone in settings where types are unknown and the system manager cannot regulate the line. We contribute to the literature of repeated games with private monitoring by providing an interesting application of settings with private signals, driven by the specific operational characteristics of a waiting line, and demonstrating the ability to sustain efficient equilibria using a contagious punishment process which depends on the queueing structure.

One of the main implications of our study is that attention should be given to the possibility of endorsing such social norms when designing a service facility, specifically the queueing area. For example, signs may be displayed such as the one used by the physician in Haifa (mentioned at the beginning of the Introduction). Alternatively, one may decide not to put ropes delineating the lines as a way of allowing possible legitimate queue jumping. In particular, the studied models can be used by the system manager to decide when an intervention may be needed to improve the system performance and customer service, and when they may be able to rely just on community enforcement.

2. Literature Review

The current study lies in the intersection of three important streams of literature: First, it is motivated by studies of social norms in service systems that explain observed behavior reported in the sociology and psychology literature of social norms in queues. Second, the framework we use to model the service system is based on queueing theory and the study of delays. Third, we use game theory tools, specifically repeated games, to analyze the conflicts customers face when deciding whether to follow the norms and the strategic interaction between them. We next survey the key papers in each of these three literatures.

Queueing models with strategic customers. This literature dates back to Naor (1969), who studied a system in which strategic customers observe the length of the queue prior to making the decision whether

to join or balk. His model contains a partial conflict of interest between the self-interested customer and the interests of the social-welfare-maximizing service provider. Naor shows that pricing can be used to achieve the first-best solution. See also Lippman and Stidham Jr (1977) and references therein for extensions of these results. The follow up literature extended the analysis along multiple dimensions. One such stream studied models where the firm offers different grades of services when the customers' cost of waiting, and possibly service rate, are private information. Mendelson and Whang (1990) and Afeche (2004) considered extensions to setting in which customers make decisions based on average waiting times prior to observing the state of the system, or when the actual state of the system is unobservable. Mendelson and Whang (1990) showed that when trying to optimize the social welfare, it is incentive compatible to prioritize according to the $c\mu$ discipline and charge customers the externalities they inflict on other customer types. Afeche (2004) extended these results to study situations in which the service organization is interested in maximizing revenues and showed that the $c\mu$ policy may not be optimal, and the firm may have to resort to non-work conserving strategies to maintain incentive compatibility. Recently, Lennon et al. (2007) studied a related problem, in which a common resource has to be allocated repeatedly among selfish customers. We refer to Hassin and Haviv (2003) for a thorough review of the literature on queues with strategic customers.

Social norms in queues.

Queues have long been recognized in the literature as social systems with social norms, with Mann (1969) being among the first papers in this regard. Mann focused on an overnight line, in which customers are waiting to purchase a limited number of tickets for a sporting event. He described the first-come first-served priority as being both a fundamental concept of queueing and a basic principle of behavior referred to as distributing justice, as it gives preferential service to those waiting longer. In contrast, our study uses the idea of distributive justice equilibria where preferential service is given to those with high total waiting cost or lower requirements. The overnight line is best modeled using the single stage game studied in Section 4, for which we show that all equilibria result in FIFO priority. Schwartz (1975) set to discover the norms regulating sequences along which people routinely satisfy their needs in service systems. The author stated that while a considerable attention has been given to the psychology of waiting, the problem is actually of a sociological nature: “competitive allocation cannot operate without institutionalization of norms defining

the limits of legitimate action, particularly in this case with regard to the legitimacy of means of attaining the goal. The problem of allocation has to do not only with how much different persons are to be given from the finite supply of goods and services, but also with the priority in which their needs are to be satisfied.”

Following the above, Milgram et al. (1986) designed an experiment to study the response of people in queues to intrusions. The authors pointed out that one of the main characteristics of queuing environments is that the maintenance of the line depends on shared knowledge of the standard of behavior appropriate in each situation. The authors also pointed to the queue as an excellent example of how individuals build a social system. The experiment involved people intruding lines in different places in New York. In these experiments, the intruder was merely saying “Excuse me, I would like to get in here”. The authors showed that intrusions may be successful, yet very likely to be rejected. The studied system resembles one in which all customers have the same waiting cost (or, at least, there is no reason within the context of the service provided for someone to have more urgent needs than others). Thus, in such systems, the only possible sustainable priority is FIFO. In a different study, Schmitt et al. (1992) tested whether a waiting line should be viewed as a social system with norms and roles or whether the behavior in queues is better explained solely by individual personal interests. The authors showed that customers are more likely to react to illegitimate intrusions than to legitimate ones. They concluded that their study presents strong evidence that the queue constitutes a social system, specifically through inherent rules and norms defining the rights and obligations of the individual queuer. Our study reconciles the sociology point of view with that of rational decision making. Using repeated game theoretic tools we show that both views lead to the same behavior, where a welfare maximizing priority is induced when players take different roles and adhere to certain norms, that are also sustained in equilibrium. This allows the group to achieve an efficient equilibrium in situations where legitimate intrusions exist, and should be distinguished from illegitimate ones, yet cannot be verified a-priori.

Larson (1987) studied how feeling of social justice affect attitude towards queues. In particular, the authors showed that features of the queueing environment, such as social injustice and feedback regarding the magnitude of delay, influence the attitude of customers. He defined the deviations from FIFO as social

injustice, but he provided very little explanation as to why this is indeed the case. Our analysis generalizes the observations made in his study by showing that FIFO is indeed the prominent and only priority scheme that can be sustained in equilibrium in certain cases, yet in others, where a legitimate reason for queue jumping exists, more efficient priority schemes can be sustained in equilibrium. Helweg-Larsen and LoMonaco (2008) studied the overnight queue for tickets to a music event. They show that the queue is indeed a social system in which customer satisfaction is influenced by norms of procedural justice. Furthermore, deviations from these norms were upsetting even when outcomes were not influenced. This stands in line with our contagious punishment which calls for punishment of deviators even when players are not affected themselves. Oberholzer-Gee (2006) reported the results of a field experiment in which randomly selected customers are offered \$10 if they let a stranger cut in line. He showed that the higher the offer, the more likely it is that individuals let someone cut in. But while a majority agreed to wait longer, only a small minority accepted the monetary reward. Our study is motivated as well by the fact that the FIFO priority rule may be inefficient, yet we explore situations in which monetary payments are not allowed or acceptable.

Repeated games

Game theory has been long recognized as the most appropriate tool to address situations in which rational players with different interests interact. Specifically, repeated games have been recognized as a way to model the evolutions and sustainability of social norms when individuals engage in long term relationships. Following this literature, we also assume that people do not follow a norm just for the sake of it, but rather because it is in their best interest and allows them to maximize their utility. While we initially use repeated games with public monitoring, we also study the robustness of our results in cases where deviations are observed only by those close to the deviator in line. Such models are closely related to the study of games with random matching or repeated games with private monitoring. We first outline a subset of the relevant papers in the literature on repeated games with private monitoring and then discuss the relevant papers within the random matching literature. We then outline our main contribution to the literature.

The folk theorem in repeated games states that in games with perfect monitoring, every individually rational outcome can be sustained in an infinitely repeated game when the players are sufficiently patient

(see Mailath and Samuelson (2006) and the references therein for a thorough review of the literature on repeated games). Kandori (1992) studied the ability to sustain the folk theorem in settings where agents change their partners over time. He used the term community enforcement to describe cases where the punishment is carried by the community and not by the one being hurt. The mechanism he assumed for transmitting information from period to period does not exist in the settings we focus on. Yamamoto (2007) and Horner and Olszewski (2006) studied models with almost perfect monitoring. They showed that the folk theorem applies when the monitoring approaches being perfect. The conditions in both papers do not apply to the private signals in our setting due to the multiplicity of types and the fact that in our model some customers do not observe deviations when these occur.

