Chapter 2
Auctions for Trading Queueing Positions

1 Introduction

A queue constitutes a "miniature social system" in which the underlying fabric that ties individuals to society also guides the relationships between those in a queue (Mann 1969). In particular, the behavioral protocol of queueing collectively endorses the notion of "property rights" (Gray 2009), a fundamental part of the social fabric. An individual's *position* in a queue is considered by its occupant as her *property* that she temporarily *owns*. Tampering with one's position in a queue amounts to taking away someone's property and may be met with strong objection: Any attempt to cut in line may be disapproved since this infringes on the "bumped" customers' perceived property rights over their waiting positions. This is one of the reasons why the first-in, first-out (FIFO) queue discipline is predominant in many services systems. The FIFO rule ensures "a direct correspondence between inputs (time spent waiting) and outcomes (preferential service)" and thus manifests a basic principle of distributive justice (Mann 1969). However, the FIFO rule disregards queue occupants' heterogeneous time-sensitivities. The system would be more efficient if more time-sensitive customers jump ahead and get served faster. To that end, service providers often sell priorities to customers. For instance, EE, one of the largest telecommunications companies in the UK, once launched a new service feature called "Priority Answer" that allowed customers to pay £0.50 to jump the queue for a service call. This new feature soon created a huge uproar and irked many customers who complained they were not being treated fairly.

What goes awry with Priority Answer is that the proceeds go to the service provider, yet a longer wait is inflicted on the non-paying customers. The misalignment would be resolved if the monetary transfer were among customers themselves: impatient customers may be willing to pay to acquire the position of less impatient customers who are potentially willing to give away their spots for monetary gains. This calls for a two-sided marketplace where customers consensually *trade* their waiting spots. Such a marketplace enables waiting customers to voluntarily swap

S. Cui et al., *Innovative Priority Mechanisms in Service Operations*, SpringerBriefs in Service Science, https://doi.org/10.1007/978-3-031-30841-3_2

positions at mutually agreed prices. Since such swaps do not influence the positions of any other customers on the wait list, no customers are forcibly pushed back without being compensated. Thus, customers can have the best of both worlds: their proprietary entitlements to waiting positions are preserved as in the FIFO system while their diverging priority preferences are accommodated, improving system efficiency.

We study in Yang et al. (2017) how trading in a queue can be organized by simple auctions in an environment where customers are privately informed about their waiting costs. We design the optimal mechanisms from three different perspectives: social welfare, the service provider's revenue, and the revenue of the trading platform (which we refer to as the intermediary) that mediates trading. While the first two perspectives are common in the queueing literature, they implicitly rely on the assumption that the trading platform is and can be managed by the service provider, which may not necessarily be true in practice. Instead, the service provider may be inclined to delegate the trading platform to an intermediary for technological reasons and reputational concerns. First and foremost, the infrastructure that facilitates trade hinges on technology (e.g., mobile apps) that typically falls beyond the expertise of the service provider. Therefore, if a specialized intermediary is responsible for developing, deploying and maintaining the platform on behalf of the service provider, the service provider will not be distracted from its core competencies. In the restaurant industry, for example, dining reservation platforms (intermediaries) are typically not fully integrated with restaurants (service providers): examples include OpenTable which charges restaurants for each reservation, and a similar dining app, Reserve, which alternatively charges customers for each booking. Second, if the service provider were to operate and conceivably profit from a resale market of waiting positions (either directly by collecting fees for trading or indirectly via surcharges in service fees), there might be a backlash from customers given the sensitive nature of queue-jumping (as in the case of Priority Answer). To the extent that this results in a loss of goodwill, the service provider would rather be detached from the trading platform and leave it to a third-party intermediary to arbitrate swaps of waiting positions. This begs the question of what is the optimal mechanism to collect fees from trading customers for an intermediary, who has the potential of raising sizable revenues once the technology is scaled up.

The problem of trading waiting positions in a queue has been studied by several papers in the extant literature. Rosenblum (1992) assumes that customers' waiting costs are public information in their trading model and that future values of transactions are ignored. Our model relaxes these two strong assumptions: customers are privately informed of their own waiting costs and take into account the expected values of future transactions when they trade. Gershkov and Schweinzer (2010) formulate a mechanism design problem of rescheduling a fixed number of players in a clearing system where there is no arrival process and trading is completed before service starts. Since all customers are present at time zero, it is not clear how the initial property rights are formed, so they study different initial allocations and show that an efficient mechanism can be implemented if the initial schedule is random ordering but not if it is deterministic like FIFO. El Haji and

Onderstal (2019) experimentally examine how human subjects trade in a queueing environment similar to Gershkov and Schweinzer (2010). They provide evidence that organizing such a time-trading market can achieve a nontrivial amount of efficiency gains. Our model incorporates the operational dynamics of a queueing system where the order of arrivals naturally gives rise to the initial allocation. It allows us to study how trading impacts customers' endogenous queue-joining behaviors. While Gershkov and Schweinzer (2010) and El Haji and Onderstal (2019) mostly focus on the efficiency of the time-trading mechanisms, our work also incorporates the perspectives of the revenue maximizing service provider and intermediary.

2 Model Setup

Consider a congested service facility, modeled as an $M/M/1$ queueing system, that faces a population of delay-sensitive customers. Customers arrive at the system according to an exogenous Poisson process with rate Λ (market size). Each customer requests one unit of service. The service times are i.i.d. samples from an exponential distribution with mean $1/\mu$. Let $\rho \doteq \Lambda/\mu$. Customers have a common valuation V for service. For a customer with delay cost rate c, who experiences waiting time w, defined as the entire duration in the system, and money m after receipt of service, her utility is $V - c \cdot w + m$. For simplicity, we normalize initial money wealth for all customers at zero, but assume that they are not budget-constrained, so $m > 0$ means a customer is a net receiver; $m < 0$, a net payer. Customers differ in their delay cost rate c. Each customer's delay cost per unit time is an i.i.d. draw from a continuous distribution with a strictly increasing cumulative distribution function F and a finite, strictly positive and continuously differentiable probability density function f over the support $c \in \Xi \triangleq [\underline{c}, \overline{c}]$ and $0 \leq \underline{c} < \overline{c} < \infty$. Customers are risk-neutral and expected utility maximizers. To exclude the case where no customers have a positive net value even if served immediately, we assume $V > \underline{c}/\mu$.

Upon arrival, customers decide whether to join the service facility to obtain service, or balk. In case they do not join, they obtain the reservation utility, which we normalize to zero. The inter-arrival time distribution, the service time distribution, the delay cost distribution f and the service value V are common knowledge. The type of each individual customer (delay cost rates c) is her private information. Customers do not observe the system state upon arrival but can estimate the expected waiting time and the expected monetary transfer.

3 Baseline Auction

In this section, we study an auction-based trading mechanism that is budget-balanced among customers: all monetary transfers are internal within customers. This auction is the building block for subsequent results about the social planner, service provider and intermediary in Sects. 4 and 5.

