

ORIGINAL ARTICLE

Designing professional services: Pricing and prioritization

Chenguang (Allen) Wu¹  | Chen Jin²  | Senthil Veeraraghavan³ 

¹Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

²Department of Information Systems and Analytics, School of Computing, National University of Singapore, Singapore

³The Operations, Information, and Decisions Department, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Correspondence

Chen Jin, Department of Information Systems and Analytics, School of Computing, National University of Singapore, Singapore.
Email: disjinc@nus.edu.sg

Handling Editor: Michael Pinedo

Funding information

Hong Kong General Research Fund, Grant/Award Number: 16506122; The Wharton School Dean's Postdoctoral Research Fund; Mack Institute Research Fund; Singapore Ministry of Education Academic Research Fund, Grant/Award Number: Tier 1 (251RES2101)

Abstract

We study the optimal design of a professional service in a mixed market of customers with heterogeneous skills and capabilities of using such service. *Expert* customers can avail of the service on their own, whereas *amateur* customers find it challenging to deploy the service and can only procure the service through an intermediary who resolves the technical issues. We develop a model that captures the essential trade-offs in such settings: heterogeneity in customer expertise, decentralization between a service provider and intermediary, and congestion due to limited capacity. We analyze how customer expertise differences drive the equilibrium outcomes under various pricing and priority schemes. We find that a sufficient base of amateur customers allows expert customers to “free-ride” under single pricing. Price discrimination can fully allay such free-riding, but it may drive prices downward. Price discrimination also favors expert customers under the First-Come-First-Served (FCFS) policy, but such preference is generally reversed under prioritization. Specifically, prioritizing amateur customers can bring revenue and welfare gains relative to the FCFS policy and a policy that prioritizes expert customers. Our results offer normative guidelines for managing professional services, clarifying regimes for price and priority discrimination, along with revenue and welfare implications.

KEYWORDS

pricing, priority, professional services, queueing game, service operations

1 | INTRODUCTION

In recent years, the service sector in the United States has accounted for a majority of total employment.¹ Rapid growth in technology sectors has created an enormous market for professional services. For example, Amazon's cloud service, Amazon Web Service (AWS), has been a more profitable channel for Amazon compared to traditional retail. According to Amazon 2021 Annual Report, even though AWS revenues were much smaller than domestic retail revenues (11% vs. 61% of overall revenues, respectively), AWS profits were higher than retail profits.²

A growing feature of such services is their extended complexity. Indeed, users have become increasingly differentiated in their capabilities to deploy the service offerings. Some

users can deploy the service and tailor it to their own use. Other users find the service offerings technically complex and would prefer “plug and play” and may often choose to work with a third-party intermediary. In the context of cloud computing, many machine learning projects are first deployed on distributed large-scale data frameworks (e.g., Apache Spark or Hadoop) before the data training process can be initiated. Tech-savvy users, familiar with computing hardware and infrastructure (e.g., CPU and GPU servers), can install and configure the parameters by themselves. Other users, lacking expertise in computer science, may contract with a third-party intermediary (e.g., Databricks³) to help them build the necessary framework to access the cloud service.

The boom in professional services, such as the industry of Infrastructure as a Service (IaaS) coupled with the ecosystem of intermediaries servicing customers of different expertise, calls for a better understanding of the optimal design of these services. We hope that our paper takes an important step

Accepted by Michael Pinedo, after two revisions.

Chenguang (Allen) Wu and Chen Jin contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs License](https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Production and Operations Management* published by Wiley Periodicals LLC on behalf of Production and Operations Management Society.

toward service design, by understanding the trade-offs in how to price and prioritize these services to customers.

In this paper, we refer to *professional services* as services for which there exists substantial heterogeneity in user capabilities and deployment. Hereafter, we refer to customers who are capable of accessing the service on their own as *expert* customers, and others whose access to the service must be established through an intermediary as *amateur* customers.⁴

Another prominent feature of professional services is that customers often have to compete for a limited shared resource, giving rise to congestion externalities. In the context of cloud computing, when there is a surge of requests on clusters, incoming jobs may experience delays in accessing busy servers based on their priority levels. Thus, customers on the cloud (either on their own or through an intermediary) impose an externality on other customers in a way that depends on their types. Given this subtle interaction, cloud operators should understand how to serve high-margin expert customers that directly procure a service as against an increasing number of amateur customers brought by the intermediary. Note that the presence of the intermediary cuts into the provider's margins in serving amateur customers, but opens access to a new customer segment that otherwise would not deploy on the cloud.

In this paper, we develop a model that captures the key features in designing professional services: heterogeneity in customers' accessibility to a service, business through intermediaries, and congestion due to limited capacity. We analyze various pricing and prioritization schemes, and identify the fraction of amateur customers in the market as a critical driver of equilibrium outcomes, prices, and revenues. We show that under single pricing and the First-Come-First-Served (FCFS) policy, the provider facing a high demand of expert users will serve these customers exclusively. When the demand of expert customers is not too high, the provider will naturally also serve amateur customers. The presence of amateur customers, however, allows expert customers to "free-ride" and receive a positive surplus. So, the presence of large demand from amateur customers is a boon to expert customers, under the single pricing scheme.

Price discrimination allays free-riding effects. As the service provider must share the reward of serving amateur customers with the intermediary, the provider prefers to serve expert customers for their high-profit margins. As a result, the provider will not plan expansive coverage to all customers unless the demand of expert customers is sufficiently low. In this case, price discrimination may drive prices for both classes downward compared to the optimal single price. This finding offers an explanation as to why cloud services such as AWS and Microsoft Azure have expanded in scale and market reach over time across customer classes and have seen reduced profit margins after expansion.

We also optimize pricing and priorities jointly. A service provider may find it tempting to prioritize expert customers under price discrimination (as they bring better revenues). However, we caution against such policies. In fact, we show that prioritizing amateur customers can bring additional rev-

enue benefits relative to the FCFS policy and a policy that prioritizes expert customers. Such benefits are driven by the wait-time reduction of amateur customers when they are prioritized, which effectively alleviates double marginalization and boosts the joining rate of the amateur segment. Of course, prioritizing amateur customers will impose a higher waiting cost on expert customers. However, the joint revenues are better, as the revenue accrued from the additional amateur customers served outweighs any reduced margins of expert customers. In a similar vein, we find that prioritizing amateur customers can also bring welfare benefits. We identify *resource utilization* as the main driver of the welfare implications.

Finally, we consider two extensions. In the first extension, we allow the intermediary to have a marginal cost in acquiring amateur customers and show that all our conclusions extend. In the second extension, we assume the intermediary's fee is exogenous. Under this assumption, we show that prioritization cannot improve revenue or welfare benefits relative to FCFS. This is because customers share the same delay sensitivity irrespective of their type. This makes priority discrimination futile under the intermediary's exogenous fee. However, when the intermediary can set its fee, the provider can use prioritization to manage double marginalization.

2 | LITERATURE REVIEW

Our work is related to the broad literature on service operations and queueing games. Naor (1969) proposes the first queueing model with strategic customers and Edelson and Hilderbrand (1975) extend Naor's framework to unobservable queues. A large body of the literature follows in this vein, such as Gilbert and Weng (1998), Mendelson and Whang (1990), and Chen and Frank (2004). See Hassin and Haviv (2003) for a comprehensive review of this literature. Recently, Chen et al. (2022) and Feldman et al. (2023) extend Naor's framework to consider a decentralized service setting similar to ours (e.g., a restaurant and a food delivery platform). They study the incentive issues between decentralized players in serving a market of heterogeneous customers. Cui et al. (2020) study the interactions between a service provider and a line-sitting intermediary that makes money from providing queueing service for its customers. It is significant that none of these papers consider using prioritization to manage double marginalization, which is critical in our study. Our work differs from these papers as we explicitly demonstrate the economic and welfare benefits of joint price and priority discrimination.

As our work concerns priority pricing, it is closely related to the literature on priority queues. This literature begins with Adiri and Yechiali (1974), Balachandran and Schaefer (1979), and Alperstein (1988), and is extended by Afeche and Sarhangian (2015), Debo and Veeraraghavan (2014), Gavirneni and Kulkarni (2016), Hassin and Haviv (2006), and Armony et al. (2021). Notably, the priority decisions in these papers are primarily driven by customers' heterogeneity in

delay sensitivities, whereas these decisions in our model are mainly due to the fact that customers are heterogeneous in their abilities of deploying the service. Specifically, a key feature of our model, the decentralization between the service provider and intermediary, is generally missing in this literature. Our work contributes to this literature by revealing the operational benefits of prioritization in combating double marginalization.

3 | MODEL

We consider a service provider facing a heterogeneous market of customers with varied capabilities of deploying a service. A fraction α of the market consists of *amateur* customers who are technically naive and must work with an intermediary to procure the service. The remaining $1 - \alpha$ fraction are *expert* customers who can procure the service on their own and tailor it to personal use. Thus, the parameter α can be interpreted as a measure of sophistication or specialization of a professional service and can be estimated from marketing research. For highly specialized professional services, amateur customers will constitute a significant portion of the market, in which case, α is likely high. We assume the parameter α is known to both the provider and intermediary (see Feldman et al., 2023 and Chen et al., 2022 for similar assumptions).

Further, customers often have to compete for a limited shared resource, giving rise to congestion externalities that adversely affect customers' service experience and their willingness to pay in the first place.⁵ To capture this effect, we model the service system as an $M/M/1$ queue (for similar modeling choices, see Jafarnejad Ghomi et al. 2019 for a comprehensive review of recent developments in applying queueing theory to model cloud computing). Specifically, in our model, customers' service requests arrive in a Poisson stream at an exogenous rate Λ and we refer to Λ as the *market size*. The service time of each request is exponentially distributed with rate μ and we refer to μ as the *service capacity*.

Each customer has a valuation V for her requested service and incurs a waiting cost c per unit of time spent in the system (including time in service). Each customer decides whether to purchase the service (throughout this paper, we interchangeably refer to customers' *purchasing* and *joining* as the same) or quit the service. Customers do not renege or abandon after they join. We normalize customers' utilities of not joining to zero. We term the joining rate of each segment as the *effective arrival rate* of that segment. To decide whether to join, expert customers, if they make a direct purchase, obtain a net utility that equals their valuation of the service, less the price the provider charges them, and less the expected waiting cost. For amateur customers, their joining decisions also depend on the intermediary's fee decision. As the main goal of this paper is to understand how a provider of a complex professional service should treat various customer segments differentially on pricing and prioritization, we abstract away the contracting issues between the provider and intermediary (see Feldman

TABLE 1 Glossary of main notation.

Symbol	Definition
V	Customers' valuations of service
α	Fraction of amateur customers
Λ	Market size
μ	Service capacity
c	Delay sensitivity
p_A, p_F	Prices charged to amateur and expert customers
λ_A, λ_F	Effective arrival rates of amateur and expert customers
R	Revenue of the service provider
Π	Social welfare

et al., 2023 and Chen et al., 2022 for a relevant discussion) and focus on a simple setting in which the intermediary generates revenues by charging amateur customers a separate fee. We present a detailed formulation of the intermediary's fee decision and amateur customers' joining decisions in Subsection 3.2.

We assume $V > c/\mu$ throughout this paper to rule out the uninteresting case in which customers are unwilling to join the service even if they are not delayed in queue. In what follows, we first consider two boundary cases of market conditions: (i) a market of expert customers only, $\alpha = 0$; (ii) and a market of amateur customers only, $\alpha = 1$. We then consider a heterogeneous customer population mixed with expert and amateur customers, $\alpha \in (0, 1)$. Table 1 summarizes the main notations in this paper.