Ellison (1994) considered the repeated prisoner's dilemma in a large population random-matching setting. He showed that cooperations can be sustained in equilibrium and be supported by a contagious punishment process. While our game is not one of random matching by itself, our punishment process echoes the one he developed. Okuno-Fujiwara and Postlewaite (1995) provided a folk theorem for random matching games while proposing a new equilibrium concept, norm equilibrium, which requires a less restrictive information assumption than common knowledge of the game. They define social norms as having two components, namely standard of behavior and a transition mapping when the standard is violated, and showed that norms allow coordination for individuals involved in conflict situations in large societies. The main differences between these papers and our model are as follows: First, our game has multiple types of players, where types are private information. Second, our game has multiple players, while all other games in this literature, but Yamamoto (2007), involve only two players. Third, the contagious punishment process we use is in the spirit of Ellison, yet, since the payoffs in each period depend on the number of players in the punishment phase, the analysis is more involved. Here we use the queueing properties to demonstrate that such contagious process is possible. Finally, in our private monitoring setting, which could be termed semi-private, one player has an accurate signal regarding a deviation, yet others do not observe any signal. This is natural in a queueing environment (e.g. bank, clinic or airports) because among the potential customer base, only the very few customers standing in the queue (on a specific day and a specific point in time) may be aware of a deviation.

3. Model

Consider a group of M customers that demand a certain service. We assume that each customer has a stream of service requests, which can be modeled as a Poisson process with rate $\frac{\lambda}{M}$. Customers are assumed to be a-priori identical. A customer obtains a value v at the end of each service and incurs a waiting cost during the time spent in the system. We assume that customers may experience different levels of urgency, depending on the circumstances. For example, in a customer service environment, the waiting cost rate will be higher in instances where the problem is urgent. In a retail banking environment, high opportunity costs may result from customers waiting to make a critical money transfer. In other settings, such as health care, customers may differ in the amount of work required to complete their request: some may only have a question or need a simple referral, and others may require a thorough check-up. Therefore while all customers are a-priori identical, service requests upon each arrival of a customer to the system may vary in terms of waiting cost rates and expected service rates. For tractability we assume two request types, denoted by H and L , and a non-degenerate Bernoulli distribution over these request types, with corresponding probabilities α_t for type t , which we also denote by α and $1 - \alpha$, respectively. Denote by c_H, c_L and μ_H, μ_L the corresponding waiting cost rates and expected service rates. Without loss of generality we label the request types so that H represents a higher expected service cost rate than L , i.e. $c_H\mu_H > c_L\mu_L$.

Upon arrival a customer may choose one of the following two options: join the end of the queue, or decide to cut the line while possibly providing an excuse². We assume that the request type is private information and can be only verified ex-post, i.e. after customers complete their service and leave the system. For example, when waiting in the line for airport security checkups, a flyer may ask to cut the queue saying that their flight is about to depart; if customers are not sincere, they would be observed in the terminal after other customers completed their service. Similarly, a customer claiming to have only a short question is detected to misrepresent his request type if he spends a significant time with the service provider. A customer in the queue can react in two possible ways to an attempt of an arriving customer to cut the line: either accept the cutting or reject it. We assume that rejecting intruders or being rejected do not carry any explicit costs. The

² We assume that customers do not balk because the service is essential. For example, passengers must go through airport security and patients must see the physician. Allowing for balking will not change the results significantly.

only consequence of rejection is that the intruding customer must then join the end of the line. We discuss the impact of such costs and the robustness of the results in the Discussion section.

The goal of this paper is to explore conditions under which various social norms with respect to queue joining or cutting may evolve and be sustained in equilibrium. Furthermore, we study under what conditions an efficient allocation of resources may arise without intervention of the system manager. Each of these questions will be first answered for a simple setting in which (i) customers' behavior history is publicly observable, and (ii) customer strategies do not depend on the state of the queue (this simple setting is analyzed both as a single stage game in Section 4 and as a repeated game in Section 5). We will then refine both assumptions to allow for queue-dependent strategies (Section 6) and private monitoring of past actions (Section 7).

4. Basic Model: Single Stage Game

In the single stage game, it is easy to show that the only equilibrium that can be sustained is one where all requests to cut the line, if any, are rejected, thus each arriving customer is placed at the end of the line. To formally state the result, denote the two possible actions for an arriving customer, i.e. joining the end of the line, and cutting/pushing, by J and P , respectively. We assume that a cutting attempt, when one occurs, begins with the customer at the end of the queue and proceeds sequentially to the next customer in front, with the intention of reaching the top of the queue. As a queue-incumbent, a customer may then either reject all attempts to cut the line, denoted by R , or accept any attempt, denoted by A . A queue-cutting attempt ends immediately after the first rejection by any customer (in our model, it will never be optimal for a customer to end a cutting attempt before the first rejection, or to accept some cutting attempts and reject others). For a cutting attempt to be beneficial, at least one incumbent must accept it. The full strategy of customer $i \in \{1, \dots, M\}$ is therefore denoted by $(E^i, I^i) \equiv (E_H^i E_L^i, I_H^i I_L^i)$, where $E_t^i \in \{J, P\}$ and $I_t^i \in \{R, A\}$ are the actions chosen when the customer request is of type $t \in \{H, L\}$.

Theorem 4.1 *A strategy profile where each customer i chooses $I^i = RR$ forms an equilibrium. The queueing dynamics that arise are equivalent to a queue in which customers are served in a First In First Out manner. Any other strategy profile in the single stage game does not form an equilibrium.*

Since request types are not observable until departure, any equilibrium that relies on giving priority to customers according to their type cannot be supported in the single stage game. This intuitive result is consistent with the observation made by Mann (1969) and others in the sociology literature that in overnight lines, the only normative behavior is to join the line according to the order of arrival. It is important to note that all cutting attempts will be rejected regardless of the specifics of the procedure by which people attempt to cut the line (e.g., whether they make several attempts or only one, and irrespective of the queue position at which they start their cutting attempt).

As discussed in the Introduction, many social norms are sustained due to community enforcement. To best study situations where such enforcement is possible, we will next consider settings in which the game is being played repeatedly.

5. Basic Model: Repeated Game with Perfect Public Monitoring

Consider a setting in which the M players require the service repeatedly. Following Cachon and Feldman (2008), we assume that M is large so that each customer is not likely to have two concurrent requests. Periods are defined such that there is a clear separation between the time required to complete the service and the length of time between periods. For example, customers fly once a month, while waiting for airport security takes on average 20 minutes. Similarly, customers may visit their branch of bank once in a quarter/month, yet spend a few minutes in line. This assumption guarantees that a customer's choice in any period affects the current period waiting cost in a way that is separable from its effect on future period discounted waiting costs. Future period payoffs and waiting costs are discounted according to a per period discount factor $\delta \in (0, 1)$.

To analyze this setting first note that a repetition of the FIFO-inducing strategy profile described for the single stage game is always an equilibrium in the repeated game. However, it is clearly inefficient from a social perspective given the distribution of expected service cost rates. The main goal of this section is to provide conditions under which an equilibrium may arise which induces self enforcing, efficient $c\mu$ priority queuing dynamics, thus providing justification for several of the examples discussed in the Introduction. That is, customers who have requests with a high expected service cost rate (H) cut the line upon arrival

to the system so that they start queueing immediately after the last customer with H -type requests, and customers who have requests with low expected service cost rate (L) join the end of the line and allow all H customers to cut the queue. This occurs despite the fact that the request type cannot be verified until after the completion of the service and departure from the system.

