Gainers and Losers in Priority Services

Alex Gershkov and Eyal Winter*

November 26, 2022

Abstract

We analyze the implications of introducing priority service on customers' welfare. In monopoly markets, introducing priority service can often decrease consumer surplus. This negative effect exists despite an increase in efficiency gains. In other words, the monopolist extracts from customers an aggregated payment higher than the total efficiency gain generated by the service and hence leaves customers worse off than when no priority is offered at all.

In duopoly markets with homogeneous customers the price competition over priority service is blocked, i.e., the equilibrium price and customers' welfare coincides with the monopoly outcome where this monopolist faces half of the market. With heterogeneous customers as well priority can reduce the aggregated consumer welfare.

On the other hand, priority service can increase customers' surplus if it expands the consumption's coverage, i.e., if it introduces new customers who would not otherwise purchase the service.

Furthermore, a market environment in which the dominant effect of customers' individual preferences is on the value of the basic good and less on the disutility of waiting tends to be more conducive to welfare improvement due to priority service.

Keywords: Externalities, priority pricing.
JEL classification: D62, D4.

1 Introduction

By priority service (PS) we refer to the option offered by service providers to customers to purchase the right to obtain priority over regular customers. We are primarily concerned with priority queues that distort "first-come first-served" queues by serving priority customers before regular ones but our results are also relevant for other environments including the case

*We are grateful to Emir Kamenica and to three anonymous referees for their helpful comments. We wish to thank Peter Cramton, Paul Klemperer, Preston McAfee, Benny Moldovanu, Jean Tirole, Alexis Walckiers, Robert Wilson, and seminar audiences at Toulouse, Oxford, ZEW, EUI and the Virtual Market Design Seminar Series for helpful discussions. We wish to thank Michael Borns for his invaluable editorial work. Gershkov wishes to thank the Israel Science Foundation for financial support. Gershkov: Department of Economics and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, and School of Economics, University of Surrey, alexg@huji.ac.il; Winter: Department of Economics and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, and Department of Economics, Lancaster University, eyal.winter@mail.huji.ac.il.

of a free public service that is complemented by a private provision of an improved service if this service dwindles the existing public resources.

PS is prevalent in many industries that involve queues, but it presents a large range of consequences for those whose waiting time is reduced as well as for those whose waiting time increases. While priority boarding and priority check-in in airlines merely grant some extra convenience for customers who purchase it, toll (express) roads and priority delivery of goods can often determine the value of the ride or the purchase. If we arrive at the meeting shortly before it ends, or if the suit is delivered after the wedding takes place, benefits go practically down to null. Private service providers in the health industry will often give priority to patients not next in the queue for the operating room if they pay extra. Such a priority queue can easily have major health consequences not only for the patients who purchase it but also for those who do not.

Another example of a priority market is front-running: fees charged by financial intermediaries for faster data transmissions and execution of trade orders. It is harmful to traders who do not pay the fee and benefits those who do. While many aspects of front-running are illegal, in reality preventing it is very difficult.

Our objective in this paper is to study the welfare effect of priority service on customers. Our main results show that priority service can be welfare-reducing for customers in some environments and welfare-enhancing in others. More specifically we show that priority service enhances the overall customers' welfare if it attracts a large set of customers to purchase the primary service relative to the case where no priority service is offered. However, in several important queuing environments this expansion does not take place. If the regular service is offered for free (as in national health service) or if the pricing of the priority service is done separately and independently of the pricing of the regular service, or if the value of the regular service is sufficiently high relative to the cost of waiting, priority service does not expand the set of customers who purchase the regular service, and introducing priority service reduces the overall customers' welfare. We also show that under certain conditions (derived later in the paper) introducing priority service can leave all customers worse off. Indeed this may happen due to the fact that priority service can result with the monopolist increasing even the price of the regular service.

To study these welfare effects we start with a simple model of priority service. Priority customers are served before regular customers. Within each of these two groups customers are served in random order. Each customer has a constant marginal waiting cost for each of the other customers standing before her/him in the queue. These costs are differential across customers and determined by a probability distribution. The service provider who knows the distribution sets the price of priority service so as to maximize its revenue.

To demonstrate the simplest manifestation of the loss in custmers' welfare due to priority service, consider two customers who purchase a certain service. Waiting to be served second incurs a cost of 1 to customer 1, and 2 for customer 2. In the absence of priority service customers are served in random order. The same applies if both customers purchase priority. If only one of them purchases it, s/he is served first. Under no priority service agents are served in a random order, and the total expected cost of waiting is $1\frac{1}{2}+2\frac{1}{2}=\frac{3}{2}$. Alternatively, if priority is offered to the more impatient customers for free, then the overall (expected) cost of waiting declines from 3/2 to 1. Hence, the efficiency gain of the priority queue is 1/2. However, the service provider can extract much more than 1/2 in equilibrium. Any price of less than 1 will be accepted by customer 2 regardless of the other customer's decision. (If customer 1 purchases the priority service then customer 2's willingness to pay for it is $2 - 2\frac{1}{2} = 1$, and if customer 1 is not a priority customer then customer 2's willingness to pay is $2\frac{1}{2} - 0 = 1$ as well.) Similarly, any price of less than 1/2 will be accepted by customer 1 regardless of the other customer's decision. If priority service does not exist the customers' overall waiting disutility is 3/2, but when the service provider determines the price of priority their total disutility is[1] $1 + 1 = 2 > 3/2$. Hence not only does the service provider levy the entire efficiency gain, but it also manages to extract an additional revenue of one half, making the customers jointly worse off.

Consider now a symmetric case where both customers' costs of waiting are 1; then any priority price below 1/2 yields that purchasing priority is the dominant strategy for both customers. The customers' waiting time in this unique equilibrium is exactly the same as in the case where none of them purchase priority. Hence, in the unique equilibrium outcome under optimal pricing the two customers transfer a total of 1 unit of money to the service provider without getting any reduction in their waiting time. If there were 100 customers all with a fixed marginal cost of waiting of 1, the unique equilibrium would have each of them transfer about 50 units of money to the service provider, without improving their expected waiting time relative to the situation where there is no priority service at all.

The type of excessive surplus extraction by the service provider that PS induces is very different from other types of surplus extraction including that generated by price discrimination (e.g., Mussa and Rosen [36]). Firstly, it builds on the negative externalities among consumers, and the fact that the "good" (i.e., the priority service) is less valuable the more customers purchase it. Secondly, because of the negative externalities among consumers, the degree of surplus extraction can be greater than the total efficiency gain (i.e., the reduction in overall cost of waiting under priority service relative to a market without priority). Finally, as we shall see, the excessive power of service providers remains also when we depart from the monopolistic market structure, and introduce competition. This again won't be the case with price discrimination of any degree. Priority service can be thought of as a form of second-degree price discrimination where consumers' valuations for the two quality goods are endogenously determined in equilibrium. While the standard second-degree price discrimination assumes consumers' willingness to pay to be exogenous (and not endogenous as in our framework), one common feature in both frameworks is that the benefits of discrimination increase with the level of expansion of the traded quantity in equilibrium.

Why in some cases can priority services treat customers so badly? The answer to this question is quite simple in the case of homogeneous customers. Here, the priority service generates no value whatsoever. Relative to the case of no priority service, what it actually does is merely transfer welfare from one customer to another, at a price that goes wholly to the service provider, without offering any compensation to the customer who has been made worse off. With heterogeneity, the priority service generates efficiency gains. Some customers whose costs of waiting are excessively high may well be better off with priority service than with no priority service, in spite of the high price they might pay for it. Nevertheless, if the price of the primary good is zero or negligible, then under a mild condition on the probability

[1] The provider either sets the price of priority to 1, and only consumer 2 buys the priority service, or the provider sets the price to 1/2 and both consumers buy the priority service. While the provider is indifferent between the two possibilities in this example, we assume that s/he sets the price to 1 and only consumer 2 buys the priority service. In the second possibility the loss in consumer welfare is even greater.

distribution (over the cost of waiting) the total welfare increase enjoyed by priority customers with high waiting costs is offset by the price they have to pay and by the loss borne by the regular customers who get later service. Observe that introducing priority service diminishes the attractiveness of the regular service since consumers of this service lose precedence to priority customers. Reducing the value of the regular service allows the provider to extract from the priority customers more than the increase in the efficiency. Hence priority service yields a negative total welfare for consumers and a bounty for the service provider.

Our argument here as well as in the examples discussed earlier relies on the assumption that the price of the regular service remains fixed and does not change due to the introduction of priority service. In the sequel our analysis also addresses the case where the price of the regular service is endogenously determined. It should be noted that in this framework priority service might yield a lower price for the primary good as part of the monopolist's optimal pricing policy. Such a decline in the price of the primary good can expand the quantity of the traded good (by introducing patient customers with low willingness to pay). As will be shown, this can increase the total customers' welfare to the extent that priority service becomes welfare-enhancing.[2] Finding sharp and elegant conditions for the gain in customers' welfare under PS turns out to be harder than finding such conditions for the loss in consumers surplus under PS. This is because the former analysis requires intricate tracking of the effect of PS on the consumption coverage and on the price of the basic good. Nevertheless, we establish conditions for the improvement of consumer surplus due to PS and provide some intuition for these conditions. We also show that under these conditions the monopolist's gain from PS remains positive. Furthermore, we use these conditions to establish some special cases with simpler and more elegant sufficiency conditions (albeit more restrictive one) that guarantee consumer welfare gain due to PS. Our results show that when the price of the regular service is endogenous then markets where the dominant effect of the consumers' type is on the value of the regular service (e.g., health services) are more conducive to consumer welfare gain when PS is introduced (relative to markets where the dominant effect is on the cost or waiting, e.g., "the right to choose" markets).

The one-dimensional pricing assumption (i.e., the provider controls the price of priority service only) is reasonable in many important applications. First, we note that many applications in which priority service is offered at a price in addition to a free service offered by the government warrants analysis under such an assumption. National health systems in the UK and Israel allow patients either to purchase priority service and to jump the queue regardless of urgency or take the standard free-of-charge health service. Here the validity of our result about a loss in consumer surplus depends on which group of consumers one refers to for welfare comparisons. If the additional revenue generated by priority service is used to improve the health service of other NHS patients or reduce the waiting time for surgeries, then the practice of PS might be welfare-enhancing. If this additional revenue collected by the NHS, for example, is used to finance R&D of new treatments or drugs that will mature in the future, then priority service reduces the welfare of present patients, and can be viewed as a tax imposed on present patients to subsidize future ones. However, if (and not unrealistically) the PS revenues substitute government funding, then the effect of PS on the consumers

[2]However, we believe that the loss in consumers welfare due to PS is more surprising than the gain in consumers welfare due to PS. When including also the profit made by the monopolist, the total gain in welfare due to PS is already known to be positive (see Wilson [40] among others). The case of loss in consumers welfare due to PS is also the case that is most relevant for policy implications.

of health services is unambiguously negative in spite of its efficiency gains.

Other environments in which the assumption of zero price for the regular service is appropriate are ones in which the pricing is completely separate from the market of the primary service, either because of a price control over the latter or because it is set by a different economic entity. Some utility companies that supply a price-controlled good offer a separate service that offers priority in answering phone calls and in earlier warnings regarding power cuts (see in https://www.enwl.co.uk/power-cuts/priority-services-register/why-become-a-priority-services-customer/). Finally, SAT prep courses offered by private companies also provide a sort of priority service totally separate from the pricing of the regular service, i.e., the test itself.

Our results apply in two major contexts. The first concerns a priority service offered by a service provider at a higher price than the regular service/good. The second involves environments in which a single market of public service splits as a result of introducing a secondary (private) market that offers the same service with an improved quality at a higher price (introducing or expanding private education in a way that consumes the professional resources of the free public education system is a primary example).[3] In both contexts consumer welfare is affected by two countervailing forces when a priority service/private service is introduced. First, efficiency gains are introduced by endogenously matching the consumers to one of the two markets based on their willingness to pay (instead of matching them randomly to a service). Second, the average quality that consumers get at the standard quality market declines, and consumers at the improved quality market pay a higher price. The first force increases the overall welfare of consumers while the second one reduces it. Our results compare the overall welfare of customers in both markets. If the price of the standard quality service remains fixed, then introducing improved quality service generates a lower total customers' welfare. In other words, the second force described above overrides the first one. If the price of the regular service is not held fixed, then the result can be reversed.

In view of some of the negative results that we present it is important to stress that our paper does not propose that PS (or the privatization of a public service) should be prohibited. It is not even suggesting that in most cases in which PS is applied its effect on customers' welfare is negative. Rather our analysis attempts to identify environments in which its effect is negative and others in which it is positive.

In the second part of the paper we study several extensions of the model (while maintaining the zero price for the primary good assumption). In the first extension we study the case of multiple priority levels.[4] A customer who purchases a priority service of level k is guaranteed to be served before any customer who purchases a lower priority level and after any customer who purchases a higher priority level. Customers of the same priority level are served in random order. In this setting more priority levels lead to more efficient scheduling. Hence, the total welfare (of all customers and the service provider) increases with the number of levels. One would therefore think that with a large number of priority levels customers would get in total some share of these growing efficiency gains, and be made better off compared with when there are no priority service at all. This is, however, not the case. We show that loss in customers' welfare due to priority service applies to any number

[3]We postpone the presentation of this setting to Appendix B.

[4]Using multiple priority levels is a common practice in shipping; e.g., Amazon offers standard vs. Prime two-day delivery vs. Prime one-day vs. Prime now (one- or two-hour delivery), and in applying for a visa; e.g., the UK offers standard, priority, and super-priority service options.

of levels of priority service. Interestingly, this result applies not just to the monopolist's optimal price but to every price that attracts some customers. Furthermore, the monopolist's profit strictly increases with the number of levels. The level of surplus extraction here can be quite staggering. If, for example, the distribution of waiting costs is uniform, then the service provider's revenue can be twice as high as the total efficiency gains (relative to the no-priority service case).

Priority services are mostly offered by de facto monopolies, mainly because they are almost always secondary to some other primary service (e.g., priority boarding is secondary to the flight, priority delivery is secondary to the product delivered, etc.[5]). Hence, priority services are susceptible to the hold-up problem.[6] Once a customer commits to a primary service provider s/he cannot purchase priority elsewhere. Nevertheless, our analysis here covers also the duopoly case, and reveals inherent barriers to competition in these markets. Our model of a priority service duopoly game is a simple two-stage Bertrand game. In stage 1 service providers decide simultaneously on the price of the priority service. In stage 2 consumers sort themselves between the two providers and between the two queues within each provider (priority and regular). The subgame-perfect equilibrium requires that no consumer be made better off by switching a provider or a queue within a provider for any prices set by the providers. Moreover, no provider can increase its revenue by changing its priority price (while taking into account equilibrium behavior by consumers following such a change).

We show that priority service in a duopoly presents an intrinsic barrier to competition. Under homogeneity (identical costs of waiting) the duopoly does not increase competition at all relative to the case of monopoly. The unique equilibrium under optimal pricing splits the set of customers equally between the two service providers and each of the providers extracts from its set of customers exactly the same revenue that it would have extracted had it served this set of customers as a monopolist. We further show that under heterogeneous costs for the parameterized class of distribution functions that even in the case of duopoly competition introducing priority services decrease aggregated customers' welfare, in sharp contrast to the outcome of a Bertrand competition for a standard good.

The intuition behind the barrier to competition that is inherent in the structure of priority service markets is quite simple. In the case of a standard good, when one provider reduces the price of the good below the price charged by its competitor, it is able to attract the other provider's customers, without fearing of losing any of its current customers. This is not necessarily the case in markets for priority services. As customers move from the more expensive provider to the less expensive one, the latter becomes more congested. As the set of priority customers grows, the priority service becomes less valuable. Some customers

[5]Whether or not excessive surplus extraction applies to these markets depends on the level of linkage between these two services from the consumers" point of view. To be more specific, consider the process of the online booking of an airline ticket (e.g., through Expedia or Kayak). Such a process facilitates a market where the price of the regular service is practically fixed. The two factors airline customers pay most attention to when booking are the airfare and the flight's itinerary i.e., departure time and connections (these and only these details are listed on the booking site before customers make their choice of flight). Only after choosing their option and going through the ticketing process are they presented with the airline priority options. Indeed, they can still opt out and check the cost of similar priority options with other airlines. Hence whether one can assume that the price of the basic good is fixed in this market depends on the tendency of customer to consider the entire package when making the final decision (Einav et al. [22] find that eBay buyers do not fully internalize shipping fees, which are displayed).

[6]For the effect of the hold-up problem on the pricing of ancillary goods see Gomes and Tirole [25].

might prefer now to join regular service and in so doing will reduce the revenue of the competing provider. Hence, markets for priority service introduce tacit collusion that requires no communication, no signals, and not even good will – just profit maximization.

The last extension of our model is deferred to Appendix A, and it involves the analysis of our model under nonlinear waiting costs. Convex homogeneous waiting costs create strategic complementarities between customers: the more customers that join the priority service, the more substantial is the cost-saving from joining the priority service for other customers, and hence it is more beneficial to join the priority service. The opposite happens with concave waiting costs. For heterogeneous costs, the main result of customers' welfare extraction holds for both convex and concave waiting costs.

Related literature

In the classic literature on rationing and priority pricing, Wilson [40] and Chao and Wilson [13] analyze welfare-maximizing properties of priority pricing in the context of markets with random shocks like electricity provision markets. Priority pricing there is used as a rationing tool for market clearing. They show equivalence in terms of the induced allocation between welfare-maximizing priority pricing and spot pricing. In particular, they show the existence of priority pricing scheme that implements the allocation that maximizes the total welfare (of consumers and producers). Furthermore, the scheme can be adjusted for redistribution of the raised revenues among the customers such that this scheme Pareto dominates random assignment. Our paper is the first paper that studies the effect of PS on consumer surplus. More precisely, we provide a counterweight to this important literature by showing that a profit-maximizing provider often yields the customers a surplus below the surplus from the random queue. Moreover, in some cases, profit-maximizing priority service may decrease the expected utility of every customer relative to a random queue. Bulow and Klemperer [11] show conditions under which regulated prices (and the appropriate rationing) decrease consumer surplus in competitive markets. Furthermore, Hall [26] uses a dynamic mechanism design approach to show that express toll lanes can be used in a Pareto-improving manner to deal with traffic congestion.