3.1 Trading Rules

Auction Format In the baseline auction, upon arrival, a customer decides whether to join the queue or not. If the customer does not join, she earns a reservation utility of 0. If the customer joins, she submits a sealed bid b that can either be "No," or a price for one unit of time. We allow customers to bid "No" to reflect that trading is voluntary and that customers can always preserve their FIFO property right. The bid b represents the least amount she wants to receive for expecting to wait one additional unit of time and also the greatest amount she is willing to pay for one unit of the expected waiting time reduced. The queue is reorganized in such a way that the arriving customer swaps positions consecutively with those who place bids strictly lower than hers. In each transaction, the customer who jumps ahead (the buyer) compensates the one who moves back (the seller) by the seller's bid price times the expected waiting time exchanged. The existing customers who submitted bids strictly higher or those who submitted "No" are not affected in their waiting position. Nor are customers who bid the same amount as the buyer. Any customers with equal bids are served FIFO amongst themselves. Note that this auction follows a "pay-as-you-overtake" paradigm, since customers' realized payment as buyers depends on the actual number of customers they overtake. For simplicity, trading is instantaneous (transactions do not take any time) and preemptive-resume (customers at the server can suspend their service and sell their spot; service is resumed when this customer reaches the server again). Customers submit a bid before observing the queue length and commit to the bid throughout their stay in the system.[1]

Illustration 1 Consider an arriving customer who joins and participates in trading by submitting a price b'. Assume that there are four other existing customers in the system. Among them, the first, the second and the fourth customer participate in trading with bids b_1, b_2 and b_4 (with $b_1 \geq b' > b_2 > b_4$), respectively. The third bids "No" and thus does not participate in trading. Thus, before the new arrival, the system can be represented by (b_1, b_2, F, b_4), where F stands for a *F*IFO customer who bids "No". Adding the arriving customer (customer 5) who bids b' to the tail of

[1] After submitting their bid, customers could see the queue length, but this would be technically irrelevant to the bidding game since the trading process goes on autopilot once bids are collected from customers.

the queue, we have (b_1, b_2, F, b_4, b'), which is not (yet) ordered. Then, the auction swaps customer 5 and customer 4, yielding (b_1, b_2, F, b', b_4). Customer 5 makes a payment of b_4/μ to customer 4. Next, the auction swaps customer 5 and customer 2, yielding (b_1, b', F, b_2, b_4). Notice that the expected wait time of the FIFO customer does not change. Customer 5 makes a payment of $2b_2/\mu$ to customer 2 (because the latter moves back by two positions). Customer 5 does not swap positions with customer 1 since customer 1 bids weakly higher and they are served FIFO. Thus, the trading process is completed. The total payment customer 5 makes to the other customers is thus $(b_4 + 2b_2)/\mu$. Similarly, customer 5 expects a compensation of b' per unit of time if she ever moves back and swaps positions with other, later arriving customers who make a higher bid than b'.

3.2 Auction Equilibrium

Strategy We focus on pure strategies specified by two functions; the joining function $J : \Xi \mapsto \{\text{join, balk}\}$ specifies which customer types join or balk, and the bid function $b : \{c|J(c) = \text{join}\} \mapsto \mathbb{R}_+ \cup \{\text{No}\}$ specifies the bid of each customer type (either a price for one unit of time or "No"). Thus the effective arrival rate to the system is $\lambda \triangleq \Lambda \int_{\underline{c}}^{\overline{c}} \mathbf{1}\{J(c) = \text{join}\} dF(c)$, where $\mathbf{1}\{X\}$ is the indicator function of condition X.

Waiting Time and Utility Given the bid function $b(\cdot)$ and the joining function $J(\cdot)$, let $W : \mathbb{R}_+ \mapsto \mathbb{R}_+$ denote the mapping from a customer's bid to her expected waiting time. Since trading does not affect any joining customer who bids "No," it is immediate that these customers' expected waiting time is equal to the mean waiting time of an $M/M/1$ system: $W(\text{No}|b, J) = \frac{1}{\mu-\lambda}$. Note that this waiting time depends on the endogenously determined λ, the aggregate arrival rate of the system, and is not impacted by any individual, infinitesimal customer's action. Since customers submit their bid up front and make a commitment during their wait, they take into account all future transactions in the expected utility (note that this is one of the key distinctions from Rosenblum 1992). We assume that customers do not discount future payments. Let $P_p : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be the function that maps a customer's bid to the total expected amount of money she *p*ays as a buyer upon arrival; and $P_r : \mathbb{R}_+ \mapsto \mathbb{R}_+$ maps a customer's bid to the total expected amount of money she *r*eceives as a seller during her stay in the system.

Thus, given $b(\cdot)$ and $J(\cdot)$, the expected utility of a joining customer of type c who bids β is

$$U(c, \beta|b, J) = \begin{cases} V - cW(\beta|b, J) - P_p(\beta|b, J) + P_r(\beta|b, J), & \beta \in \mathbb{R}_+ \\ V - \frac{c}{\mu-\lambda}, & \beta = \text{No}. \end{cases} \tag{2.1}$$

Customer Equilibrium A symmetric pure-strategy Nash equilibrium is defined by the following conditions:

$$b(c) \in \arg\max_{\beta \in \mathbb{R}_+ \cup \{\text{No}\}} U(c, \beta | b, J), \quad \forall c \in \{c | c \in \Xi,\ J(c) = \text{join}\} \tag{2.2a}$$

$$U(c, b(c) | b, J) \geq 0, \quad \forall c \in \{c | c \in \Xi,\ J(c) = \text{join}\} \tag{2.2b}$$

$$U(c, \beta | b, J) \leq 0, \quad \forall c \in \{c | c \in \Xi,\ J(c) = \text{balk}\},\ \forall \beta \in \mathbb{R}_+ \cup \{\text{No}\}. \tag{2.2c}$$

Condition (2.2a) states that for all the joining customers, the best response of the equilibrium bid function should be itself. Condition (2.2b) ensures that all joining customers get nonnegative expected utility and (2.2c) specifies that the balking customers in equilibrium have no incentive to join the system since their expected utility would not turn positive regardless of what she bids.

An equilibrium is said to achieve *efficiency* or be an *efficient schedule* if $b(c)$ is strictly increasing in c whenever $J(c) =$join. If this holds, customers are effectively prioritized by the $c\mu$ rule.

It is immediate that there is a trivial equilibrium: all joining customers submit "No". Thus, nobody participates in trading and customers are served FIFO. This equilibrium holds in all auction settings in this paper. We analyze other equilibria that realize gains from trade. We indicate the equilibrium in the baseline auction by means of a superscript B.

Theorem 1 (Full Trading, Separating Equilibrium) *Under the baseline auction, there exists an equilibrium in which*

(i) $J^B(c) =$*join for* $c \in [\underline{c}, \tilde{c}]$ *(and balk otherwise) with* $\tilde{c} \leq \overline{c}$*, i.e.,* $\lambda^B = \Lambda F(\tilde{c})$*;*
(ii) *the equilibrium bid function is strictly increasing and given by*

$$b^B(c; \tilde{c}) = c + \frac{\int_c^{\tilde{c}} (F(\tilde{c}) - F(s))^2 W^e(s; \tilde{c}) ds}{(F(\tilde{c}) - F(c))^2 W^e(c; \tilde{c})}, \quad c \in [\underline{c}, \tilde{c}]$$

where $W^e(c; \tilde{c}) = \frac{1}{\mu[1-\rho(F(\tilde{c})-F(c))]^2}$ *is the time customer* c *expects to wait given* $\tilde{c}$*;*
(iii) *the equilibrium expected utility of the joining customers,* $U(c, b^B(c; \tilde{c}))$*, is convex decreasing in* c*. Either* $\tilde{c}$ *uniquely solves* $U(\tilde{c}, b^B(\tilde{c}; \tilde{c})) = 0$ *or* $\tilde{c} = \overline{c}$ *if there is no solution.*

We illustrate Theorem 1 in Fig. 2.1. Unless otherwise stated, we use the parameters in Table 2.1 for numerical illustrations throughout the paper.

Theorem 1 suggests that customers follow a threshold policy in their joining decisions, and they balk if their waiting cost is high, i.e., c is greater than the cutoff value $\tilde{c}$. We henceforth use $\overline{W}(\tilde{c}) \triangleq \frac{1}{\mu - \Lambda F(\tilde{c})}$ to denote the expected FIFO waiting time. In this equilibrium, however, all the joining customers participate in trading.