To demonstrate that such behavior can arise in equilibrium under certain conditions, we design a strategy profile incorporating grim trigger actions, i.e. punishing actions that are triggered once some customer deviates. Toward this end, it is important to note that the only possible punishment is a FIFO-inducing strategy, whereby all customers reject any queue-cutting behavior, consequently each arriving customer is placed at the end of the line. This is true because the punishment strategies must form an equilibrium of the single stage game, which induces FIFO by Theorem 4.1. Note also that, in particular, any form of punishment inflicting different treatments to different types of requests requires the punisher to be able to distinguish between these types, which is assumed impossible. Moreover, priority rules such as LIFO or random orders cannot be used as punishments as they require customers to let others cut the line against their interest.

We assume that the punishment strategies are anonymous, i.e. punishment strategies do not target a specific customer. It is also important to note though that when a $c\mu$ priority can be sustained in equilibrium, it can be done regardless of whether the punishment is collective or whether it targets the deviating customer, i.e. whether the strategies are anonymous or not.

Before turning to state the result, we need to introduce some notations. Denote by W_t^P and W_t^J the expected waiting time in the system experienced by a type $t \in \{H, L\}$ customer when pushing and joining the end of the line, respectively, and when all other customers follow the $c\mu$ -inducing strategy (PJ, RA) (i.e., type H jumping the queue to after the last H customer and rejecting other customers cutting requests, and type L joining the end of the queue and accepting cutting requests). Also, let $V^{c\mu}$ be the long term expected discounted utility when all customers follow the $c\mu$ -inducing strategy in all periods. Let V^{FIFO} be the long term expected discounted payoff when each customer i follows a FIFO-inducing strategy in each period, i.e. $I^i = RR$.

Theorem 5.1

(i) *The strategy in which each customer i chooses (PJ, RA) if this choice was maintained by all customers in all previous periods, and punishes by choosing $I^i = RR$ otherwise, is an equilibrium if and only if*

$$\frac{\delta}{1-\delta} \geq \frac{c_L (W_L^J - W_L^P)}{\alpha c_H \left(D^{FIFO} + \frac{1}{\mu_H} - W_H^P \right) + (1-\alpha) c_L \left(D^{FIFO} + \frac{1}{\mu_L} - W_L^J \right)}, \quad (1)$$

where

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

is the delay in the queue under FIFO priority.

(ii) *If $\mu_H = \mu_L$, (1) simplifies to*

$$\frac{\delta}{1-\delta} \geq \frac{c_L}{c_H - c_L} \frac{1}{\alpha(1-\alpha)}. \quad (2)$$

The theorem shows when an efficient priority scheme can be sustained in equilibrium as a result of customers' willingness to forgo the immediate benefits from cutting the line for the future benefits of enjoying a reduced expected service cost when they need it, i.e. when their waiting cost rate is high or when they require quick resolutions. In particular, this happens when the waiting cost rate differential between the two types of requests is sufficiently large, as this implies a high future relative benefit. We can also observe that for such an equilibrium to be sustained, customers should have a significant likelihood of having both types of request. To see this note that the future benefit from reduced expected service time is influenced by two opposing forces: on the one hand, a low likelihood of high-type reduces the likelihood of benefitting from receiving high priority, but on the other hand a high likelihood of high-type reduces the realized benefit, as the relative advantage compared to other customers is diminished. Therefore, it is the intermediate values of high-type likelihood that lead to the highest future benefits, implying that the likelihood of both types should be significant. Consequently, when all the factors described above leading to high future benefits are weak, an efficient equilibrium can be sustained only if the customers are very patient (i.e. δ is very high). Furthermore, for any combination of cost parameters and fraction of high-type requests, the efficient equilibrium can be sustained as long as customers are sufficiently patient.

When a $c\mu$ inducing equilibrium is sustainable, the queueing dynamics are identical to those in which the system is managed with two separate queues: one queue for high-type customers and one for low-type customers. Each line is managed according to a FIFO priority, yet when the server becomes available, customers from the low-type queue will be served only when the high-type queue is empty.

Finally, in understanding the impact of the equilibrium strategies, note that maximization of social welfare by matching between high priority and high expected service cost rate is supported by the threat of punishment by a priority rule under which customers with high- or low-type requests are indistinguishable.

6. Queue Length Dependent Strategies

Until this point in the paper, we studied settings in which customers make decisions regarding cutting and joining the queue a-priori before fully realizing the exact state of the queue. In particular, we showed that under certain conditions, an efficient priority regime can be sustained in equilibrium without the intervention of a service manager or the ability to observe the identity of the different customers. We next explore whether such a regime can be sustained when queue length dependent strategies are accounted for. Furthermore, we study in this section whether other priority schemes, beyond FIFO and $c\mu$, are sustainable in equilibrium. We show that both questions are related. In particular, when characterizing queue length dependent equilibrium priority schemes, we identify FIFO as the worst priority scheme that can be obtained by self-interested customers. To some extent, just as we show that welfare maximizing priority schemes may arise in equilibrium using community enforcement, we also bound the negative impact that selfish behavior can have on the system.

In answering which priority schemes are sustainable, we can argue that regimes that depend on randomization between absolute priority schemes cannot be supported in equilibrium. This is true because under such regimes it would be very hard to distinguish between a customer that misrepresented their type and a customer that randomized.

Since randomization among absolute priority rules cannot be supported in equilibrium, we concentrate on queue length dependent priority schemes whereby customers decide whether to push or join the line upon arrival to the system, depending on the observed queue length. Denote by q_t the number of type

$t \in \{H, L\}$ customers observed upon arrival of a customer to the system (without the arriving customer), and let $q = (q_H, q_L)$. Note that despite our assumption that types are not observed before customers leave the system, the customers' behavior in equilibrium reveals their type. Thus an arriving high-type customer can know q by observing his place in the queue following a cutting attempt. We say that a priority scheme is generated by queue length dependent threshold strategies if a customer's strategy gives priority to one class over the other only when the queue length is below some pre-specified threshold \bar{q} , i.e. if for each customer i , $E^{i,\bar{q}}(q) = PJ$ and $I^{i,\bar{q}}(q) = RA$ when $q \equiv (q_H, q_L) \leq (\bar{q}_H, \bar{q}_L)$, and $I^{i,\bar{q}}(q) = RR$ otherwise. We denote the arising priority scheme by $c\mu\bar{q}$. Similarly to the notation used in the previous section, denote by $W_L^{P,c\mu\bar{q}}$ and $W_L^{J,c\mu\bar{q}}$ the expected waiting time in the system experienced by a type L customer when pushing and joining the end of the line, respectively, and when all other customers follow the $c\mu\bar{q}$ strategy. Also, let $V^{c\mu\bar{q}}$ denote the long term expected discounted utility when all customers follow the $c\mu\bar{q}$ inducing strategy in all periods.

We next show that for such systems, our result regarding the ability to sustain priority schemes more efficient than FIFO using community enforcement continues to hold.