The comparison of rationing and market allocation in the context of consumer surplus goes back to Weitzman [39]. Recently, Dworzsac, Kominers, and Akbarpour [20] addressed a more general mechanism design problem of reallocation of resources in an environment where agents differ in their valuation for an object and money. However this strand of literature looks at markets without externalities, whereas the main novelty of our paper is its focus on markets with externalities.

Hassin and Haviv [28] provide an excellent survey of models on queueing. In Chapter 4 they deal with different models of priority. While they illustrate some models of monopolistic service providers, they don't illustrate the welfare impact of such policies. Moreover, they don't provide analyses of competition between the providers. Haviv and Winter [29] study a queuing model with stochastic arrival and show that the optimal pricing of priority service requires discrimination between agents even when they are identical and belong to the same priority.

Deneckere and McAfee [18] analyzed the possibility to price differentiate by introducing inferior (damaged) goods. Decreasing the quality of the basic good allows charging higher price of the better good. Notice that in our setting the value of the services is specified endogenously, in equilibrium. In some sense, introducing priority service creates "equilibrium damaged" good – because of the negative externalities the values of the damaged (regular service) and non-damaged (priority service) goods are specified by the consumers' choices in equilibrium.

Mechanism design literature on queueing started with Dolan [19].[7] This paper extends the classic characterization of Vickrey, Clark, and Groves mechanisms to the queueing environment. Follow-up analyses have considered different cost structures and evaluated their implications on implementation of the first-best efficient allocation, while satisfying budget balancedness (for a recent survey of this literature see Chun, Mitra, and Mutuswami [14]). However, this literature does not analyze the effect of priority services on customers' surplus. Glazer and Hassin [24] analyzed effect of stable transfer schemes on consumers utilities.

Another related paper is Hoppe, Moldovanu and Ozdenoren [30] who consider a two-sided market with heterogeneous, privately informed agents. Agents in the two sides of the markets are complementary to one another in generating the joint surplus. They first announce their type to a planner who then, depending on the reports, forms pairs and extracts payments. Our (monopoly) priority service model can be reformulated as a two sided market where one of the sides is passive. The main objective of Hoppe et al [30] paper is to show that assortative matching doesn't provide substantial improvement relative to coarse matching (which divides each side of the market to two sections, High and Low and then randomly matches each section in one market to the corresponding section in the other one). However that paper doesn't deal with our main issue, which is the comparison between random matching of agents to service slots (no priority service) and priority service (that can be interpreted as a coarse matching in a one-sided market). Moldovanu, Sela, and Shi [35] analyzed a contest model with externalities. However, there are several differences between that model and ours. In contrast to our model setup, status does not add efficiency, but only redistributes wealth among participants. In addition, the goal of the designer is to maximize the total aggregated efforts.

Duopoly price competition between service providers in queueing is analyzed in Luski [33] and Levhari and Luski [32]. They analyze different models in which each provider with limited capacity decides on the price of its services and faces a stream of randomly arriving customers. The main question studied in these papers is the existence of a symmetric equilibrium in which both providers charge the same price. For analysis with more than two providers (including a continuum of providers) see Reitman [37]. These studies do not address the question of the effect of priority service on the consumers. Moreover, there is no natural counterpart of this question in the models studied in these papers.

It is well known that in the case of congestion (or in the case of externality in consumption in general), the revenue-maximizing non-discriminating monopolist sets welfare-maximizing price if faced with customers with homogeneous costs; see Edelson [21]. De Borger and Van Dender [17] show that a duopoly market equilibrium with linear demand and homogeneous costs has a higher than socially optimal congestion level.[8] Moreover, Acemoglu and Ozdaglar

[7] In the mechanism design literature that followed Dolan [19], the environment analyzed in this paper is called sequencing problem, while by queueing setup usually called dynamic setup with stochastic arrival of new customers.

[8] De Borger and Van Dender [17] also analyze the capacity choices of providers, which we do not address. They show that while the monopolist chooses the socially optimal capacity, in the case of duopoly market, the equilibrium levels of capacity are lower than is socially optimal.

[1] show that increasing competition in congested markets can reduce efficiency. These papers do not address the main question of our paper, which is the effect of priority on customers' surplus. In addition, the main difference between congested markets and markets for priority is that in the models of congested markets the derived demand stems from the comparison between participation in the congested market and staying out, hence customers who stay out of the market impose no externalities, whereas in the priority markets, the derived demand for priority follows from comparison between priority and regular service, and hence the customers who do not acquire priority service impose externalities on other customers in regular service. In addition, the value of regular service is specified endogenously and indirectly in the priority markets.

For the case of a single server provider our model can be formally regarded as a model of contracting with externalities (see Segal [38] and Winter [41]). While most of this literature concern environments of complete information ours is not. Hence we prefer to think of our model as one describing markets rather than a relationship between a principal and an agent. Indeed, our section on competition cannot be embedded in the framework of contracting.

This paper is organized as follows. After presenting the basic illustration and the model in the next two sections, Section 4 shows the impact of priority pricing on consumer surplus. Section 5 generalizes the model to markets where the monopolist charges optimal prices for the regular and priority services. Section 6 extends the analysis to multiple priority classes. Competition between service providers is analyzed in Section 7. Section 8 concludes. Appendix A shows that the main conclusion of the model remains even if the customers have nonlinear waiting costs. Appendix B illustrates a model of a private market introduced into a public service. Most proofs are presented in Appendices C - E. More technical proofs and derivations are presented in the Supplementary Material Appendix.

2 Illustration

We illustrate in two very simple examples the implication of a revenue-maximizing priority provider on customers' surplus. Assume first that there is a clientele consisting of n homogeneous customers with the same waiting costs per service normalized to 1. There is a single service provider (monopolist) with capacity normalized to 1 per period. The overall waiting cost of all customers is $\frac{n(n-1)}{2}$, and hence the aggregated utility of the customers without priority pricing is $-\frac{n(n-1)}{2}$, while the average utility is $-\frac{(n-1)}{2}$. Assume now that the provider introduces a priority service. The monopolist announces a price p. Customers, after observing the price for priority decide whether to join the priority service or to consume the regular, free service. A customer who acquires priority service is served before the regular customers, while within each category (both priority and regular) the service order is random.[9] Assuming n^p other consumers acquire priority service, the expected utility of customer i is

$$-p - \frac{n^p}{2}$$

[9] One can think about other, more sophisticated contracts in which the monopolist applies price discrimination and makes individual offers and guarantees a specific queue position in the case of acceptance and another, very unfavorable specific position in the case of rejection. Such contracts, while theoretically interesting, are not practically appealing in many situations. Furthermore, we assume that the provider cannot artificially keep the server busy while not providing service to the waiting customers.

if i acquires priority service, and

$$-n^p - \frac{n - n^p - 1}{2} = -\frac{n + n^p - 1}{2}$$

if i doesn't acquire priority service and hence is served as a regular customer. Therefore, joining the priority service improves individual utility by $\frac{n-1}{2}$ independently of the action of the other agents. Put differently, by joining the priority service, customer i overcomes on average $\frac{n-1}{2}$ other customers independently of the size of priority and regular queues.[10] The actions of the other customers specify the composition of these $\frac{n-1}{2}$ customers, but it is irrelevant for i's decision whether to join the priority service. The next proposition characterizes the equilibrium in such markets.

Proposition 1 *In the case of homogenous customers, if $p < \frac{n-1}{2}$ all customers have a dominant strategy to join the priority service, while if $p > \frac{n-1}{2}$ all customers have a dominant strategy to join the regular, non-priority service. If $p = \frac{n-1}{2}$ all customers are indifferent between joining and not joining, independently of the choices of the other customers. The unique equilibrium outcome in the game with homogeneous customers is when the provider sets the price $p = \frac{n-1}{2}$ and all customers join the priority service.*

In this game, the monopolist is able to extract $\frac{n(n-1)}{2}$ from the customers without offering them anything since their expected waiting time remains the same. Hence the mere existence of a market for priority makes agents worse off. So the aggregated utility of the customers is given by $-\frac{n(n-1)}{2} - \frac{n(n-1)}{2}$: the aggregated waiting costs $-\frac{n(n-1)}{2}$ of the priority service exactly as without priority service, and $\frac{n(n-1)}{2}$ a total transfer to the priority service provider that corresponds to extraction of the customers' surplus.[11]

We shall now illustrate a similar result in the case where the consumers differ in their waiting costs, and hence introducing priority service increases efficiency. Assume that there is a continuum of consumers of mass 1. Consumers aren't homogeneous and face different waiting costs. Consumers' waiting costs per unit of time are given by a uniform distribution on the interval $[0, 1]$. There is a single provider with capacity normalized to 1 per period. If no priority service is offered, the allocation is random, and the expected waiting time is $1/2$, which is equivalent to the aggregated consumers' surplus of $-1/4$ since the expected cost of waiting is $1/2$ per unit of time. If the provider introduces a priority service at price p, then consumers with high waiting costs pay the price p and join the priority service to save time, while the consumers with low waiting costs join the regular, free service. The utility of the consumer with waiting cost c who decides to join the regular service is given by

$$-p - c\frac{1 - c^*}{2},$$

where c^* is the waiting cost of the consumer who is indifferent between these two options. If the consumer with waiting cost c joins the regular queue, his utility is $-c\left(1 - \frac{c^*}{2}\right)$. As

[10] For instance, if only one other customer joins the priority service, buying priority decreases waiting costs from $\frac{n}{2}$ to $\frac{1}{2}$, while if all other customers join the priority service, buying priority decreases the waiting time from $n-1$ to $\frac{n-1}{2}$.

[11] A similar observation appears in Hassin and Haviv [28] p. 85 and credited to a private communication with Murali Agastya from 2001.

the consumer with cost c^* is indifferent between these two options, the waiting cost of this consumer depends on the price of the priority service and is given by $c^* = 2p$. Moreover, the surplus of consumers with waiting costs below c^* is $-c\left(1 - \frac{c^*}{2}\right)$, while the surplus of consumers with waiting costs above c^* is $-p - c\frac{1-c^*}{2} = -\frac{c^*}{2} - c\frac{1-c^*}{2}$. Hence, the aggregated consumer surplus is

$$-\frac{1}{4} - \frac{c^*}{4} + \frac{(c^*)^2}{4}.$$

The latter expression is strictly lower than $-1/4$ unless $c^* = 0$ or $c^* = 1$. In other words, whenever the provider offers a price such that both the regular and priority services attract a non-trivial share of consumers, the aggregate consumers' surplus decreases as a result of introducing the priority service. The revenue-maximizing provider sets $c^* = 1/2$ and the customers' surplus equals $-5/16$; however, every cutoff (not only the revenue-maximizing one) decreases the consumers' surplus.

3 Model

In this section we describe the benchmark model. A single provider faces a continuum of customers of mass 1.[12] Assume that customers are heterogeneous with respect to per-unit waiting time. More precisely, the distribution of customers' waiting costs per unit of time is given by distribution function F on support $[0, \bar{c}]$ with $\bar{c} < \infty$ and density $f(c) > 0$ for any $c \in [0, \bar{c}]$. Hence, a customer with a per unit of time waiting cost of $c \in [0, \bar{c}]$ who gets service at time t and pays p has a utility of $-p - tc$. The provider can serve at each instant a single customer. We normalize the service time of a mass of m of consumers to be exactly m units of time, and the cost of the provider to be zero.

While some priority services can be analyzed as completely separate, independent products (like medical procedures or loans of expensive equipment), many priority services are bundled with other products and in fact play only a secondary role in the product bundle, like priority boarding or first-class quick delivery. Hence, acquiring the major component of the bundle essentially locks the customer with the priority provider. Therefore, we start our analysis with a single service provider – a monopolist. We later extend it to a competitive environment.

We analyze a simple market interaction in which the provider at the first stage offers a non-discriminatory price p for its priority service. Customers then decide whether to acquire the offered priority service at price p, which gives them priority over regular customers. The service sequencing within each category is random.

4 Equilibrium Consumer Welfare

When the monopolist sets price p for its priority service, this price separates the clientele into two categories[13]: regular customers and priority customers. The price of the regular service is fixed and normalized to zero (See Section 5 for discussion of the provider that sets optimal

[12] Analyzing a continuum of customers allows us to refrain from integer problems of the queues' sizes.

[13] Depending on the price, some categories may be empty.

prices for both regular and priority services). The marginal customer $c^*(p)$ is indifferent between joining the priority service and the regular one, i.e.,

$$-p - c^*(p)\frac{1-F(c^*(p))}{2} = -c^*(p)\left(1-\frac{F(c^*(p))}{2}\right) \Leftrightarrow c^*(p) = 2p.$$

Therefore, if the monopolist sets a price p for its priority service with $\frac{\bar{c}}{2} \geq p \geq 0$, customers with waiting costs $c \geq c^*(p) \equiv 2p$ buy priority service and customers with waiting costs $c < c^*(p) \equiv 2p$ refrain from buying it, and consume the regular service.

To establish the effect of priority service on the customers' surplus, notice that without priority service the customers' welfare, assuming the random assignment of the queue positions, is

$$-\int_0^{\bar{c}} \frac{c}{2} f(c)dc = -\frac{\mathbb{E}(c)}{2}.$$

With priority, assuming that types above $c^*(p)$ acquire priority service and types below $c^*(p)$ get served after priority customers, the customers' welfare is[14]

$$\begin{aligned} & -\int_0^{c^*(p)} c\left(1-F(c^*(p))+\frac{F(c^*(p))}{2}\right)f(c)\,dc + \int_{c^*(p)}^{\bar{c}}\left(-p-c\frac{1-F(c^*(p))}{2}\right)f(c)\,dc \\ = & -\frac{1-F(c^*(p))}{2}\mathbb{E}(c) - \int_0^{c^*(p)}\frac{c}{2}f(c)\,dc - \int_{c^*(p)}^{\bar{c}}\frac{c^*(p)}{2}f(c)\,dc. \end{aligned}$$

Therefore, introduction of priority is detrimental to customers if and only if

$$F(c^*(p))\,\mathbb{E}(c) < \int_0^{c^*(p)} cf(c)\,dc + c^*(p)\left(1-F(c^*(p))\right). \tag{1}$$

Introduction of the priority service has two effects on the customers' welfare that work in opposite directions. (1) Without priority service the allocation is completely random, unrelated to the real waiting costs. From the efficiency perspective we would like to have the allocation of the slots depend on the waiting costs. Introducing a positive price for priority splits the market into two segments, whereby the customers with higher waiting costs get service first. Hence introducing priority service creates gains from the improved efficiency in allocation. McAfee [34] shows that splitting the market into two submarkets can gain a very substantial part of the fully efficient assignment.[15] (2) However, the provider can expropriate (at least part of) the created surplus via the payment for priority. The next proposition shows that under a rather weak condition the second effect dominates and introducing priority service is necessarily detrimental to customers' surplus. In other words, the provider extracts from the customers more than the created gains from the improved efficient allocation.

Proposition 2 *Assume that $\mathbb{E}(c) - \frac{1-F(c)}{f(c)}$ changes sign only once from negative to positive. Then the customers' welfare if priority service is not available is higher than it would be if priority service were offered.*

[14] Recall that to induce the division into the two categories with the indifferent type of c^*, the monopolist sets a priority price of $p(c^*) = c^*/2$.

[15] Hoppe, Moldovanu and Ozdenoren [30] extended McAfee's analyses to markets with incomplete informations.

Proof. We show that for any $c \in (0, \overline{c})$ it must be the case that

$$F(c)\,\mathbb{E}(c) < \int_0^c cf(c)\,dc + c(1 - F(c)).$$

To see this, observe that for $c = 0$ both sides of the inequality are 0, while for $c = \overline{c}$ both sides of the inequality are equal to $E(c)$. The derivative of the left-hand side of the inequality with respect to c is $f(c)\,\mathbb{E}(c)$, while the derivative of the right-hand side of the inequality with respect to c is $cf(c) + 1 - F(c) - cf(c) = 1 - F(c)$. The derivative of the right-hand side is greater if and only if

$$f(c)\,\mathbb{E}(c) < 1 - F(c) \Longleftrightarrow \mathbb{E}(c) - \frac{1 - F(c)}{f(c)} < 0.$$

The condition of the proposition implies that there exists $\tilde{c}$ s.t. for all $c < \tilde{c}$ the derivative of the right-hand side is greater than the derivative of the left-hand side, while for all $c > \tilde{c}$ the derivative of the left-hand side is greater than the derivative of the right-hand side. ■

A sufficient condition for $\mathbb{E}(c) - \frac{1-F(c)}{f(c)}$ to cross zero only once, from below, is that F has an increasing failure rate (satisfies the IFR property): i.e., $\frac{1-F(c)}{f(c)}$ is decreasing.[16]

Remark 1 *One natural solution to the excessive market power of the provider is to restrict the number of priority slots that the provider is eligible to sell. Yet, as the last proposition shows, unless the provider is completely precluded from selling priority services such a restriction is not able to eliminate the negative effect of priority pricing on the customers' surplus. In fact the proof of the last proposition shows that introducing priority service (even without necessarily using the revenue-maximizing price) always decreases the customers' surplus if the distribution satisfies the IFR property. Yet a restriction on the number of priority slots can grant a certain control over the size of the customers' welfare loss (relative to no priority service) to those who purchase the regular service and if the number of such slots is small so is the welfare loss.*

Remark 2 *The reader might find the situation of priority service to be akin to second degree price discrimination (SDPD) but the two differ in multiple aspects in terms of both intuition and formal analysis. Priority service builds on negative externalities among the customers, which plays no role in SDPD. Priority customers are offered a different good than those opting to be non-priority customers, and the willingness to pay for the improved good depends on the number of people who consume it. This dependence exacerbates the loss of surplus for customers relative to SDPD.*

In addition to analyzing the effect of priority pricing on the aggregated consumers' surplus, we can analyze its effect on individual types in terms of expected utilities (before realization of randomization of the allocation within every priority group). Clearly all types below c^* are worse off after the introduction of priority service: although these types find it optimal not to buy priority service, they must suffer from delays in getting served as now they will be served in random order after all priority customers are served. Since type c^* is indifferent between joining priority service and getting regular service, this type is worse off as well,

[16] IFR is a rather weak property. Most well-known distributions satisfy it.

and by continuity some types above c^* are also worse off. However, there may be types with sufficiently high waiting costs that are better off after the introduction of priority service. The utility of type $c > c^*$ before introducing priority service is $-c/2$, while after introducing the service it is

$$-p - c\frac{1 - F(c^*(p))}{2} = -\frac{c^*(p)}{2} - c\frac{1 - F(c^*(p))}{2}.$$

This implies that if priority service attracts a substantial mass of consumers, the time-saving element of the priority service may be limited and insufficient to compensate for the disutility from paying the price of the service. The next proposition shows the formal conditions for this comparison.