3.1 | Market of expert customers

In a market of expert customers only ($\alpha = 0$), each expert customer will directly purchase the service, so that the presence of the intermediary is not relevant. As all customers have the same valuation for the service and incur the same expected waiting cost after joining, the service provider can fully extract the surplus of joining customers (Hassin & Haviv, 2003, chapter 3). In equilibrium, all customers receive a zero surplus whether they join or not, and thus are indifferent between joining and not joining.

Formally, given a service price p charged to expert customers, the effective arrival rate λ_0 of expert customers solves⁶

$$V - \frac{c}{\mu - \lambda_0} - p = 0. \quad (1)$$

Anticipating (1), the provider selects price p to maximize his revenue

$$\max_p p\lambda_0(p),$$

with $\lambda_0(p)$ solving (1). The following result characterizes the optimal price for such a market.

Theorem 1 (Anand et al., 2011, Proposition 1). *In a market of expert customers, the provider's optimal price and the resulting effective arrival rate are:*

$$p_0^* = V - \frac{c}{\mu - \lambda_0^*}, \text{ where } \lambda_0^* = \begin{cases} \Lambda, & \text{if } \Lambda \leq \hat{\lambda}_0, \\ \hat{\lambda}_0, & \text{if } \Lambda > \hat{\lambda}_0, \end{cases}$$

and

$$\hat{\lambda}_0 := \mu - \sqrt{c\mu/V}. \quad (2)$$

The proofs of the results in Sections 3–5 can be found in Appendix B, and the proofs of the results in Section 6 are in Appendix C in the E-companion. Theorem 1 gives the threshold of market size $\hat{\lambda}_0$ that differentiates the provider's coverage of expert customers. When the market size Λ is small, congestion is less important and the provider opts to cover the entire market. As Λ grows and exceeds $\hat{\lambda}_0$, congestion substantially cuts into the provider's margin, forcing the provider to only partially cover the market. This leads to a fixed effective arrival rate $\hat{\lambda}_0$ despite a further growing market size.

We remark that in a market of expert customers, the price charged by a revenue-maximizing service provider, as given in Theorem 1, is also socially optimal. This is because each customer receives a zero surplus whether she joins or not. This implies that the revenue collected by the service provider is the same as the social welfare, so that his revenue-maximizing price is also socially optimal (see Hassin & Haviv, 2003, chapter 3 for similar observations).

3.2 | Market of amateur customers

In a market full of amateur customers ($\alpha = 1$), customers find the service technically complicated for direct use and must work with a professional intermediary to access the service. In the context of cloud computing, Databricks helps amateur users customize their computing infrastructure (e.g., parameter configuration) to support the normal operations of their computing projects. For simplicity, we abstract away the contracting issues between the service provider and intermediary and focus on a simple setting in which the intermediary charges a separate fee s to each amateur customer for providing the connection service.⁷ For example, Databricks charges users an additional hourly rate of \$0.07 for the standard Data Engineering Light service that connects users to Microsoft Azure clusters. In this case, an amateur customer's net utility of purchasing the service through the intermediary would equal her valuation of the service, reduced by the cost of three items: the price charged by the provider, the fee tolled by the intermediary, and the expected waiting cost.

We model the interaction between the service provider, intermediary, and amateur customers as a three-stage game. First, the provider sets a price p charged to each ama-

teur customer. The intermediary then sets a fee s to serve each amateur customer. Finally, amateur customers decide whether to purchase the service through the intermediary. Our model captures the decentralization between the provider and intermediary in serving amateur customers.

We use backward induction to solve the game. In optimality (of the intermediary's fee s), each amateur customer receives a zero surplus whether she joins or not. Thus, the effective arrival rate of amateur customers satisfies $\lambda_1(s) = \mu - \frac{c}{V-p-s}$. Then, given fixed p , the intermediary optimizes its revenue by selecting fee s ,

$$\max_s s\lambda_1(s) = s \left(\mu - \frac{c}{V-p-s} \right). \quad (3)$$

The one-to-one correspondence between s and λ_1 allows us to convert the intermediary's optimization problem (3) to an equivalent one that optimizes over λ_1 ,

$$\max_{\lambda_1 < \min\{\mu, \Lambda\}} \left(V - \frac{c}{\mu - \lambda_1} - p \right) \lambda_1. \quad (4)$$

Then, anticipating (4), the provider solves

$$\max_p p\lambda_1(p),$$

with $\lambda_1(p)$ solving (4). The following result characterizes the provider's optimal price for such a market.

Theorem 2. *In a market of amateur customers, there exists a threshold $\hat{\lambda}_1 < \hat{\lambda}_0$, where $\hat{\lambda}_0$ is defined in (2), such that the provider's optimal price and the resulting effective arrival rate are*

$$p_1^* = V - \frac{c\mu}{(\mu - \lambda_1^*)^2}, \text{ where } \lambda_1^* = \begin{cases} \Lambda, & \text{if } \Lambda \leq \hat{\lambda}_1, \\ \hat{\lambda}_1, & \text{if } \Lambda > \hat{\lambda}_1. \end{cases}$$

Recall that in a market of expert customers, the provider can fully extract the surplus of joining customers. However, in a market of amateur customers, the provider has to share the reward of serving amateur customers with the intermediary. Thus, the presence of the intermediary cuts into the provider's profit margins. Nevertheless, the provider's optimal market coverage has a similar threshold structure, but this time, the threshold to induce full coverage is lower than the corresponding threshold in a market of expert customers, as congestion is more critical in the former amateur market with decreased profit margins. Thus, full market coverage is optimal in this market only when the market size is even smaller.

We next discuss the welfare implications of the two markets. As no consumer surplus is retained in either market, the social welfare $\Pi(\lambda) = [V - cW(\lambda)]\lambda$ is the sum of the provider's and intermediary's respective revenues. To compare the provider's prices, revenues, and social welfare, let R_0^*

and R_1^* denote the provider's optimal revenues under $\alpha = 0$ and $\alpha = 1$, respectively.

Proposition 1. *We have*

- (i) $p_0^* > p_1^*$, $R_0^* > R_1^*$;
- (ii) $\lambda_0^* \geq \lambda_1^*$ and $\Pi(\lambda_0^*(p_0^*)) \geq \Pi(\lambda_1^*(p_1^*))$, with inequalities being strict when $\Lambda > \hat{\lambda}_1$.

As evinced, the necessity of relying on an intermediary to reach amateur customers puts the provider at risk, creating double marginalization that adversely affects the provider's revenues. Moreover, the same double marginalization reduces market coverage too, leading to a lower utilization and decreased social welfare.

3.3 | Heterogeneous market of expert and amateur customers

In this section, we consider a heterogeneous market mixed with expert and amateur customers. For example, AWS and Microsoft Azure serve both tech-savvy customers who build their own computing infrastructure and amateur customers who hire Databricks to connect them to the cloud. In such markets, the joining decisions of expert and amateur customers create congestion externalities that mutually affect each other. In this section, we focus on the FCFS policy. (We study non-FCFS queueing policies in Section 5.) We next formulate the interaction between two customer segments under FCFS. To this end, we utilize the one-to-one correspondence between the intermediary's fee s and the joining rate of amateur customers λ_A (in a similar spirit to (3) and (4)) and formulate the interaction between the intermediary and expert customers as follows.

Definition 1. Let $\mathbf{p} = (p_A, p_F)$ denote the prices charged to amateur and expert customers, and $(\lambda_A(\mathbf{p}), \lambda_F(\mathbf{p}))$ denote the effective arrival rates of amateur and expert customers. We say that the effective arrival rate pair $(\lambda_A(\mathbf{p}), \lambda_F(\mathbf{p}))$ is a Subgame Perfect Equilibrium (SPE) under FCFS if the following holds:

- (1) (Intermediary's best response) $\lambda_A(\mathbf{p}) \in \arg \max_{0 \leq \lambda < \min\{\mu - \lambda_F(\mathbf{p}), \alpha \Lambda\}} (V - \frac{c}{\mu - \lambda_F(\mathbf{p}) - \lambda} - p_A)\lambda$;
- (2) (Expert customers' best response)
 - (a) if $\mu > (1 - \alpha)\Lambda + \lambda_A(\mathbf{p})$ and $V - \frac{c}{\mu - (1 - \alpha)\Lambda - \lambda_A(\mathbf{p})} - p_F \geq 0$, then $\lambda_F(\mathbf{p}) = (1 - \alpha)\Lambda$;
 - (b) if $\mu > (1 - \alpha)\Lambda + \lambda_A(\mathbf{p})$ and $V - \frac{c}{\mu - (1 - \alpha)\Lambda - \lambda_A(\mathbf{p})} - p_F < 0$ and $V - \frac{c}{\mu - \lambda_A(\mathbf{p})} - p_F > 0$, then $\lambda_F(\mathbf{p})$ is such that $V - \frac{c}{\mu - \lambda_F(\mathbf{p}) - \lambda_A(\mathbf{p})} - p_F = 0$;
 - (c) if $\mu < (1 - \alpha)\Lambda + \lambda_A(\mathbf{p})$ and $V - \frac{c}{\mu - \lambda_A(\mathbf{p})} - p_F \geq 0$, then $\lambda_F(\mathbf{p})$ is such that $V - \frac{c}{\mu - \lambda_F(\mathbf{p}) - \lambda_A(\mathbf{p})} - p_F = 0$.

To understand Definition 1, note that the payoffs of expert and amateur customers from joining the service are

$$\begin{cases} U_F = V - p_F - cW_F(\lambda_A, \lambda_F), \\ U_A = V - p_A - cW_A(\lambda_A, \lambda_F) - s, \end{cases}$$

where $W_F(\lambda_A, \lambda_F)$ and $W_A(\lambda_A, \lambda_F)$ are the expected wait times of expert and amateur customers under effective arrival rates (λ_A, λ_F) . Under FCFS, the expected wait times are

$$W_F(\lambda_A, \lambda_F) = W_A(\lambda_A, \lambda_F) = \begin{cases} \frac{1}{\mu - (\lambda_A + \lambda_F)} & \text{if } 0 \leq \lambda_A + \lambda_F < \mu, \\ \infty & \text{otherwise.} \end{cases}$$

We next explain the best-response formulations of different players in Definition 1. First, given the provider's prices (p_A, p_F) and the effective arrival rate of expert customers $\lambda_F(\mathbf{p})$, the intermediary's best response is solved by optimizing the effective arrival rate of amateur customers, given its one-to-one correspondence with the intermediary's fee s . Second, given the effective arrival rate of amateur customers $\lambda_A(\mathbf{p})$, the joining decisions of expert customers depend on their wait times. In case (a), capacity is ample and wait times are short, so that all expert customers join the service and receive a positive surplus. In cases (b) and (c), capacity is limited and wait times are excessively long if all expert customers join the service, so that only a portion of them can afford to join and receive a zero surplus.

Our next analysis focuses on a simple *single pricing* strategy by enforcing $p_A = p_F = p$. Under this strategy, it is easy to see that the equilibrium in Definition 1 is *incentive compatible*. On the one hand, expert customers will purchase the service directly from the provider because pretending to be amateur customers will incur an additional yet unnecessary payment to the intermediary. (Recall the queueing cost is the same for both types under FCFS.) On the other hand, amateur customers are not technically sophisticated to imitate expert customers and they must purchase the service through the intermediary.

Now, for any price $p \geq 0$, one can show that there exists a unique SPE $(\lambda_A(p), \lambda_F(p))$ satisfying Definition 1. Anticipating this, the provider selects price to maximize revenue:

$$\begin{aligned} \max_p \quad & p[\lambda_A(p) + \lambda_F(p)] \\ \text{s.t.} \quad & (\lambda_A(p), \lambda_F(p)) \text{ satisfies Definition 1.} \end{aligned} \quad (5)$$

To solve the provider's revenue-optimization problem (5), we first make an observation on the structure of SPE under single pricing.