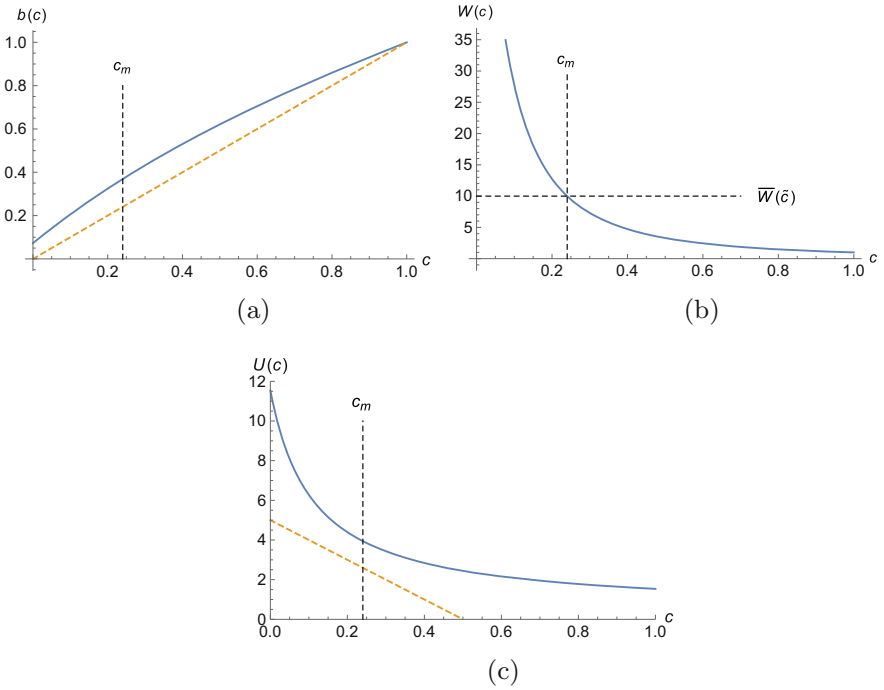

Fig. 2.1 The baseline auction. *Note.* The solid curves are the bidding function in (**a**), the expected waiting time in (**b**), and the expected utility under the auction. The dashed lines are a 45-degree line in (**a**), and the expected utility if customers are served FIFO under arrival rate λ^B in (**c**). In this example, $\tilde{c} = \overline{c} = 1$. (**a**) Biding function. (**b**) Waiting time. (**c**) Utility

Table 2.1 Model primitives for numerical illustrations

V	Λ	μ	F
5	0.9	1	$U[0, 1]$

Most importantly, the equilibrium bid function is strictly increasing and the expected waiting time is strictly decreasing in c, implying that the budget-balanced baseline auction implements an efficient schedule: any two customers with different waiting costs trade waiting spots with one another so that customers are prioritized in decreasing order of their waiting cost. The resulting expected wait time is illustrated in Fig. 2.1b.

Achieving allocative efficiency via a budget-balanced trading mechanism under private information is a nontrivial result under individual rationality in trading (cf. Myerson and Satterthwaite 1983). As illustrated in Fig. 2.1c, under this equilibrium, trading makes all the joining customers better off, i.e., their expected utility exceeds the utility they would get if they bid "No" and wait FIFO. The intuition is the following. Prior to trading, all customers expect FIFO waiting time, so the initial waiting time allocation is the same. This is analogous to having equal shares before partnership dissolution (cf. Cramton et al. 1987). When trading starts, customers

may buy from existing customers, and may also sell to future arriving customers. The countervailing incentives as both buyers and sellers offset each other, making an efficient schedule possible.

We also highlight that one favorable feature of the budget-balanced auction is that it has very simple rules that do not require knowledge of customer valuations like V and F. An efficient schedule is automatically achieved by customers themselves. While the auction, in principle, allows for an arriving customer to overtake another customer who chooses not to participate in trading (without influencing her expected waiting time, e.g., customer 5 overtakes the FIFO customer in Illustration 1), this case never happens in equilibrium as all joining customers trade, i.e., any customer who is overtaken gets compensated at an agreed-upon price in equilibrium.

As illustrated in Fig. 2.1a, all the joining customers overbid, i.e., $b^B(c) > c$, except that customer $\tilde{c}$ bids truthfully, i.e., $\lim_{c\to\tilde{c}} b^B(c) = \tilde{c}$ ($\tilde{c} = \overline{c}$ in this example). This is a consequence of the auction rule that the seller's bid dictates the trading price in each transaction. Thus, customers have an incentive to inflate their bid as a seller to gain more revenue. As illustrated in Fig. 2.1c, customers' expected utility is downward sloping in their waiting costs, implying that the more impatient a customer is, the less she gains from joining the system; whereas the convexity of the curve implies that customers with extreme waiting costs (either very high or very low) have the most gains from trading relative to being served FIFO whereas customers with medium waiting cost reap the least relative gains. Customers with very low waiting cost favor trading since they are most willing to sell spots for money, i.e., $P_p(b^B(\underline{c})) = 0$ and $P_r(b^B(\underline{c})) > 0$; while those with high waiting cost benefit from trading since they are most willing to pay to skip the line, i.e., $P_r(b^B(\tilde{c})) = 0$ and $P_p(b^B(\tilde{c})) > 0$. Both incentives are weak for customers in the middle. In particular, there is one type of customers, c_m, who expects exactly the same waiting time as if she were served FIFO, i.e., $W^e(c_m; \tilde{c}) = \overline{W}(\tilde{c})$. Because of the convexity of the utility curve, c_m is also the type of customers whose gain in trading relative to FIFO is the smallest. Still, she strictly prefers trading to FIFO since $P_r(b^B(c_m)) > P_p(b^B(c_m))$, i.e., the amount she expects to receive exceeds the amount she expects to pay. In a nutshell, two types of customers warrant special attention: the one with the least patience who would be the most sensitive to joining; and the one with moderate patience who would be the most sensitive to trading.

4 Social Welfare and Service Provider's Revenue

In this section, we study how social welfare and the service provider (SP)'s revenue can be maximized using the trading mechanism proposed in Sect. 3.

4.1 Social Optimization

Definition 1 The maximum social welfare SW is determined by:

$$SW = \max_{\tilde{c}\in\Xi:\Lambda F(\tilde{c})<\mu} \Lambda F(\tilde{c})V - \Lambda \int_{\underline{c}}^{\tilde{c}} cW^e(c;\tilde{c})dF(c). \tag{2.3}$$

The socially optimal arrival rate $\lambda^{SW} = \Lambda F(\tilde{c}^{SW})$, where $\tilde{c}^{SW}$ is the maximizer of (2.3).

Definition 1 formalizes the concept of the social optimum in a centralized system where the social planner can dictate the arrival rate and scheduling policy. First, it is socially optimal to serve customers with the smallest waiting costs for any arrival rate and scheduling policy; thus, the social optimum requires a threshold joining policy, which coincides with the equilibrium structure of the baseline auction. Second, for any arrival rate, it is socially optimal to prioritize customers by the $c\mu$ rule, which is achieved by the equilibrium structure of our trading mechanism. It is natural to ask whether the baseline auction as a decentralized mechanism implements the social optimum. Proposition 1 indicates the answer is negative in general.

Proposition 1 *$\lambda^B \geq \lambda^{SW}$ with equality if and only if $\lambda^B = \lambda^{SW} = \Lambda$. The social planner can achieve SW by running the baseline auction with an admission fee $p^{SW} = \int_{\underline{c}}^{\tilde{c}^{SW}} c\left[1 - \frac{F(c)}{F(\tilde{c}^{SW})}\right]\left[-\frac{\partial W^e(c;\tilde{c}^{SW})}{\partial c}\right]dc$.*

Although the baseline auction achieves the "right" service order (efficiency), it does not attain the socially optimal arrival rate in general: in particular, customers with high waiting cost who should otherwise balk in social optimum join the system under the trading mechanism. This runs counter to the well-known result for the typical priority auction as in Kleinrock (1967) which is shown to be self-regulating in both the arrival rate and service order (Hassin 1995). The problem with the trading mechanism is that unlike in the priority auction, customers do not fully internalize the negative externalities inflicted on others. They do pay for the cost imposed on the existing customers if they jump over them; in fact, they are over-penalized in our auction since the trading price overstates the seller's waiting cost. However, they are not held accountable for the cost imposed on future arrivals; worse still, they can even earn rents on their waiting spots for future customers to buy. The inability to achieve the maximum social welfare is similarly found in the bilateral trading model in Myerson and Satterthwaite (1983), but takes a different form. There, maximizing social welfare is synonymous with achieving ex-post efficiency due to a fixed number of traders (one buyer and one seller). Their system is afflicted by the lack of ex-post efficiency, hence a loss in social welfare. Our queueing system attains efficiency for a given arrival rate, but customers' joining decisions are endogenous, precisely because of which, the system suffers from over-joining, again engendering a loss in social welfare. Fundamentally, this loss in social welfare is symptomatic of

Table 2.2 Comparisons of different mechanisms

	Baseline	Socially optimal	SP revenue maximizing	FIFO pricing
Social welfare	2.942	3.106	2.921	2.121
SP's revenue	0	2.121	2.381	2.011
Admission fee	0	2.592	3.439	3.505
Arrival rate	0.9	0.818	0.692	0.574
Percentage loss in SW	5.29%	0.00%	5.95%	31.72%

the presence of property rights: customers take the FIFO waiting time as their initial property and thus do not internalize the externalities inflicted on those who arrive later.