Theorem 6.1 *Any priority scheme generated by queue length dependent threshold strategies is sustainable in equilibrium for sufficiently patient customers (sufficiently large δ).*

The above theorem shows that for low δ that do not permit sustaining $c\mu$, other, less efficient priority schemes are always sustainable, induced by queue-dependent strategies $c\mu\bar{q}$ with a sufficiently low threshold \bar{q} . Among these regimes, the most efficient one becomes closer to $c\mu$ in terms of social welfare as customers become more patient. Note that all of these equilibria are more efficient than FIFO. In Appendix B we derive a sufficient condition on δ for sustaining the $c\mu\bar{q}$ strategies priority schemes in equilibrium. One can also generalize the theorem by showing that for sufficiently high δ it is possible to sustain equilibrium strategies in which for every q_L there exists a threshold $\bar{q}_H(q_L)$ such that high-type customers push whenever the queue length is below this threshold (as long as the threshold $\bar{q}_H(q_L)$ is uniformly bounded from above). Note also that, using similar arguments, one can show that strategies that prioritize only when the queue is long are also sustainable in equilibrium.

7. Repeated Game with Private Monitoring

A main assumption made in previous sections was of perfect public monitoring, i.e. once a customer deviates, all customers know it. This assumption may not be realistic in some settings. We now turn to a model with private monitoring. In this model, a deviating customer is detected by only one other customer located next in line, for whom a public announcement to all other customers at once is impossible. We establish conditions under which an efficient equilibrium is still sustainable. Note that it is natural to assume in our setting that only a small subset of the entire population observes a deviation when one occurs, and this subset cannot immediately inform all other customers (in the Conclusions section we discuss settings in which more than one customer observe a deviation).

We will construct an equilibrium in which each customer i 's strategy has the following two stages, between which the play may keep alternating.

Stage 1 Play $c\mu$, i.e. $(E^i, I^i) \equiv (PJ, RA)$, if no type L deviation from this choice was observed by the customer at the end of the previous period; Otherwise switch to Stage 2.

Stage 2 Punish by playing $I^i = RR$; If a public signal is observed, then switch back to Stage 1 in the next period; Otherwise, remain in Stage 2.

Note that in the efficient Stage 1, the two customer types act differently and are treated differently. In particular, observing a customer with type H request will always lead to continuing in Stage 1 in the next period, while observing an ex-post deviation by a customer with type L request will lead to a switch to Stage 2. In the punishment Stage 2 however, both types play the same strategy. Note also that by definition, the public signal, as used in Ellison (1994), must be observed simultaneously by all customers, thus allowing all customers to coordinate switching back to Stage 1. The signal itself may or may not have any substantive meaning in the context of the queue, so for example it could be the hire of a new employee serving the customers, but it could also be the event of a sunny day, etc. We assume that the probability of the public signal occurring in a period is the same for all periods and is commonly known among all customers. Note that a setting in which there is no such mechanism is a special case. The dynamics implemented by these equilibrium strategies is analogous to ideas in the literature on repeated games with private monitoring, where punishments are contagious and develop gradually. A customer who observes a deviation joins those

who punish, regardless of whether he observed an actual deviation from stage 1 or a punishment, i.e., a customer playing stage 2.

Before turning to the construction, we need to introduce some notation. Denote by ϕ the probability of not observing a public signal in a given period. Let $W_t^P(k)$ be the waiting time of a customer i of type $t \in \{H, L\}$ who plays $E_t^i = P$ and $I_t^i = R$ when starting the period at Stage 2 together with $k \in \{1, \dots, M-1\}$ other customers also at Stage 2 (i.e., already observed a deviation), where the remaining customers are still at Stage 1. Similarly, let $W_t^J(k)$ be the waiting time of a customer i of type $t \in \{H, L\}$ who plays $E_t^i = J$ and $I_t^i = A$ when k other customers are at Stage 2. Let $U^P(k) = v - \alpha c_H W_H^P(k) - (1 - \alpha) c_L W_L^P(k)$ be the Stage 2 ex-ante immediate payoff of a customer i who plays $E_t^i = P$ and $I_t^i = R$ when k other customers are at Stage 2. We make the following two assumptions.

Assumption 7.1 $U^P(k)$ is decreasing in k .

Assumption 7.2 $W_t^J(k) - W_t^P(k)$ is increasing in k for all $t \in \{H, L\}$.

Assumption 7.1 is natural as it says that the Stage 2 ex-ante immediate payoff of a customer who cuts the line decreases with the number of other customers also at Stage 2. Assumption 7.2 says that the gain from cutting the line is increasing the more other customers are also doing so. These assumptions are satisfied, for example, for M/M/1 queues with equal service rates for both types.

Proposition 7.1 *If the queue can be described as M/M/1 and $\mu_H = \mu_L$, then Assumptions 7.1 and 7.2 are satisfied.*

We are now ready to state the main result of this section.

Theorem 7.1 *Under assumptions 7.1 and 7.2, the strategy profile where each customer plays the two stage strategy described above forms an equilibrium if*

$$(I) \quad \frac{c_L(W_L^J(0) - W_L^P(0))}{V^{c\mu}(1 - \delta) - U^P(1)} \leq \frac{\delta}{1 - \delta\phi},$$

$$\text{and (II)} \quad \frac{c_L(W_L^J(1) - W_L^P(1))}{U^P(1) - U^P(M-1)} \geq \frac{\delta\phi}{1 - \delta\phi\alpha}.$$

For the $M/M/1$ queue with $\mu_H = \mu_L = \mu$, these conditions simplify to

$$\frac{\frac{\frac{\lambda}{\mu-\lambda} \frac{c_L}{\mu-\alpha\lambda}}{\frac{\alpha c_H + (1-\alpha)c_L}{\mu-\bar{\alpha}(1)\lambda} - \frac{\alpha(c_H-c_L)}{\mu-\alpha\lambda} - \frac{c_L}{\mu-\lambda}}}{\frac{\lambda}{\mu-\lambda} \frac{c_L}{\mu-\bar{\alpha}(1)\lambda}} \leq \frac{\delta}{1-\delta\phi} \text{ and}$$

$$(\alpha c_H + (1-\alpha)c_L) \left[\frac{1}{\mu-\bar{\alpha}(M-1)\lambda} - \frac{1}{\mu-\bar{\alpha}(1)\lambda} \right] \geq \frac{\delta\phi}{1-\delta\phi\alpha},$$

where $\bar{\alpha}(k) \equiv \alpha + (1-\alpha)\frac{k}{M}$.

The main implication of this theorem is that an efficient priority scheme can be sustained in equilibrium even when public monitoring is not possible. It is important to note that for an efficient equilibrium to be sustained, the customer's patience must be of intermediate level: while condition (I), for Stage 1, requires customers to be sufficiently patient, a too high level of patience may hinder the ability to sustain condition (II) thus making the punishment too weak. The latter may occur because a customer may prefer not to punish to avoid eventually reaching the low payoffs of FIFO dynamics. Note also that the rate of contagion depends on the proportion of each type – the higher the proportion of high-type requests, the slower the contagion since only low-type customers can ‘infect’ others.

Condition (II) in the theorem is the most compact sufficient condition ensuring that, irrespective of the number of customers in Stage 2, a customer in Stage 2 prefers to punish. As constructed in the proof, this condition can be replaced by $M-1$ conditions, one for each state $k \in \{1, \dots, M-1\}$, each ensuring punishment continuation given the number k of customers in Stage 2.

Interestingly, unlike most repeated games (even with private monitoring) in which every individually rational solution can be supported by an appropriate scheme for some discount rates, this is not the case here. This is due to the specific nature of the queueing game, which entails different payoffs to the players along the punishment phase. In particular, during the early steps of Stage 2, when only a few customers punish, the punishment is not too severe, thus making the impact of punishment weaker in comparison to the case of perfect monitoring.