Lemma 1. *Consider single pricing and FCFS policy. For any price p , if $\lambda_A(p) > 0$ under an SPE, then $\lambda_F(p) = (1 - \alpha)\Lambda$.*

The intuition of Lemma 1 can be explained as follows. If $\lambda_A(p) > 0$, namely, a nonnegligible portion of amateur customers purchase the service through the intermediary, then the intermediary should generate a non-negative revenue from serving these customers. This implies the intermediary's fee $s = V - \frac{c}{\mu - \lambda_A(p) - \lambda_F(p)} - p \geq 0$. The inequality must be strict because if otherwise, $V - \frac{c}{\mu - \lambda_A(p) - \lambda_F(p)} - p = 0$, then the intermediary's fee s is zero. In this case, the intermediary can be better off by increasing s and reducing $\lambda_A(p)$ by a small amount. This will generate a positive revenue for the intermediary, contradicting the optimality of $\lambda_A(p)$ in Definition 1. Thus, it must hold that $V - \frac{c}{\mu - \lambda_A(p) - \lambda_F(p)} - p > 0$. As the provider charges a single price to all customers and joining the service incurs the same expected waiting cost for all customers, it follows that expert customers can receive a strictly positive surplus from joining. Hence, all of them will join so that $\lambda_F(p) = (1 - \alpha)\Lambda$.

Lemma 1 suggests that under single pricing, the provider must serve expert customers entirely before serving any amateur customers. It also suggests that the presence of amateur customers allows expert customers to *free ride* and retain a positive surplus that otherwise would be fully extracted in a market without amateur customers. As we will demonstrate shortly, the free-riding of expert customers increases system congestion and adversely affects the efficacy of pricing as a tool to regulate congestion.

Lemma 1 provides a structural property of SPE under single pricing. Using this property, we divide all equilibria into two types based on whether amateur customers are served.

Definition 2. Let p^* be the optimal prices and $(\lambda_A(p^*), \lambda_F(p^*))$ be the corresponding effective arrival rates of amateur and expert customers, respectively.

- (i) The equilibrium $(p^*, \lambda_A(p^*), \lambda_F(p^*))$ is a *Type I Equilibrium* if $\lambda_A(p^*) = 0$.
- (ii) The equilibrium $(p^*, \lambda_A(p^*), \lambda_F(p^*))$ is a *Type II Equilibrium* if $\lambda_A(p^*) > 0$.

By Definition 2, the provider serves expert customers exclusively in a Type I equilibrium and serves both customer types in a Type II equilibrium. Intuitively, a Type I equilibrium cannot be sustained when there are insufficient expert customers. The following result formalizes this intuition.

Theorem 3. Consider single pricing and FCFS.

- (i) A *Type II equilibrium* occurs if and only if

$$\alpha > \underline{\alpha}, \text{ where } \underline{\alpha} > 1 - \mu/\Lambda \text{ satisfies} \tag{6}$$

$$\frac{V}{c} = \frac{\mu + (1 - \underline{\alpha})\Lambda}{[\mu - (1 - \underline{\alpha})\Lambda]^2}.$$

In this case, there exists a threshold $\hat{\alpha}_A \in (0, \mu/\Lambda - (1 - \alpha))$ such that the provider's optimal price and the

corresponding effective arrival rates of amateur and expert customers are given by

$$p^* = V - \frac{c[\mu - (1 - \alpha)\Lambda]}{[\mu - (1 - \alpha)\Lambda - \lambda_A^*]^2},$$

$$\lambda_A^* = \begin{cases} \alpha\Lambda, & \text{if } \alpha \leq \hat{\alpha}_A \\ \hat{\alpha}_A\Lambda, & \text{if } \alpha > \hat{\alpha}_A \end{cases} \text{ and } \lambda_F^* = (1 - \alpha)\Lambda.$$

- (ii) A *Type I equilibrium* occurs when Condition (6) does not hold. In this case, the provider's optimal price and the corresponding effective arrival rate of expert customers are given by

$$p^* = V - \frac{c}{\mu - \lambda_F^*},$$

$$\lambda_F^* = \begin{cases} (1 - \alpha)\Lambda, & \text{if } (1 - \alpha)\Lambda \leq \hat{\lambda}_0, \\ \hat{\lambda}_0, & \text{if } (1 - \alpha)\Lambda > \hat{\lambda}_0, \end{cases}$$

where $\hat{\lambda}_0$ is defined in (2).

Theorem 3 fully characterizes the provider's optimal single price under FCFS. The presence of the intermediary cuts into the provider's margin of serving amateur customers, but it also opens access to a segment that otherwise cannot be reached. Charging a lower price to cover both segments has the potential to expand market coverage, but it will also introduce the adverse free-riding effect. The trade-off between double marginalization, market coverage, and congestion externality forms the crux of Theorem 3.

Theorem 3 identifies the fraction of amateur customers, α , as a critical driver of equilibrium outcomes. Specifically, there exists a threshold $\underline{\alpha}$ that differentiates the provider's coverage strategies. For α below this threshold, the provider serves expert customers exclusively and sets a high price to fully screen out amateur customers. This leads to a Type I equilibrium. For α above this threshold, the demand of expert customers is very slim and it is better to expand coverage and serve both segments. A Type II equilibrium then emerges. In this latter case, there is another threshold, $\hat{\alpha}_A$, that further differentiates the provider's coverage strategy for the amateur segment. Specifically, for $\underline{\alpha} < \alpha < \hat{\alpha}_A$, all amateur customers are served, and thus, the entire market is covered. For $\alpha > \hat{\alpha}_A$, amateur customers are only partially served.

We present a graphical illustration of single pricing in Figure 1, where we normalize both the delay sensitivity c and service rate μ to 1. We vary the service valuation $V \in [1, 13]$, the fraction of amateur customers $\alpha \in [0, 1]$, and market size $\Lambda \in \{1/2, 2\}$. To explain the notations in the figure, we use $\lambda_e := \lambda_A + \lambda_F$ to denote the total effective arrival rates of both types.

Under fixed c and μ , Condition (6) partitions the parameter space (α, V) into two separate regions, with a Type I

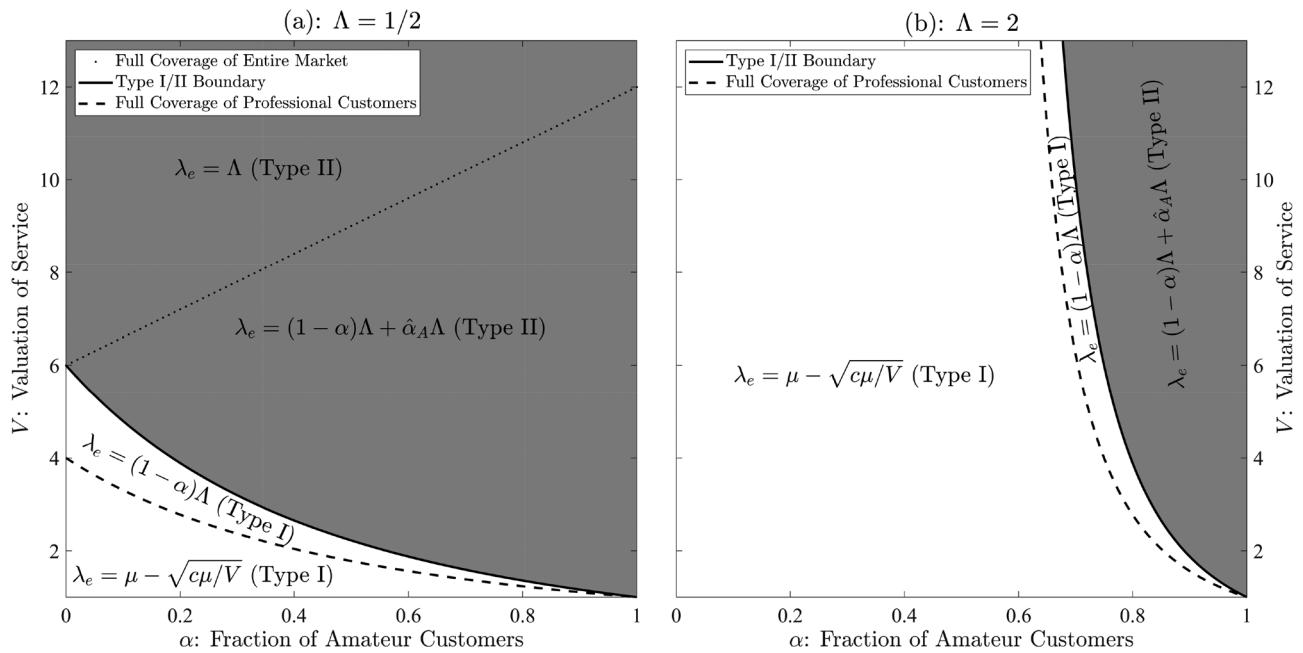


FIGURE 1 Equilibrium under single pricing.

(white area) and Type II equilibrium (dark area) being the dominant equilibrium in each region. The solid curve depicts the switching boundary. Each region is further partitioned into two subregions differentiated by the provider's coverage strategy for certain customer segments. In the Type I region where expert customers are served exclusively, the dashed curve represents the boundary between full and partial coverage of expert customers. In the Type II region where both expert and amateur customers are served, the dotted curve represents the boundary between full and partial coverage of amateur customers.

4 | PRICE DISCRIMINATION

In the previous section, we focused on single pricing and identified the free-riding of expert customers when amateur customers are served. In this section, we consider type-based price discrimination and show that this advanced pricing scheme can fully allay free-riding. To implement type-based price discrimination, it requires identifying the type of each incoming customer. In the context of cloud computing, AWS and other major cloud operators have provided special entries for intermediaries such as Databricks. These entries are generally different from those of direct users who access the cloud without using these intermediaries.⁸ Thus, whether one uses the special portal of Databricks to access the cloud will reveal the type of that customer, and this information can be used for price discrimination.⁹

Under price discrimination, let p_A and p_F denote the prices charged to amateur and expert customers, respectively. All customers are served under the FCFS policy as

before. Similar to the case of single pricing, one can show that there exists a unique SPE satisfying Definition 1 under price discrimination. The resulting SPE under the optimal price discrimination is *incentive compatible*. Specifically, amateur customers have limited technical sophistication to imitate expert customers and expert customers do not have an incentive to mimic amateur customers either.

To explain this latter point, note that the revenue-maximizing intermediary will set s such that amateur customers receive a zero surplus from joining the service. Price discrimination also ensures that expert customers receive a zero surplus if they make a direct purchase from the provider. So, expert customers are indifferent between purchasing the service from the provider and through the intermediary. In this case, we assume that all expert customers stick to their "expert" identities with probability one. There are two reasons to enforce such pure-strategy behaviors. First, if instead, expert customers play a mixed strategy, we can consider a symmetric equilibrium in which all expert customers select each option with the same probability. The switching expert customers will adjust the market composition and allow α to increase. The resulting equilibrium under the new market composition with an increased number of *effective* amateur customers (including real amateur customers who are technically incapable and expert customers who pretend to be amateur) can be derived using Theorem 4 by treating the remaining expert segment as if they were playing a pure strategy. One can show that the provider's optimal revenues are decreasing in the number of effective amateur customers. In this sense, stimulating a direct purchase among all expert customers (see Lemma 2) is indeed optimal for the provider. Second, stimulating direct purchases among expert customers can be achieved by charging p_F slightly less than

the optimal p_F^* in Theorem 4, leaving a tiny but positive surplus to these customers. This will incentivize all expert customers to make a direct purchase and the resulting revenue will be sufficiently close to the provider's optimal revenue given in Theorem 4.