Proposition 3 *Assume that F first-order stochastically dominates the uniform distribution on $[0, \bar{c}]$. Then all types of consumers are worse off after the introduction of priority service while extreme types $\{0, \bar{c}\}$ are merely indifferent.*

Proof. All types below the cutoff c^* are worse off as they get served later. Types above c^* get the service earlier, but now pay $p = \frac{c^*}{2}$. Type $c > c^*$ is worse off if and only if

$$\begin{aligned} -\frac{c^*}{2} - c\frac{1 - F(c^*)}{2} &< -\frac{c}{2} \iff \\ cF(c^*) &< c^*. \end{aligned}$$

F first-order stochastically dominating the uniform distribution on $[0, \bar{c}]$ implies that for any c^* it holds that $F(c^*) \leq \frac{c^*}{\bar{c}}$. Hence, for any $c^* \in (0, \bar{c})$ and any $c \in (c^*, \bar{c})$, we have

$$cF(c^*) \leq c\frac{c^*}{\bar{c}} < c^*.$$

■

5 Endogenous Pricing of the Regular Service

We have assumed thus far that the price of the regular service for which priority is offered is exogenous and fixed. In the Introduction we listed several important priority markets in which this is the case and argued that this assumption holds to a certain extent also in other markets where the booking procedure separates the decision regarding the regular service from the decision regarding priority. Nevertheless, we provide here also the analysis of the case where the price of the regular service is determined endogenously by the monopoly, and identify conditions under which our main finding applies here as well. This analysis is also meant to highlight the fact that priority service can be welfare-enhancing for customers and we demonstrate this fact.

For simplicity, we remain within the one-dimensional type framework. We consider two scenarios here. In both scenarios we assume that higher types derive higher values from the service and bear higher waiting costs. In both scenarios we normalize the utility from nonparticipation to zero. In the first one, we assume that the effect of the types on values dominates the effect on costs, and hence the overall utility is increasing in types. In this scenario, the customers that have the lowest incentives to enter the market are low-value and low-waiting cost customers. Examples of such a scenario include some (non-essential) health services[17] and retail delivery services. In the second scenario, the effect on the cost is the dominant one, and hence the overall utility is decreasing in types. In this scenario, the customers that have the lowest incentives to enter are high-value and high-waiting cost customers. Examples of such a scenario include services that give priority to consumers with "the right to choose," such as services that allow customers to choose apartments in a multi-unit building.[18] These two scenarios lead to different types of pricing policies: in the first scenario, the marginal consumer is a low-value, low-cost type, and all higher types join the market and acquire some type of the service. In the second scenario the marginal consumer is a high-value, high-cost consumer and all types with lower waiting costs join the market and acquire some type of the service. We describe here in greater detail the analysis of the first case. For the second case we give a summary of the results, while deferring the detailed analysis to Appendix C.

5.1 Low-type exclusion

Assume that the types are one-dimensional, but that the type affects both the cost of waiting and the value of the service. That is, the utility of a customer of type θ from getting the service, waiting t units of time, and paying price p is

$$v(\theta) - c(\theta)\, t - p,$$

where we assume that $v(\theta) > 0$, $c(\theta) > 0$, $v'(\theta) > 0$, $c'(\theta) \geq 0$, and $v'(\theta) - c'(\theta) > 0$. Assume further that $v''(\theta) \leq 0$ and $c''(\theta) \geq 0$. Under these assumptions low types stay out of the market, while high types join it. Types distribute according to a CDF F with support normalized to $[0, \bar{\theta}]$ with $f(\theta) > 0$ for any $\theta \in [0, \bar{\theta}]$. Further assume that the distribution of types satisfies IFR (increasing failure rate); i.e., $\frac{1-F(\theta)}{f(\theta)}$ is decreasing.

We start our comparison with the case of a provider that offers regular service only and chooses its price optimally, to maximize its revenue. Denote by p the price of the service. Given a price p there is a cutoff type θ^* such that types $\theta \geq \theta^*$ acquire the regular service and types $\theta < \theta^*$ remain outside of the market.[19]

$$\underset{0}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{no service}}\underset{\theta^*}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{regular service}}\underset{\bar{\theta}}{|}$$

The utility of type θ from joining the market is

$$v(\theta) - c(\theta)\,\frac{1 - F(\theta^*)}{2} - p.$$

[17] In the case of health services normalizing the utility of not getting service to zero is problematic. In this case the value from the service is given by the difference in the individual utility between getting and not getting the medical treatment.

[18] Ashenfelter and Genesove [7] analyze auction data in such markets, and a random priority for the right to choose has also been used recently in Israel for its affordable housing program. It is reasonable to expect that priority service in such environments enhances consumer welfare, as the value of the primary good is closely entangled with the cost of choosing late. Customers who have strong preferences for specific units may reasonably avoid purchasing an apartment in such buildings unless they are offered priority in choosing their desired unit.

[19] Assumption $v'(\theta) - c'(\theta) > 0$ guarantees that if some type finds it optimal to purchase the service, then all higher types also join the market.

For type θ^* to be indifferent between joining the service and staying out of the market, it must be the case that[20]

$$v(\theta^*) - c(\theta^*)\frac{1-F(\theta^*)}{2} = p.$$

The monopolist chooses the price (or the cutoff type) to maximize its revenues:

$$R^{RS}(\theta^*) = p(1-F(\theta^*)) = (1-F(\theta^*))\left(v(\theta^*) - c(\theta^*)\frac{1-F(\theta^*)}{2}\right)$$

while the consumer surplus is

$$\int_{\theta^*}^{1}\left(v(\theta) - c(\theta)\frac{1-F(\theta^*)}{2} - p\right) f(\theta)\, d\theta. \tag{2}$$

If the provider offers both priority and regular services, then the monopolist sets two prices, $\Pi \geq \pi$ and the customers are divided into three groups (with $\bar{\theta} \geq \theta'' \geq \theta' \geq 0$). Types $\theta \geq \theta''$ acquire priority service at price Π, types $\theta \in [\theta', \theta'')$ acquire regular service at price π. Types $\theta < \theta'$ remain out of the market.

$$\underset{0}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{no service}}\underset{\theta'}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{regular service}}\underset{\theta''}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-}_{\text{priority service}}\underset{\bar{\theta}}{|}$$

The utility of type θ from getting priority service is

$$v(\theta) - c(\theta)\frac{1-F(\theta'')}{2} - \Pi,$$

while her/his utility from regular service is

$$v(\theta) - c(\theta)\left[1 - F(\theta'') + \frac{F(\theta'') - F(\theta')}{2}\right] - \pi.$$

Type θ'' is indifferent between purchasing the regular and the priority service, and assumption $c' > 0$ guarantees that all types above θ'' strictly prefer priority to regular service, while the types below θ'' prefer regular to priority service. Similarly, type θ' is indifferent between regular service and remaining outside of the market. Assumption $v' - c' > 0$ guarantees that types above θ' prefer regular service to getting no service at all, while types below θ' prefer to stay out of the market. That is, for given prices Π and π, the cutoff θ'' satisfies

$$\begin{aligned} v(\theta'') - c(\theta'')\frac{1-F(\theta'')}{2} - \Pi &= v(\theta'') - c(\theta'')\left[1 - \frac{F(\theta'') + F(\theta')}{2}\right] - \pi \Leftrightarrow \\ c(\theta'')\frac{1-F(\theta')}{2} &= \Pi - \pi, \end{aligned}$$

while the cutoff θ' satisfies

$$v(\theta') - c(\theta')\left[1 - \frac{F(\theta'') + F(\theta')}{2}\right] = \pi.$$

[20] The above assumptions guarantee uniqueness of the indifferent type for any p.

Observe that θ'' doesn't depend on $v(\theta'')$ since both the priority and the regular customers get served. The provider's profits are

$$R^{PS}(\theta', \theta'') = \Pi(1 - F(\theta'')) + \pi(F(\theta'') - F(\theta')) = \pi(1 - F(\theta')) + (\Pi - \pi)(1 - F(\theta''))$$

and the consumer surplus is

$$\int_{\theta'}^{\theta''} \left(v(\theta) - c(\theta) \left(1 - \frac{F(\theta'') + F(\theta')}{2} \right) - \pi \right) f(\theta)\, d\theta + \int_{\theta''}^{\overline{\theta}} \left(v(\theta) - c(\theta) \frac{1 - F(\theta'')}{2} - \Pi \right) f(\theta)\, d\theta. \tag{3}$$

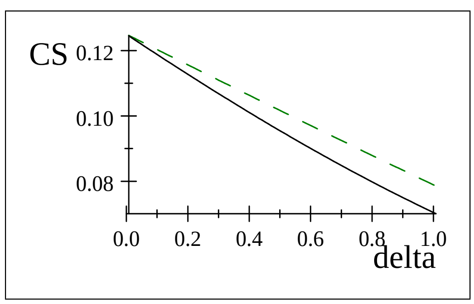
We first show that introducing priority service always (strictly) increases the provider's revenues.

Lemma 1 *Introducing priority service increases the provider's revenues.*

The proof of this lemma shows that setting the price such that a very small group of customers purchase the priority service, while keeping the overall size of the clientele fixed generates higher profits relative to the case where only the regular service is offered.

We start by demonstrating the positive effect of PS in cases where it expands the consumption coverage. One particularly interesting example that falls within the chosen setting is the case where the waiting costs take the form of discounting and hence are proportional to the value of the object.

Example 1 *Assume that $v(\theta) = \theta$ and $c(\theta) = \delta\theta$, where $\delta \in (0, 1)$ is a discount factor which is common to all customers. Furthermore, assume that θ distributes uniformly on support $[0, 1]$.*



Dashed line – consumer surplus if both the regular and priority services are offered. Solid line – consumer surplus if only the regular service is offered.

Introducing priority service in this case improves the consumer surplus. Introduction of the priority service has multiple effects on consumers' utilities: (1) the provider (optimally) decreases the price of the regular service and as a result attracts an additional set of consumers; (2) as a result of this price decrease the utility of some consumers that acquired the regular service also goes up despite the longer waiting time, and (3) the high-type consumers get higher utility, because the time saving compensates the payment for priority service. Note that as δ grows the efficiency gain grows as well, and the figure of example 1 shows that the increase in consumer surplus due to PS grows as well.

The next example illustrates that in some cases introduction of priority services can make all consumers better off.

Example 2 *Assume there are two sets of customers: (1) High-value and high-waiting cost and (2) low-value and low-waiting cost. Customers within each group are homogeneous. More specifically, consider the following value and cost functions:*

$$v(\theta) = \begin{cases} v_2 & \text{if } \theta > \tilde{\theta} \\ v_1 & \text{if } \theta \leq \tilde{\theta} \end{cases}, c(\theta) = \begin{cases} c_2 & \text{if } \theta > \tilde{\theta} \\ c_1 & \text{if } \theta \leq \tilde{\theta} \end{cases}$$

with $v_2 > v_1$, $c_2 > c_1$, and $v_2 - c_2 > v_1 - c_1$ with[21] $\alpha = F\left(\tilde{\theta}\right)$*. In this example we choose the parameters such that (1) without the priority service the provider finds it optimal to offer the regular service to customers with types $\theta > \tilde{\theta}$ only, and so types $\theta \leq \tilde{\theta}$ get no service; (2) with priority service, the monopolist optimally offers priority service to types $\theta > \tilde{\theta}$ and regular service to types $\theta \leq \tilde{\theta}$. We will show that customers with types $\theta > \tilde{\theta}$ are strictly better off as a result of introducing priority service. In this example the utility of types $\theta \leq \tilde{\theta}$ remains the same; however, if we replace functions v and c with strictly monotone functions that approximate the original functions, then the payoffs of all types except for* 0 *increase as well.*

If no priority service is offered, and the principal charges for its regular service price p such that

$$p = v_2 - c_2 \frac{1-\alpha}{2}$$

then only types $\theta > \tilde{\theta}$ acquire it. The provider's revenues are $(1-\alpha)\left(v_2 - c_2 \frac{1-\alpha}{2}\right)$. If the regular service is offered for all agents, the provider gets revenues of $v_1 - \frac{c_1}{2}$. Hence, the provider finds it optimal to offer the regular service to agents with $\theta > \tilde{\theta}$ if and only if $(1-\alpha)\left(v_2 - c_2 \frac{1-\alpha}{2}\right) > v_1 - \frac{c_1}{2}$.

If the provider offers both the priority and the regular service, then it is optimal to charge $\pi = v_1 - c_1\left(1 - \frac{\alpha}{2}\right)$ for the regular service and types $\theta \leq \tilde{\theta}$ acquire this service, while the priority service is offered for price Π that satisfies

$$\begin{aligned} v_2 - c_2 \frac{1-\alpha}{2} - \Pi &= v_2 - c_2\left(1 - \frac{\alpha}{2}\right) - \pi \\ &\iff \Pi = v_1 - c_1\left(1 - \frac{\alpha}{2}\right) + \frac{c_2}{2}. \end{aligned}$$

Only types $\theta > \tilde{\theta}$ acquire priority service and the revenues are

$$\alpha\pi + (1-\alpha)\Pi = v_1 - c_1\left(1 - \frac{\alpha}{2}\right) + c_2 \frac{1-\alpha}{2}.$$

It is an optimal scheme for the provider if

$$v_1 - c_1\left(1 - \frac{\alpha}{2}\right) + c_2 \frac{1-\alpha}{2} > (1-\alpha)\left(v_2 - c_2 \frac{1-\alpha}{2}\right) > v_1 - \frac{c_1}{2}. \tag{4}$$

[21] While this example doesn't satisfy strict monotonicity and differentiability requirements of v, c and $v - c$, these functions can be approximated by monotone and differentiable functions. Since all the inequalites are strict, the derivations should hold also for "closely approximated" monotone and differentiable functions.

Observe that as a result of introduction of priority service, the utility of agents with types $\theta > \tilde{\theta}$ increases from zero to[22] $v_2 - v_1 - (c_2 - c_1)\left(1 - \frac{\alpha}{2}\right) > v_2 - v_1 - (c_2 - c_1) > 0$.

We proceed with our analysis by identifying conditions with optimal full consumption coverage (with and without PS), and provide necessary and sufficient conditions under which PS increases consumer surplus. Our next proposition shows when the consumers will benefit from the improvement in efficiency due to introduction of priority service.

Proposition 4 *Assume that $\left(v'(0) - c'(0)\frac{1}{2}\right) \leq f(0)(v(0) - c(0))$ and R^{RS} is concave*[23]*. Then introducing priority service increases consumer surplus if and only if*

$$F\left(\theta''\right)\mathbb{E}\left[c(\theta)\right] - \int_0^{\theta''} c(\theta) f(\theta)\, d\theta - \left(1 - F\left(\theta''\right)\right)\left(c\left(\theta''\right) - c(0)\right) \geq 0 \tag{5}$$

where θ'' is the solution of[24]

$$c\left(\theta''\right) - c(0) = c'\left(\theta''\right)\left(\frac{1 - F\left(\theta''\right)}{f\left(\theta''\right)}\right).$$

The conditions of the last proposition ($\left(v'(0) - c'(0)\frac{1}{2}\right) \leq f(0)(v(0) - c(0))$ and concavity of R^{RS}) guarantee that the provider optimally sets $\theta^* = \theta' = 0$. In other words, the monopolist does not want to exclude any consumers from the service whether with or without PS. To see this observe that if $\theta^* > 0$ then reducing θ^* has two effects on the provider's revenue. On the one hand, it decreases the price $p = v(\theta^*) - c(\theta^*)\frac{1-F(\theta^*)}{2}$, but on the other hand it increases the set of consumers that get the service and pay p. Condition $\left(v'(0) - c'(0)\frac{1}{2}\right) \leq f(0)(v(0) - c(0))$ together with concavity of R^{RS} guarantees that the second effect dominates and the provider sets the highest price that includes all the customers. The intuition for $\theta' = 0$ is similar.

Inequality (5) specifies the effect of priority service under the assumptions of full coverage, and can be divided into two parts. First, by comparison of (2) and (3) we get that the expression

$$F\left(\theta''\right)\mathbb{E}\left[c(\theta)\right] - \int_0^{\theta''} c(\theta) f(\theta)\, d\theta > 0$$

describes the overall saving of waiting costs due to the improved efficiency of the allocation. Second, by comparison of (2) and (3) and the definition of prices described above the expression $\left(1 - F\left(\theta''\right)\right)\left(c\left(\theta''\right) - c(0)\right) > 0$ describes the increase in the overall payment of the consumers as a result of the introduction of priority service. The provider chooses θ'' optimally to maximize the second effect $\left(1 - F\left(\theta''\right)\right)\left(c\left(\theta''\right) - c(0)\right)$.

If the regulator could control the size of the set of priority customers, then the next Lemma shows that making the priority service very exclusive decreases consumer surplus (relative to no PS), while making it very inclusive enhances consumer surplus.

[22] There always exists a range of parameters that satisfy (4). For instance, for $\alpha = 0.5$, $v_1 = 2$, $c_1 = 1$, $c_2 = 2$, any $v_2 \in (3.5, 4)$ satisfy all the inequalities.

[23] The concavity of R^{RS} is guaranteed if in addition to the assumed conditions on v, c, and the failure rate $f(\theta)/(1 - F(\theta))$ we further assume that F is convex and $v(0) - c(0)/2 \geq 0$. As the condition of the proposition requires $0 < \left(v'(0) - c'(0)\frac{1}{2}\right) \leq f(0)(v(0) - c(0))$, assuming convexity of F guarantees that R^{RS} is concave.