7.1. Numerical Illustration

We now turn to a numerical study of the above results when players engage in a repeated game with private monitoring. Our goal is to illustrate scenarios where an efficient equilibrium exists, but also to demonstrate

cases in which it does not exist and discuss the reasons. Consider a service with an aggregate arrival rate of $\lambda = 0.5$ customers per unit of time. Service times are exponentially distributed with mean $1/\mu = 1$. Assume that there are $M = 10$ customers and thus each customer arrives with requests at a rate of 0.05 requests per period. Note that even though M is not too large, λ is sufficiently small to justify our assumption that each customer is not likely to have two concurrent requests. Further assume that customers receive value of \$100 from the service and that the public randomization occurs with $\phi = 0.9$. We outline several examples, demonstrating the effect of varying the waiting costs and the customer discount factor. Table 1 is derived based on the computations outlined in the proofs of Proposition 7.1 and Theorem 7.1.

Consider Example I, where $c_H = \$10$ per unit of time, and $c_L = \$0.2$ per unit of time. As shown in Table 1, one can sustain an equilibrium that induces a $c\mu$ priority scheme when δ around 0.6, and it cannot be sustained in higher or lower levels. When the patience level is at 0.5 and below, the first stage condition is violated and thus customers with low-type requests have incentive to deviate and not to adhere to the $c\mu$ priority rule. When the patience levels are too high, as discussed above, Stage 2 becomes more difficult to sustain since customers who are forward looking prefer to stop the punishment contagion as the utility from future periods is diminishing.

When the cost of waiting for high-type requests is cut in half to $c_H = \$5$ per unit of time, (see Example II) the same behavior is observed, yet for an efficient equilibrium to be sustained the patience levels have to be higher than before. The intuition is that with lower cost for high-type requests, the difference between the type H and type L costs decreases, making the punishment less severe.

Similarly, when the service rate is doubled to $\mu = 2$ (see Example III), the set of patience levels over which $c\mu$ can be sustained through community enforcement is widened. The intuition here is that when the utilization drops it is easier to sustain the contagious punishment (since customers have weaker incentives to stop it) as well as to sustain stage 1: customers have weaker incentives to break the norm when the immediate benefits are not as valuable as when the queue utilization is high.

Table 1 Sustainability of an efficient priority rule for different patience levels

δ	Example I		Example II		Example III	
	$c\mu$	Deviation	$c\mu$	Deviation	$c\mu$	Deviation
0.1	No	stage 1	No	stage 1	No	stage 1
0.2	No	stage 1	No	stage 1	No	stage 1
0.3	No	stage 1	No	stage 1	No	stage 1
0.4	No	stage 1	No	stage 1	No	stage 1
0.5	No	stage 1	No	stage 1	No	stage 1
0.6	Yes		No	stage 1	No	stage 1
0.7	No	Stage 2	No	stage 1	Yes	
0.8	No	Stage 2	Yes		Yes	
0.9	No	stage 2	Yes		Yes	
0.95	No	stage 2	No	stage 2	No	
0.99	No	stage 2	No	stage 2	No	stage 2

8. Conclusion and Discussion

8.1. Extensions

The models studied in this paper make several assumptions regarding the type of operational system and information structure. Our analysis is valid in the case where the service system can be described using a M/M/1 queue. This assumption can be relaxed while maintaining the result that efficient priority rules can be sustained in equilibrium under perfect monitoring as long as customers are sufficiently patient. One may derive analytical conditions as long as analytical formulae exist for the queueing system. Similar results will hold for the model with private monitoring as long as Assumptions 7.1 and 7.2 hold for the specific queueing system.

Another assumption that may be relaxed is one pertaining to the likelihood of detecting a deviating customer. In the models above we assumed that a deviating customer will always be detected, at least by the last customer bypassed. One may relax this assumption and assume instead that a deviation is detected only in a certain probability. For example, envision a deviating customer with service time large in expectation but short in realization. Given the observed service time, the customer standing behind will use Bayesian updating to determine the likelihood of deviation, leading to a possibly small but positive probability of concluding that no deviation has occurred. In general, this can be modeled by allowing the detection probability to be smaller than 1. This will not break the main result that efficient priority rules can be sustained in equilibrium as long as customers are sufficiently patient. A reduction in the likelihood of detection will

make deviations more profitable, thus making the efficient strategy more difficult to sustain. In other words, to sustain an efficient priority rule in equilibrium, customers will have to be more patient than in a model with accurate detection.

A different assumption, made in the model with private monitoring, is that only one customer observes the deviation. This assumption is consistent with the fact that in a queueing environment only a small subset of the entire customer base is present at any point in time, thus only a small subset may detect a deviation when one occurs. One can imagine a situation in which all customers present in line detect a deviation when one occurs. It can be shown that in this case the sufficient conditions for the $c\mu$ to be sustained in equilibrium are weaker because the punishment process evolves faster towards FIFO when a deviation occurs, thus making the deviations less beneficial.

Finally, assume that the arriving customers belong to different societies. While all customers may have high and low-type requests, each society is characterized by a different urgency level, i.e. the likelihood of having a high-type request may be different (which affects also the ex-ante expected service cost rate). For simplicity, suppose that the societies are identical in all other respects, including the service requirements. In this case, one can show that all of the results of the paper continue to hold with adjusted parameters. In particular, a $c\mu$ equilibrium will be sustainable under conditions similar to those provided for ex-ante homogenous customers, with the only modification being that the proportion of high-type requests must take into account the variance among the societies. It follows that members of some societies would tend to cut lines more often than others, because these members are more likely to have high-type requests (low service content or high waiting costs).

8.2. Discussion

The paper is motivated by the empirical observation that in many cases some customers cut in line and others allow such behavior. The main goal of the paper is to demonstrate that such behavior can be explained on rational individual grounds, while being socially and mutually beneficial to all sides in the long run. Our main results (Theorems 5.1, 6.1 and 7.1) state that under certain conditions such behavior may arise in equilibrium. These results do not preclude the existence of other equilibrium priority schemes, in particular

a no-cutting priority where all customers are served according to the order of their arrival. Instead, our results should be interpreted as saying that queue cutting may arise as a rational and desirable social norm.

Waiting lines vary in their norms pertaining to whether queue jumping is legitimate and allowed. We show that when players engage in a repeated game, queue jumping may be part of the social norms in queues irrespective of whether monitoring is perfect or not and regardless of whether the strategies are queue dependent or not. Under perfect monitoring, the more patient the customers are, the more likely such norms may be sustained in equilibrium. When only private monitoring is possible, a more refined retaliation process is required, leading to the necessity of intermediate customer patience levels. These results show that community enforcement can lead to norms that allow occasional queue jumping, a value enhancing outcome in the long run due to the ability of customers with urgent needs or low work content to obtain priority.