The provider then selects prices $\mathbf{p} = (p_A, p_F)$ to maximize revenue

$$\begin{aligned} \max_{\mathbf{p}} \quad & p_A \lambda_A + p_F \lambda_F \\ \text{s.t.} \quad & (\lambda_A(\mathbf{p}), \lambda_F(\mathbf{p})) \text{ satisfies Definition 1.} \end{aligned} \quad (7)$$

To solve the provider's optimization problem, we provide a structural property of the optimal price discrimination.

Lemma 2. Consider price discrimination and FCFS.

- (i) If $(1 - \alpha)\Lambda < \mu$ and the optimal prices \mathbf{p}^* yield $\lambda_A^* > 0$, then it must hold that $\lambda_F^* = (1 - \alpha)\Lambda$ in optimality.
- (ii) If $(1 - \alpha)\Lambda \geq \mu$, then the optimal prices \mathbf{p}^* must yield $\lambda_A^* = 0$.

Similar to Lemma 1, Lemma 2 suggests that if any amateur customers are served under price discrimination ($\lambda_A^* > 0$), then expert customers should have been served entirely ($\lambda_F^* = (1 - \alpha)\Lambda$). In other words, the provider will serve expert customers exclusively before switching to serve any amateur customers. To understand this result, note that by charging different prices, the provider can fully extract the surplus of expert customers, eliminating free-riding that emerged under single pricing. Serving amateur customers, however, is less profitable as a portion of their surplus must be shared with the intermediary.

Lemma 2 allows us to narrow down the search region of feasible prices in (7). Specifically, we only need to consider prices such that either amateur customers are screened out ($\lambda_A = 0$) or expert customers are served completely ($\lambda_F = (1 - \alpha)\Lambda$). We characterize the optimal prices in the following result.

Theorem 4. Consider price discrimination and FCFS.

- (i) A Type II equilibrium occurs if and only if

$$\alpha > 1 - \hat{\lambda}_0/\Lambda, \quad (8)$$

where $\hat{\lambda}_0$ is defined in (2). In this case, there exists a threshold $\hat{\alpha}_d \in (0, \mu/\Lambda - (1 - \alpha))$ such that the provider's optimal price and the corresponding effective arrival rates of amateur and expert customers are given by

$$\begin{aligned} p_A^* &= V - \frac{c[\mu - (1 - \alpha)\Lambda]}{[\mu - (1 - \alpha)\Lambda - \lambda_A^*]^2} \quad \text{and} \\ p_F^* &= V - \frac{c}{\mu - (1 - \alpha)\Lambda - \lambda_A^*}, \end{aligned}$$

where

$$\lambda_A^* = \begin{cases} \alpha\Lambda, & \text{if } \alpha \leq \hat{\alpha}_d, \\ \hat{\alpha}_d\Lambda, & \text{if } \alpha > \hat{\alpha}_d, \end{cases} \quad \text{and } \lambda_F^* = (1 - \alpha)\Lambda.$$

- (ii) A Type I equilibrium occurs when Condition (8) does not hold. The provider's optimal price and the corresponding effective arrival rate of expert customers are given by

$$p_F^* = V - \frac{c}{\mu - \lambda_F^*}, \quad p_A^* \geq p_F^*,$$

$$\lambda_F^* = \hat{\lambda}_0, \quad \text{where } \hat{\lambda}_0 \text{ is defined in (2).}$$

Some observations are in order. First, we can rewrite (8) as $(1 - \alpha)\Lambda < \hat{\lambda}_0$, that is, the size of expert customers should not exceed $\hat{\lambda}_0$. To explain this condition, recall that in a market of expert customers only, the effective arrival rate of expert customers is capped at $\hat{\lambda}_0$ (cf. Theorem 1). Recall also that in a market with both expert and amateur customers, the provider favors expert customers for their high margins under price discrimination. So, if the demand of expert customers exceeds $\hat{\lambda}_0$, it is optimal to serve these customers up to the threshold $\hat{\lambda}_0$, forgo the excess demand and screen out remaining amateur customers.

Thus, a Type II equilibrium can only emerge when the fraction of amateur customers is above $(1 - \hat{\lambda}_0/\Lambda)$. Recall that a Type II equilibrium emerges under single pricing when α exceeds $\underline{\alpha}$ (cf. Theorem 3). It then follows that, when the fraction of amateur customers falls in a range, that is, between $(1 - \hat{\lambda}_0/\Lambda)$ and $\underline{\alpha}$, the provider serves both segments under price discrimination but only the expert segment under single pricing. Hence, a Type II equilibrium emerges in a wider range of parameter space under price discrimination. Therefore, price discrimination reduces free-riding and creates efficient market coverage.

In Figure 2, we give a graphical illustration of the equilibrium outcomes under price discrimination. Using the same parameters as those under single pricing, we compute the new equilibria under price discrimination. The solid and dashed curves represent the switching boundaries between Type I and Type II regimes under single pricing and price discrimination, respectively. We find that the Type II regime expands as the provider switches from single pricing to price discrimination. Moreover, under a small market size ($\Lambda = 1/2$ in Panel a), price discrimination increases the chance of full market coverage.

Not surprisingly, price discrimination can improve the provider's revenue by pulling more pricing levers. However, prices may also drop under price discrimination. Specifically, we find cases in which a provider who initially charges a single price chooses to lower prices for both segments as he switches to type-based price discrimination.

Proposition 2 (Single pricing vs. price discrimination).

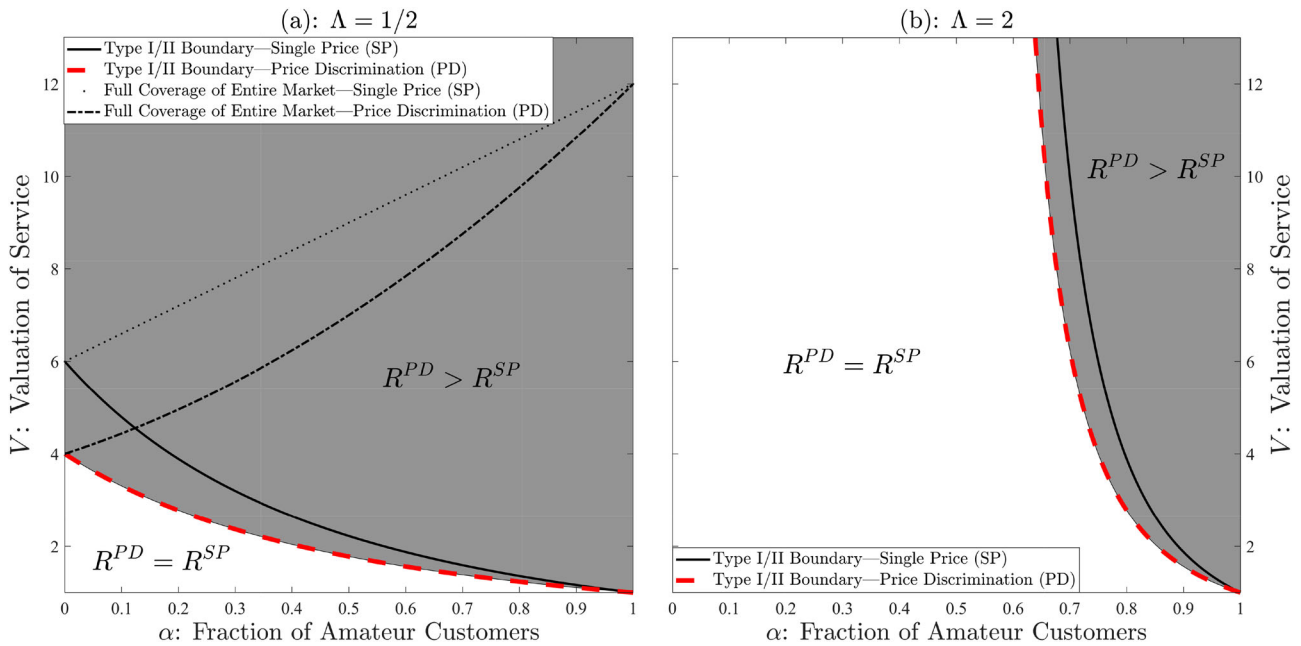


FIGURE 2 Equilibrium comparison between single pricing and price discrimination. [Color figure can be viewed at wileyonlinelibrary.com]

- (i) The provider's revenue under price discrimination is strictly higher than that under single pricing when $\alpha > 1 - \hat{\lambda}_0/\Lambda$. Otherwise, they are equal.
- (ii) Under price discrimination, the provider charges a higher price to expert customers than amateur customers, $p_F^* > p_A^*$. Compared with the optimal price p^* under single pricing, if α satisfies $\max\{\underline{\alpha}, \hat{\alpha}_A\} < \alpha \leq \hat{\alpha}_d$, then there exists $\tilde{\alpha} < \alpha$ such that

$$\begin{cases} p^* > p_F^* > p_A^*, & \text{if } \hat{\alpha}_A < \tilde{\alpha}, \\ p_F^* > p^* > p_A^*, & \text{if } \hat{\alpha}_A > \tilde{\alpha}, \end{cases}$$

where $\underline{\alpha}$ and $\hat{\alpha}_A$ are defined in Theorem 3(i) and $\hat{\alpha}_d$ is defined in Theorem 4(i).

To explain Proposition 2, note that when the demand of expert customers exceeds $\hat{\lambda}_0$ (the white region in Figure 2), the provider will serve expert customers exclusively up to $\hat{\lambda}_0$ under both pricing schemes, effectively reducing price discrimination to single pricing. Thus, price discrimination can only outperform single pricing in a Type-II equilibrium (the dark-shaded region in Figure 2). In this case, the price charged to amateur customers is lower than that charged to expert customers, $p_A^* < p_F^*$. This is due to the fact that the reward of serving amateur customers must be shared with the intermediary, whereas the reward of serving expert customers can be extracted exclusively.

However, it is intriguing that in some circumstances, price discrimination drives prices downward for both expert and amateur customers compared to the optimal single price. In this case, customers, irrespective of their types, can find a better price under price discrimination. To understand this

result, recall that $\hat{\alpha}_A$ and $\hat{\alpha}_d$ are the thresholds that differentiate the provider's coverage strategy for the amateur segment. Specifically, when the fraction of amateur customers is below these thresholds, the provider will serve amateur customers entirely and thus fully cover the market. Now, under single pricing, as the price charged to both types is identical, expert and amateur customers are pooled together and contribute to a common demand, which is too large to be served all especially when $\hat{\alpha}_A$ is small ($\hat{\alpha}_A < \tilde{\alpha}$), partially due to the free-riding effect of expert customers. This prompts the provider to charge a higher single price to regulate system congestion. In contrast, under price discrimination, each type is priced separately and their demands are not very tightly pooled. The provider can focus on improving the aggregate throughput by charging a lower price to each type ($p^* > p_F^* > p_A^*$) without causing excessive congestion.

Indeed, we find that over time, highly profitable cloud services such as AWS and Microsoft Azure have all expanded in scale and market reach, but only to see their profit margins drop after expansion. Our results provide a tentative explanation for this trend.

5 | PRICING AND PRIORITIZATION

In previous sections, we focused on analyzing the provider's optimal pricing strategy under FCFS. In this section, we propose another operational instrument to manage service systems: resource reallocation using *priorities*. For example, in the context of cloud computing, low-priority jobs can be interrupted and preempted by high-priority ones (Costa et al., 2018).¹⁰ Motivated by this observation, we consider a revenue-maximizing provider that jointly optimizes over

prices and priorities. (We allow type-based differentiation in both prices and priorities.)