[24] If there are multiple solution to the following equation, the optimal cutoff θ'' is one of these solutions.

Lemma 2 *Assume that* $\left(v'(0) - c'(0)\frac{1}{2}\right) \leq f(0)(v(0) - c(0))$ *and* R^{RS} *is concave. Assume furthermore that* $c'(0) \geq 0$ *is close to zero; then there exists* $0 < \check{\theta} < \hat{\theta} < \bar{\theta}$ *such that*

1. *setting* $\theta'' < \check{\theta}$ *increases consumer surplus, while*
2. *setting* $\theta'' > \hat{\theta}$ *decreases consumers' surplus,*

relative to the case where only regular service is offered.

Under the conditions of Proposition 4, if the monopolist doesn't offer the priority service, the optimal price of the regular service is

$$v(0) - \frac{c(0)}{2} = p,$$

while the price of the regular service if the priority service is offered is

$$v(0) - c(0)\left[1 - \frac{F(\theta'')}{2}\right] = \pi.$$

Hence, introducing priority service necessarily (weakly) reduces the price of the regular service.

The condition of Proposition 4 combines restrictions on both the cost of waiting function c and the distribution of types F. The next example illustrates when this condition holds in a parametrized class of waiting cost functions.

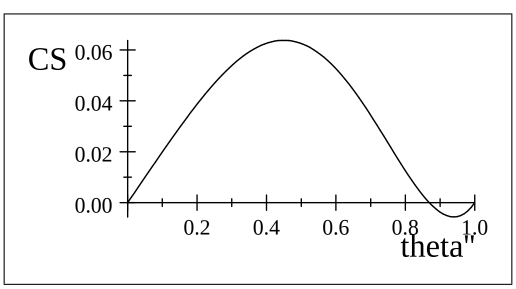
Example 3 *Consider* $c(\theta) = \theta^k$ *for* $k > 0$ *and the uniform distribution of* θ*. We note that*

$$c(\theta'') - c(0) = c'(\theta'')\left(\frac{1 - F(\theta'')}{f(\theta'')}\right) \Longleftrightarrow \theta'' = \frac{k}{k+1}$$

The difference in the consumer surplus is given by

$$F(\theta'')\,\mathbb{E}[c(\theta)] - \int_0^{\theta''} c(\theta) f(\theta)\, d\theta - \left(1 - F(\theta'')\right)\left(c(\theta'') - c(0)\right)$$
$$= \frac{1}{k+1}\frac{k}{k+1}\left[1 - \left(\frac{k}{k+1}\right)^{k-1} - \left(\frac{k}{k+1}\right)^{k}\right].$$

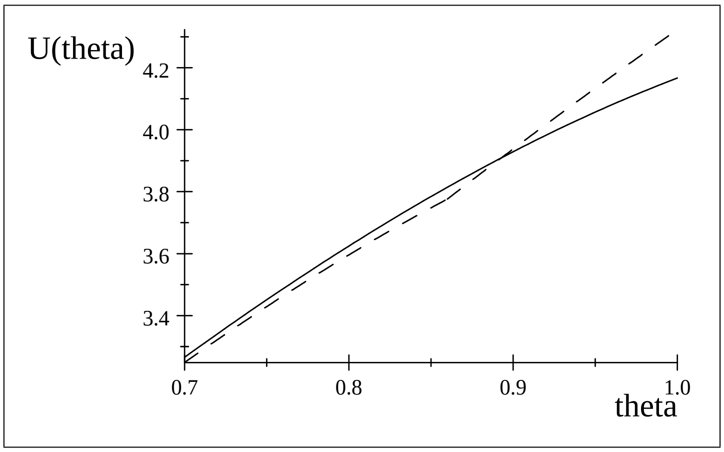
The last expression is negative for $k = 1, 2$ *and positive for* $k > 3$.



Effect of introducing priority service on consumer surplus as a function of θ'' *in the case of full coverage, uniform distribution, and cost function* $c(\theta) = \theta^4$.

Note that this example is consistent with Lemma 2. If θ'' is low (the PS is inclusive), then PS raises the consumer surplus. But if PS is sufficiently exclusive (θ'' is close to 1), then PS reduces the consumer surplus. Yet, the reduction in the consumer surplus would not be substantial if θ'' were sufficiently close to 1. The monopoly's optimal cutoff is $\theta'' = 0.8$.

The next plot shows the utility of different types of agents when it is assumed that $v(\theta) = 6 + 5\theta - \frac{1}{3}\theta^3$. *In particular, it shows that sufficiently high types of agents benefit from introducing priority service, while lower types are worse off, albeit not very substantially. See the discussion of this comparison at the end of this section.*



Solid line shows utility of different types of agents in the regime if only the regular service is offered, while the dashed line shows the utility of different types of agents if both the regular and the priority service is offered with a profit-maximizing cutoff of $\theta'' = 0.86$.

Summarizing the scenario of low-type exclusion we note that under this scenario PS can increase the aggregate consumer surplus even when consumption expansion does not take place. Furthermore, under consumption expansion the increase in consumer surplus can be such that all consumers will benefit from priority service due to a significant decline in the price of the regular service. As we shall see in the sequel the other scenario of high-type exclusion is much less conducive to the improvement of consumer welfare through priority service. Roughly speaking, it only allows this welfare to improve when PS expands consumption. Without such expansion PS can even lead to cases with welfare loss to each and every consumer due to the increase in the price of both services, i.e., regular and priority. Furthermore, under this scenario (and unlike the former one) even when consumption expands and the overall consumer surplus increases due to introducing PS, some consumers remain worse off.

5.2 High-type exclusion

In this part we assume that the individual type has a stronger effect on the costs than on the value from obtaining the service. That is, we assume that $v'(\theta) \geq 0$, $c'(\theta) > 0$ and $v'(\theta) - c'(\theta) < 0$. We keep the normalization of the utility from nonparticipation to zero. The implication of the assumption that $v'(\theta) - c'(\theta) < 0$ is that the types that prefer to stay out of the market are now high types.

If the provider offers only the regular service at price p, then types below a cutoff type θ^* acquire the service, while types above θ^* stay out of the market. The cutoff type is given by[25]

$$v(\theta^*) - c(\theta^*)\frac{F(\theta^*)}{2} = p.$$

The provider's profits are $pF(\theta^*)$ are given by

$$R^{RS}(\theta^*) = \left(v(\theta^*) - c(\theta^*)\frac{F(\theta^*)}{2}\right)F(\theta^*),$$

while the consumer surplus is given by

$$\begin{aligned} &\int_0^{\theta^*}\left(v(\theta) - c(\theta)\frac{F(\theta^*)}{2} - p\right)f(\theta)\,d\theta \\ = &\int_0^{\theta^*}\left(v(\theta) - c(\theta)\frac{F(\theta^*)}{2} - v(\theta^*) + c(\theta^*)\frac{F(\theta^*)}{2}\right)f(\theta)\,d\theta. \end{aligned}$$

If the provider offers both the priority and the regular service, then the monopolist sets two prices $\Pi > \pi$ and the consumers are divided into three groups (with $\bar{\theta} \geq \theta^{\Diamond} \geq \theta'' \geq 0$). Types $\theta \geq \theta^{\Diamond}$ remain out of the market, types $\theta \in [\theta'', \theta^{\Diamond})$ acquire priority service at price Π. Types $\theta < \theta''$ acquire regular service at price[26] π.

One specific setting that belongs to the case of high-type exclusion is the one where the value of the service is constant and independent of type, i.e., $v(\theta) = \tilde{v}$. In the Supplementary Material we provide a derivation for the following proposition.

Proposition 5 *Assume that the waiting costs are distributed according to the distribution function $F(\theta) = \theta^k$ with $c(\theta) = \theta$ where $k > 0$ and $\theta \in [0,1]$, and the value of the service is $0 < \tilde{v} \leq 1$ is constant and identical for all agents. Then the aggregated consumer surplus is higher in the case where the priority service is offered.*

In the proof of Proposition 5 we show that introducing priority service increases the coverage of the market. That is, whenever the market with the regular service only does not serve the entire market, then introducing priority service enlarges the set of served consumers. This result is reminiscent of the classic result on the effect of price discrimination on output (see Armstrong and Vickers [6] and Armstrong [5], among others).

In Appendix C we replicate the main results that we illustrate in Section 5.1 for the current setup with high-type exclusion. In particular, we show that the provider indeed gets higher profits if she offers both the priority and the regular services, and that introducing the priority service reduces the exclusion of the consumers, i.e., $\theta^{\Diamond} \geq \theta^*$.

Similar to Proposition 4 we show here the conditions under which introducing priority service improves consumers' surplus if the market has full coverage. As in the case of Proposition 4, the condition combines restrictions on both the cost function and on the distribution of types. However, as Corollary 2 shows, in contrast to the previous case of low-type exclusion, in the current case introducing priority service reduces consumer surplus under relatively general conditions.

[25] If for some p there exist multiple solutions to this equation, then we adopt a standard assumption from mechanism design that the service provider's optimal solution is selected.

[26] We adopt here notation for θ'' to be similar to the case of low-type exclusion: θ'' denotes a type that is indifferent between the regular and priority services.

Proposition 6 *Assume that* $\left(v'\left(\bar{\theta}\right)-\frac{c'\left(\bar{\theta}\right)}{2}\right)+\left(v\left(\bar{\theta}\right)-c\left(\bar{\theta}\right)\right)f(\bar{\theta})\geq 0$ *and* $R^{RS}\left(\theta^{*}\right)$ *is concave. Then introducing priority service is surplus-enhancing for consumers if and only if*

$$F\left(\theta''\right)\mathbb{E}\left[c\left(\theta\right)\right]-\int_{0}^{\theta''}c\left(\theta\right)f\left(\theta\right)d\theta-F\left(\theta''\right)\left(c\left(\bar{\theta}\right)-c\left(\theta''\right)\right)\geq 0 \tag{6}$$

where θ'' *solves*

$$c\left(\bar{\theta}\right)-c\left(\theta''\right)=c'\left(\theta''\right)\frac{F\left(\theta''\right)}{f\left(\theta''\right)}.$$

We now present a relatively simple condition that is sufficient for PS to reduce consumer surplus. For this result we need to introduce a slightly different assumption on the distribution of types.

Definition 1 *The distribution function* F *satisfies a decreasing reversed failure rate (DRFR) if* $F(\theta)/f(\theta)$*is increasing in*[27] $\theta\in\left[0,\bar{\theta}\right]$.

Corollary 2 *Assume that* $\left(v'\left(\bar{\theta}\right)-\frac{c'\left(\bar{\theta}\right)}{2}\right)+\left(v\left(\bar{\theta}\right)-c\left(\bar{\theta}\right)\right)f(\bar{\theta})\geq 0$ *and* $R^{RS}\left(\theta^{*}\right)$ *is concave. If* $c''\left(\theta\right)\geq 0$ *and* F *satisfies DRFR, then introducing priority service decreases consumer surplus.*

One possible interpretation of Proposition 5 and Corollary 2 is that for the priority service to increase consumer surplus, it must be the case that the market is not fully covered, and hence introducing priority service increases service inclusion by adding to the coverage those who are the most impatient (who are also those who value the basic good the most). This result contrasts with Section 5.1 where the dominant effect of the type was assumed to be on the value of the basic good, and where we have shown that consumer surplus improvement due to PS can take place also without expanding consumption.

The conditions of the last proposition guarantee that the monopolist will set the prices such that there will be full coverage of the market. It is worthwhile to understand the difference between Propositions 4 and 6. If the indifferent type is θ'' then the difference between the consumer surplus in these two settings stems from differences in the increase of the monetary payments as a result of the introduction of priority service: in the setting of Proposition 4 the increase is $\left(1-F\left(\theta''\right)\right)\left(c\left(\theta''\right)-c\left(0\right)\right)$, while in the setting of Proposition 6 it is $F\left(\theta''\right)\left(c\left(\bar{\theta}\right)-c\left(\theta''\right)\right)$. In the setting of Proposition 6 the marginal consumer is type $\bar{\theta}$, and hence introducing priority services makes every priority customer worse off relative to the case of no priority service (while type $\bar{\theta}$ is indifferent between these two cases). Customers of the regular service get later service, sometimes at a lower price (see the next proposition), yet the price reduction of the regular service, if any, is insufficient to increase the overall consumer surplus. In fact, as the next proposition shows, in this setting, the price of the regular service may even increase. By contrast, in the setting of Proposition 4 the marginal type is 0, and hence introducing priority service, if sufficiently inclusive, increases aggregated

[27] Since $\left(\frac{f(c)}{F(c)}\right)'=\left(\ln F(c)\right)''$, DRFR is equivalent to the log-concavity of F. See Bagnoli and Bergstrom [8] for a discussion of log-concave distribution functions.

consumer surplus as many priority consumers get earlier service and the price they are paying isn't too high. The customers of the regular service have relatively low waiting costs, and so their utility is not effected substantially (while type 0 gets the same utility as without priority service). See Example 3 for an illustration of this point.

Our next proposition shows that (under some conditions) even the price of the regular service may increase as a result of the introduction of priority service. In this case the prices of both the regular and the priority service will be higher than the price of the regular service if priority service is not provided, i.e., $\Pi > \pi \geq p$, and hence all types of agents are worse off as a result of the introduction of priority service: while type $\bar{\theta}$ is indifferent, lower types that acquire priority service are worse off because the saved time in the priority service is insufficient to compensate for the higher price of priority service. All types of agents that consume the regular service are also worse off, because they get later service and pay a higher price as a result of the introduction of priority service. We note that such a dramatic loss in consumer welfare cannot take place under the assumptions we had in Section 5.1, where the dominant effect of the type is on the value of the regular service.

Proposition 7 *Assume that* $\left(v'\left(\bar{\theta}\right) - \frac{c'\left(\bar{\theta}\right)}{2}\right) + \left(v\left(\bar{\theta}\right) - c\left(\bar{\theta}\right)\right) f(\bar{\theta}) \geq 0$ *and* $R^{RS}\left(\theta^*\right)$ *is concave. If* $c\left(\theta\right)/F\left(\theta\right)$ *is non-decreasing then introducing priority service makes all but type* $\bar{\theta}$ *consumers strictly worse off.*

Remark 3 *This model allows us also to analyze the provider's incentives to invest in server capacities. Assume that the individual time of the service is* τ *instead of 1, and the provider can, at some cost, increase its capacity which in turn reduces the service time* τ*. In this case, under the conditions of Propositions 4 and 6, the introduction of priority service decreases the incentives to invest in capacities. Intuitively, priority service becomes less valuable if the provider offers low waiting time (and hence the regular service is relatively fast), and hence it doesn't allow the provider to charge a high premium for priority service. Hence, if the provider offers priority service, he is more reluctant to invest in increasing capacity.*

Our analysis thus far has examined the prevalence of the loss in consumer surplus due to priority service with and without endogenous pricing for the regular service. We now continue by focusing on two other aspects of priority services. One concerns the presence of multiple levels of priority service, and the other concerns the extent to which competition can alleviate the loss in consumer welfare caused by priority services. We provide this analysis by assuming the price of the regular service to be exogenous. As both these aspects require an extension of the model along different dimensions, adding the endogeneity of this price would render the model intractable. Yet, at least as far as competition is concerned, the intuition for the forces that limit it seems to apply also in the endogenous case.

6 Multiple Priority Levels

In this section we analyze the case of multiple priority level and show that also here the seller can extract more than the total benefits he provides with the priority service. Very often providers offer more than one priority level.[28] We adopt here the same assumptions on the

[28] In addition to regular and priority service, providers often also offer super-priority service that gives priority over all other categories.

consumers' utilities as in Section 3. Assume now that the provider sets k priority classes with prices $0 = p_1 < p_2 < ... < p_{k-1} < p_k$, where buying priority level l means that the customer will be served after all the customers who buy higher priority classes, $\{l+1,..,k\}$.[29] There exists the basic service which is provided for free. The customers from the same priority class are served in random order. Any list of k ordered prices divides the market into k categories (some may be empty). This division is specified by the cutoff types.[30] The cutoff type c_i, $i \in \{1, ..., k-1\}$ is indifferent between being in the class i and paying price p_i, on the one hand, and being served in priority class $i+1$ and paying price p_{i+1}, on the other. That is,

$$-p_i - \left[1 - F(c_i) + \frac{F(c_i) - F(c_{i-1})}{2}\right] c_i = -p_{i+1} - \left[1 - F(c_{i+1}) + \frac{F(c_{i+1}) - F(c_i)}{2}\right] c_i,$$

which implies that[31]

$$p_{i+1} = p_i + \frac{F(c_{i+1}) - F(c_{i-1})}{2} c_i.$$

This equation illustrates the trade-off that the marginal customer with type c_i faces when contemplating what priority level to choose (either level i or $i+1$). The additional costs that this customer has to pay to be upgraded to the higher priority level is $p_{i+1} - p_i$, and the time saving associated with this upgrade is $\frac{F(c_{i+1}) - F(c_{i-1})}{2}$.

This recursive specification of prices allows us to write them as

$$p_i = \sum_{j=1}^{i-1} \frac{F(c_{j+1}) - F(c_{j-1})}{2} c_j. \tag{7}$$

Due to this recursive structure the change in the cutoff c_i affects the prices of all higher priority categories. More precisely, assume that the cutoff of category i, c_i decreases by ϵ to $c_i - \epsilon$. It switches some customers from priority class i to a higher class $i+1$. Such a change increases the size of priority class $i+1$ and so all customers in this, now larger class are willing to pay less than with the cutoff c_i. This shift decreases the size of priority class i. Moreover, since more customers now belong to the higher priority classes than with cutoff c_i, customers in priority class i are willing to pay less than before the shift. However, the increased size of class $i+1$ allows the provider to charge class $i+2$ a higher amount as the option to join class $i+1$ has become less valuable. Such a domino effect recursively influences the prices of all the higher priority classes. In the optimal mechanism the monopolist chooses cutoffs that exactly balance these two effects, i.e., the decrease in the revenues from priority class $i+1$ and the increase in the revenues from customers of higher priority classes.

---

[29] Restricting attention to strictly monotone prices is without loss, as if there exists l such that $p_l \leq p_{l-1}$, then it means that at least one category is empty and there exists a sequence of monotone prices (with less categoies) that generates the same allocation and the same utilities to all participants.