Like Naor (1969), an intuitive remedy to over-joining is to charge an admission fee p^{SW}. This fee can be interpreted as what the service provider charges for accessing the service facility, and thus it applies to all joining customers regardless of their trading decision. It is important to recognize that the admission fee only alters customers' joining incentives, but not their trading incentives since it decreases their utility if they trade just as much as it does the utility obtained from waiting FIFO. To the extent that all money flows are viewed as internal transfers, Proposition 1 shows that an appropriate admission fee can restore the social optimum. Charging a single admission fee and running the baseline auction for trading, this mechanism is *outcome equivalent* to the aforementioned priority auction that regulates both the arrival rate and service order (Hassin 1995), but customers' perception can be quite different. In our mechanism, the service provider charges a flat fee for admission and customers sort out the right service order by themselves through trading. Moreover, joining customers can also opt out of the auction and maintain their FIFO position, but it just so happens that they all voluntarily trade in equilibrium. The second and third column in Table 2.2 illustrate that charging the admission fee can reduce the arrival rate and eliminate the 5.39% social welfare loss in the baseline auction.

4.2 Service Provider's Revenue Maximization

Given the admission fee p and the baseline auction, the service provider's long-run average revenue is $p\lambda(p)$, where $\lambda(p)$ is the arrival rate under p. We show this structure raises the optimal revenue for the service provider under some technical assumptions we will introduce presently. Thus, finding the revenue-maximizing optimal mechanism reduces to pinning down the optimal admission fee.

Before we proceed, we define *virtual type functions* and assume they are monotone.

Definition 2 Denote by $f_r(\cdot)$ and $f_p(\cdot, \tilde{c})$ the receivers' and payers' virtual type functions, respectively:

$$f_r(c) \triangleq c + \frac{F(c)}{f(c)}, \qquad f_p(c; \tilde{c}) \triangleq c - \frac{F(\tilde{c}) - F(c)}{f(c)}.$$

ASSUMPTION 1 $\frac{df_r(c)}{dc} > 0$ *and* $\frac{df_p(c;\tilde{c})}{dc} > 0$ *for all* $c \in [\underline{c}, \tilde{c}]$ *and* $\tilde{c} \in [\underline{c}, \overline{c}]$.

Assuming monotone virtual type functions is common in the mechanism design literature. Monotone virtual types are satisfied by many common probability distributions, such as the uniform, normal, logistic and power function distributions, and the gamma and Weibull distributions with shape parameters greater than or equal to 1; any log-concave distribution has this property (Bagnoli and Bergstrom 2005).

The service provider is not bound by the form of mechanism we introduce (a flat admission fee plus the baseline auction). For example, it could revise the auction rule so as not to induce strict priority. Proposition 2 indicates, nevertheless, that it is optimal under Assumption 1 for the service provider to appeal to the same mechanism structure as the social planner does. The only difference is that the service provider should set a higher admission fee.

Proposition 2 *The service provider maximizes revenue by setting a price* $p^M > p^{SW}$ *and running the baseline auction.*

Given the mechanism structure, if the service provider's only lever were the admission fee, then it should be intuitive that the service provider would set a higher fee than is socially optimal. Naor (1969) has a similar result in a different queueing context. As a monopolist, the service provider would command a higher price than the efficient level to maximize its own revenue. Proposition 2 reveals that even if the service provider has more levers, it should stick to mechanisms that implement strict priority. The monotone virtual types in Assumption 1 guarantee that the service provider has the same incentive as the social planner in prioritizing customers. Otherwise, the service provider would prefer pooling, i.e., serving a class of customers by the FIFO rule despite differences in their waiting costs (cf. Katta and Sethuraman 2005).

Note that as one of the many implementations of the service-provider's optimal direct mechanism, the proposed trading mechanism in Proposition 2 is outcome equivalent to a priority auction with an optimally determined reserve price (cf. Lui 1985). Yet unlike the priority auction, there is no price discrimination by the service provider: all the payments generated in the baseline auction are transfers among customers; still, the same optimal revenue is achieved. We highlight that our proposed trading mechanism, albeit not the unique implementation of the optimal mechanism, is rather simple and that the flat admission fee is only for accessing the service facility, not for gaining priority, so it does not have the unfair connotation like the Priority Answer feature offered by EE.

Table 2.2 illustrates the service provider's optimal trading mechanism in column four, and the optimal pricing of a FIFO queue in column five. The FIFO price, p^F, is defined by

$$(p^F, \tilde{c}^F) = \arg\max_{(p,\tilde{c})} \; p\Lambda F(\tilde{c}), \quad \text{s.t. } V - p - \tilde{c}W(\tilde{c}) = 0. \tag{2.4}$$

Denote the FIFO revenue by $\Pi^F = p^F \Lambda F(\tilde{c}^F)$. While it is immediate that the trading mechanism outperforms FIFO pricing in its revenue performance (2.381 vs. 2.011), it is not clear how the admission fees in the two scenarios, p^M and p^F, compare. Since the exclusive source of revenue in both scenarios is the admission fee, one might expect the service provider who shifts from FIFO to trading to increase this price to extract more revenue. This intuition is correct if the full market is already captured by FIFO pricing, i.e., $\tilde{c}^F = \overline{c}$, but in general, the direction of the service provider's price adjustment is ambiguous. Table 2.2 shows a possibility that the service provider decreases the price (from 3.505 to 3.439) and achieves a higher revenue through a higher throughput. A lower price might be more palatable to customers and make them more receptive of the trading platform.

5 Trading Through an Intermediary

In this section, we study a setting in which the service provider delegates trading to a revenue-maximizing intermediary. The key distinction between the service provider and the intermediary is that the intermediary can only charge customers for using the trading platform (e.g., a trade participation fee), but not for access to the service facility (e.g., an admission fee). Since a high trade participation fee will make trading less attractive and eventually deter some customers from trading altogether, the intermediary's fee-structure will potentially affect customers' trading incentives.

5.1 Baseline Auction with a Trade Participation Fee

We start by considering a benchmark trading mechanism where an arriving customer must pay the intermediary an upfront trade participation fee H; then, the baseline auction is run as before. We refer to this as an "H auction." The intermediary's revenue is $H\lambda^T(H)$, where $\lambda^T(H)$ is the arrival rate of the customers who trade given H. By definition, trading customers are a subset of joining customers, i.e., $\lambda^T(H) \leq \lambda$ for any H.

Recall that in the baseline auction (where $H = 0$), all joining customers are strictly better off by participating in trading. Thus, the equilibrium structure identified in Proposition 1 remains valid if H is slightly positive. It is easy to see

that if the trade participation fee H is too high, then trading will no longer be favored over FIFO. Hence, there exists a threshold value $\overline{H}$ such that the equilibrium structure identified in Theorem 1 is preserved and all joining customers voluntarily trade ($\lambda^T(H) = \lambda$) if and only if $H \leq \overline{H}$.

Definition 3 $\overline{H}$ is such that $\lambda^T(H) = \lambda$ if and only if $H \leq \overline{H}$.