Note that priorities are *relative* treatments. Prioritizing one segment will reduce the wait times of that segment, but will also increase the wait times of other segments de-prioritized. We thus analyze two scenarios, each giving *preemptive* priorities to a specific customer segment. In both scenarios, it is incentive compatible for expert and amateur customers to truthfully reveal their types. Amateur customers have limited technical sophistication and must work with an intermediary to access the service. Expert customers are indifferent between purchasing the service directly from the provider and pretending to be amateur (each yielding a zero surplus under price discrimination). As done in Section 4, we assume that all expert customers purchase directly from the provider with probability one.

When preemptive prioritization is allowed, the expected wait times of the high-priority class only depend on the joining rate of that class, whereas the expected wait times of the low-priority class depend on the joining rates of both classes. In a two-class queueing system that prioritizes class i over class j , the expected wait times of each class are given by¹¹

$$W_i(\lambda) = \frac{1}{\mu - \lambda_i}, \quad W_j(\lambda) = \frac{\mu}{(\mu - \lambda_i)(\mu - \lambda_i - \lambda_j)},$$

where $\lambda = (\lambda_i, \lambda_j)$.

5.1 | Prioritizing expert customers

We established in previous sections that the provider prefers to serve expert customers for their high margins under price discrimination. This intuition extends to the current setting with priorities. We first consider prioritizing expert customers and formulate the provider's revenue-maximization problem as follows:

$$\begin{aligned} & \max_{p_A, p_F} p_A \lambda_A + p_F \lambda_F \\ \text{s.t.} \quad & V - \frac{c}{\mu - \lambda_F} - p_F = 0, \\ & 0 \leq \lambda_F \leq (1 - \alpha)\Lambda, \\ & \lambda_A = \arg \max_{0 \leq \lambda < \min\{\mu - \lambda_F, \alpha\Lambda\}} \left(V - \frac{c\mu}{(\mu - \lambda_F)(\mu - \lambda_F - \lambda)} - p_A \right) \lambda. \end{aligned} \tag{9}$$

We make an observation on the structure of the optimal prices.

Lemma 3. *Consider price discrimination and prioritizing expert customers. If $(1 - \alpha)\Lambda < \mu$ and the optimal prices \mathbf{p}^* yield $\lambda_A^* > 0$, then it must hold that $\lambda_F^* = (1 - \alpha)\Lambda$. If $(1 - \alpha)\Lambda \geq \mu$, then the optimal prices must yield $\lambda_A^* = 0$.*

In the context of Lemmas 1 and 2 with FCFS being the queueing discipline, the provider will serve expert customers entirely before switching to serving any amateur customers. The same structure is preserved when expert customers are prioritized, extending the intuition from Lemma 2. Specifically, price discrimination allows the provider to fully extract the surplus of expert customers, whereas the rewards of serving amateur customers must be shared with the intermediary.

Theorem 5. *Consider price discrimination and prioritizing expert customers.*

(i) *A Type II equilibrium occurs if and only if*

$$\alpha > 1 - \hat{\lambda}_0/\Lambda, \tag{10}$$

where $\hat{\lambda}_0$ is defined in (2). In this case, there exists a threshold $\hat{\alpha}_d^{F-Pri} \in (0, \mu/\Lambda - (1 - \alpha))$ such that the provider's optimal prices and the corresponding effective arrival rates of amateur and expert customers are given by

$$\begin{aligned} p_A^* &= V - \frac{c\mu}{[\mu - (1 - \alpha)\Lambda - \lambda_A^*]^2} \quad \text{and} \\ p_F^* &= V - \frac{c}{\mu - (1 - \alpha)\Lambda}, \end{aligned}$$

where

$$\lambda_A^* = \begin{cases} \alpha\Lambda, & \text{if } \alpha \leq \hat{\alpha}_d^{F-Pri}, \\ \hat{\alpha}_d^{F-Pri}\Lambda, & \text{if } \alpha > \hat{\alpha}_d^{F-Pri}, \end{cases} \quad \text{and } \lambda_F^* = (1 - \alpha)\Lambda.$$

(ii) *A Type I equilibrium occurs when Condition (10) does not hold. In this case, the provider's optimal prices and the corresponding effective arrival rate of expert customers are given by*

$$\begin{aligned} p_F^* &= V - \frac{c}{\mu - \lambda_F^*}, \quad p_A^* \geq p_F^*, \\ \lambda_F^* &= \hat{\lambda}_0, \quad \text{where } \hat{\lambda}_0 \text{ is defined in (2).} \end{aligned}$$

The equilibrium outcomes in Theorem 5 when expert customers are prioritized exhibit a similar threshold structure as those under FCFS. The threshold also matches that in Theorem 4. The optimal prices, however, depend on the queueing policy. The subtlety in prices under different queueing policies stems from the fact that FCFS places an endogenous waiting cost on each segment determined by decisions of both classes, whereas in priority queues, the joining decisions of the high-priority class are determined solely within that class independently from the decisions of the low-priority class.

5.2 | Prioritizing amateur customers

We next consider prioritizing amateur customers and formulate the provider's revenue-optimization problem as follows:

$$\begin{aligned} & \max_{p_A, p_F} p_A \lambda_A + p_F \lambda_F \\ & \text{s.t. } \lambda_A \in \arg \max_{0 \leq \lambda < \min\{\mu, \alpha \Lambda\}} \left(V - \frac{c}{\mu - \lambda} - p_A \right) \lambda, \\ & \quad V - p_F - \frac{c\mu}{(\mu - \lambda_A)(\mu - \lambda_A - \lambda_F)} = 0, \\ & \quad 0 \leq \lambda_F \leq (1 - \alpha)\Lambda. \end{aligned} \quad (11)$$

We characterize the optimal prices in the following result.

Theorem 6. Consider price discrimination and prioritizing amateur customers.

(i) A Type II equilibrium occurs if and only if

$$\alpha > 1 - \hat{\lambda}_0 / \Lambda, \quad (12)$$

where $\hat{\lambda}_0$ is defined in (2). In this case, there exists a threshold $\hat{\alpha}_d^{A-Pri} \in (0, \mu/\Lambda - (1 - \alpha))$ such that the provider's optimal prices and the corresponding effective arrival rates of amateur and expert customers are given by

$$\begin{aligned} p_A^* &= V - \frac{c\mu}{(\mu - \lambda_A^*)^2} \quad \text{and} \\ p_F^* &= V - \frac{c\mu}{(\mu - \lambda_A^*)[\mu - (1 - \alpha)\Lambda - \lambda_A^*]}, \end{aligned}$$

where

$$\lambda_A^* = \begin{cases} \alpha\Lambda, & \text{if } \alpha \leq \hat{\alpha}_d^{A-Pri}, \\ \hat{\alpha}_d^{A-Pri}\Lambda, & \text{if } \alpha > \hat{\alpha}_d^{A-Pri}, \end{cases} \quad \text{and } \lambda_F^* = (1 - \alpha)\Lambda.$$

(ii) A Type I equilibrium occurs when Condition (12) does not hold. In this case, the provider's optimal prices and the corresponding effective arrival rate of expert customers are given by

$$p_F^* = V - \frac{c}{\mu - \lambda_F^*}, \quad p_A^* \geq p_F^*,$$

$$\lambda_F^* = \hat{\lambda}_0, \quad \text{where } \hat{\lambda}_0 \text{ is defined in (2).}$$

Theorem 6 shows that the equilibrium outcomes when amateur customers are prioritized exhibit a similar threshold structure as those under other queueing policies. Specifically, the switching boundaries between Type I and Type II regimes

are identical in priority queues irrespective of which segment is prioritized. This result is because different priority policies are considered jointly with price optimization. Thus, the thresholds to induce a Type I equilibrium are identical in Theorems 4, 5, and 6. The provider's revenues are the same in a Type I equilibrium irrespective of the queueing policy because amateur customers are screened out and expert customers are served exclusively. However, when both segments are served in a Type II equilibrium, the provider's revenues will heavily depend on the queueing policy, as reallocating wait times across segments will readjust the profit margins of each segment. We address this issue in the next section.

5.3 | Comparison between priority and FCFS

We next examine the provider's priority preference by comparing the provider's optimal revenues under three queueing policies (FCFS, prioritizing expert customers, and prioritizing amateur customers). As serving expert customers entails a higher profit margin, one may expect that prioritizing these customers can improve the revenue by reducing their wait times and creating even higher profit margins. However, contrary to this expectation, we show in the next result that prioritizing expert customers generates the lowest revenue among all three queueing policies.

Formally, let R^{F-Pri} , R^{A-Pri} , and R^{FCFS} denote the provider's optimal revenues by prioritizing expert customers, prioritizing amateur customers, and serving all customers under FCFS, respectively.

Proposition 3 (Optimal pricing and prioritization). Given $\hat{\lambda}_0$ defined in (2), we have

- (i) $R^{A-Pri} > R^{FCFS} > R^{F-Pri}$ when $\alpha > 1 - \hat{\lambda}_0 / \Lambda$; $R^{A-Pri} = R^{FCFS} = R^{F-Pri}$ when $\alpha \leq 1 - \hat{\lambda}_0 / \Lambda$;
- (ii) $p_F^* > p_A^*$ when the provider prioritizes expert customers and $p_A^* > p_F^*$ when the provider prioritizes amateur customers.

Recall that the provider in principle prefers to serve expert customers under price discrimination. Thus, when there is a sufficient base of expert customers, the provider will serve them exclusively up to $\hat{\lambda}_0$.

This implies that reallocating wait times through prioritization is only relevant in a Type II equilibrium, which emerges when the demand of expert customers is not too high, $\Lambda(1 - \alpha) < \hat{\lambda}_0$. Proposition 3 further shows that in this equilibrium, prioritization does not favor expert customers even though price discrimination does.

To explain such discrepancy, recall that prioritization is optimized jointly with price discrimination. Prioritizing expert customers, while creating higher profit margins of expert customers, increases the waiting cost of amateur customers. The effect of the increased waiting cost is further amplified by a revenue-maximizing intermediary that adjusts

fees in response to the de-prioritization of its customers. This exacerbates double marginalization and prompts amateur customers to join the service at a rate significantly lower than that under FCFS. The latter effect can dominate the increased margins of expert customers, leading to inferior performance of this queueing policy.

In a similar spirit, there can be surprising revenue benefits by prioritizing amateur customers. Such benefits are driven by the wait-time reduction of amateur customers which effectively alleviates double marginalization and boosts the joining rate of amateur customers. The provider then can charge a higher price to the amateur segment ($p_A^* > p_F^*$) despite an adjusted fee tolled by the intermediary. Of course, prioritizing amateur customers poses a higher waiting cost on expert customers and reduces their profit margins. The joint outcome of these effects is that the revenue accrued from the newly served amateur customers outweighs the reduced margins of expert customers, allowing this policy to achieve the best revenue among all three queueing policies.

The above findings hint at a unique perspective of double marginalization rooted in the decentralization in serving amateur customers. They also demonstrate how the provider can use prioritization to manage double marginalization for better revenues. It is significant that in the current setting, the intermediary's fees are *endogenous* and will change with the provider's priority policy. To demonstrate the implications of endogenous fees, we will consider in Subsection 6.2 an extension with the intermediary's fee being *exogenous*. We will show in that extension that the provider can no longer use prioritization to improve revenue. Thus, the intermediary's adaptivity to the provider's queueing policy is critical to the comparison result in Proposition 3.

We give a numerical comparison of three queueing policies in Figure 3 for various values of V and Λ . As before, we normalize the delay sensitivity c and service rate μ to 1. We plot the provider's revenues and social welfare (under the provider's optimal prices) in the upper and lower rows of Figure 3. In all cases, the vertical axis denotes the *percentage gain or loss* of implementing prioritization relative to FCFS. We find that prioritizing amateur customers (solid lines) increases the provider's revenue and social welfare, whereas prioritizing expert customers (dashed lines) decreases the provider's revenue and social welfare. In all cases, the effect of prioritization is most pronounced when the fraction of amateur customers is intermediate.