[30] Choosing cutoffs is equivalent to choosing quantities. It is well known that there is no difference between a monopolist that optimally chooses prices and a monopolist that optimally chooses quantities.

[31] We use notation of $c_0 = 0$, and $c_k = \bar{c}$.

The provider's revenues are given by

$$
\begin{aligned}
R &= \sum_{i=2}^{k} p_i \left[F\left(c_i\right) - F\left(c_{i-1}\right)\right] \\
&= \sum_{i=2}^{k} \left[F\left(c_i\right) - F\left(c_{i-1}\right)\right] \sum_{j=1}^{i-1} \frac{F\left(c_{j+1}\right) - F\left(c_{j-1}\right)}{2} c_j = \sum_{i=1}^{k-1} \left(1 - F\left(c_i\right)\right) \frac{F\left(c_{i+1}\right) - F\left(c_{i-1}\right)}{2} c_i.
\end{aligned}
$$

One may expect that when the number of priority classes is large, then since the resulting allocation is more efficient than in case of random allocation, consumers are able to keep part of the increase in the total efficiency. However, as our next result shows, this is not the case.

Proposition 8 *Assume that F satisfies the IFR property. Then for any number of priority classes $k > 1$, the customers' welfare if priority service is not offered is higher than if priority service is offered.*

The proof of the last proposition is recursive: for any allocation cutoffs with $l-1$ priority classes $\{c_1, ...c_{l-2}\}$ when $c_{l-2} < \bar{c}$, adding another priority class that splits the highest interval of waiting costs $[c_{l-2}, \bar{c}]$ into two, i.e., having l priority classes with allocation cutoffs $\{c_1, ...c_{l-1}\}$, always decreases the aggregated consumers' surplus. Repeating this argument k times, each time adding another priority class, implies the result.

The previous proposition does not use the profit maximization arguments of the provider. However, as the next proposition shows, the provider's profits are strictly increasing in the number of classes.

Proposition 9 *Assume that F satisfies the IFR property; then, the provider's profits are strictly increasing in the number of priority classes.*

The next example shows the optimal mechanism in the case of k priority classes where the waiting costs are distributed according to the uniform distribution on $[0, 1]$.

Example 4 *Assume a uniform distribution of waiting costs with support $[0, 1]$. If there are k categories, the optimal price for priority category $i \in \{1, ..., k\}$ is*

$$p_i = \frac{i\left(i-1\right)}{2k^2}.$$

The provider's optimal revenues and the corresponding customers' surplus are

$$
\begin{aligned}
R &= \frac{1}{6}\left[1 - \frac{1}{k^2}\right], \\
CS &= -\frac{1}{3} + \frac{1}{12k^2}.
\end{aligned}
$$

Therefore, in the limit (when $k \to \infty$) the customers' surplus goes to $-1/3$. Recall that without any priority, the customers' surplus was $-\mathbb{E}(c)/2 = -1/4$ (which corresponds to $k = 1$ in the expressions above). Hence in the limit, the total efficiency gain in the equilibrium allocation is $1/12$ (the total waiting costs decrease from $-1/4$ to $-1/6$) while the service provider gets twice! as much revenue $(1/6)$.

7 Competition

As we argued earlier, the market of priority is best described as a monopoly due to the hold-up problem arising from the fact that priority is typically offered by the same provider who provides the primary good to which the customer has already committed him/herself. Yet the exploitative nature of priority service presents itself also in a more competitive environment. To show this, we will now study a model of competition between two identical providers. Each provider is able to serve the entire market and has a cost normalized to zero. We shall show that with homogeneous costs not only it is the case that priority service reduces customers' welfare but also that, in spite of the fact that providers can compete over the price of priority, the equilibrium price ends up being identical to the monopoly price (with the appropriate adjustment for the increase in market service capacity). In the case of heterogeneous costs, we show that for a large class of the distribution functions for any possible prices of the priority these providers can set, introducing priority service reduces customers' welfare. We further show in an example that even outside of this class of the distributions, the equilibrium prices of priority service reduce customers' welfare.

We assume that two providers at the first stage simultaneously choose prices p_1 and p_2 for their priority services. At the second stage customers decide whether to join a queue of provider 1 or provider 2 and whether to buy the priority service of that provider or to get the regular service. Our assumption regarding the duopoly market is that the price of the primary service (for which priority is offered) is already fixed and identical for the two providers. This assumption can be interpreted as the outcome of a Bertrand competition over the primary service. It is also very relevant for services that are offered for free or at a fixed, regulated price (such as health insurance under national schemes) and customers pay only for add-ons and priority services.

7.1 Homogeneous Costs

We first assume that all customers have the same (linear) waiting costs, normalized to 1. Denote by $n_i^p(p_1, p_2)$ the share of customers who acquire priority services from service provider i if the prices of the providers for their priority services are p_1 and p_2, and denote by $n_i^{np}(p_1, p_2)$ the share of customers who join the regular service of provider i. The total share of customers of provider i is $n_i = n_i^p + n_i^{np}$, where $n_1 + n_2 = 1$. Like in the monopoly case, if

$$p_i \leq \frac{n_i}{2},$$

then the customers of provider i prefer priority service to regular service of that provider. Unlike the monopoly case, however, each customer has more options, as he may join the service of the other provider.

We show that in a unique pure strategy subgame-perfect equilibrium the providers set the prices of $\left(\frac{1}{4}, \frac{1}{4}\right)$ and the customers are divided such that $n_1^p = n_2^p = 1/2$. Thus, in equilibrium all customers get the priority service and each provider essentially gets the monopoly profits from half of the market.

Proposition 10 *In a unique pure strategy subgame-perfect equilibrium, prices are* $\left(\frac{1}{4}, \frac{1}{4}\right)$ *and the customers are divided such that* $n_1^p = n_2^p = 1/2$.

Our proof of the last proposition is insightful as it reveals the forces that cripple competition in a market of priority service. In contrast to a standard good in a market of priority service, a price cut does not guarantee a provider a larger clientele as the expansion of the clientele will make the service less attractive to existing customers who might prefer to move to the regular service and save on priority charges. More specifically, at the first stage of the proof we show how the clientele chooses between the providers and between their services for any possible pair of prices, p_1 and p_2. At the second stage we derive the optimal responses of each provider and characterize the equilibrium prices.

It is interesting to notice that although the service providers compete à la Bertrand, the equilibrium outcome is very different from the standard Bertrand competition with a perfectly competitive price. While increasing the price above 1/4 would lead to losing all priority customers similar to a Bertrand competition, decreasing the price would not attract all the customers to that provider, and this deviation would lead to an outcome very different from that of the standard Bertrand competition. Imagine that provider 1 deviates and undercuts its competitor by offering a price of $\frac{1}{4}-\varepsilon$. Observe that if in this case some share of customers move from provider 2 to the cheaper provider 1, it would take away all the priority service customers of provider 2 – as now it would hold that $p_2 > \frac{n_2}{2}$ as $p_2 = \frac{1}{4}$ and $n_2 < \frac{1}{2}$ – leaving provider 2 with only regular customers. In such a case provider 1 cannot serve in its priority service even half of the market, and, clearly, having only regular customers at provider 2 and only priority customers at provider 1 is not part of a subgame equilibrium**.** In the proof of the above proposition we show that undercutting the competitor and charging a price of $\frac{1}{4} - \varepsilon$ **does not change** the priority clientele of that provider and only causes a division of the competitor's clientele into priority and regular customers such that each customer gets exactly the same utility (as a priority customer at either of the providers or as a regular customer of provider 2). This argument essentially shows that prices $\left(\frac{1}{4}, \frac{1}{4}\right)$ are part of an equilibrium. The formal proof will show that it is the unique one.

7.2 Heterogeneous Costs

Now we assume that the two competitive providers are facing heterogeneous customers with different linear waiting costs. We adopt the assumptions that were introduced in Section 3 regarding the distribution of the individual waiting costs. Denote by p_1 and p_2 the prices for the priority service of providers 1 and 2, respectively. Assume without loss of generality that at the first stage the providers set prices such that $p_1 \geq p_2$. We first characterize the customers' optimal choices for a given pair of priority prices (p_1, p_2). We then analyze the optimal prices set by the providers.

Assume the following equilibrium structure: share $n_1^p(p_1, p_2)$ of customers with the highest waiting costs (with cost parameters between $c_1^*(p_1, p_2)$ and $\overline{c}$) choose priority service from provider 1, share $n_2^p(p_1, p_2)$ of customers with relatively high costs (with cost parameters between $c_2^*(p_1, p_2)$ and $c_1^*(p_1, p_2)$) choose priority service from provider 2 with $n_1^p \leq n_2^p$. Moreover, share $n_1^{np}(p_1, p_2)$ of customers with waiting costs below $c_2^*(p_1, p_2)$ get regular service from provider 1, and share $n_2^{np}(p_1, p_2)$ of customers with waiting costs below $c_2^*(p_1, p_2)$ get regular service from provider 2.

$$\underset{0}{|}\underbrace{-\,-}_{n_1^{np}+n_2^{np}}\underset{c_2^*}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-}_{n_2^{p}}\underset{c_1^*}{|}\underbrace{-\,-\,-\,-\,-\,-\,-}_{n_1^{p}}\underset{\bar{c}}{|}$$

For any p_1 and p_2 with $p_1 \geq p_2$ the customers' decisions satisfy the following conditions:

(1) A customer with a waiting cost of c_1^* is indifferent between provider 1's and 2's priority service[32]:

$$-p_1 - c_1^* \frac{1 - F(c_1^*)}{2} = -p_2 - c_1^* \frac{F(c_1^*) - F(c_2^*)}{2}.$$

(2) A customer with a waiting cost of c_2^* is indifferent between provider 2's priority service and any regular service

$$-p_2 - c_2^* \frac{F(c_1^*) - F(c_2^*)}{2} = -c_2^* \left[F(c_1^*) - F(c_2^*) + \frac{F(c_2^*) - n_1^{np}}{2} \right].$$

(3) The expected waiting time in both providers' regular service is the same:

$$1 - F(c_1^*) + \frac{n_1^{np}}{2} = F(c_1^*) - F(c_2^*) + \frac{F(c_2^*) - n_1^{np}}{2}.$$

Our first result here shows that these three conditions imply unique division of the consumers between the services and the providers.

Lemma 3 *For any $p_1 \geq p_2$, there exists a unique equilibrium division of the customers into priority and regular customers; i.e., there exist unique $c_1^* \geq c_2^* \geq 0$ that satisfy requirements (1)–(3).*

Moreover, we show that a zero price cannot be a part of the equilibrium strategy of the competition game.

Proposition 11 *There is no pure strategy equilibrium with $p_i = 0$ for $i \in \{1, 2\}$.*[33]

The reason that equilibrium prices in both duopoly models (with both homogeneous and heterogeneous customers) are very different from the standard competitive equilibrium in the Bertrand competition is the negative externalities that customers impose on each other. By joining a queue, each customer prolongs the waiting time of all customers in this queue. These negative externalities generate market power to the service providers, which reduces competition: in the duopoly competition case (in both cases of homogeneous and heterogeneous customers), charging a lower price cannot increase the priority customer base substantially, since in the case of an increase in the number of priority customers, this service becomes less valuable to other customers.

[32] In the case of a symmetric equilibrium with $p_1 = p_2$, this condition implies that $1 - F(c_1^*) = F(c_1^*) - F(c_2^*)$. That is, each provider has the same share of priority customers.

[33] This argument applies also to the setup in which the competitors set the prices of both of their services endogenously.

We show next that in a big, parameterized class of distribution functions competition over priority services makes customers' welfare lower than completely random allocation - without any priority service. That is, despite the competition, even in case of heterogeneous customers the providers are able to extract from the consumers payments in excess of the increase in the allocative efficiency.

Proposition 12 *Assume that $F(c) = c^{\mu}$ with $\mu \geq 1$. Then the customers' welfare if priority service is not available is higher than if priority service is offered by the providers.*

For $\mu \geq 1$ any positive prices for priority service decrease the consumers surplus, not only the optimal, equilibrium prices. However, as our next example shows for $\mu < 1$ the equilibrium prices are such that the consumers' welfare is lower than in the market without priority service. For its derivation see Supplementary Material appendix.

Example 5 *Assume now that the distribution of types is given by $F(c) = \sqrt{c}$ for $c \in [0,1]$. Then the equilibrium cutoffs are $c_1 = 0.67336$ and $c_2 = 0.34744$. Hence, the providers set prices $p_1 = 0.0998$ and $p_2 = 0.08237$. The sets of priority customers of both providers are $n_1^p = 0.17941$ and $n_2^p = 0.23114$, respectively. The profits of providers 1 and 2 are $\pi_1 = 0.0179$ and $\pi_2 = 0.019$, respectively. Customers' surplus equals to -0.0878, while if priority service is not available customers' surplus is*

$$-\frac{\int_0^1 \frac{1}{2}\sqrt{s}ds}{4} = -0.083.$$

Hence introducing priority service lowers customers' surplus.

8 Conclusion

In this paper we analyzed the welfare effects of priority service on consumers and show that priority service may allow a service provider to extract in revenue more than the efficiency gains, yielding an overall net loss to the consumers. Furthermore, we show that it can also harm competition between service providers due to an implicit deterrence mechanism whereby attracting more customers by a price reduction may create congestion that will induce other consumers to leave. Our primary results show that this excessive extraction takes place when the price of the regular service is fixed and the pricing game involves only the priority service, and we argue that several important real-life markets are consistent with this market condition.

Our primary focus on the case of loss in customers welfare due to the presence of priority service is driven by the fact that we find such excessive extraction of surplus surprising, and because the case of a loss in customers' welfare is the more relevant case for policy making. However, priority service can enhance customers welfare in some market environments. As we have shown, such enhancement is likely to take place when the priority service substantially expands the customers coverage by attracting new customers who due to their waiting cost would not consume without it. Even when the customers coverage does not expand due to priority service it can still increase customers' welfare and we provide sufficient conditions under which this can take place.

The contribution of this paper for public policy is in highlighting the fact that the loss in customers' welfare due to priority service is possible and can be rather prevalent. If a policy maker wishes to control for such a loss in customers welfare it is important to identify the market conditions that are conducive to the excessive extraction of revenue and to the impediments to competition in markets of priority services. Any such policy will also need to identify the losers and gainers that PS might create in order to design a compensation scheme that will improve customers welfare and even bring about a Pareto improvement. Gershkov and Schweinzer [23] showed conditions under which such policy is implementable using equivalence between queueing problem and partnership resolution problem (see Cramton et al. [16]).

There are several reasons why regulators might find intervention in priority service justified. Firstly, unlike markets where the impediment to competition is caused by an exogenous constraint, such as a capacity constraint (see Compte, Jenny, and Rey [15]), in PS markets it is generated by a trading practice of the firms themselves (similarly to the trading practice of exclusive dealing; see Bernheim and Whinston [9]). Secondly, as we have shown, PS facilitates (tacit) collusion in oligopolistic markets of priority service and impedes competition. Thirdly, PS provides negative incentives to invest in innovations to reduce service time and make lines shorter. The general effect of impediments to competition on investment is in general quite ambiguous (see Aghion, Harris, and Vickers [4], Aghion, Harris, Howitt, and Vickers [3], and Aghion, Dewatripont, and Rey [2]), but in our framework it is more straightforward: if firms can make money out of selling PS to customers who suffer from waiting in long and slow lines, and can sustain this extra revenue through a tacit collusion, why should they invest in innovation that might reduce the waiting cost of their regular service?

Unlike exclusive dealing, which is prohibited by competition authorities, PS need not be prohibited even under market conditions that induce an overall loss in customers' welfare. Remedies that alleviate and even eliminate such loss are likely to exist. Designing a regulatory policy that achieves this goal is challenging, and this is not the prime purpose of our paper. Some measures may turn out to be ineffective: a cap on priority service prices, for example, would hardly be successful. It would clearly increase their demand and hence impose harsher consequences on a smaller group of customers who can't even afford to pay the capped price. An alternative policy would be to limit the number of consumers who receive priority service. This policy would guarantee a cap on the welfare loss for customers of the regular service at the expense of some gains from trade in priority rights. But neither of these policies can guarantee that priority service will enhance the overall customer surplus and, as we show, under a mild assumption on the distribution of the individual costs, they will not. The one type of policy that can guarantee it is one that facilitates trade in priority rights. In fact, such a policy guarantees not only that the total aggregate consumer surplus from the priority service is positive but also that priority service constitutes a Pareto improvement, i.e., by making all consumers and the service provider better off relative to the benchmark of no priority service. To facilitate such a policy, it would be required that the service provider price only the regular service. All upgrades to priority service would then be auctioned out with an incentive-compatible mechanism that transfers money between customers whereby priority customers compensate regular ones (see Kittsteiner and Moldovanu [31]).

9 Appendix A. Non-linear Cost

The utility function of the customers as presented in our model is derived from the customers' waiting costs, which are assumed to be linear in the waiting time. In this appendix we show that our results are robust to this assumption by studying the cases in which these waiting costs are either concave or convex. A priori it is not clear which of these two scenarios is more adequate for representing queueing disutility. On the one hand, cost functions in economics are typically assumed to be convex. However, the convexity of a cost function is very intuitive only in the context of production as it reflects the idea that low-hanging fruits that are picked first are less costly to pick than those that remain after the early ones are gone. In queues, however, the shape of the cost function depends either on individuals' mental discomfort from waiting, or on the way opportunities disappear due to delays. It seems to us that one can come up with arguments in favor of both concavity and convexity in both interpretations. Hence we consider both options.

