Queueing Systems (2025) 109:29
https://doi.org/10.1007/s11134-025-09956-z

Bundle pricing of congested services

Chenguang (Allen) Wu[1] · **Luyi Yang**[2]

Received: 7 August 2024 / Revised: 4 September 2025 / Accepted: 10 October 2025

Abstract

Bundle pricing is commonly adopted by service firms managing multiple congestion-prone service facilities. Under bundle pricing, the firm sells all services as a single package. This scheme is in contrast to à la carte pricing, whereby the firm sells each service separately. The existing theory generally sees bundling as being more lucrative when the marginal cost of production is low. However, little is known about how bundling compares to à la carte pricing in service systems with delay-sensitive customers, despite the prevalence of both practices. Our paper compares these two pricing schemes in congested service systems. We find that the classical prescription can be reversed in such congested service settings even in the absence of any marginal cost of service provision. Specifically, bundling generates less revenue than à la carte pricing when the potential arrival rate of customers is high relative to service capacity or when customers are highly delay-sensitive relative to their valuation of services. Moreover, the relative revenue difference between the two pricing schemes is non-monotone in either the potential arrival rate or delay sensitivity, with the percentage revenue loss from suboptimally practicing bundle pricing being the most substantial when the potential arrival rate or delay sensitivity is intermediate. From an operational perspective, bundle pricing results in higher (resp. lower) capacity utilization and thus more (resp. less) system congestion than à la carte pricing when the potential arrival rate is low (resp. high). For customers, bundling generates higher consumer surplus when the potential arrival rate is low or high, but may generate lower consumer surplus when the potential arrival rate is intermediate. Our results offer normative guidance to service firms considering these two pricing strategies and shed light on their operational and welfare implications.

Keywords Service operations · Pricing · Bundling · Queueing

✉ Chenguang (Allen) Wu
allenwu@ust.hk

Luyi Yang
luyiyang@berkeley.edu

[1] Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Hong Kong, China

[2] Haas School of Business, University of California, Berkeley, USA

Published online: 25 October 2025

1 Introduction

Should a firm offering multiple services individually price each of its services or sell them together as a bundle? What role does congestion—a common phenomenon in service systems—play in a firm's choice between these two pricing strategies? To put the questions into context, consider, for example, amusement parks. Some, such as Disneyland and Universal Studios, with bigger brand names and more resources to scale capacity, adopt *bundle pricing*, charging a single admission fee that grants access to all the rides within a park, whereas others, local parks in particular, such as Knoebels in Elysburg, Pennsylvania and Family Kingdom in Myrtle Beach, South Carolina, adopt *à la carte pricing*, where guests pay for each attraction they visit.

Examples of both pricing schemes abound in service-oriented businesses. For instance, à la carte pricing is commonly used by carnivals and fairs [31], whereas bundle pricing is exemplified by CityPASS, which combines various sightseeing tours into a ticket package [14], and the Buffet of Buffets pass, which entitles one to dine in multiple Las Vegas buffet restaurants [38]. Museums vary in whether they sell admission to various exhibitions as a bundle or separately. Hotel resorts and cruise lines may offer all-inclusive spa packages that cover massages, acupuncture, skin care, nail care, etc., or they may let guests purchase these services on an individual basis. Gas stations sometimes bundle fuel and car wash and sometimes choose not to do so.

"To bundle or not to bundle" is an age-old question that has long attracted attention of academics and practitioners alike, yet one unique feature that distinguishes the examples above is the presence of congestion-driven delay, i.e., due to the stochastic nature of the underlying processes, customers often have to wait before receiving their desired services, which, in turn, diminishes the appeal of those services and reduces customers' willingness to pay. Incorporating this congestion effect has subtle implications for the firm's pricing strategies. For a given price, how much a customer waits depends on how many other customers are present, and therefore the amount of congestion generated in the system is endogenously determined by customers' own interactions. This implies a firm aimed to maximize revenue not only faces the usual price-demand trade-off, but also must appropriately pull the pricing lever to regulate congestion. In particular, under bundle pricing, contrary to settings without congestion, it may be rational for delay-sensitive customers to buy the bundle but forgo certain services if they find the wait times at those facilities too long.

The extant bundling literature [1, 7, 21] often identifies bundling as being more lucrative when the marginal cost of production is low. The rationale is that it reduces customers' valuation heterogeneity, thereby enabling more efficient surplus extraction. How is this insight from the literature affected by the presence of congestion in service systems? What is the impact of congestion on the revenue gap between the two pricing schemes? What are the operational and welfare implications?