Revenue comparison

First, when full market coverage is optimal ($\Lambda = 1/2$, corresponding to the two leftmost boxplots), the provider's optimal revenues under three queueing policies can vary significantly when the valuation V is low. In this case, because the market is fully covered, the law of work conservation implies that the total waiting cost is constant irrespective of the queueing pol-

icy. When V is low, the waiting cost of each segment becomes a critical factor of the segment's profit margin. In this case, reallocating wait times across segments through prioritization can lead to significant revenue improvement.

Second, when full market coverage is not feasible under a large market size ($\Lambda = 2$, corresponding to the two rightmost boxplots), the provider's revenues under three queueing policies can vary significantly when both the valuation V and fraction of amateur customers α are high. This is because, a large α implies a low demand of expert customers and this leads to a Type II equilibrium. A higher V further implies higher total joining rates, leading to considerable system congestion. In this case, reallocating wait times through prioritization is effective in regulating congestion and can lead to significant revenue benefits.

Social welfare comparison

The social welfare is computed by summing up the provider's and intermediary's respective revenues, as no consumer surplus is retained under price discrimination. We find that prioritization has a similar effect on social welfare as it does on the provider's revenues. Specifically, prioritizing amateur customers increases social welfare, whereas prioritizing expert customers decreases social welfare. The driving mechanism of social welfare, however, is very different from the revenue metric.

Welfare gains from prioritizing amateur customers emerge from the fact that prioritizing amateur customers allows this segment to join the service at a significantly higher rate. This brings the total joining rates closer to the socially optimal level. So, welfare improvement from prioritization is a result of *better resource utilization*. Clearly, such benefits will only materialize when both types are served in equilibrium and this can only occur when the demand of expert customers is not too high.

Panel f presents a scenario with a high valuation V and small market size Λ . In this scenario, the market is fully covered under all three queueing policies for any market composition $\alpha \in [0, 1]$. Panel f shows that the provider's optimal revenues can vary significantly across three queueing policies, whereas social welfare is a constant irrespective of the queueing policy. To explain the latter result, notice that an alternative way of computing social welfare is to subtract the valuations of joining customers by their waiting costs. As the market is fully covered, the law of work conservation implies that the total waiting cost does not depend on the queueing policy, so that social welfare is insensitive to the queueing policy too.

6 | EXTENSION

In this section, we consider two extensions to our main model. Specifically, we consider a positive marginal cost and an exogenous fee for the intermediary in Subsections 6.1 and

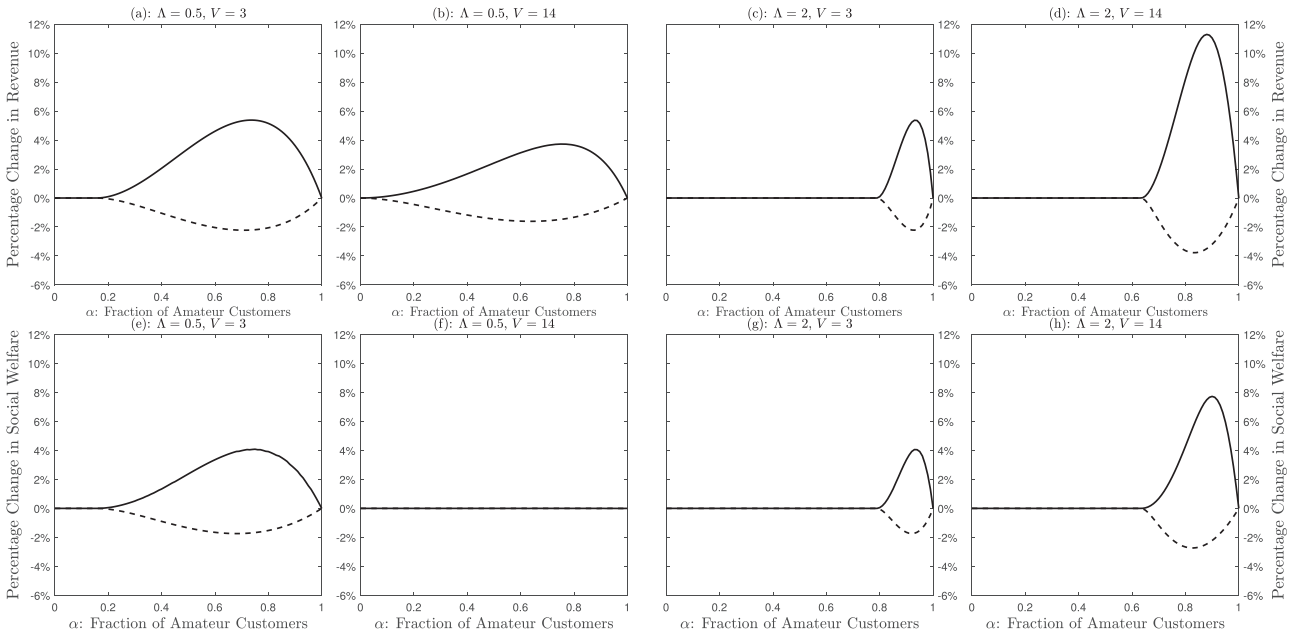


FIGURE 3 Percentage change in revenue and social welfare of prioritizing amateur customers (solid line) and prioritizing expert customers (dashed line) relative to First-Come-First-Served.

6.2, respectively. In each extension, we relax one assumption in the main model while keeping all other assumptions fixed.

6.1 | Marginal cost of intermediary

In the main model, we assumed that the intermediary incurs a zero marginal cost in serving amateur customers. This may be a reasonable assumption for cloud intermediaries with fixed facility costs (e.g., developing a platform and configuring connection parameters). Nevertheless, we acknowledge that in reality, acquiring new customers can be costly (e.g., marketing and advertising). We now analyze a model in which the intermediary incurs a marginal cost h in acquiring a new amateur customer.

To incorporate this cost h into our analysis, we revise the intermediary's best response in Definition 1 to

$$\lambda_A(\mathbf{p}) \in \arg \max_{0 \leq \lambda < \min\{\mu - \lambda_F(\mathbf{p}), \alpha\Lambda\}} \left(V - \frac{c}{\mu - \lambda_F(\mathbf{p}) - \lambda} - p_A - h \right) \lambda$$

when solving the SPE under FCFS. We next present the result of price discrimination. (For brevity, we leave the result of single pricing to Appendix A in the E-Companion.)

Theorem 4'. Consider price discrimination and FCFS.

- (1) If $h \geq [V - \frac{c}{\mu - (1-\alpha)\Lambda}]^+$, then only a Type I equilibrium can occur. The provider's optimal prices and the corresponding effective arrival rates of amateur and expert

customers are given by

$$p_A^* \geq p_F^* = V - \frac{c}{\mu - \lambda_F^*}, \lambda_A^* = 0 \text{ and}$$

$$\lambda_F^* = \begin{cases} (1-\alpha)\Lambda, & \text{if } (1-\alpha)\Lambda \leq \mu - \sqrt{c\mu/V}, \\ \mu - \sqrt{c\mu/V}, & \text{if } (1-\alpha)\Lambda > \mu - \sqrt{c\mu/V}. \end{cases} \quad (13)$$

- (2) If $0 \leq h < [V - \frac{c}{\mu - (1-\alpha)\Lambda}]^+$, then the following holds.

- (i) A Type II equilibrium occurs if and only if

$$\alpha > 1 - \hat{\lambda}_h/\Lambda, \quad (14)$$

where $\hat{\lambda}_h = \mu - \sqrt{\frac{c\mu}{V-h}}$. Further, define $x^* := \hat{x} \mathbb{1}_{\{\mu \leq \Lambda\}} + \min\{\hat{x}, \bar{x}\} \mathbb{1}_{\{\mu > \Lambda\}}$, where $\bar{x} := \frac{\sqrt{c[\mu - (1-\alpha)\Lambda]}}{\mu - \Lambda}$ and \hat{x} is the unique solution to

$$\frac{V-h}{x^2} - 2x\sqrt{\frac{\mu - (1-\alpha)\Lambda}{c}} + \frac{\mu - 2(1-\alpha)\Lambda}{\mu - (1-\alpha)\Lambda} = 0$$

$$\text{for } x \in \left(\sqrt{\frac{c}{\mu - (1-\alpha)\Lambda}}, \infty \right).$$

Then, the provider's optimal prices and the corresponding effective arrival rates of amateur and

expert customers are given by

$$\begin{cases} p_A^* = V - h - x^{*2}, & \lambda_A^* = \mu - (1 - \alpha)\Lambda \\ & - \frac{\sqrt{c[\mu - (1 - \alpha)\Lambda]}}{x^{*2}}, \\ p_F^* = V - \frac{c}{\mu - (1 - \alpha)\Lambda - \lambda_A^*}, & \lambda_F^* = (1 - \alpha)\Lambda. \end{cases}$$

(ii) A Type I equilibrium occurs when Condition (14) does not hold. The provider’s optimal prices and effective arrival rates of amateur and expert customers are given by (13).

Intuitively, when the intermediary’s marginal cost is prohibitively high, the intermediary will stop serving amateur customers, closing the provider’s access to the amateur segment and making the market effectively composed of expert customers only. So, the amateur segment is only relevant when the intermediary’s marginal cost is not too high. In this case, the positive marginal cost increases the threshold of α in (14) (relative to the case of zero marginal cost in (8)). In other words, under a positive marginal cost, it requires a higher fraction of amateur customers in the market for a Type II equilibrium to sustain. The marginal cost raises the intermediary’s fee, which intensifies double marginalization and reduces the provider’s willingness to serve amateur customers. Thus, a Type II equilibrium will not emerge unless the demand of expert customers is even smaller than that required in the case of zero marginal cost.

We compare the provider’s optimal prices under single pricing and price discrimination and find that $p^* > p_F^*$ can only hold when h is sufficiently small. As noted, the intermediary’s marginal cost h intensifies double marginalization. Thus, the provider finds it harder to efficiently target both segments without generating excessive congestion. As a result, when h is large, the provider has to set $p_F^* > p^*$ under price discrimination in order to regulate system congestion.

We next consider the provider’s joint optimization of pricing and prioritization. When expert customers are prioritized, we revise the intermediary’s best response in (9) to

$$\lambda_A = \arg \max_{0 \leq \lambda < \min\{\mu - \lambda_F, \alpha\Lambda\}} \left(V - \frac{c\mu}{(\mu - \lambda_F)(\mu - \lambda_F - \lambda)} - p_A - h \right) \lambda.$$

We are able to fully characterize the provider’s optimal prices. Not surprisingly, when amateur customers are de-prioritized, the marginal cost must be even lower for the intermediary to serve any amateur customers than under FCFS.

Theorem 5’. Consider price discrimination and prioritizing expert customers.

(1) If $h \geq [V - \frac{c\mu}{[\mu - (1 - \alpha)\Lambda]^2}]^+$, then only a Type I equilibrium can occur. The provider’s optimal prices and the corresponding effective arrival rates of amateur and expert customers are given by

$$p_A^* \geq p_F^* = V - \frac{c}{\mu - \lambda_F^*}, \lambda_A^* = 0 \text{ and}$$

$$\lambda_F^* = \begin{cases} (1 - \alpha)\Lambda, & \text{if } (1 - \alpha)\Lambda \leq \mu - \sqrt{c\mu/V}, \\ \mu - \sqrt{c\mu/V}, & \text{if } (1 - \alpha)\Lambda > \mu - \sqrt{c\mu/V}. \end{cases} \tag{15}$$

(2) If $0 \leq h < [V - \frac{c\mu}{[\mu - (1 - \alpha)\Lambda]^2}]^+$, then the following holds.