We assume that waiting t units of time creates disutility of $g(t) \geq 0$, where $g(t)$ is increasing, bounded and differentiable. We first analyze the homogeneous customers. There is mass of size 1 of customers. We allow the disutility function g to be either convex or concave and we further assume that $g(0) = 0$. Assume that the provider sets a price of p for priority. If share s of the customers joins the priority service, the expected utility of a marginal customer in the priority service is

$$-p - \frac{\int_0^s g(t)\,dt}{s},$$

while the expected utility of the marginal customer in the regular service is

$$-\frac{\int_s^1 g(t)\,dt}{1-s}.$$

In case of linear waiting costs, the benefits from joining the priority service are independent of the number of other customers who join the service, and it is the reason for customers to have a dominant action. This is not the case with nonlinear costs. While joining the priority service still improves the averaged position by $1/2$ independently of the action of the other customers, in the case of nonlinear value of time the effect of such an improvement on a customer's utility depends on the actions of the other customers (or on the sizes of priority queue and regular queue). The benefits from joining the priority service if share s of customers joins are

$$B(s) = -\frac{\int_0^s g(t)\,dt}{s} + \frac{\int_s^1 g(t)\,dt}{1-s}.$$

Lemma 4 *Assume that g is concave, then $B(s)$ is decreasing. If g is convex, then $B(s)$ is increasing.*

Proof. Observe that

$$B'(s) = \frac{-g(s)}{1-s} + \frac{\int_s^1 g(t)\,dt}{(1-s)^2} - \frac{g(s)}{s} + \frac{\int_0^s g(t)\,dt}{s^2}.$$

Assume that g is concave. Since g is differentiable, the mean value theorem implies that for any $t < s$ there exists $z(t) \in (t, s)$ such that $g(t) = g(s) - g'(z(t))(s-t)$. Similarly, for any $t > s$ there exists $z(t) \in (s, t)$ such that $g(t) = g(s) + g'(z(t))(t-s)$. Plugging the expressions for $g(t)$ gives

$$\begin{aligned} B'(s) &= \frac{-g(s)}{1-s} + \frac{\int_s^1 (g(s) + g'(z(t))(t-s))\,dt}{(1-s)^2} - \frac{g(s)}{s} + \frac{\int_0^s (g(s) - g'(z(t))(s-t))\,dt}{s^2} \\ &= \frac{\int_s^1 g'(z(t))(t-s)dt}{(1-s)^2} - \frac{\int_0^s g'(z(t))(s-t)dt}{s^2} < \frac{\int_s^1 g'(s)(t-s)dt}{(1-s)^2} - \frac{\int_0^s g'(s)(s-t)dt}{s^2} \\ &= g'(s)\left[\frac{\int_s^1 (t-s)dt}{(1-s)^2} - \frac{\int_0^s (s-t)dt}{s^2}\right] = 0, \end{aligned}$$

where the inequality follows from the concavity of g, which implies that g' is decreasing. For convex g we get the opposite inequality. ■

Hence, for concave g the benefits from joining decrease with the share of customers that join the priority service. Therefore, for p such that[34]

$$B(0) \geq p \geq B(1),$$

a share s of customers that satisfies

$$p = \frac{\int_s^1 g(t)\,dt}{1-s} - \frac{\int_0^s g(t)\,dt}{s} \tag{8}$$

join the priority service. The next proposition generalizes the characterization of the equilibrium of linear case to concave and convex waiting costs.

Proposition 13 *Assume that g is concave, then in a unique subgame perfect equilibrium outcome all customers join the priority service and the optimal price is*

$$p^* = g(1) - \int_0^1 g(t)\,dt.$$

Assume that g is convex; then the optimal price is $p^* = g(1) - \int_0^1 g(t)\,dt$ and all customers join the priority service.

Proof. We start with concave waiting costs $g(t)$. The providers' profits are $sp(s)$, where $p(s)$ is given by (8). Therefore, the optimal share is

$$\max_s s\left(\frac{\int_s^1 g(t)\,dt}{1-s} - \frac{\int_0^s g(t)\,dt}{s}\right).$$

[34] We define $B(0) = \lim_{s\to 0} B(s) = \int_0^1 g(t)\,dt$ and $B(1) = \lim_{s\to 1} B(s) = g(1) - \int_0^1 g(t)\,dt$.

The derivative with respect to s is

$$
\begin{aligned}
& \frac{\int_s^1 g\,(t)\,dt}{1-s} - \frac{\int_0^s g\,(t)\,dt}{s} + s\left(-\frac{g(s)}{1-s} + \frac{\int_s^1 g\,(t)\,dt}{(1-s)^2} - \frac{g(s)}{s} + \frac{\int_0^s g\,(t)\,dt}{s^2}\right) \\
= \; & \frac{\int_s^1 g\,(t)\,dt}{1-s}\left(1+\frac{1}{1-s}\right) - g(s)\left(1+\frac{1}{1-s}\right) = \frac{1}{(1-s)^2}\left(\int_s^1 g\,(t)\,dt - g(s)(1-s)\right) \\
= \; & \frac{1}{(1-s)^2}\int_s^1 \left(g\,(t) - g(s)\right) dt > 0,
\end{aligned}
$$

where the inequality follows from monotonicity of g. Therefore, the optimal share is $s^* = 1$. Hence, the optimal price is

$$
p^* = p(1) = \lim_{s=1}\left(\frac{\int_s^1 g\,(t)\,dt}{1-s} - \frac{\int_0^s g\,(t)\,dt}{s}\right) = \lim_{s=1}\frac{-g\,(s)}{-1} - \int_0^1 g\,(t)\,dt = g\,(1) - \int_0^1 g\,(t)\,dt.
$$

Assume now convex g. Since Lemma 4 implies that $B(s)$ is monotone, any price $p \in [B(0), B(1)]$ will lead all customers to join the priority service. Therefore, the optimal price is $p^* = B(1)$. ■

At the optimal price, similarly to the linear cost with homogeneous case, all customers join the priority service and hence the total waiting cost of these customers is the same as that of the customers without priority service, and so the transfer to the service provider is just the customers' surplus extraction.

There is a substantial difference between the convex and concave waiting cost cases. While in the case of concave g the equilibrium is unique and is in (weakly) dominant strategies, in the case of convex waiting costs the more customers that join the priority service, the more substantial is time-saving effect on the utility from joining it. Therefore, for convex g, if $p < B(0)$, it is optimal for all customers to join the priority service, and if $p > B(1)$ it is optimal for customers to use the regular service. For $p \in (B(0), B(1))$ we have three possible equilibria: (1) all customers join priority service, (2) all customers join regular service, (3) a share s of customers where s satisfy

$$
p = \frac{\int_s^1 g\,(t)\,dt}{1-s} - \frac{\int_0^s g\,(t)\,dt}{s}
$$

join the priority service. Adopting the standard assumption of mechanism design for choosing the best equilibrium for the seller/mechanism designer allows us to conclude that the provider will set the price of $g(1) - \int_0^1 g\,(t)\,dt$, which is the highest price that attracts all customers to the priority service.

We introduce heterogeneity of the weighting costs into the nonlinear cost model by assuming a specific functional form of the cost. Assume that the disutility from waiting t periods of time is ct^{μ} where c is the individual cost parameter that follows the assumptions introduced in Section 3 and $\mu > 0$. For different values of μ this functional form allows for both convexity and concavity of the cost in waiting time. Similar to the linear cost case, by setting the price of p for its priority service, the provider divides the customers into two categories: customers with a higher cost parameter who join priority service and customers with a lower cost parameter who join the regular service. For a given price p, the expected utility of type c from joining priority service if all types with parameters $c \geq c^*$ join the priority service is given by

$$-p - c\frac{\int_0^{1-F(c^*)} t^\mu dt}{1 - F(c^*)} = -p - c\frac{(1 - F(c^*))^\mu}{\mu + 1},$$

while that type's expected utility from the regular service is

$$-c\frac{\int_{1-F(c^*)}^1 t^\mu dt}{F(c^*)} = -c\frac{1 - (1 - F(c^*))^{\mu+1}}{(\mu + 1) F(c^*)}.$$

Therefore, for a given price p, types above c^* join the priority service, while types below c^* join the regular service, where c^* solves[35]

$$\begin{aligned} p &= \frac{c^*}{\mu + 1}\left(\frac{1 - (1 - F(c^*))^{\mu+1}}{F(c^*)} - (1 - F(c^*))^\mu\right) \\ &= \frac{c^*}{\mu + 1}\frac{1 - (1 - F(c^*))^\mu}{F(c^*)}. \end{aligned}$$

The provider's problem is to choose the cutoff that maximizes its expected profits:

$$\max_{c^o \in [0,\bar{c}]} c^o \frac{1 - F(c^o)}{F(c^o)} \frac{1 - (1 - F(c^o))^\mu}{\mu + 1}.$$

The next proposition generalizes Proposition 2 to nonlinear waiting costs.

Proposition 14 *Assume that F satisfies the IFR property. Assume further that $\mu > 0$. Then the customers' welfare if priority service is not available is higher than if priority service is offered.*

Proof. Similarly to the linear case we will show that for any price of the priority service, and not necessarily the optimal one, the customers' welfare if priority service is offered is lower than if this service is not offered. If the provider sets the price that induces cutoff $c^* \in [0, \bar{c}]$ that divides the customers into two categories, the total customers' welfare is

$$\begin{aligned} & -\int_0^{c^*} c\frac{\int_{1-F(c^*)}^1 t^\mu dt}{F(c^*)} f(c)dc - \int_{c^*}^{\bar{c}} \left(p + c\frac{\int_0^{1-F(c^*)} t^\mu dt}{1 - F(c^*)}\right) f(c)dc \\ = & -\frac{1 - (1 - F(c^*))^{\mu+1}}{(\mu + 1) F(c^*)} \int_0^{c^*} cf(c)dc - c^* \frac{1 - F(c^*)}{F(c^*)} \frac{1 - (1 - F(c^*))^\mu}{\mu + 1} - \frac{(1 - F(c^*))^\mu}{\mu + 1} \int_{c^*}^{\bar{c}} cf(c)dc. \end{aligned}$$

The customers' welfare if no priority service is offered is

$$-E(c) \int_0^1 t^\mu dt = -\frac{E(c)}{\mu + 1}.$$

[35] While it may not follow immediately from the equation below, there is at most one indifference type for any price.

We show that

$$-\frac{1-(1-F(c^*))^{\mu+1}}{(\mu+1)F(c^*)}\int_0^{c^*} cf(c)dc - c^*\frac{1-F(c^*)}{F(c^*)}\frac{1-(1-F(c^*))^{\mu}}{\mu+1} - \frac{(1-F(c^*))^{\mu}}{\mu+1}\int_{c^*}^{\bar{c}} cf(c)dc \leq -\frac{E(c)}{\mu+1} \tag{9}$$

for any $c^* \in [0,\bar{c}]$ and $\mu \geq 0$. Rearranging (9) gives

$$\left(-\frac{1-(1-F(c^*))^{\mu+1}}{F(c^*)} + (1-F(c^*))^{\mu}\right)\int_0^{c^*} cf(c)dc - c^*\frac{1-F(c^*)}{F(c^*)}\left(1-(1-F(c^*))^{\mu}\right)$$
$$-(1-F(c^*))^{\mu}\int_0^{\bar{c}} cf(c)dc \leq -E(c).$$

We can rewrite the last inequality as

$$-\frac{1-(1-F(c^*))^{\mu+1}-(1-F(c^*))^{\mu}F(c^*)}{F(c^*)}\int_0^{c^*} cf(c)dc -$$
$$c^*\frac{1-F(c^*)}{F(c^*)}\left(1-(1-F(c^*))^{\mu}\right) \leq -\left(1-(1-F(c^*))^{\mu}\right)E(c)$$

$$\begin{aligned} -\frac{1-(1-F(c^*))^{\mu}}{F(c^*)}\int_0^{c^*} cf(c)dc - c^*\frac{1-F(c^*)}{F(c^*)}\left(1-(1-F(c^*))^{\mu}\right) &\leq -\left(1-(1-F(c^*))^{\mu}\right)E(c) \\ -\frac{\int_0^{c^*} cf(c)dc}{F(c^*)}\int_0^{c^*} cf(c)dc - c^*\frac{1-F(c^*)}{F(c^*)} &\leq -E(c), \end{aligned}$$

which is independent of μ. Since we showed the inequality for the linear case ($\mu = 1$) in Proposition 2, this inequality holds for any $\mu > 0$. ■

Similarly to the linear case described in Proposition 2, the last proposition holds for any price such that the two classes are nonempty, and not necessarily the optimal price.

10 Appendix B. Private Provision of Service

In this part we show that a similar phenomenon of consumers' loss of welfare applies in a different economic environment where no priority service is introduced, but rather a better service takes resources from a common pool of resources. This situation well describes the allocation of teacher and student placements in a school system where the introduction of a private school attracts the better teachers that otherwise would remain in the public schools.

To illustrate this point we consider the following model. Assume that there exists a population of consumers with parameters/types $v \sim F[0,\bar{v}]$ that characterize their needs for service. There are k providers of the service with heterogeneous qualities where the quality of provider i is denoted by t_i when $t_1 > t_2 > \cdots > t_k > 0$. If a consumer with need v_i is allocated to a provider of quality t_j the consumer's utility is $v_i t_j$. Each provider can serve share $1/k$ of all consumers.

As a benchmark we first consider the case where there is no private provision, and hence the allocation of the consumers to the providers is random. In this case the random allocation between the consumers and the providers generates the expected welfare of consumers of $\bar{t}E[v]$ where $\bar{t} = \frac{t_1+t_2+\cdots+t_k}{k}$ is the average quality of the providers.

Now assume that the provider has quality t_1 and charges p. We can think about a monopoly firm that sells the service and is engaged in a contract with the single provider of quality t_1. Other providers remain in the public service and do not charge for the service. The capacity of each provider is unchanged and it is $1/k$ of the market. The price p is set such that the market clears, that is, a share of $1/k$ of the consumers are willing to pay the price p. Other consumers get randomly allocated providers from the set of remaining ones $t_2 > \cdots > t_k > 0$. The cutoff type v^* (the type that is indifferent between joining the private provider and remaining in the public system) satisfies

$$-p + v^* t_1 = v^* \hat{t}$$

where $\hat{t} = \frac{t_2+\cdots+t_k}{k-1}$. That is,

$$-p = v^* \left(\hat{t} - t_1\right).$$

Moreover, the market clearing implies that $1 - F(v^*) = 1/k$. The next proposition (proven in online appendix) shows that the introduction of a private service that uses pool of joint resources decreases the consumers' welfare.

Proposition 15 *Assume that $\mathbb{E}(v) - \frac{1-F(v)}{f(v)}$ changes sign only once from negative to positive; then the consumers' welfare if private service is not offered is higher than if private service is offered.*

As with Proposition 2, IFR assumption is sufficient for the single change of sign requirement of the last Proposition. Since the technical proof and the intuition of Proposition 15 are very similar to Proposition 2, we do not replicate all the remaining results we illustrate for priority service to private service setup.

The consumers' loss of welfare that emerges in Proposition 2 and Proposition 15 relies on the assumption that the additional revenue generated by the priority service is not invested in increasing capacity. In the two applications of health services and education discussed above, the validity of this assumption is a political economy question and depends on many other factors. Nevertheless, it is reasonable to assume that the adjustment of capacity has a long-run effect and that in the short run it is more limited and constrained.

10.1 Appendix C. Proofs for Section 5

10.1.1 Low-type exclusion.

Proof of Lemma 1. Assume that in the regime without priority service the provider optimally sets a price of p for its service. First, observe that setting $p = \pi$ and

$$c\left(\bar{\theta}\right) \frac{1 - F(\theta^*)}{2} + p = \Pi$$

implies that $\theta'' = \bar{\theta}$ and $\theta' = \theta^*$. Hence, the regime with priority service replicates the same revenue as the revenues in the regime with regular service only. The derivative of the provider's expected revenue (with respect to θ'') is

$$\frac{\partial R^{PS}(\theta', \theta'')}{\partial \theta''} = \left(1 - F(\theta')\right) \left(\frac{f(\theta'')}{2} c(\theta') + c'(\theta'') \frac{1 - F(\theta'')}{2} - c(\theta'') \frac{f(\theta'')}{2} \right).$$

This derivative at $\theta'' = \bar{\theta}$ equals to

$$\left(1 - F\left(\theta'\right)\right) \frac{f\left(\bar{\theta}\right)}{2} \left(c\left(\theta'\right) - c\left(\bar{\theta}\right)\right) < 0,$$

where the inequality follows because of the monotonicity of c and the assumption that $f(\theta) > 0$. Continuity guarantees that the derivative remains negative in some interval to the left of $\theta = \bar{\theta}$. ■

We prove Proposition 4 in the following chain of lemmata.

We first show that introducing priority service expands the clientele.

Lemma 5 *Assume that R^{RS} is concave. In the optimal mechanism it holds that $\theta' \leq \theta^*$.*

Proof. If the provider offers the regular service only, then the profits are given by

$$(1 - F(\theta^*)) \left(v(\theta^*) - c(\theta^*) \frac{1 - F(\theta^*)}{2}\right).$$

The optimal cutoff must satisfy

$$\left(v'(\theta^*) - c'(\theta^*) \frac{1 - F(\theta^*)}{2}\right) (1 - F(\theta^*)) - f(\theta^*) \left(v(\theta^*) - c(\theta^*)(1 - F(\theta^*))\right) \leq 0,$$

where the inequality holds if $\theta^* = 0$. Otherwise it holds as equality. We can rewrite the last inequality as

$$\left(v'(\theta^*) - c'(\theta^*) \frac{1 - F(\theta^*)}{2}\right) \frac{1 - F(\theta^*)}{f(\theta^*)} \leq v(\theta^*) - c(\theta^*)(1 - F(\theta^*)).$$

The concavity of R^{RS} guarantees a unique solution θ^*.