For convenience, we refer to the auction where $H = \overline{H}$ as the "$\overline{H}$ auction." In our running numerical example, at $H = \overline{H} = 1.342$, $U(c)$ in Fig. 2.1c would be tangent to the FIFO line, and the intermediary's revenue is 1.208. If $H > \overline{H}$, some customers with medium waiting costs (since they benefit the least from trading) will find trading too costly and thus refuse to trade by submitting "No," and this would lead to $\lambda^T(H) < \lambda$. The revenue-maximizing intermediary's is in a conundrum. On one hand, if it would like to get all joining customers to trade, its fee is bounded above by $\overline{H}$. One the other hand, if the intermediary wants to charge more aggressively (above $\overline{H}$), it must bear the cost of being unable to collect the fee from some joining customers: a direct loss of revenues via a decreased trading volume, plus, an indirect loss via a lower arrival rate as the non-trading (FIFO) customers downgrade the expected utility of those who trade.

To resolve this conundrum, we enrich the baseline auction with two trade-restriction prices that enable the intermediary to charge above $\overline{H}$ while still inducing voluntary trading of all joining customers. We shall show this is the optimal trading mechanism for the intermediary.

5.2 *Augmented Auction: Trading Rules and a Motivating Example*

Auction Format The augmented auction contains two trade restriction parameters $\underline{R}$ and $\overline{R}$ ($\underline{R} \leq \overline{R}$) in addition to the trade participation fee H. The trading rule is the same as before except that *if both customers' bids are within the interval* $[\underline{R}, \overline{R}]$, *they are barred from trading with one another and are served FIFO*. However, if only one of the two customers' bids are within $[\underline{R}, \overline{R}]$, trade still occurs between the two. This auction is referred to as an "$(H, \underline{R}, \overline{R})$ auction."

Illustration 2 Consider the illustrative scenario in Sect. 3 and assume that $b_4 < \underline{R} < b_2 < b' \leq b_1 < \overline{R}$. As before, the system prior to trading is represented by (b_1, b_2, F, b_4, b'). Only customer 4 and 5 swap positions, and the system after trading is represented by (b_1, b_2, F, b', b_4). Note that despite the fact that $b_2 < b'$, customers 2 and 5 do *not* swap positions since both of their bids fall in $[\underline{R}, \overline{R}]$.

Table 2.3 shows that when $H = 1.510$, $\underline{R} = 0.257$ and $\overline{R} = 0.425$, the intermediary's revenue would be 1.352 in the $(H, \underline{R}, \overline{R})$ auction, 11.9% higher than the revenue that would be achieved in the $\overline{H}$ auction. Note that the trade participation fee H in the augmented auction is higher than $\overline{H}$, yet all joining customers sign up for trading, which can be verified by recognizing the revenue (1.352) is equal to H (1.510) times λ (0.896).

Table 2.3 The intermediary's optimal $(H, \underline{R}, \overline{R})$ auction and the $\overline{H}$ auction

	Revenue	H	$\underline{R}$	$\overline{R}$	λ
Optimal augmented auction	1.352	1.510	0.257	0.425	0.896
$\overline{H}$ auction	1.208	1.342	–	–	0.9

5.3 Auction Equilibrium

To generate insights into how the augmented auction with trade restriction benefits the intermediary, we derive the equilibrium for the case when trading is free ($H = 0$, budget-balanced among customers), and compare that with the budget-balanced baseline auction. We indicate the equilibrium in the augmented auction by means of a superscript A. With a slight abuse of notation, we use $U(c, \beta)$ to denote the expected utility of customer c who bids β in the equilibrium of the augmented auction (including the trade participation fee).

Theorem 2 (Full Trading, Partial Pooling Equilibrium) *Under the augmented auction with given $\underline{R}$ and $\overline{R}$, when $H = 0$, there exists an equilibrium in which:*

(i) $J^A(c) =$*join for* $c \in [\underline{c}, \tilde{c})$ *(and balk otherwise);*

(ii) *the equilibrium bid function is weakly increasing and given by*

$$b^A(c; c_r, c_p, \tilde{c}) = \begin{cases} c + \frac{\int_c^{c_r} (F(\tilde{c}) - F(s))^2 W^e(s; \tilde{c}) ds + K(\underline{R}, c_r, c_p, \tilde{c})}{(F(\tilde{c}) - F(c))^2 W^e(c; \tilde{c})}, & c \in [\underline{c}, c_r) \\ \overline{R}, & c \in [c_r, c_p] \\ b^B(c; \tilde{c}), & c \in (c_p, \tilde{c}] \end{cases} \tag{2.5}$$

where constant $K(\underline{R}, c_r, c_p, \tilde{c}) = (\underline{R} - c_r)(F(\tilde{c}) - F(c_r))^2 W^e(c_r; \tilde{c})$ *and* $c_r, c_p, \tilde{c} \in \Xi$ *with* $\underline{c} \le c_r \le c_p \le \tilde{c}$ *are a solution to the following equations:*

$$[U(c_r, \underline{R}) - U(c_r, \overline{R})][c_r - \underline{c}][c_r - \tilde{c}] = 0 \tag{2.6a}$$

$$[U(c_p, b^A(c_p^+; c_r, c_p, \tilde{c})) - U(c_p, \overline{R})][c_p - \underline{c}][c_p - \tilde{c}] = 0 \tag{2.6b}$$

$$U(\tilde{c}, b^A(\tilde{c}; c_r, c_p, \tilde{c}))[\tilde{c} - \overline{c}] = 0; \tag{2.6c}$$

(iii) *the expected waiting time for customer* $c \in [\underline{c}, \tilde{c}]$ *is*

$$W^A(c; c_r, c_p, \tilde{c}) = \begin{cases} W^e(c; \tilde{c}), & \forall\, c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}] \\ \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c_r))(1-\rho F(\tilde{c})+\rho F(c_p))}, & \forall c \in [c_r, c_p]. \end{cases} \tag{2.7}$$

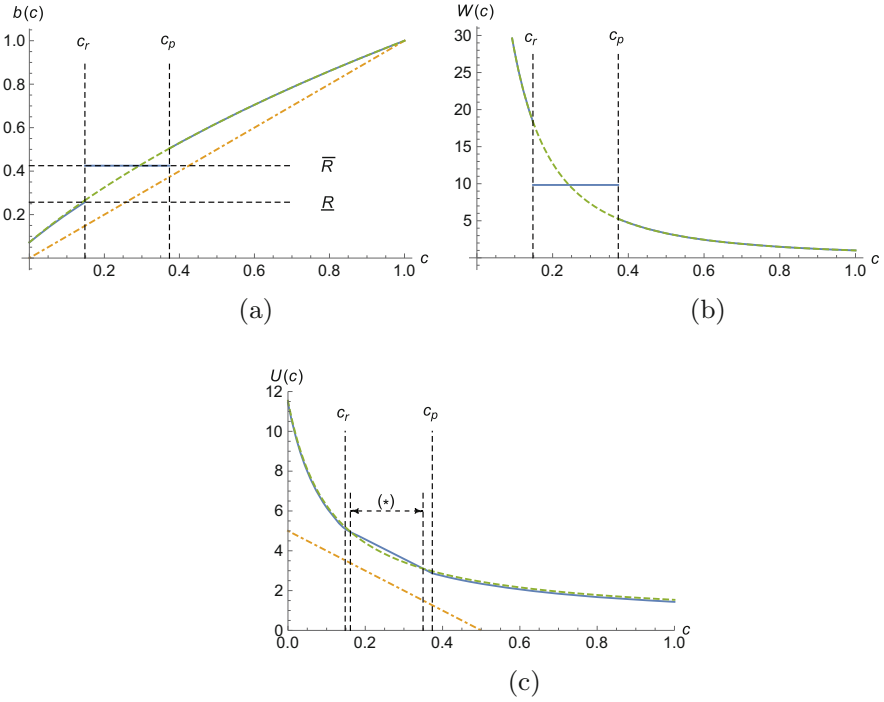

Fig. 2.2 The augmented auction. *Note.* $H = 0$, $\underline{R} = 0.257$, $\overline{R} = 0.425$. $\tilde{c} = 1$, $c_r = 0.148$, $c_p = 0.373$. The blue solid curves correspond to the equilibrium properties of the augmented auction; the green dashed curve, the baseline auction as in Fig. 2.1. The orange dot-dashed line is the 45-degree line in (**a**); the expected utility if customers are served FIFO under the same equilibrium arrival rate in (**c**). $(*)$ indicates the subset of customer types in the pooling segment that receive a higher expected utility than in the baseline auction. (**a**) Bidding function. (**b**) Waiting time. (**c**) Utility