(i) A Type II equilibrium occurs if and only if

$$\alpha > 1 - \hat{\lambda}_h/\Lambda, \tag{16}$$

where $\hat{\lambda}_h = \mu - \sqrt{\frac{c\mu}{V-h}}$. Further, define $x^* := \hat{x} \mathbb{1}_{\{\mu \leq \Lambda\}} + \min\{\hat{x}, \bar{x}\} \mathbb{1}_{\{\mu > \Lambda\}}$, where $\bar{x} := \frac{\sqrt{c\mu}}{\mu - \Lambda}$ and \hat{x} is the unique solution to

$$\frac{(V - h)\sqrt{c\mu}}{x^2} - 2x[\mu - (1 - \alpha)\Lambda] + \sqrt{c\mu} = 0,$$

for $x \in \left(\frac{\sqrt{c\mu}}{\mu - (1 - \alpha)\Lambda}, \infty \right)$.

The provider’s optimal prices and the corresponding effective arrival rates of amateur and expert customers are given by

$$p_A^* = V - h - x^{*2}, \lambda_A^* = \mu - (1 - \alpha)\Lambda - \frac{\sqrt{c\mu}}{x^*},$$

$$p_F^* = V - \frac{c}{\mu - (1 - \alpha)\Lambda}, \lambda_F^* = (1 - \alpha)\Lambda.$$

(ii) A Type I equilibrium occurs when Condition (16) does not hold. The provider’s optimal prices and effective arrival rates of amateur and expert customers are given by (15).

We are unable to characterize the provider’s optimal prices when amateur customers are prioritized. We numerically compute the optimal prices under this prioritization scheme and compare the resulting revenue and social welfare with those under other queueing policies. Figure 4 presents the comparison results for various values of $h \in \{0.05, 0.5, 1, 2\}$ under $\Lambda = 1/2$ and $V = 3$, which are largely in line with those in our main model with zero marginal cost.

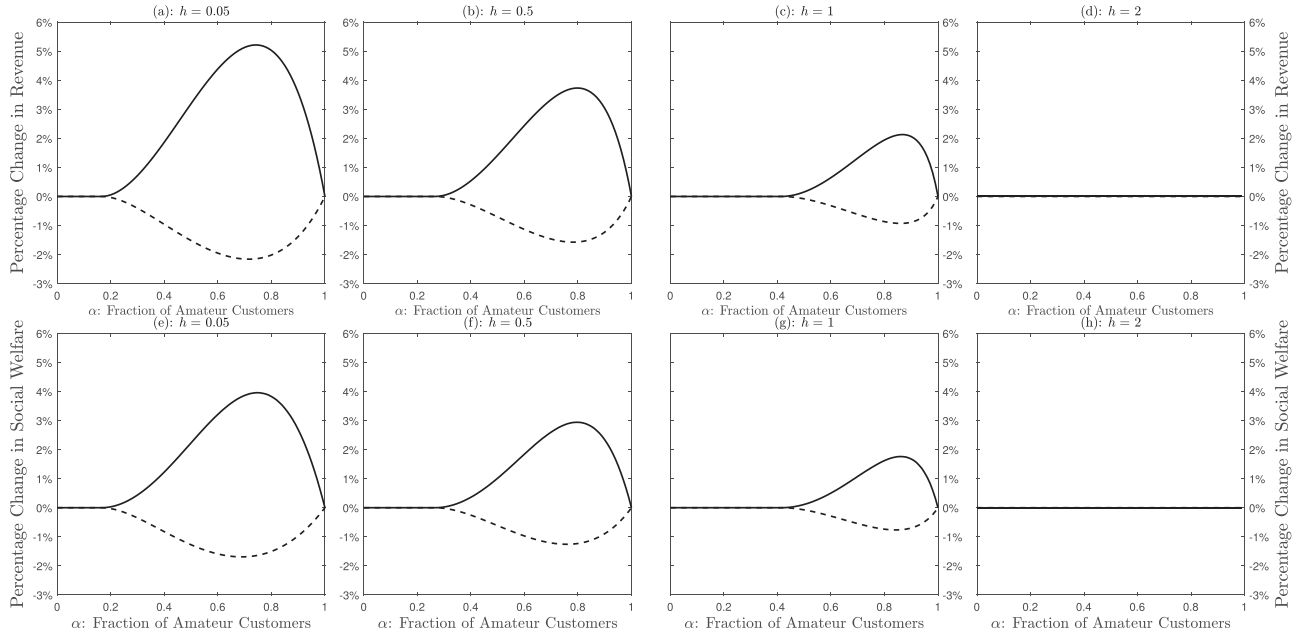


FIGURE 4 Percentage change in revenue and social welfare of prioritizing amateur customers (solid line) and prioritizing expert customers (dashed line) relative to First-Come-First-Served: $\Lambda = 1/2$, $V = 3$.

Specifically, prioritizing amateur customers (solid lines) increases the provider's revenue and social welfare, whereas prioritizing expert customers (dashed lines) decreases the provider's revenue and social welfare. Moreover, the effect of prioritization is most pronounced when the fraction of amateur customers is intermediately high. Hence, the insights in our main model extend to intermediaries with different levels of cost parameters.

6.2 | Exogenous fee of intermediary

In our main model, we considered a monopoly intermediary that optimally selects its fee. Under this assumption, we demonstrated the provider's opposing preferences of customer segments in implementing price and priority discrimination. To better illustrate double marginalization as the critical driver, in this extension we consider a price-taking intermediary that charges an *exogenous* fee s to amateur customers. In general, an exogenous fee could be a joint outcome of numerous factors such as competition, regulations, and business norms. We abstract away these details and focus on analyzing how the intermediary's exogenous fee will affect the provider's optimal pricing and prioritization strategies.

At a high level, the exogenous fee of the intermediary cuts into the provider's profit margin in serving amateur customers. This is equivalent to reducing the valuations of amateur customers by s , creating a market effectively composed of two customer segments with the same delay sensitivity but different valuations. Thus, the joining behaviors of the two segments are very similar to those in the literature on multi-class customers (e.g., Hassin & Haviv, 2006). However, customers in our model share *the same delay sensitivity*

irrespective of their types. This feature has important implications for the provider's priority preference and allows us to generate new results different from those in the main model. We elaborate below.

We first characterize customers' joining decisions under FCFS. The intermediary cannot control the joining rate of amateur customers under an exogenous fee s ; instead, this rate is endogenized by the mutual decisions of both segments. Accordingly, we revise Definition 1 to characterize the new equilibrium.

Definition 3 (Exogenous fee of intermediary). Let $\mathbf{p} = (p_A, p_F)$ denote the prices charged to amateur and expert customers, and $(\lambda_A(\mathbf{p}), \lambda_F(\mathbf{p}))$ denote the effective arrival rates of amateur and expert customers. We say that, the effective arrival rate pair $(\lambda_A(\mathbf{p}), \lambda_F(\mathbf{p}))$ is an SPE under FCFS if the following are satisfied:

- (1) (Amateur customers' best response)
 - (a) if $\mu > \alpha\Lambda + \lambda_F(\mathbf{p})$ and $V - \frac{c}{\mu - \alpha\Lambda - \lambda_F(\mathbf{p})} - p_A - s > 0$, then $\lambda_A(\mathbf{p}) = \alpha\Lambda$;
 - (b) if $\mu > \alpha\Lambda + \lambda_F(\mathbf{p})$ and $V - \frac{c}{\mu - \alpha\Lambda - \lambda_F(\mathbf{p})} - p_A - s \leq 0$ and $V - \frac{c}{\mu - \lambda_A(\mathbf{p})} - p_A - s \geq 0$, then $\lambda_A(\mathbf{p})$ is such that $V - \frac{c}{\mu - \lambda_A(\mathbf{p})} - p_A - s = 0$;
 - (c) if $\mu < \alpha\Lambda + \lambda_F(\mathbf{p})$ and $V - \frac{c}{\mu - \lambda_F(\mathbf{p})} - p_A - s \geq 0$, then $\lambda_A(\mathbf{p})$ is such that $V - \frac{c}{\mu - \lambda_A(\mathbf{p}) - \lambda_F(\mathbf{p})} - p_A - s = 0$.
- (2) (Expert customers' best response)
 - (a) if $\mu > (1 - \alpha)\Lambda + \lambda_A(\mathbf{p})$ and $V - \frac{c}{\mu - (1 - \alpha)\Lambda - \lambda_A(\mathbf{p})} - p_F > 0$, then $\lambda_F(\mathbf{p}) = (1 - \alpha)\Lambda$;

- (b) if $\mu > (1 - \alpha)\Lambda + \lambda_A(\mathbf{p})$ and $V - \frac{c}{\mu - (1 - \alpha)\Lambda - \lambda_A(\mathbf{p})} - p_F \leq 0$ and $V - \frac{c}{\mu - \lambda_A(\mathbf{p})} - p_F \geq 0$, then $\lambda_F(\mathbf{p})$ is such that $V - \frac{c}{\mu - \lambda_F(\mathbf{p}) - \lambda_A(\mathbf{p})} - p_F = 0$;
- (c) if $\mu < (1 - \alpha)\Lambda + \lambda_A(\mathbf{p})$ and $V - \frac{c}{\mu - \lambda_A(\mathbf{p})} - p_F \geq 0$, then $\lambda_F(\mathbf{p})$ is such that $V - \frac{c}{\mu - \lambda_F(\mathbf{p}) - \lambda_A(\mathbf{p})} - p_F = 0$.

Using Definition 3, we characterize the provider’s optimal prices under single pricing and price discrimination, respectively. We relegate these results to Appendix A in the E-Companion and only present their comparison in the following result.

Proposition 4. Consider FCFS and an exogenous fee s of the intermediary.

- (i) The firm’s revenue under price discrimination is strictly higher than that under single pricing when $\alpha > 1 - \hat{\lambda}_s/\Lambda$, where $\hat{\lambda}_s = \mu - \sqrt{\frac{c\mu}{V-s}}$; otherwise, the firm’s revenues under price discrimination and single pricing are equal.
- (ii) When it is optimal to serve both types under pricing discrimination, it holds that

$$p_F^* = p_A^* + s > p_A^* = p^*,$$

where p_F^* and p_A^* are the optimal prices charged to expert and amateur customers, and p^* is the optimal price under single pricing.

Similar to our main model, even when the intermediary’s fee is exogenous, the provider prefers to serve expert customers for their high margins. So, a Type II equilibrium in which both types are served will only emerge when the demand of expert customers is sufficiently low. However, the threshold $1 - \hat{\lambda}_s/\Lambda$ under an exogenous fee is higher than that in our main model. In other words, for $\alpha \in (1 - \hat{\lambda}_0/\Lambda, 1 - \hat{\lambda}_s/\Lambda)$, the provider will serve both segments under the intermediary’s endogenous fee but only the expert segment under the intermediary’s exogenous fee. This is because, for α slightly above $1 - \hat{\lambda}_0/\Lambda$, expert customers constitute a dominant portion of served customers. This allows the provider to target only a small portion of amateur customers and charge them a higher price. In our main model, the intermediary then has to lower its fee in order to attract the amateur segment. However, when the intermediary’s fee is exogenous and held fixed (at a relatively high level), it fails to invite more amateur customers. This leads to a lower joining rate of amateur customers, making a Type II equilibrium less likely.

When comparing the provider’s optimal prices under price discrimination and single pricing, we find that the prices charged to both types under price discrimination always exceed the optimal single price. This is in contrast to Proposition 2 established under the intermediary’s endogenous fee that suggests a mixed price comparison between the two

pricing schemes. Interestingly, the optimal price discrimination under an exogenous fee can have a very simple form: the price charged to amateur customers is set equal to the optimal single price, and the price charged to expert customers is higher by an amount of s . To explain, recall that customers are only differentiated by their *effective* valuations under the intermediary’s exogenous fee, with V for expert customers and $V - s$ for amateur customers. Under single pricing, the free-riding of expert customers prevails. If the provider decides to serve both segments, the expert segment would be treated as if they had the same valuations as amateur customers, $V - s$. Now, under price discrimination, the provider can fully extract the surplus of expert customers and this can be achieved by increasing the price charged to expert customers by s .