If the provider offers both the regular and the priority service, then the provider's profits are

$$\begin{aligned}
&\Pi\left(1 - F\left(\theta''\right)\right) + \pi\left(F\left(\theta''\right) - F\left(\theta'\right)\right) = \pi\left(1 - F\left(\theta'\right)\right) + (\Pi - \pi)\left(1 - F\left(\theta''\right)\right) \\
&= \left(v\left(\theta'\right) - c\left(\theta'\right)\left[1 - \frac{F\left(\theta''\right) + F\left(\theta'\right)}{2}\right]\right)\left(1 - F\left(\theta'\right)\right) + c\left(\theta''\right) \frac{1 - F\left(\theta'\right)}{2}\left(1 - F\left(\theta''\right)\right) \\
&= \left(1 - F\left(\theta'\right)\right)\left(v\left(\theta'\right) - c\left(\theta'\right)\left[\frac{1 - F\left(\theta''\right)}{2} + \frac{1 - F\left(\theta'\right)}{2}\right] + c\left(\theta''\right) \frac{1 - F\left(\theta''\right)}{2}\right).
\end{aligned}$$

The optimal cutoff θ' must satisfy

$$\begin{aligned}
&\left(1 - F\left(\theta'\right)\right)\left(v'\left(\theta'\right) - c'\left(\theta'\right)\left[\frac{1 - F\left(\theta''\right)}{2} + \frac{1 - F\left(\theta'\right)}{2}\right] + c\left(\theta'\right) \frac{f\left(\theta'\right)}{2}\right) \\
&- f\left(\theta'\right)\left(v\left(\theta'\right) - c\left(\theta'\right)\left[\frac{1 - F\left(\theta''\right)}{2} + \frac{1 - F\left(\theta'\right)}{2}\right] + c\left(\theta''\right) \frac{1 - F\left(\theta''\right)}{2}\right) \leq 0,
\end{aligned}$$

where the inequality holds if $\theta' = 0$. We can rewrite it as

$$\begin{aligned}
& \left(v'\left(\theta'\right) - c'\left(\theta'\right)\left[\frac{1-F\left(\theta''\right)}{2} + \frac{1-F\left(\theta'\right)}{2}\right]\right)\frac{1-F\left(\theta'\right)}{f\left(\theta'\right)} \\
\leq \quad & v\left(\theta'\right) - c\left(\theta'\right)\left(1-F\left(\theta'\right)\right) + \left(c\left(\theta^{''}\right) - c\left(\theta'\right)\right)\frac{1-F\left(\theta''\right)}{2}.
\end{aligned} \tag{10}$$

Since for any $\theta^{''} \geq \theta$ it holds that

$$\begin{aligned}
& v\left(\theta\right) - c\left(\theta\right)\left(1-F\left(\theta\right)\right) + \left(c\left(\theta^{''}\right) - c\left(\theta\right)\right)\frac{1-F\left(\theta''\right)}{2} \\
\geq \quad & v\left(\theta\right) - c\left(\theta\right)\left(1-F\left(\theta\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
& \left(v'\left(\theta\right) - c'\left(\theta\right)\left[\frac{1-F\left(\theta''\right)}{2} + \frac{1-F\left(\theta\right)}{2}\right]\right)\frac{1-F\left(\theta\right)}{f\left(\theta'\right)} \\
\leq \quad & \left(v'\left(\theta\right) - c'\left(\theta\right)\frac{1-F\left(\theta\right)}{2}\right)\frac{1-F\left(\theta\right)}{f\left(\theta\right)},
\end{aligned}$$

we have $\theta' \leq \theta^*$. ■

The next lemma shows that under the conditions of the proposition we have full coverage of the market.

Lemma 6 *If $\left(v'\left(0\right) - c'\left(0\right)\frac{1}{2}\right) \leq f\left(0\right)\left(v\left(0\right) - c\left(0\right)\right)$ and R^{RS} is concave, then $\theta^* = \theta' = 0$.*

Proof. The previous lemma implies that it is enough to check that $\theta^* = 0$. The FOC in the problem with the regular service only boils down to

$$\left(v'\left(0\right) - c'\left(0\right)\frac{1}{2}\right) - f\left(0\right)\left(v\left(0\right) - c\left(0\right)\right) \leq 0.$$

The concavity of R^{RS} guarantees that it is indeed a global maximum. ■

Proof of Proposition 4. With $\theta^* = 0$ we can express the consumer surplus if the provider offers the regular service only as

$$\begin{aligned}
& \int_0^{\bar{\theta}}\left(V\left(\theta\right) - V\left(0\right) - \frac{c\left(\theta\right) - c\left(0\right)}{2}\right)f\left(\theta\right)d\theta \\
= \quad & \mathbb{E}\left[V\left(\theta\right) - \frac{c\left(\theta\right)}{2}\right] - \left[V\left(0\right) - \frac{c\left(0\right)}{2}\right].
\end{aligned}$$

If the provider offers both the regular and the priority service, then the consumer surplus is given by

$$\int_0^{\bar{\theta}} \left(V(\theta) - V(0) - (c(\theta) - c(0)) \frac{1 - F(\theta'')}{2} \right) f(\theta)\, d\theta - \int_0^{\theta''} (c(\theta) - c(0)) \frac{1}{2} f(\theta)\, d\theta$$

$$- \int_{\theta''}^{\bar{\theta}} \left(c(\theta'') - c(0)\right) \frac{1}{2} f(\theta)\, d\theta$$

$$= \mathbb{E}\left[V(\theta) - \frac{c(\theta)}{2} \right] - \left[V(0) - \frac{c(0)}{2} \right] + \int_0^{\bar{\theta}} (c(\theta) - c(0)) \frac{F(\theta'')}{2} f(\theta)\, d\theta$$

$$- \int_0^{\theta''} (c(\theta) - c(0)) \frac{1}{2} f(\theta)\, d\theta - \int_{\theta''}^{\bar{\theta}} \left(c(\theta'') - c(0)\right) \frac{1}{2} f(\theta)\, d\theta.$$

Hence the introduction of priority service is welfare enhancing for consumers if and only if

$$\int_0^{\bar{\theta}} (c(\theta) - c(0)) \frac{F(\theta'')}{2} f(\theta)\, d\theta - \int_0^{\theta''} (c(\theta) - c(0)) \frac{1}{2} f(\theta)\, d\theta - \int_{\theta''}^{\bar{\theta}} \left(c(\theta'') - c(0)\right) \frac{1}{2} f(\theta)\, d\theta > 0.$$

We can rewrite the last inequality as

$$-\left(c(\theta'') - c(0)\right)\left(1 - F(\theta'')\right) + F(\theta'') \int_{\theta''}^{\bar{\theta}} c(\theta) f(\theta)\, d\theta - \left(1 - F(\theta'')\right) \int_0^{\theta''} c(\theta) f(\theta)\, d\theta > 0.$$

■

Proof of Lemma 2. Observe that if either $\theta'' = 0$ or $\theta'' = \bar{\theta}$, then the left-hand side of (5) equals zero. This is not surprising as it means that essentially there is only one service class, while the other is empty. The sign of the derivative of the left-hand side of (5) coincides with the sign of

$$\mathbb{E}[c(\theta)] - c(0) - c'(\theta'') \frac{1 - F(\theta'')}{f(\theta'')}.$$

Hence the derivative is positive if either θ'' is sufficiently high (close to $\bar{\theta}$) or if θ'' is very low and close to zero (assuming $c'(0)$ is low). This implies that priority service is beneficial for the customers if θ'' is low (the vast majority of the consumers are priority consumers, and there is a very small set of regular customers with very low waiting costs). However, if θ'' is very high, and hence there is very small set of priority customers, then priority service reduces the consumer surplus. ■

10.1.2 High-type exclusion.

In this part of the Appendix we derive the results shown in Section 5.2. We start with analyzing the case where only the regular service is offered. Then we proceed to analyze the case where both the regular and the priority service are offered.

Regular service only

Without priority service, given a price p, there is a cutoff type θ^* such that types $\theta \geq \theta^*$ acquire the regular service and types $\theta < \theta^*$ remain outside of the market.

$$\underset{0}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{regular service}}\underset{\theta^*}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{no service}}\underset{\bar{\theta}}{|}$$

The utility of type θ from joining is

$$v(\theta) - c(\theta) \frac{F(\theta^*)}{2} - p.$$

For type θ^* to be indifferent between joining the service and staying out of the market, it is the case that[36]

$$v(\theta^*) - c(\theta^*) \frac{F(\theta^*)}{2} = p.$$

The provider's profits are

$$pF(\theta^*) = \left(v(\theta^*) - c(\theta^*) \frac{F(\theta^*)}{2} \right) F(\theta^*).$$

Hence, the optimal cutoff θ^* satisfies

$$\left(v'(\theta^*) - c'(\theta^*) \frac{F(\theta^*)}{2} - c(\theta^*) \frac{f(\theta^*)}{2} \right) F(\theta^*) + \left(v(\theta^*) - c(\theta^*) \frac{F(\theta^*)}{2} \right) f(\theta^*) \geq 0,$$

where the inequality is strict if $\theta^* = \bar{\theta}$. We can rewrite the last inequality as

$$\left(v'(\theta^*) - c'(\theta^*) \frac{F(\theta^*)}{2} \right) F(\theta^*) + \left(v(\theta^*) - c(\theta^*) F(\theta^*) \right) f(\theta^*) \geq 0.$$

The consumer surplus is given by

$$\begin{aligned} & \int_0^{\theta^*} \left(v(\theta) - c(\theta) \frac{F(\theta^*)}{2} - p \right) f(\theta)\, d\theta \\ = & \int_0^{\theta^*} \left(v(\theta) - c(\theta) \frac{F(\theta^*)}{2} - v(\theta^*) + c(\theta^*) \frac{F(\theta^*)}{2} \right) f(\theta)\, d\theta. \end{aligned}$$

Priority and regular services

If the service provider offers both the priority and the regular service, then the monopolist sets two prices $\Pi > \pi$ and the customers are divided into three groups (with $\bar{\theta} \geq \theta^{\diamondsuit} \geq \theta'' \geq 0$). Types $\theta \geq \theta^{\diamondsuit}$ remain out of the market, types $\theta \in [\theta'', \theta^{\diamondsuit})$ acquire priority service at price Π. Types $\theta < \theta''$ acquire regular service at price π.

$$\underset{0}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{regular service}}\underset{\theta''}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-\,-}_{\text{priority service}}\underset{\theta^{\diamondsuit}}{|}\underbrace{-\,-\,-\,-\,-\,-\,-\,-}_{\text{no service}}\underset{\bar{\theta}}{|}$$

The utility of type θ from getting the priority service is

$$v(\theta) - c(\theta) \frac{F\left(\theta^{\diamondsuit}\right) - F\left(\theta''\right)}{2} - \Pi,$$

while the utility from the regular service is

$$v(\theta) - c(\theta) \left[F\left(\theta^{\diamondsuit}\right) - F\left(\theta''\right) + \frac{F\left(\theta''\right)}{2} \right] - \pi.$$

[36] In case of multiple solutions to this equation we adopt a standard assumption from mechanism design that the provider's optimal (highest) cutoff is selected.

Hence, type $\theta^{\Diamond}$ satisfy

$$v\left(\theta^{\Diamond}\right)-c\left(\theta^{\Diamond}\right)\frac{F\left(\theta^{\Diamond}\right)-F\left(\theta''\right)}{2}=\Pi,$$

and type θ' satisfy

$$\begin{aligned}
v\left(\theta''\right)-c\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)-F\left(\theta''\right)}{2}-\Pi &= v\left(\theta''\right)-c\left(\theta''\right)\left[F\left(\theta^{\Diamond}\right)-\frac{F\left(\theta''\right)}{2}\right]-\pi \Leftrightarrow \\
c\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)}{2} &= \Pi-\pi.
\end{aligned}$$

The monopolist's profits in this case are

$$\begin{aligned}
\Pi\left[F\left(\theta^{\Diamond}\right)-F\left(\theta''\right)\right]+\pi F\left(\theta''\right) &= \Pi\left[F\left(\theta^{\Diamond}\right)-F\left(\theta''\right)\right]+\left[\Pi-c\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)}{2}\right]F\left(\theta''\right) \\
&= \left[v\left(\theta^{\Diamond}\right)-c\left(\theta^{\Diamond}\right)\frac{F\left(\theta^{\Diamond}\right)-F\left(\theta''\right)}{2}\right]F\left(\theta^{\Diamond}\right)-c\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)}{2}F\left(\theta''\right).
\end{aligned}$$

Lemma 7 *The designer always prefers to offer priority service.*

Proof. Assume that θ^{*} is chosen optimally by the service provider if she is allowed to offer regular service only. If the provider is allowed to offer both the priority and the regular service, then setting $\theta''=0$ and $\theta^{\Diamond}=\theta^{*}$ replicates the same revenue as in the case where only the regular service is offered. Observe that the derivative of the provider's profits with respect to θ'' equals

$$\frac{f\left(\theta''\right)}{2}c\left(\theta^{\Diamond}\right)F\left(\theta^{\Diamond}\right)-c'\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)}{2}F\left(\theta''\right)-c\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)}{2}f\left(\theta''\right).$$

This derivative at $\theta''=0$ equals to $\frac{f(0)}{2}c\left(\theta^{\Diamond}\right)F\left(\theta^{\Diamond}\right)-c\left(0\right)\frac{F\left(\theta^{\Diamond}\right)}{2}f\left(0\right)>0$ where the inequality follows from the monotonicity of c. Hence introducing second type of the service increases the provider's profits. ■

The optimal cutoffs chosen by the provider satisfy the following conditions:

(1) θ' must satisfy

$$\begin{aligned}
c\left(\theta^{\Diamond}\right)\frac{f\left(\theta''\right)}{2}F\left(\theta^{\Diamond}\right)-c'\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)}{2}F\left(\theta''\right)-c\left(\theta''\right)\frac{F\left(\theta^{\Diamond}\right)}{2}f\left(\theta''\right) &= 0\Leftrightarrow \\
c\left(\theta^{\Diamond}\right)-c\left(\theta''\right) &= c'\left(\theta''\right)\frac{F\left(\theta''\right)}{f\left(\theta''\right)}.
\end{aligned}$$

(2) θ'' must satisfy

$$\begin{aligned}
&\left[v'\left(\theta^{\Diamond}\right)-c'\left(\theta^{\Diamond}\right)\frac{F\left(\theta^{\Diamond}\right)-F\left(\theta''\right)}{2}-c\left(\theta^{\Diamond}\right)\frac{f\left(\theta\Diamond\right)}{2}\right]F\left(\theta^{\Diamond}\right) \\
&+f\left(\theta^{\Diamond}\right)\left[v\left(\theta^{\Diamond}\right)-c\left(\theta^{\Diamond}\right)\frac{F\left(\theta^{\Diamond}\right)-F\left(\theta''\right)}{2}\right]-c\left(\theta''\right)\frac{f\left(\theta^{\Diamond}\right)}{2}F\left(\theta''\right)\geq 0,
\end{aligned}$$

where the strict inequality holds if $\theta^{\diamondsuit} = \bar{\theta}$. We can rewrite it as

$$\left[v'\left(\theta^{\diamondsuit}\right) - c'\left(\theta^{\diamondsuit}\right)\frac{F\left(\theta^{\diamondsuit}\right) - F\left(\theta''\right)}{2}\right] F\left(\theta^{\diamondsuit}\right)$$
$$+f\left(\theta^{\diamondsuit}\right)\left[v\left(\theta^{\diamondsuit}\right) - c\left(\theta^{\diamondsuit}\right)F\left(\theta^{\diamondsuit}\right) + \left(c\left(\theta^{\diamondsuit}\right) - c\left(\theta''\right)\right)\frac{F\left(\theta''\right)}{2}\right] \geq 0.$$

Lemma 8 *Assume that $R^{RS}\left(\theta^*\right)$ is concave. Then in the optimal mechanism it holds that $\theta^{\diamondsuit} \geq \theta^*$.*

Observe that for any $\theta > \theta''$ we have

$$\begin{aligned} & \left(v'\left(\theta\right) - c'\left(\theta\right)\frac{F\left(\theta\right) - F\left(\theta''\right)}{2}\right) F\left(\theta\right) + f\left(\theta\right)\left(v\left(\theta\right) - c\left(\theta\right)F\left(\theta\right) + \left(c\left(\theta\right) - c\left(\theta''\right)\right)\frac{F\left(\theta''\right)}{2}\right) \\ \geq & \left(v'\left(\theta\right) - c'\left(\theta\right)\frac{F\left(\theta\right)}{2}\right) F(\theta) + \left(v\left(\theta\right) - c\left(\theta\right)F\left(\theta\right)\right) f(\theta). \end{aligned}$$

The concavity of $R^{RS}\left(\theta^*\right)$ implies that $\theta^{\diamondsuit} \geq \theta^*$.