In our study, we provided an economic and operational rationale for such norms. One may link these results to cultural characteristics. Hofstede's Power Distance Index (HPDI), introduced in Hofstede (1983), is a common measure of the extent to which members in a society accept and expect an unequal power distribution. In cultures with low HPDI (e.g. Australia, Austria, Denmark, Ireland, Israel, New Zealand), people expect and accept more consultative or democratic power relations, e.g. by relating to each other more as equals regardless of formal positions. In cultures with high HPDI (e.g. Malaysia), the less powerful accept autocratic or paternalistic power relations. Thus HPDI measures people's perception of power differences, rather than measuring the objective power distribution. While our model did not account for such differences, one may view α , the probability of a customer having a type H request, as a proxy for such a measure. Most of the conditions in the paper are derived from the point of view of a customer with a type L request, anticipation a future type H request with probability α . As we have seen, the more balanced the likelihood of being both types (i.e. the closer it is to 0.5), the easier it is to satisfy the conditions for $c\mu$ to be sustained in equilibrium. It is not surprising then that some of the examples discussed in the paper arise in countries with low HPDI index (such as Israel). Future work should attempt to test the above predictions empirically and experimentally.

References

- Adiri, I., U. Yechiali. 1974. Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research* **22**(5) 1051–1066.
- Afeche, P. 2004. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delaying tactics Working paper, Kellogg School of Management, Northwestern University.
- Cachon, G.P., P. Feldman. 2008. Pricing Services Subject to Congestion: Charge Per-Use Fees or Sell Subscriptions? *Operations and Information Management, The Wharton School, University of Pennsylvania* .
- Cox, D. R., W. L. Smith. 1961. *Queues*. Methuen (London) and Wiley (New York).
- Ellison, G. 1994. Cooperation in the prisoner's dilemma with anonymous random matching. *The Review of Economic Studies* **61**(3) 567–588.
- Federgruen, A., H. Groenevelt. 1988. Characterization and optimization of achievable performance in general queueing systems. *Operations Research* **36**(5).
- Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.
- Helweg-Larsen, M., B.L. LoMonaco. 2008. Queuing Among U2 Fans: Reactions to Social Norm Violations1. *Journal of Applied Social Psychology* **38**(9) 2378–2393.
- Hofstede, G. 1983. The cultural relativity of organizational practices and theories. *Journal of international business studies* **14**(2) 75–89.
- Horner, J., W. Olszewski. 2006. The folk theorem for games with private almost-perfect monitoring. *Econometrica* **74**(6) 1499–1544.
- Kandori, M. 1992. Social norms and community enforcement. *The Review of Economic Studies* **59**(1) 63–80.
- Larson, R.C. 1987. Perspectives on queues: social justice and the psychology of queueing. *Operations Research* **35**(6) 895–905.
- Lennon, CG, JM McGowan, KY Lin. 2007. A game-theoretic model for repeated assignment problem between two selfish agents. *Journal of the Operational Research Society* **59**(12) 1652–1658.
- Lippman, S.A., S. Stidham Jr. 1977. Individual versus social optimization in exponential congestion systems. *Operations Research* **25** 233–247.
- Mailath, G.J., L. Samuelson. 2006. *Repeated games and reputations: long-run relationships*. Oxford University Press, USA.
- Mann, L. 1969. Queue culture: The waiting line as a social system. *American Journal of Sociology* **75**(3) 340–354.
- Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38**(5) 870–883.
- Milgram, S., H.J. Liberty, R. Toledo, J. Wackenhut. 1986. Response to intrusion into waiting lines. *Journal of Personality and Social Psychology* **51**(4) 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**(3) 427–440.
- Okuno-Fujiwara, M., A. Postlewaite. 1995. Social norms and random matching games. *Games and Economic Behavior* **9**(1) 79–109.
- Parsons, T. 1955. *The social system*. Psychology Press.
- Schmitt, B.H., L. Dube, F. Leclerc. 1992. Intrusions into waiting lines: Does the queue constitute a social system?. *Journal of Personality and Social Psychology* **63**(5) 806.
- Schwartz, B. 1975. *Queuing and waiting*. University of Chicago Press.
- Yamamoto, Y. 2007. Efficiency results in N player games with imperfect private monitoring. *Journal of Economic Theory* **135**(1) 382–413.

Appendix A: Proofs

Proof of Theorem 4.1: Given that all customers choose $I^i = RR$, any customer is indifferent between joining or cutting the queue since a queue-cutting attempt will be rejected anyway, and there is no incentive for any type to deviate and accept an attempt to cut the line because that would increase their waiting time. Therefore such a strategy profile forms an equilibrium in weakly dominant strategies. Clearly if all cutting attempts are rejected, then the queue operates based on First In First Out. Now consider any other strategy profile, i.e. suppose that some type t of some customer i chooses $I_t^i = A$. On the one hand, if there exists some type t' of some other customer i' that chooses $E_{t'}^{i'} = P$, then customer i of type t , who has positive probability of encountering that type, prefers to deviate to $I_t^i = R$ to avoid being pushed. On the other hand, if all types t' of all customers i' other than i choose $E_{t'}^{i'} = J$, then each of them has positive probability of encountering type t of customer i and so prefers to deviate to $E_{t'}^{i'} = P$ to gain from that encounter. In either case the strategy profile is not an equilibrium. ■

Proof of Theorem 5.1: First, note that off the equilibrium path, i.e. after a deviation was observed, the threat to punish by rejecting all cutting attempts is credible. Indeed, in this case all customers are supposed to turn to rejecting all cutting attempts from this point on. Since all customers do so, a single customer has no incentive to deviate and allow cutting. Thus, once all customers start punishing, it is incentive compatible to continue doing so in all future interactions. We now turn to the equilibrium path. It is clear that customers with high-type requests have no incentive to deviate since they can only lose by joining the line at the end or accepting a cutting request. Moreover, such customers are indifferent on equilibrium path between accepting or rejecting a cutting attempt. This is true because the queue dynamics among high-type customers are FIFO after rejecting and LIFO (Last In First Out)³ after accepting, where both are work conserving priority rules resulting in identical expected workload (see Federgruen and Groenevelt (1988)), and since high-type customers are homogeneous, also in identical expected waiting time. We can thus focus on low-type requests. Such customers can hope to gain only on the equilibrium path, in which case they face a clear

³ When all customers cut the line and accept all cutting attempts, the last customer to join is the first to be served (even though a cutting attempt begins at the end of the line). A similar reasoning holds when applied only to high type customers.

trade-off. They have incentive to improve the position in the queue by misidentifying as having high-type requests in the current period, yet doing that forgoes future utilities that might have been earned again as a high-type. Note that we may ignore partial deviations in which the customer joins the line but then rejects any cutting request (J, R) , or cuts the line but then accepts other cutting requests (P, A) . This is true because a full deviation (P, R) is always preferable to a partial one (to be precise, the customer strictly prefers (P, R) to (J, R) , but is indifferent between (P, R) and (P, A) following the discussion about a type H request at the beginning of this paragraph).

A customer with a low-type request will join the end of the line and accept all cutting requests if and only if

$$v - c_L W_L^J + \delta V^{c\mu} \geq v - c_L W_L^P + \delta V^{\text{FIFO}},$$

where the left hand side is the long term expected discounted payoff when L joins the end of the queue and continues according to the $c\mu$ -inducing strategy, and the right hand side is the long term expected discounted payoff when L cuts the line, which results in triggering the FIFO-inducing strategy in all future periods. The above can be rewritten as

$$\frac{\delta}{1 - \delta} \geq \frac{c_L (W_L^J - W_L^P)}{(1 - \delta)(V^{c\mu} - V^{\text{FIFO}})}. \quad (3)$$

Since the expected waiting time for a type t customer under FIFO is $D^{\text{FIFO}} + \frac{1}{\mu_t}$ (see, e.g., Federgruen and Groenevelt (1988)), the difference between the two priority rules in terms of the per period long term expected discounted payoff is

$$\begin{aligned} & (1 - \delta)(V^{c\mu} - V^{\text{FIFO}}) \\ &= [v - \alpha c_H W_H^P - (1 - \alpha)c_L W_L^J] - \left[v - \alpha c_H \left(D^{\text{FIFO}} + \frac{1}{\mu_H} \right) - (1 - \alpha)c_L \left(D^{\text{FIFO}} + \frac{1}{\mu_L} \right) \right]. \end{aligned}$$

Substituting into (3) and simplifying leads to condition (1), which completes the proof of part (i).