Comparing Theorem 2 with Theorem 1 shows the effects of the trade restriction parameters, $\underline{R}$ and $\overline{R}$. While the bid function $b^A(\cdot)$ in (2.5) is still strictly increasing in $[\underline{c}, c_r] \cup (c_p, \tilde{c}]$, it is flat in $[c_r, c_p]$: these customers all bid $\overline{R}$ and thus do not trade with one another (see Fig. 2.2a). As a result, the waiting time schedule is no longer efficient since these customers all expect the same waiting time despite their different waiting costs (see Fig. 2.2b). Consequently, there is pooling of customers in $[c_r, c_p]$, who are served as a single FIFO class. As shown in the expression of $W^A(\cdot)$ in (2.7), for any given arrival rate, the expected waiting time for customers in $[\underline{c}, c_r) \cup (c_p, \tilde{c}]$ is still the efficient waiting time, i.e., trading allows these customers to be strictly prioritized over any other joining customer with lower waiting cost.

Similar to the baseline auction, adding $\underline{R}$ and $\overline{R}$ to the budget balanced auction does not discourage any joining customers from voluntarily participating in trading. As shown in Fig. 2.2c, all joining customers are better off by participating in trading than submitting "No." One noticeable difference of customers' expected utility in

the augmented auction is that it decreases linearly in c for $c \in [c_r, c_p]$. This is by the linearity assumption of waiting costs. Customers in the pooling segment $[c_r, c_p]$ differ in their waiting costs but choose to bid the same amount and thus expect to wait the same amount of time and pay/receive the same amount of money.

5.4 Optimal Auction Parameters and Structure

When $H = 0$ in the augmented auction (the intermediary has no revenue), all customers strictly prefer trading. A slightly positive H will not alter this preference and the equilibrium bidding behavior in the augmented auction insofar as all joining customers trade. We show that this is indeed the optimal structure the intermediary wants to implement. Furthermore, the $(H, \underline{R}, \overline{R})$ auction is an optimal mechanism for the intermediary given the optimal auction parameters.

Theorem 3 (Optimality of the Augmented Auction) *The* $(H^*, \underline{R}^*, \overline{R}^*)$ *is an optimal mechanism for the intermediary with:*

$$H^* = \frac{\Pi(c_r^*, c_p^*, \tilde{c}^*)}{\Lambda F(\tilde{c}^*)} \tag{2.8a}$$

$$\overline{R}^* = \frac{c_p^* \overline{W}(\tilde{c}^*) + [\rho b^R(c_p^*; c_r^*, c_p^*, \tilde{c}^*)(F(\tilde{c}^*) - F(c_p^*)) - c_p^*] W^e(c_p^*; \tilde{c}^*)}{\rho(F(\tilde{c}^*) - F(c_r^*))\overline{W}(\tilde{c}^*)} \tag{2.8b}$$

$$\underline{R}^* = \frac{c_r^* W^e(c_r^*; \tilde{c}^*) - c_r^* \overline{W}(\tilde{c}^*) + \rho \overline{W}(\tilde{c}^*)[F(\tilde{c}^*) - F(c_p^*)]\overline{R}^*}{\rho W^e(c_r^*; \tilde{c}^*)[F(\tilde{c}^*) - F(c_r^*)]} \tag{2.8c}$$

where $\Pi(c_r, c_p, \tilde{c}) = -\Lambda \int_{\underline{c}}^{c_r} \left(W^e(c; \tilde{c}) - \overline{W}(\tilde{c})\right) f_r(c) dF(c) + \Lambda \int_{c_p}^{\tilde{c}} \left(\overline{W}(\tilde{c}) - W^e(c; \tilde{c})\right) f_p(c; \tilde{c}) dF(c)$ *and* $c_r^*, c_p^*, \tilde{c}^*$ *solve the following optimization problem:*

$$\max_{c_r, c_p, \tilde{c} \in \Xi : c_r \le c_p \le \tilde{c}} \quad \Pi(c_r, c_p, \tilde{c}) \tag{2.9a}$$

$$\text{s.t.} \quad \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))(1 - \rho F(\tilde{c}) + \rho F(c_p))} = \overline{W}(\tilde{c}) \tag{2.9b}$$

$$V - \int_{c_p}^{\tilde{c}} W^e(c; \tilde{c}) dc - c_p \overline{W}(\tilde{c}) \ge 0. \tag{2.9c}$$

The resulting equilibrium structure is the same as identified in Theorem 2. In particular,

$$U\left(c, b^A\left(c; c_r^*, c_p^*, \tilde{c}^*\right)\right) = V - c\overline{W}(\tilde{c}^*), \quad \forall c \in \left[c_r^*, c_p^*\right]. \tag{2.10}$$

Theorem 3 determines the optimal parameters $(H^*, \underline{R}^*, \overline{R}^*)$ by reverse engineering. Instead of characterizing the equilibrium outcome by solving (2.6a)–(2.6c) for $c_r, c_p, \tilde{c}$ under any given auction parameters $(H, \underline{R}, \overline{R})$, we first determine what the optimal outcome should be by obtaining $c_r^*, c_p^*, \tilde{c}^*$ from the optimization problem (2.9a)–(2.9c) and then determine the optimal auction parameters $(H^*, \underline{R}^*, \overline{R}^*)$ that can implement the optimal outcome using (2.8a)–(2.8c), where (2.8b) and (2.8c) are obtained from shuffling terms of (2.6a) and (2.6b).

Combining (2.9b) and $W^A(\cdot)$ in (2.7) implies that in the optimal auction, the expected waiting time for customers in $[c_r^*, c_p^*]$ is equal to the FIFO waiting time $\overline{W}(\tilde{c})$ (see Fig. 2.3b). Furthermore, (2.10) suggests that these customers' expected utility is equal to what they would get if they just bid "No" and wait FIFO (see Fig. 2.3c). This does not imply they do not trade at all: they still swap positions with customers *outside the pool* by selling their spot to higher bidders and buying positions from lower bidders, but on average trading does not realize any gains. The fact that they trade is crucial to achieving an efficient expected waiting time for