We next consider the provider’s joint optimization of pricing and prioritization. Unlike our main model, the following result shows that the benefits of prioritization will not carry through under the intermediary’s exogenous fee. In other words, it is sufficient to serve all customers under FCFS. To state the result, recall that we use R^{F-Pri} , R^{A-Pri} , and R^{FCFS} to denote the provider’s optimal revenues by prioritizing expert customers, prioritizing amateur customers, and serving all customers under FCFS, respectively.

Proposition 5. Under price discrimination and an exogenous fee of the intermediary, it always holds that $R^{A-Pri} = R^{FCFS} = R^{F-Pri}$.

Proposition 5 can be further strengthened: the provider’s optimal revenues under price discrimination are always the same under any non-idling work-conserving queueing policy. To understand the result, let λ_A and λ_F be the effective arrival rates of amateur and expert customers, respectively. Under a queueing policy π , let $W_A^\pi(\lambda)$ and $W_F^\pi(\lambda)$ denote the expected wait times of amateur and expert customers. As the provider can fully extract the surplus of each segment under price discrimination, the provider’s optimal revenue under policy π is

$$\begin{aligned} R^\pi(\lambda) &= \lambda_F [V - cW_F^\pi(\lambda)] + \lambda_A [V - s - cW_A^\pi(\lambda)] \\ &= \lambda_F V + \lambda_A (V - s) - c [\lambda_F W_F^\pi(\lambda) + \lambda_A W_A^\pi(\lambda)]. \end{aligned}$$

Note that the last term $\lambda_F W_F^\pi(\lambda) + \lambda_A W_A^\pi(\lambda)$ is a constant under any non-idling work-conserving policy,

$$\lambda_F W_F^\pi(\lambda) + \lambda_A W_A^\pi(\lambda) = \frac{\lambda_A + \lambda_F}{\mu - \lambda_A - \lambda_F}.$$

This implies that the provider needs only optimize over λ to find the optimal revenue.

The key to the above argument is that customers share the *same* delay sensitivity irrespective of their types. In this sense, the traditional $c\mu$ rule (e.g., Wolff, 1989, chapter 5) that relies on differentiation in at least one of customers’ delay sensitivities and service rates simply reduces to FCFS. In

other words, customers' heterogeneity in delay sensitivity is critical to the success of priority discrimination (e.g., Hasin & Haviv, 2006) when customers have fixed effective valuations.

Our main model, however, allows the intermediary to adjust its fee in response to the provider's queueing policy. This adaptivity changes the effective valuations of the amateur segment and brings prioritization benefits that do not materialize when the intermediary is unable to optimize its fee.

7 | CONCLUSION

With the growth in technology, many professional service offerings have become increasingly complex. This creates a chasm among users in their capabilities to deploy the service. Such user heterogeneity has important implications for the optimal design of professional services. In particular, users' onboarding experience can influence a service provider's pricing and prioritization strategy. Our work explores these concomitant issues, through a model that integrates user heterogeneity in skill sets with a classic framework of service operations.

We found that the presence of amateur customers allows expert customers to *free ride* under single pricing. To allay free-riding, the service provider can engage in type-based price discrimination. Our key guidance was on how a provider should use prioritization to enhance price differentiation. Despite the preference of expert customers under price discrimination, we showed that it is actually optimal to de-prioritize them. We also uncovered the welfare implications of prioritization and showed that prioritizing amateur customers can generate the highest social welfare.

We believe that the optimal design of professional services is a rich and complex problem that involves many strands of exploration. For instance, our paper does not consider competition between service providers and between intermediaries. Extending our framework to study competitive settings would generate potentially new policy recommendations. There are also contracting issues between the service provider and intermediary that we did not explore in this paper (e.g., Chen et al., 2022; Feldman et al., 2023). Integrating incentive issues with the service provider's pricing and priority decisions will advance our understanding of professional services. Another possible direction to pursue is the provider's information strategy at the customer level that has been proved effective for revenues and social welfare (e.g., Hu et al., 2017). A joint study of the provider's optimal pricing, prioritization, and information strategies constitutes an interesting direction for future study.

Finally, although our work is motivated by professional services and we introduce our model using cloud computing as the main backdrop, we believe that our results can also provide guidance for other relevant make-to-order sys-

tems with channel conflicts due to user heterogeneity in their accessibility to a product or service. A prominent example is restaurants that take both offline orders from dine-in customers and online orders from food delivery platforms. Another relevant example could be make-to-order manufacturers that serve different regions, operating direct channels in their home region and indirect channels (e.g., through a retailer) in other regions.


ACKNOWLEDGMENTS

The authors thank the department editor Michael Pinedo, a senior editor, and two anonymous reviewers for constructive feedbacks in the review process. The authors also thank Fazil Pac and Will Wang for conversations and input on service industry contracts. Chenguang (Allen) Wu acknowledges support from the Hong Kong General Research Fund (Grant Number: 16506122). Chen Jin acknowledges the Singapore Ministry of Education Academic Research Fund Tier 1 (251RES2101), The Wharton School Dean's Postdoctoral Research Fund, and Mack Institute Research Fund.

ORCID

Chenguang (Allen) Wu  <https://orcid.org/0000-0002-2528-0286>

Chen Jin  <https://orcid.org/0000-0001-9940-0757>

Senthil Veeraraghavan  <https://orcid.org/0000-0003-3143-6686>

ENDNOTES

¹ See <https://www.bls.gov/emp/tables/employment-by-major-industry-sector.htm>

² See <https://ir.aboutamazon.com/annual-reports>.

³ Databricks is a data analytics agency integrated with Amazon Web Services and Microsoft Azure via Apache Spark. Databricks itself does not own a large-scale computing cluster. See <https://docs.microsoft.com/en-us/azure/databricks/what-is-azure-databricks>.

⁴ We employ the dictionary definition of "amateur" in the strictest sense of "one lacking in experience and expertise in an art or science," as it relates to the specific infrastructural details of the service.

⁵ In the context of cloud computing, existing computing resources can be limited relative to the huge demand during peak hours and this can cause an "insufficient capacity" issue. When it happens, users attempting the busy servers will get an "insufficient instance capacity" error (see https://aws.amazon.com/premiumsupport/knowledge-center/emr-cluster-failed-capacity-quota/?nc1=h_ls). In this case, AWS will recommend users to "wait a few minutes, and then try to launch the cluster again" (see <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/troubleshooting-launch.html#troubleshooting-launch-capacity>). Likewise, Google Cloud will display to users a "resource exhausted" error (see <https://cloud.google.com/compute/docs/troubleshooting/troubleshooting-vm-creation>) and Microsoft Azure will display to users an "allocation failed" error (see <https://docs.microsoft.com/en-us/troubleshoot/azure/virtual-machines/allocation-failure>).

⁶ In general, we use λ_α to denote the total joining rate of expert and amateur customers when amateur customers constitute an α fraction of the market size.

⁷ It is common that the service provider (e.g., AWS) and intermediary (e.g., Databricks) charge users hourly rates. However, because customers' processing times are often random, it is reasonable to assume that customers make their purchase decisions by computing the *expected* payments to the provider and intermediary in the form of a lump sum fee.

- ⁸Evidence can be found from various sources. AWS: see <https://aws.amazon.com/quickstart/architecture/databricks/>; Google Cloud: see <https://cloud.google.com/databricks/>; Microsoft Azure: see <https://azure.microsoft.com/en-us/services/databricks/#capabilities>.
- ⁹Our communication with industry practitioners has also confirmed this fact. We are informed that Google cloud has implemented fairly different prices for direct clients and for intermediaries (Google cloud terms these intermediaries as “Value Added Resellers/Partners”).
- ¹⁰Our communication with industry practitioners confirms that customers get different priority and preemption rates based on their classes and on how they access the service.
- ¹¹The within-class service discipline is still FCFS and high-priority arrivals preempt low-priority jobs both in service and in queues.

REFERENCES

- Adiri, I., & Yechiali, U. (1974). Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research*, 22(5), 1051–1066.
- Afeche, P., & Sarhangian, V. (2015). *Rational abandonment from priority queues: Equilibrium strategy and pricing implications*. Columbia Business School Research Paper 15–93.
- Alperstein, H. (1988). Note: Optimal pricing policy for the service facility offering a set of priority prices. *Management Science*, 34(5), 666–671.
- Anand, K. S., Paç, M. F., & Veeraraghavan, S. (2011). Quality–speed conundrum: Trade-offs in customer-intensive services. *Management Science*, 57(1), 40–56.
- Armony, M., Roels, G., & Song, H. (2021). Pooling queues with strategic servers: The effects of customer ownership. *Operations Research*, 69(1), 13–29.
- Balachandran, K., & Schaefer, M. E. (1979). Class dominance characteristics at a service facility. *Econometrica: Journal of the Econometric Society*, 47(2), 515–519.
- Chen, H., & Frank, M. (2004). Monopoly pricing when customers queue. *IIE Transactions*, 36(6), 569–581.
- Chen, M., Hu, M., & Wang, J. (2022). Food delivery service and restaurant: Friend or foe? *Management Science*, 68(9), 6539–6551.
- Costa, B. G., Reis, M. A. S., Araújo, A. P., & Solis, P. (2018). Performance and cost analysis between on-demand and preemptive virtual machines. Working paper. <https://pdfs.semanticscholar.org/d665/3407a76a8f8b5ff7e3e031d18c3802ab3e0d.pdf>
- Cui, S., Wang, Z., & Yang, L. (2020). The economics of line-sitting. *Management Science*, 66(1), 227–242.
- Debo, L., & Veeraraghavan, S. (2014). Equilibrium in queues under unknown service times and service value. *Operations Research*, 62(1), 38–57.
- Edelson, N. M., & Hilderbrand, D. K. (1975). Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society*, 43(1), 81–92.
- Feldman, P., Frazelle, A. E., & Swinney, R. (2023). Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination. *Management Science*, 69(2), 812–823.
- Gavrieni, S., & Kulkarni, V. G. (2016). Self-selecting priority queues with burr distributed waiting costs. *Production and Operations Management*, 25(6), 979–992.
- Gilbert, S. M., & Weng, Z. K. (1998). Incentive effects favor nonconsolidating queues in a service system: The principal–agent perspective. *Management Science*, 44(12 pt 1), 1662–1669.
- Hassin, R., & Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, vol. 59. Springer Science & Business Media.
- Hassin, R., & Haviv, M. (2006). Who should be given priority in a queue? *Operations Research Letters*, 34(2), 191–198.
- Hu, M., Li, Y., & Wang, J. (2017). Efficient ignorance: Information heterogeneity in a queue. *Management Science*, 64(6), 2650–2671.
- Jafarnejad Ghomi, E., Rahmani, A. M., & Qader, N. N. (2019). Applying queue theory for modeling of cloud computing: A systematic review. *Concurrency and Computation: Practice and Experience*, 31(17), e5186.
- Mendelson, H., & Whang, S. (1990). Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations Research*, 38(5), 870–883.
- Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica: Journal of the Econometric Society*, 37(1), 15–24.
- Wolff, R. W. (1989). *Stochastic modeling and the theory of queues*. Pearson College Division.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wu, C., Jin, C., & Veeraraghavan, S. (2023). Designing professional services: Pricing and prioritization. *Production and Operations Management*, 32, 2578–2595. <https://doi.org/10.1111/poms.13996>