We now prove part (ii). Since both expected service rates are identical (and denoted by μ), both types have identical expected waiting time when taking the same action under $c\mu$. Denote the expected waiting time when they join the queue at the end by W^J and when the push by W^P . Since both priority rules $c\mu$ and FIFO are work conserving, they provide identical ex-ante expected workload (see Federgruen and Groenevelt (1988)), and moreover identical ex-ante expected waiting time, as the expected service rates are identical. Therefore $\alpha W^P + (1 - \alpha)W^J = D^{\text{FIFO}} + \frac{1}{\mu}$. Thus the denominator of the right hand side of (1) simplifies to

$$\begin{aligned} & \alpha c_H [\alpha W^P + (1 - \alpha)W^J - W^P] + (1 - \alpha)c_L [\alpha W^P + (1 - \alpha)W^J - W^J] \\ &= \alpha c_H (1 - \alpha)(W^J - W^P) - (1 - \alpha)c_L \alpha (W^J - W^P) \\ &= \alpha(1 - \alpha)(W^J - W^P)(c_H - c_L). \end{aligned}$$

Substituting into (1), we obtain (2). ■

Proof of Theorem 6.1 : Fix a strategy $c\mu\bar{q}$, where $\bar{q} = (\bar{q}_H, \bar{q}_L)$. When the queue length is above the threshold, a push by any customer will be rejected by all customers, making the push unprofitable. A customer with a high type request has no incentive to deviate. We thus focus on low type requests. For such a priority scheme to be sustainable it must be the case that for all $q = (q_L, q_H) \leq (\bar{q}_H, \bar{q}_L)$,

$$v - c_L W_L^{J, c\mu\bar{q}}(q) + \delta V^{c\mu\bar{q}} \geq v - c_L W_L^{P, c\mu\bar{q}}(q) + \delta V^{\text{FIFO}},$$

which is equivalent to

$$c_L [W_L^{J,c\mu\bar{q}}(q) - W_L^{P,c\mu\bar{q}}(q)] \leq \delta [V^{c\mu\bar{q}} - V^{\text{FIFO}}]. \quad (4)$$

Since the expected waiting times are finite over the region $q \leq \bar{q}$ and the region itself is bounded, the left hand side of (4) is bounded over the region, and the bound is constant in δ . Moreover, the right hand side is unboundedly increasing in δ because it equals $\frac{\delta}{1-\delta}$ times the positive difference between the stage game payoff under $c\mu\bar{q}$ and FIFO. Thus, there exists $\bar{\delta} \in (0, 1)$ such that for all $\delta \in (\bar{\delta}, 1)$, (4) is satisfied for all $q \leq \bar{q}$. ■

Proof of Proposition 7.1: For an $M/M/1$ queue with service rate μ we have

$$U^P(k) = v - (\alpha c_H + (1 - \alpha)c_L)W^P(k).$$

Upon arrival to the system, a pushing customer jumps in front of all other customers in the queue that do not push, thus we can disregard their contribution to the arrival rate. The effective arrival rate observed by a pushing customer is therefore $\bar{\alpha}(k) = \alpha + (1 - \alpha)\frac{k}{M}$, therefore $W^P(k) = \frac{1}{\mu - \bar{\alpha}(k)\lambda}$. Since $\bar{\alpha}(k)$ is increasing in k , so is $W^P(k)$, therefore $U^P(k)$ is decreasing in k , completing the proof for Assumptions 7.1. For the proof for Assumption 7.2, note that since the ex-ante expected waiting time in the system, $\frac{1}{\mu - \lambda}$, equals $\bar{\alpha}(k)W^P(k) + [1 - \bar{\alpha}(k)]W^J(k)$, we can solve for $W^J(k)$ and find that

$$\begin{aligned} W^J(k) - W^P(k) &= \frac{1}{1 - \bar{\alpha}(k)} \left[\frac{1}{\mu - \lambda} - \frac{\bar{\alpha}(k)}{\mu - \bar{\alpha}(k)\lambda} \right] - \frac{1}{\mu - \bar{\alpha}(k)\lambda} \\ &= \frac{1}{1 - \bar{\alpha}(k)} \left[\frac{1}{\mu - \lambda} - \frac{1}{\mu - \bar{\alpha}(k)\lambda} \right] \\ &= \frac{\lambda}{\mu - \lambda} \frac{1}{\mu - \bar{\alpha}(k)\lambda}, \end{aligned}$$

which implies Assumption 7.2 because $\bar{\alpha}(k)$ is increasing in k . ■

Proof of Theorem 7.1: Denote by $f(k|H)$ (respectively, $f(k|L)$) the expected discounted continuation payoff for a type H (respectively, L) customer when $k \in \{1, \dots, M - 1\}$ other customers are at Stage 2. Let $f(k) = \alpha f(k|H) + (1 - \alpha)f(k|L)$ be the corresponding ex-ante payoff. We analyze the incentives of a customer at each stage.

Stage 1: Clearly a customer with a high-type request has no incentive to deviate. A customer i with type L request prefers playing $E_L^i = J$ and $I_L^i = A$ when

$$v - c_L W_L^J(0) + \delta V^{c\mu} \geq v - c_L W_L^P(0) + \delta f(1),$$

which is equivalent to

$$\delta(V^{c\mu} - f(1)) \geq c_L(W_L^J(0) - W_L^P(0)). \quad (5)$$

The analysis of Stage 2 will allow us to compute $f(1)$, which will then be used to show that condition (I) is sufficient for (5).

Stage 2: We analyze the evolution of the punishment process depending on the customer request type to verify that it is subgame perfect. When the proposed strategies form an equilibrium, the expected discounted continuation payoff is

$$f(k|H) = v - c_H W_H^P(k) + \delta(\phi f(k) + (1 - \phi)V^{c\mu}), \text{ and}$$

$$f(k|L) = v - c_L W_L^P(k) + \delta(\phi f(\min\{k+1, M-1\}) + (1-\phi)V^{c\mu}).$$

Therefore we have

$$f(k) = U^P(k) + \alpha\delta\phi f(k) + (1-\alpha)\delta\phi f(\min\{k+1, M-1\}) + \delta(1-\phi)V^{c\mu}. \quad (6)$$

Yet, for this to hold it must be that for all $k = 1, \dots, M-1$, a customer i with type L request will opt to play $E_L^i = P$ and $I_L^i = R$, i.e.

$$v - c_L W_L^P(k) + \delta(\phi f(\min\{k+1, M-1\}) + (1-\phi)V^{c\mu}) \geq v - c_L W_L^J(k) + \delta(\phi f(k) + (1-\phi)V^{c\mu}),$$

which is equivalent to

$$c_L(W_L^J(k) - W_L^P(k)) \geq \delta\phi(f(k) - f(\min\{k+1, M-1\})). \quad (7)$$