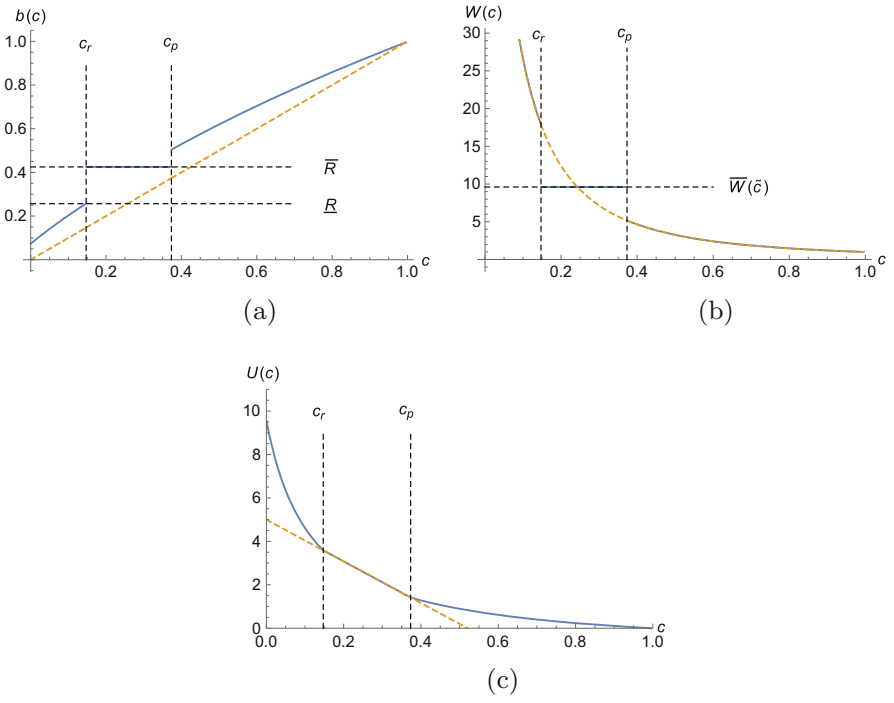

Fig. 2.3 The optimal $(H, \underline{R}, \overline{R})$ auction. *Note.* $H = 1.510$, $\underline{R} = 0.257$, $\overline{R} = 0.425$. $\tilde{c} = 0.995$, $c_r = 0.147$, $c_p = 0.373$. The solid curve corresponds to the equilibrium properties of the auction; the dashed curve is the 45-degree line in (**a**); the efficient waiting time function in (**b**); the expected utility if customers are served FIFO under the same equilibrium arrival rate in (**c**). (**a**) Bidding function. (**b**) Waiting time. (**c**) Utility

customers outside the pool $[\underline{c}, c_r^*) \cup (c_p^*, \tilde{c}^*]$. The augmented auction is designed to only prohibit trading within the pool.

In brief, the optimal auction has the following two features:

(1) All joining customers participate in trading.
(2) The schedule is efficient in $[\underline{c}, c_r^*) \cup (c_p^*, \tilde{c}^*]$ but customers in $[c_r^*, c_p^*]$ expect FIFO waiting time.

Theorem 4 (Optimality of Trade Restriction) *The optimal auction should always have* $\underline{R}^* < \overline{R}^*$*, and thus* $c_r^* < c_p^*$.

Theorem 4 shows that the intermediary would like to restrict trading to a certain extent to fully exploit its control over the trading channel. As a result, it sets $\overline{R}^*$ strictly above $\underline{R}^*$ so that pooling occurs in the intermediary's optimal auction and the schedule is not efficient. This extends the classical result in Myerson and Satterthwaite (1983) about the intermediary's trade restriction incentives in bilateral trading to a queueing context where customers can be both buyers and sellers. After all, the intermediary is a monopolist who wants to restrict output under the efficient level to command a higher price. Note that if the service provider operates the trading platform as in Sect. 4.2, it can simply exercise its monopoly power by charging a higher admission fee (which reduces arrivals). By contrast, the intermediary's trade participation fee cannot be forced upon customers even after they join the system, so setting a higher fee as an intermediary should be done in a more nuanced way that reduces the amount of trading among customers and creates a pooling segment. As we argue in Sect. 5.3, trade restriction generates value to customers in the middle who are most sensitive to trading by letting them avoid undesirable trades. This value added, in turn, passes on to the intermediary by enabling it to charge a higher trade participation fee (we shall formally establish this in Corollary 1).

Let $\lambda^* = \Lambda F(\tilde{c}^*)$ be the optimal effective arrival rate under the $(H^*, \underline{R}^*, \overline{R}^*)$ auction. Let λ^{FIFO} be the effective arrival rate if all joining customers are served FIFO. In the FIFO system, a customer of type c receives an expected utility of $V - c\overline{W}(\tilde{c}^{\text{FIFO}})$, where $\tilde{c}^{\text{FIFO}}$ satisfies $\lambda^{\text{FIFO}} = \Lambda F(\tilde{c}^{\text{FIFO}})$. Let $\lambda^{\overline{H}}$ be the effective arrival rate if the intermediary charges $\overline{H}$ in the H auction. Likewise, $\tilde{c}^{\overline{H}}$ is defined such that $\lambda^{\overline{H}} = \Lambda F(\tilde{c}^{\overline{H}})$. Proposition 3 orders the three arrival rates.

Proposition 3 $\lambda^{FIFO} \le \lambda^* \le \lambda^{\overline{H}}$*. In particular, if not all customers join in the optimal auction, i.e.,* $\lambda^* < \Lambda$*, then* $\lambda^{FIFO} < \lambda^* < \lambda^{\overline{H}}$.

In the optimal augmented auction, customers with high waiting costs who would otherwise balk in a FIFO system may now join and participate in trading because this option to trade makes them better off than if they are served FIFO. On the other hand, as compared to the $\overline{H}$ auction, the pooling segment in the optimal mechanism diminishes the appeal of trading for customers with high waiting costs since they have to pay more to get the same priority; thus fewer customers join. The optimal mechanism has a lower arrival rate than the $\overline{H}$ auction, but it raises more

revenue, which must stem from a higher trade participation fee. This is summarized in Corollary 1.

Corollary 1 *The optimal trade participation fee in the augmented auction is strictly above $\overline{H}$, i.e., $H^* > \overline{H}$.*

Trade restriction resolves the intermediary's conundrum: it always charges strictly above $\overline{H}$, yet all joining customers choose to pay this trade participation fee. Combining Theorem 4, Proposition 3 and Corollary 1 shows that the gain from charging a higher trade participation fee above $\overline{H}$ in the augmented auction overshadows the cost of a resulting lower arrival rate for the intermediary.

5.5 The Value of Trading vs. FIFO

Now, we turn to the service provider's pricing decision in the presence of the intermediary. In a FIFO queue, the service provider would set a revenue-maximizing admission fee, p^F, as formalized in (2.4). We compare this to a setting where the service provider invites the revenue-maximizing intermediary to mediate the trading platform. The service provider is completely detached from the trading platform and does not contract with the intermediary due to technological reasons and reputational concerns discussed in Sect. 1. The service provider and the intermediary play a sequential game: the service provider first sets an admission fee p^T and then the intermediary's implements its optimal trading mechanism. Corollary 2 shows that the service provider is always better off by inviting the intermediary.

Corollary 2 *The intermediary running the trading platform strictly increases the service provider's revenue relative to FIFO.*

Next, we numerically quantify, under a variety of input parameters, how the intermediary's trading mechanism impacts the service provider's pricing decision in Table 2.4. We also compare the service provider's revenue with the intermediary's trading platform relative to a FIFO queue in Table 2.5. In all numerical trials, we fix the service rate to be 1 and the waiting cost to be uniformly distributed between 0 an 1 as in Table 2.1. We vary the service value V between 2 and 10, and the market size Λ between 0.1 and 3. Denote the percentage difference between the service provider's FIFO price, p^F, and price under trading, p^T, by Δ_p; the percentage difference between the service provider's revenue under FIFO, Π^F and revenue under trading, Π^T, by Δ_Π.

$$\Delta_p = \frac{p^T - p^F}{p^F} \times 100\%, \qquad \Delta_\Pi = \frac{\Pi^T - \Pi^F}{\Pi^F} \times 100\%.$$

Table 2.4 shows that it is in general ambiguous how the service provider should adjust its price when the intermediary implements the trading platform. Table 2.5 shows the service provider's revenue improvement. In terms of its pricing behavior,

Table 2.4 Percentage change in price Δ_p

Λ	$V=2$	$V=3$	$V=4$	$V=5$	$V=6$	$V=7$	$V=8$	$V=9$	$V=10$
0.1	−0.73%	1.16%	0.76%	0.56%	0.45%	0.37%	0.32%	0.28%	0.25%
0.3	−1.49%	−1.73%	3.70%	2.66%	2.08%	1.71%	1.45%	1.26%	1.11%
0.5	−1.78%	−1.86%	−1.81%	−0.09%	6.39%	5.11%	4.26%	3.65%	3.19%
0.7	−1.86%	−1.78%	−1.60%	−1.42%	−1.19%	−0.98%	−0.82%	−0.59%	−0.09%
0.9	−1.83%	−1.63%	−1.39%	−1.01%	−0.87%	−0.57%	−0.50%	−0.29%	−0.23%
1.1	−1.76%	−1.68%	−1.31%	−0.87%	−0.52%	−0.34%	−0.18%	0.03%	0.13%
2	−1.59%	−0.73%	−0.32%	0.00%	0.38%	0.48%	0.65%	0.81%	0.87%
3	−0.79%	−0.15%	0.27%	0.50%	0.81%	0.87%	1.10%	1.20%	1.21%