The consumer surplus equals

$$\int_0^{\theta''}\left(v\left(\theta\right) - c\left(\theta\right)\left[F\left(\theta^{\diamondsuit}\right) - \frac{F\left(\theta''\right)}{2}\right] - \pi\right) f\left(\theta\right)d\theta + \int_{\theta''}^{\theta^{\diamondsuit}}\left(v\left(\theta\right) - c\left(\theta\right)\frac{F\left(\theta^{\diamondsuit}\right) - F\left(\theta''\right)}{2} - \Pi\right) f\left(\theta\right)d\theta.$$

Proposition 16 *Assume that $\left(v'\left(\bar{\theta}\right) - \frac{c'\left(\bar{\theta}\right)}{2}\right) + \left(v\left(\bar{\theta}\right) - c\left(\bar{\theta}\right)\right)f(\bar{\theta}) \geq 0$ and $R^{RS}\left(\theta^*\right)$ is concave. Then introducing priority service is consumer surplus-improving if and only if*

$$-\left(c\left(\bar{\theta}\right) - c\left(\theta''\right)\right)F\left(\theta''\right) - \left(1 - F\left(\theta''\right)\right)\int_0^{\theta''} c\left(\theta\right)f\left(\theta\right)d\theta + F\left(\theta''\right)\int_{\theta''}^{\bar{\theta}} c\left(\theta\right)f\left(\theta\right)d\theta \geq 0,$$

where θ'' solves

$$c\left(\bar{\theta}\right) - c\left(\theta''\right) = c'\left(\theta''\right)\frac{F\left(\theta''\right)}{f\left(\theta''\right)}.$$

Proof. Observe that if $\left(v'\left(\bar{\theta}\right) - \frac{c'\left(\bar{\theta}\right)}{2}\right) + \left(v\left(\bar{\theta}\right) - c\left(\bar{\theta}\right)\right)f(\bar{\theta}) \geq 0$ and $R^{RS}\left(\theta^*\right)$ is concave, then in the problem where only regular service is offered the monopolist sets $\theta^* = \bar{\theta}$. Moreover, Lemma 8 implies that $\theta^{\diamondsuit} = \bar{\theta}$. If only the regular service is offered, then the consumer surplus is given by

$$\begin{aligned} & \int_0^{\bar{\theta}}\left(v\left(\theta\right) - \frac{c\left(\theta\right)}{2} - v\left(\bar{\theta}\right) + \frac{c\left(\bar{\theta}\right)}{2}\right) f\left(\theta\right)d\theta \\ = \; & E\left(v\left(\theta\right) - \frac{c\left(\theta\right)}{2}\right) - v\left(\bar{\theta}\right) + \frac{c\left(\bar{\theta}\right)}{2}. \end{aligned}$$

However, if the seller offers both the regular and priority service, then the consumer surplus equals

$$\begin{aligned}
& \int_0^{\theta''} \left(v(\theta) - c(\theta) \left[1 - \frac{F(\theta'')}{2} \right] - \pi \right) f(\theta)\, d\theta + \int_{\theta''}^{\bar{\theta}} \left(v(\theta) - c(\theta) \frac{1 - F(\theta'')}{2} - \Pi \right) f(\theta)\, d\theta \\
= & \int_0^{\bar{\theta}} \left(v(\theta) - \frac{c(\theta)}{2} \right) f(\theta)\, d\theta - \int_0^{\theta''} c(\theta) \frac{1 - F(\theta'')}{2} f(\theta)\, d\theta + \int_{\theta''}^{\bar{\theta}} c(\theta) \frac{F(\theta'')}{2} f(\theta)\, d\theta \\
& - \left(\Pi - \frac{c(\theta'')}{2} \right) F(\theta'') - \Pi \left(1 - F(\theta'')\right) \\
= & \; E\left(v(\theta) - \frac{c(\theta)}{2} \right) - \frac{1 - F(\theta'')}{2} \int_0^{\theta''} c(\theta) f(\theta)\, d\theta + \frac{F(\theta'')}{2} \int_{\theta''}^{\bar{\theta}} c(\theta) f(\theta)\, d\theta - \Pi + \frac{c(\theta'')}{2} F(\theta'') \\
= & \; E\left(v(\theta) - \frac{c(\theta)}{2} \right) - v(\bar{\theta}) + \frac{c(\bar{\theta})}{2} \\
& - c(\bar{\theta}) \frac{F(\theta'')}{2} - \frac{1 - F(\theta'')}{2} \int_0^{\theta''} c(\theta) f(\theta)\, d\theta + \frac{F(\theta'')}{2} \int_{\theta''}^{\bar{\theta}} c(\theta) f(\theta)\, d\theta + \frac{c(\theta'')}{2} F(\theta'').
\end{aligned}$$

Hence, priority service is consumer welfare-improving if and only if

$$- \left(c(\bar{\theta}) - c(\theta'') \right) F(\theta'') - \left(1 - F(\theta'')\right) \int_0^{\theta''} c(\theta) f(\theta)\, d\theta + F(\theta'') \int_{\theta''}^{\bar{\theta}} c(\theta) f(\theta)\, d\theta \geq 0,$$

where θ'' solves

$$c(\bar{\theta}) - c(\theta'') = c'(\theta'') \frac{F(\theta'')}{f(\theta'')}.$$

■

Corollary 3 *Assume that* $\left(v'(\bar{\theta}) - \frac{c'(\bar{\theta})}{2} \right) + \left(v(\bar{\theta}) - c(\bar{\theta}) \right) f(\bar{\theta}) \geq 0$ *and* $R^{RS}(\theta^*)$ *is concave. If* $c'' \geq 0$ *and* F *satisfies DRFR, then introducing priority service decreases consumer surplus.*

Proof of Corollary. Consider expression

$$- \left(c(\bar{\theta}) - c(\theta'') \right) F(\theta'') - \left(1 - F(\theta'')\right) \int_0^{\theta''} c(\theta) f(\theta)\, d\theta + F(\theta'') \int_{\theta''}^{\bar{\theta}} c(\theta) f(\theta)\, d\theta.$$

Observe that if either $\theta'' = 0$ or $\theta'' = \bar{\theta}$ then the expression equals to zero. The derivative of the left-hand side of the equality with respect to θ'' is

$$\begin{aligned}
& c'(\theta'') F(\theta'') - \left(c(\bar{\theta}) - c(\theta'') \right) f(\theta'') + f(\theta'') \mathbb{E}[c(\theta)] - c(\theta'') f(\theta'') \\
= & \; c'(\theta'') F(\theta'') - c(\bar{\theta}) f(\theta'') + f(\theta'') \mathbb{E}[c(\theta)] \\
= & \; f(\theta'') \left[c'(\theta'') \frac{F(\theta'')}{f(\theta'')} - c(\bar{\theta}) + \mathbb{E}[c(\theta)] \right].
\end{aligned}$$

Observe that $c'' \geq 0$ and DRFR implies that the expression in the parenthesis is increasing and changes sign once, from negative to positive. Hence, for $\theta'' \in \left(0, \bar{\theta}\right)$, we have that

$$-\left(c\left(\bar{\theta}\right)-c\left(\theta''\right)\right)F\left(\theta''\right)-\left(1-F\left(\theta''\right)\right)\int_{0}^{\theta''}c\left(\theta\right)f\left(\theta\right)d\theta+F\left(\theta''\right)\int_{\theta''}^{\bar{\theta}}c\left(\theta\right)f\left(\theta\right)d\theta<0.$$

■

Proof of Proposition 7. We show that under the conditions of the proposition we have that $\Pi > \pi \geq p$. To show that $\pi \geq p$ observe that under the conditions of the proposition we have $\theta^{\Diamond} = \theta^{*} = \bar{\theta}$ and hence

$$\begin{aligned}\pi &\geq p \Longleftrightarrow v\left(\bar{\theta}\right)-c\left(\bar{\theta}\right)\frac{F\left(\bar{\theta}\right)-F\left(\theta''\right)}{2}-\frac{c\left(\theta''\right)}{2}>v\left(\bar{\theta}\right)-c\left(\bar{\theta}\right)\frac{F\left(\bar{\theta}\right)}{2}\\ &\Leftrightarrow c\left(\bar{\theta}\right)>\frac{c\left(\theta''\right)}{F\left(\theta''\right)}.\end{aligned}$$

The monotonicity of $c\left(\theta\right)/F\left(\theta\right)$ completes the proof. ■

10.2 Appendix D. Proofs for Section 7

Duopoly with heterogeneous customers.

Proof of Lemma 3. Assume, by way of contradiction, that there are two different pairs (c_1^*, c_2^*) and $(c_1^{*\prime}, c_2^{*\prime})$ that both satisfy the indifference conditions for the same pair of prices p_1 and p_2. If $c_1^* = c_1^{*\prime}$ (that is, if in both equilibria the same type is indifferent between the priority services of both providers), then $c_2^* = c_2^{*\prime}$, as otherwise the utility of the cutoff type c_1^* from choosing the priority service of provider 2 will not be the same in the two equilibria, and since the utility from choosing the priority service of provider 1 is the same, this type will not be indifferent between the two priority services of the two providers. This contradicts our assumption that $c_1^* = c_1^{*\prime}$.

Assume now that $c_1^* < c_1^{*\prime}$ (the case of $c_1^* > c_1^{*\prime}$ is similar). The last inequality implies that the utility from joining the priority service of provider 1 is higher (for all types) in equilibrium $(c_1^{*\prime}, c_2^{*\prime})$ than in equilibrium (c_1^*, c_2^*) since the prices are the same, but in equilibrium $(c_1^{*\prime}, c_2^{*\prime})$ fewer customers join the priority service of provider 1 than in the equilibrium given by (c_1^*, c_2^*) (as $F(c_1^*) < F(c_1^{*\prime})$). As $c_1^* < c_1^{*\prime}$, it follows that, in equilibrium (c_1^*, c_2^*), type $c_1^{*\prime}$ prefers the priority service of provider 1 to the priority service of provider 2. Since in equilibrium $(c_1^{*\prime}, c_2^{*\prime})$ type $c_1^{*\prime}$ is indifferent between the two priority services of the two providers, it must be the case that the utility of joining the priority service of provider 2 is higher in equilibrium $(c_1^{*\prime}, c_2^{*\prime})$ than in the equilibrium given by (c_1^*, c_2^*). Therefore,

$$F\left(c_1^{*\prime}\right)-F\left(c_2^{*\prime}\right)<F\left(c_1^*\right)-F\left(c_2^*\right).$$

This inequality, together with $F(c_1^*) < F(c_1^{*\prime})$, implies that $F(c_2^{*\prime}) > F(c_2^*)$. It further implies that $c_2^{*\prime} > c_2^*$. However, it also implies that the utility from non-priority services is lower in equilibrium $(c_1^{*\prime}, c_2^{*\prime})$ than in equilibrium (c_1^*, c_2^*), since in $(c_1^{*\prime}, c_2^{*\prime})$ fewer agents join the priority services of both providers. This in turn implies that the type that is indifferent between the priority service of provider 2 and the regular service of any provider must be lower in equilibrium (c_1^*, c_2^*) than in equilibrium $(c_1^{*\prime}, c_2^{*\prime})$. This contradicts our assumption that $c_2^{*\prime} > c_2^*$. ■

Proof of Proposition 11. Condition (3) implies that

$$n_1^{np} = 2F\left(c_1^*\right) - 1 - \frac{F\left(c_2^*\right)}{2}.$$

Plugging it into condition (2) gives us

$$p_2 = c_2^* \left[\frac{1 - F\left(c_1^*\right)}{2} + \frac{F\left(c_2^*\right)}{4}\right].$$

By condition (1) we have

$$\begin{aligned} p_1 &= p_2 + c_1^* \left[\frac{F\left(c_1^*\right) - F\left(c_2^*\right)}{2} - \frac{1 - F\left(c_1^*\right)}{2}\right] \\ &= \frac{c_2^* - c_1^*}{2} + \left(c_1^* - \frac{c_2^*}{2}\right)\left[F\left(c_1^*\right) - \frac{F\left(c_2^*\right)}{2}\right]. \end{aligned}$$

Hence, for any $p_1 \geq p_2$ the market for priority services will be divided as follows: customers with waiting costs above c_1^* join priority service of provider 1, customers with waiting costs in the interval $[c_2^*, c_1^*]$ join the priority service of provider 2, where c_1^* and c_2^* solve

$$p_1 - p_2 = c_1^* \left[\frac{F\left(c_1^*\right) - F\left(c_2^*\right)}{2} - \frac{1 - F\left(c_1^*\right)}{2}\right] \tag{11}$$

and

$$p_2 = c_2^* \left[\frac{1 - F\left(c_1^*\right)}{2} + \frac{F\left(c_2^*\right)}{4}\right]. \tag{12}$$

Assume that one of the providers, say provider 2, sets $p_2 = 0$. We will show that provider 1 should not set its price for priority service to 0. More precisely, provider 1 can increase its profits by setting $p_1 > 0$. First observe that from (12) it follows that $c_2^* = 0$. Plugging it into (11) gives us

$$p_1 = c_1^* \left[F\left(c_1^*\right) - \frac{1}{2}\right].$$

Hence, setting $p_1 = 0$ implies that $c_1^* = F^{-1}\left(\frac{1}{2}\right)$, that is, provider 1 serves half of the market. While setting a price $p_1 \in \left(0, \frac{\bar{c}}{2}\right)$ the provider will serve a positive measure of the market at a positive price, and hence gets a positive profit.

Now we show that $p_2 = 0$ and $p_1 > 0$ is not an equilibrium. We show that provider 2 can increase its profit by setting a positive price. If $p_1 < \frac{\bar{c}}{4}$, then setting $p_2 = p_1$ implies that all customers with waiting costs above $4p_1$ will acquire priority service from one of the providers and provider 2 will get a profit of $p_1 \frac{1-F(4p_1)}{2} > 0$. If $p_1 \geq \frac{\bar{c}}{4}$, then, since $0 \leq \frac{1-F\left(c_1^*\right)}{2} \leq \frac{1}{2}$ and both c_1^* and c_2^* are continuous in p_2 and p_1, there exists $p_2 > 0$ close enough to zero (and so $p_2 < p_1$) such that the resulting cutoff type c_2^* satisfies $0 < c_2^* < c_1^*$, and hence provider 2 gets a positive profit. ■

References

[1] Acemoglu, D., and A. Ozdaglar (2007), "Competition and Efficiency in Congested Markets," *Mathematics of Operations Research,* **32**, 1-31.

[2] Aghion, P., Dewatripont, M., and P. Rey (1999), "Competition, Financial Discipline, and Growth," *Review of Economic Studies*, **66**, 825-852.

[3] Aghion, P., Harris, C., Howitt, P.,and J. Vickers (2001), "Competition, Imitation and Growth with Step-by-Step Innovation," *Review of Economic Studies*, **68**, 467-492.

[4] Aghion, P., Harris, C., and J. Vickers (1997), "Competition and Growth with Step-by-Step Innovation: An Example," *European Economic Review, Paper and Proceedings,* **41**, 771–782.

[5] Armstrong, M. (2008), "Price Discrimination," in *Handbook of Antitrust Economics*, Ed P. Buccirossi, MIT Press.

[6] Armstrong, M., and J. Vickers (1991), "Welfare Effects of Price Discrimination by a Regulated Monopolist," *Rand Journal of Economics*, **22**, 571-580.

[7] Ashenfelter, O and D. Genesove (1992) "Testing Price Anomalies in Real-Estate Auctions" *The American Economic Review Papers and Proceedings* **82**, 501-505.

[8] Bagnoli, M., and T. Bergstrom (2005), "Log-concave Probability and Its Applications," *Economic Theory* **26**, 445-469.

[9] Bernheim, D., and M. Whinston (1998), "Exclusive Dealing," *Journal of Political Economy*, **106**, 64-103.

[10] Budish, E., Cramton, P., and J. Shim (2015), "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *Quarterly Journal of Economics,* **130**, 1547–1621.

[11] Bulow, J., and P. Klemperer (2012), "Regulated Prices, Rent Seeking, and Consumer Surplus," *Journal of Political Economy* **120**, 160-186.

[12] Bulow, J., and J. Roberts (1989), "The Simple Economics of Optimal Auctions," *Journal of Political Economy,* **97**, 1060-1090.

[13] Chao H., and Wilson R. (1987), "Priority Service: Pricing, Investment and Market Organization," *American Economic Review,* **77**, 899-916.

[14] Chun, Y., Mitra, M., and S. Mutuswami (2019), "Recent Developments in the Queueing Problem," *TOP* (Spanish Society of Statistics and Operations Research), 27, 1-23.

[15] Compte, O., Jenny, F., and P. Rey (2002), "Capacity Constraints, Mergers and Collusion," *European Economic Review*, **46**, 1-29.

[16] Cramton, P., Gibbons, R., and P. Klemperer (1987), "Dissolving a Partnership Efficiently," *Econometrica*, **55**, 615-632.

[17] De Borger, B., and K. Van Dender (2006), "Prices, Capacities and Service Levels in a Congestable Bertrand Duopoly," *Journal of Urban Economics*, **60**, 264-283.

[18] Deneckere, R., and P. McAfee (1996), "Damaged Goods," *Journal of Economics and Management Strategy,* **5**, 149-174.

[19] Dolan, R. (1978), "Incentives Mechanisms for Priority Queuing Problems," *The Bell Journal of Economics,* **9**, 421-436.

[20] Dworczak, P. Kominers, S. D., and M. Akbarpour (2021), "Redistribution Through Markets," *Econometrica*, **89**, 1665-1698.

[21] Edelson, N (1971), "Congestion Tolls under Monopoly," *American Economic Review,* **61**, 873-882.

[22] Einav, L., Kuchler, T., Levin, J., and N. Sundaresan (2015) "Assessing Sale Strategies in Online Markets Using Matched Listings," *American Economic Journal: Microeconomics* **7,** 215–247.

[23] Gershkov, A. and P. Schweinzer (2010), "When Queueing is Better than Push and Shove," *International Journal of Game Theory*, **39**, 409-430.

[24] Glazer, A. and R. Hassin (1986), "Stable Priority Purchasing in Queues," *Operations Research Letters* **4**, 285-288.

[25] Gomes, R., and J. Tirole (2018), "Missed Sales and the Pricing of Ancillary Goods," *The Quarterly Journal of Economics*, **133**, 2097-2169.

[26] Hall, J (2018) "Pareto Improvements from Lexus Lanes: The Effects of Pricing a Portion of the Lanes on Congested Highways," *Journal of Public Economics*, **158**, 113-125.

[27] Harrington, J. (2017), *The Theory of Collusion and Competition Policy*, The MIT Press.

[28] Hassin R., and Haviv, M (2003), *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, Springer.

[29] Haviv, M. and E. Winter (2020), "An Optimal Mechanism Charging for Priority in a Queue," *working paper.*

[30] Hoppe, H., Moldovanu, B., and E. Ozdenoren (2011), "Coarse Matching with Incomplete Information," *Economic Theory*, 47, 73-104.

[31] Kittsteiner, T., and B. Moldovanu (2005), "Priority Auctions and Queue Disciplines that Depend on Processing Time," *Management Science*, **51**, 236-248.

[32] Levhari, D. and I. Luski (1978), "Duopoly Pricing and Waiting Lines," *European Economic Review,* **11**, 17-35.

[33] Luski, I (1976), "On Partial Equilibrium in a Queuing System with Two Servers," *Review of Economic Studies*, **43**, 519-525.

[34] McAfee, P. (2002), "Coarse Matching," *Econometrica,* **70**, 2025-2034.

[35] Moldovanu, B., Sela, A., and X.Shi (2007), "Contest for Status," *Journal of Political Economy*, **115**, 338-363.

[36] Mussa, M. and S. Rosen (1978), "Monopoly and Product Quality," *Journal of Economic Theory*, **18**, 301-317.

[37] Reitman, D (1991), "Endogenous Quality Differentiation in Congested Markets," *Journal of Industrial Economics*, **39**, 621-647.

[38] Segal, I. (2003), "Coordination and Discrimination in Contracting with Externalities: Divide and Conquer?," *Journal of Economic Theory*, **113**, 147-181.

[39] Weitzman, M. (1977), "Is the Price System or Rationing More Effective in Getting a Commodity to Those Who Need it Most?" *Bell Journal of Economics*, **8**, 517-524.

[40] Wilson, R. (1989), "Efficient and Competitive Rationing," *Econometrica,* **57**, 1-40.

[41] Winter, E. (2004), "Incentives and Discrimination," *American Economic Review*, **94**, 764-773.