Consider first the case where all other customers are at Stage 2, i.e. $k = M-1$. Inequality (7) holds because the customer's action cannot increase the number of other customers at Stage 2 in the next period, and because $W_L^J(M-1) > W_L^P(M-1)$. Moreover, (6) simplifies to

$$f(M-1) = U^P(M-1) + \delta\phi f(M-1) + \delta(1-\phi)V^{c\mu}.$$

Solving for $f(M-1)$ and letting $\bar{U}^P(M-1) = U^P(M-1)$, we have

$$f(M-1) = \frac{1}{1-\delta\phi} [\bar{U}^P(M-1) + \delta(1-\phi)V^{c\mu}].$$

Consider now the induction hypothesis that $f(k) = \frac{1}{1-\delta\phi} [\bar{U}^P(k) + \delta(1-\phi)V^{c\mu}]$ for $k = M-1, M-2, \dots, M-l$ and $1 \leq l \leq M-1$, where $\bar{U}^P(k) \equiv \frac{1-\delta\phi}{1-\delta\phi\alpha} U^P(k) + \frac{\delta\phi(1-\alpha)}{1-\delta\phi\alpha} \bar{U}^P(k+1)$. Note that $\bar{U}^P(k)$ is a convex combination of $U^P(k)$ and $\bar{U}^P(k+1)$, thus a convex combination of $U^P(k), U^P(k+1), \dots, U^P(M-1)$. By the induction hypothesis and using (6) for $k = M-(l+1)$, we have that

$$\begin{aligned} f(k) &= \frac{1}{1-\delta\phi\alpha} [U^P(k) + \delta\phi(1-\alpha)f(k+1) + \delta(1-\phi)V^{c\mu}] \\ &= \frac{1}{1-\delta\phi\alpha} \left[U^P(k) + \frac{\delta\phi(1-\alpha)}{1-\delta\phi} [\bar{U}^P(k+1) + \delta(1-\phi)V^{c\mu}] + \delta(1-\phi)V^{c\mu} \right] \\ &= \frac{1}{1-\delta\phi\alpha} \left[U^P(k) + \frac{\delta\phi(1-\alpha)}{1-\delta\phi} \bar{U}^P(k+1) + \delta(1-\phi)V^{c\mu} \left(\frac{\delta\phi(1-\alpha)}{1-\delta\phi} + 1 \right) \right] \\ &= \frac{1}{1-\delta\phi} \left[\frac{1-\delta\phi}{1-\delta\phi\alpha} U^P(k) + \frac{\delta\phi(1-\alpha)}{1-\delta\phi\alpha} \bar{U}^P(k+1) + \delta(1-\phi)V^{c\mu} \right] \\ &= \frac{1}{1-\delta\phi} [\bar{U}^P(k) + \delta(1-\phi)V^{c\mu}] \end{aligned}$$

for all $k \in \{1, \dots, M-1\}$. Using the above,

$$f(k) - f(k+1) = \frac{1}{1-\delta\phi} [\bar{U}^P(k) - \bar{U}^P(k+1)] = \frac{1}{1-\delta\phi\alpha} [U^P(k) - \bar{U}^P(k+1)].$$

Since, by Assumption 7.1, $U^P(k)$ is decreasing in k ,

$$U^P(k) \geq \bar{U}^P(k) \geq \bar{U}^P(k+1).$$

Therefore

$$\frac{1}{1-\delta\phi\alpha} [U^P(k) - U^P(M-1)] \geq f(k) - f(k+1).$$

Thus for all $k = 1, \dots, M - 1$, a sufficient condition for (7) is

$$c_L(W_L^J(k) - W_L^P(k)) \geq \frac{\delta\phi}{1 - \delta\phi\alpha} [U^P(k) - U^P(M - 1)]. \quad (8)$$

Under Assumptions 7.1 and 7.2, since the difference $W_L^J(k) - W_L^P(k)$ is increasing in k and $U^P(k)$ is decreasing in k , (II) is a sufficient condition for (8). It remains to show that condition (I) is sufficient for (5). Note that using our derivation of $f(1)$ we have

$$\delta(V^{c\mu} - f(1)) = \frac{\delta}{1 - \delta\phi} [V^{c\mu}(1 - \delta) - \bar{U}^P(1)].$$

Since $U^P(k)$ is decreasing, (5) is implied by

$$\frac{\delta}{1 - \delta\phi} [V^{c\mu}(1 - \delta) - U^P(1)] \geq c_L(W_L^J(0) - W_L^P(0)),$$

establishing the sufficiency of (I).

The simplification of conditions (I) and (II) for the M/M/1 queue with $\mu_H = \mu_L = \mu$ is derived using the same substitutions as in the proof of Proposition 7.1. ■

Appendix B: Queue-Dependent Strategies

In this Appendix we derive a sufficient condition on δ for sustaining in equilibrium the priority scheme generated by the $c\mu\bar{q}$ strategies according to the analysis presented in Section 6. Assume that M is large enough so that the arrival process can be approximated by a Poisson process with rate $\frac{\lambda}{M}$. Our analysis echoes the one done by Adiri and Yechiali (1974). Since we are interested in a sufficient condition, we can use appropriate bounds on the two sides of (4) and use monotonicity to show that only a region boundary condition is relevant. Denoting $\underline{\mu} \equiv \min\{\mu_H, \mu_L\}$ and $\bar{\rho} \equiv \frac{\lambda}{\underline{\mu}}$, an upper bound on the expected waiting time of a customer with type L request is $W_L^{J,c\mu\bar{q}}(q) \leq \frac{q_H + q_L + 1}{\underline{\mu}} \frac{\bar{\rho}}{1 - \bar{\rho}}$, because for a low priority customer who joins the system with n customers, an upper bound for the expected waiting time is n times the length of one busy period, where the latter is bounded from above by $\frac{\bar{\rho}}{\underline{\mu}(1 - \bar{\rho})}$, the length of one busy period in a M/M/1 queue. Similarly, a lower bound $\underline{V}^{c\mu\bar{q}}(q)$ on the long term expected discounted payoff of playing $c\mu\bar{q}$, can be achieved by taking the above upper bound instead of the expected waiting time of a type L customer. Since the expected waiting time of a pushing L -type customer is $\frac{q_H}{\mu_H} + \frac{1}{\mu_L}$, we can conclude that a sufficient condition for (4) is $\frac{q_H + q_L + 1}{\underline{\mu}} \frac{\bar{\rho}}{1 - \bar{\rho}} - \frac{q_H}{\mu_H} - \frac{1}{\mu_L} \leq \delta (V^{c\mu\bar{q}}(q) - V^{\text{FIFO}})$. Since the above has to hold for all $q \leq \bar{q}$, we have the following sufficient condition.

$$\begin{aligned} \delta &\geq \left[\frac{\bar{q}_L}{\underline{\mu}} \frac{\bar{\rho}}{1 - \bar{\rho}} + \bar{q}_H \left(\frac{1}{\underline{\mu}} \frac{\bar{\rho}}{1 - \bar{\rho}} - \frac{1}{\mu_H} \right) + \left(\frac{1}{\underline{\mu}} - \frac{1}{\mu_L} \right) \right] \frac{1}{V^{c\mu\bar{q}}(q) - V^{\text{FIFO}}} \text{ if } \frac{\bar{\rho}}{1 - \bar{\rho}} > \frac{\mu}{\mu_H}, \text{ and} \\ \delta &\geq \left[\frac{\bar{q}_L}{\underline{\mu}} \frac{\bar{\rho}}{1 - \bar{\rho}} + \left(\frac{1}{\underline{\mu}} - \frac{1}{\mu_L} \right) \right] \frac{1}{V^{c\mu\bar{q}}(q) - V^{\text{FIFO}}} \text{ otherwise.} \end{aligned}$$