Table 2.5 Percentage change in revenue Δ_Π

Λ	$V=2$	$V=3$	$V=4$	$V=5$	$V=6$	$V=7$	$V=8$	$V=9$	$V=10$
0.1	1.65%	1.16%	0.76%	0.56%	0.45%	0.37%	0.32%	0.28%	0.25%
0.3	3.99%	5.21%	3.70%	2.66%	2.08%	1.71%	1.45%	1.26%	1.11%
0.5	5.55%	6.89%	7.82%	7.32%	6.39%	5.11%	4.26%	3.65%	3.19%
0.7	6.67%	7.97%	8.80%	9.36%	9.74%	10.01%	10.20%	10.33%	8.18%
0.9	7.49%	8.71%	9.42%	9.87%	10.15%	10.33%	10.45%	10.51%	10.55%
1.1	8.11%	9.22%	9.83%	10.18%	10.38%	10.49%	10.54%	10.56%	10.55%
2	9.64%	10.28%	10.51%	10.56%	10.53%	10.45%	10.35%	10.24%	10.13%
3	10.28%	10.55%	10.53%	10.40%	10.24%	10.07%	9.90%	9.73%	9.56%

the numerical instances can be divided into three cases. *Case 1*: In the top right corner of Table 2.4, when the service value is high and the market size is small, the service provider's price goes up. This corresponds to the case when the full market is captured in the FIFO queue. As we argued following Corollary 2, trading allows the service provider to raise its price. This can be verified by recognizing that in those instances the relative price change is equal to the relative revenue change shown in Table 2.5 as the arrival rate is unaffected. *Case 2*: In the bottom right corner of Table 2.4, when the service value is high and the market size is also large, the service provider's price rises again. However, in these instances, the system does not capture the full market, and the arrival rate is also changed as a result of trading. We observe that the revenue change is higher than the price change now, which implies that trading allows the service provider to both command a higher price and lure more customers. *Case 3*: In the rest of the instances, the service provider offers a price cut, so that the revenue increase is solely attributed to a higher arrival rate. Here the FIFO queue does not capture the full market. As Proposition 3 suggests, even if the service provider sticks to its original price, it will enjoy a higher revenue since more customers join when the trading platform is in place. However, the service provider responds by actually decreasing its price to get an even higher arrival rate. As we suggest in Sect. 4.2, a lower price might be more favorable to customers from a behavioral perspective, and this may facilitate the promotion of the intermediary's trading platform.

Somewhat strikingly, in our numerical study, when the market is not fully captured (cases 2 and 3), the magnitude of the price change is, in fact, quite small: less than 2% in all those instances; yet the revenue change is quite sizable by comparison (about 10% in many instances). The implication is that the intermediary's trading platform can potentially be a seamless built-in for the service provider: the service provider does not need to worry about running the auction itself; it does not even need to significantly alter its price as a response of the new platform (which is valuable especially when the menu cost is high). The bottom line is that the intermediary increases the service provider's revenue relative to a FIFO system, and improves system efficiency. These are the intermediary's value propositions to the service provider with either revenue or welfare considerations. Of course, there is a natural double marginalization problem in our setup where both the service provider and the intermediary are monopolists. Theoretically, the service provider would earn an even higher revenue if it operated the trading platform by itself as in Sect. 4.2. Practically, this may not be in the service provider's best interest for technological reasons and reputational concerns previously stated. Vertical integration would achieve the maximum joint revenue as in Sect. 4.2, but this usually involves efforts expended on negotiation, coordination and contracting. In this regard, an intermediary on a separate platform should probably be good enough in practice.

6 Conclusion Remarks

This chapter analyzes a congested service system in which customers are privately informed about their waiting cost and trade their waiting positions on a trading platform. We design the optimal mechanisms that maximize social welfare, the service provider's revenue, and the revenue of the intermediary that develops and manages the trading platform, respectively. We find that while both the social planner and the service provider want customers to trade as much as possible (inducing the $c\mu$ rule), the intermediary *restricts* trading among customers (pooling) to maximize its own revenue. In particular, a budget-balanced baseline auction leads to a higher arrival rate than is socially desirable and thus an admission fee must be levied to maximize social welfare. By comparison, the revenue-maximizing service provider would charge a higher admission fee than the social planner would. For practical reasons, the service provider may wish to delegate the trading platform to a revenue-maximizing intermediary. To that end, we propose an augmented auction with a trade participation fee *and* two trade restriction prices. Compared to the baseline auction with a trade participation fee only, the intermediary can charge a higher fee in the optimal auction and still have *all* joining customers voluntarily participate in trading. We show that the intermediary's trading mechanism always strictly improves the service provider's revenue relative to a FIFO system despite the intermediary's revenue-maximizing nature. This is a potentially powerful sales

argument the intermediary can make to convince the service provider of installing the platform.

One practical concern for introducing the trading marketplace is the rise of speculative behavior. One is the arrival of "scalpers" who game the system by selling their spots for money without actually receiving the service. These customers are typically time-insensitive and ascribe low valuation to the service itself and thus would not join the system otherwise. We will study queue-scalping in Chap. 4. Another related phenomenon is "line-sitting" whereby real customers hire line-sitters to wait in line on their behalf and swap in only when line-sitters approach the head of the line. We will study line-sitting in Chap. 3. On the one hand, the presence of speculators does not violate other customers' property rights since such swaps are still one-to-one substitution. On the other hand, these customers are likely to renege before entering the service, appropriating pecuniary gains that might otherwise be captured by the service provider. In principle, the up-front trade participation fee in the intermediary's optimal auction should deter some speculative customers. Additionally, the platform can act as a gatekeeper that closely monitors any suspicious trading activities and bans unscrupulous customers from using the trading platform if necessary. This further justifies the importance of the trading platform (intermediary) in mediating transactions.

References

Bagnoli M, Bergstrom T (2005) Log-concave probability and its applications. Economic Theory 26(2):445–469

Cramton P, Gibbons R, Klemperer P (1987) Dissolving a partnership efficiently. Econometrica 55(3):615–632

El Haji A, Onderstal S (2019) Trading places: an experimental comparison of reallocation mechanisms for priority queuing. J Econ Manag Strateg 28(4):670–686

Gershkov A, Schweinzer P (2010) When queueing is better than push and shove. Int J Game Theory 39(3):409–430

Gray K (2009) Property in a queue. In: Alexander GS, Penalver EM (eds.) Property and community. Oxford University Press, New York, pp 165–195

Hassin R (1995) Decentralized regulation of a queue. Manag Sci 41(1):163–173

Katta A, Sethuraman J (2005) Pricing strategies and service differentiation in queues: a profit maximization perspective. Working paper, Columbia University, New York

Kleinrock L (1967) Optimum bribing for queue position. Oper Res 15(2):304–318

Lui FT (1985) An equilibrium queueing model of bribery. J Polit Econ 93(4):760–781

Mann L (1969) Queue culture: the waiting time line as a social system. Am J Sociol 75(3):340–354

Myerson RB, Satterthwaite MA (1983) Efficient mechanisms for bilateral trading. J Econ Theory 29(2):265–281

Naor P (1969) The regulation of queue size by levying tolls. Econometrica 37(1):15–24

Rosenblum DM (1992) Allocation of waiting time by trading in position on a G/M/S queue. Oper Res 40:S338–S342

Yang L, Debo L, Gupta V (2017) Trading time in a congested environment. Manag Sci 63(7):2377–2395