To address these research questions, we develop a queueing model in which a monopoly firm managing two service facilities faces a market of delay- and price-sensitive customers. Customers are interested in obtaining at most one unit of each

service, but their valuations of the two services differ, and vary across individuals.[1] In à la carte pricing, the firm sets a separate price for each service, whereas in bundle pricing, the firm sets a single price for the two services combined. Based on the pricing scheme with the associated price(s), and the expected delay at each facility, customers first make their purchasing decisions (i.e., under à la carte pricing, which service to purchase, if any; under bundle pricing, whether to purchase the bundle) and then (particularly in the case of bundle pricing) decide on which service to use conditional on a purchase. These purchasing and visiting decisions, in turn, determine the expected delays in equilibrium.

We find that while bundle pricing continues to generate more revenue than à la carte pricing when the potential arrival rate of customers is low relative to capacity and customers' delay sensitivity is low relative to their valuations of the services, the reverse is true (i.e., bundle pricing generates less revenue than à la carte) when either the potential arrival rate is high or customers are highly delay-sensitive. So, how can the presence of congestion undermine the revenue advantage of bundle pricing? On the one hand, bundling reduces customers' valuation dispersion, causing them to cede more information rent to the firm; on the other hand, it also implies that the right tail of the valuation distribution becomes thinner, i.e., demand from high-valuation customers falls. Nevertheless, high-valuation customers are exactly the segment the firm crucially relies on in the presence of heavy congestion because delay costs drive away low-valuation customers. Under a considerable congestion effect, the benefit of reduced valuation dispersion is outweighed by the shrinkage of the high-valuation segment, making bundle pricing less lucrative than à la carte pricing.

In this vein, the delay cost customers incur can be broadly interpreted as an implicit marginal cost for the firm. In fact, high marginal costs are a known detriment to the profitability of bundle pricing [1, 21, 40]. Recognizing this analog both connects and contrasts our results with the bundling literature. However, marginal costs of products are exogenously specified, whereas delay costs in services are endogenously determined as an equilibrium outcome. This distinction has two ramifications: (1) Customers directly internalize delay cost in service settings but not marginal cost in product settings; (2) the delay cost is controlled by the firm's price and varies with pricing scheme adopted, but the marginal cost does not share these properties.

Difference (1) above drives our results on the relative revenue difference between the two pricing schemes, which we find to be non-monotone in either the potential arrival rate or delay sensitivity, with the percentage revenue loss from suboptimally implementing bundle pricing (when the optimal scheme is à la carte) being the most substantial when the potential arrival rate or delay sensitivity is intermediate; nevertheless, the revenue gap closes as the potential arrival rate gets very high or when customers get excessively delay-sensitive. This contrasts the prediction from the bundling literature, which would indicate the performance of bundle pricing worsens (in terms of the percentage revenue loss relative to à la carte) as marginal cost increases. One important implication of this result is that when the potential arrival rate is very high or when customers are highly delay-sensitive, firms in practice may still prefer

[1] One of our key results on the revenue comparison between bundle pricing and à la carte [Theorem 1-(1)] is established under fairly general valuation distributions, but for tractability, other results are developed under a uniform valuation distribution.

bundle pricing as its negligible (percentage) revenue loss may be outweighed by its simplicity in implementation.

Difference (2) above motivates us to compare the total price and the resulting capacity utilization (as a proxy for system congestion). We find that the optimal bundle price is less than the total price one would pay to visit both facilities under à la carte pricing. Thus, bundling effectively enables a price discount. When the potential arrival rate is low, such a discount attracts more customers and translates into higher capacity utilization (and thus longer expected delay for each service). However, when the potential arrival rate is high, despite the discount effect, bundle pricing only results in lower capacity utilization (and thus shorter expected delay for each service) than à la carte pricing because bundling, in this case, targets only customers who highly value *both* services and is more effective in regulating congestion.

We also study the impact of bundling on consumer surplus. We find that when the potential arrival rate is sufficiently low or sufficiently high, bundle pricing generates more consumer surplus than à la carte pricing. Combined with the earlier result on revenue comparison, it implies that bundling can be a win–win situation for both the revenue-maximizing service provider and customers when the potential arrival rate and delay sensitivity are both low. Notably, when the potential arrival rate gets sufficiently high, while bundle pricing and à la carte pricing are almost identical in their revenue performance, consumer surplus under bundling is orders of magnitude higher than that under à la carte, again because bundle pricing sells only to customers with high valuations of both services. However, when the potential arrival rate is intermediate, consumer surplus can be lower under bundle pricing.

We consider several model extensions, including asymmetric valuations, asymmetric capacities and endogenous capacity. In addition to demonstrating robustness, we also generate additional insights. For instance, we find that the asymmetries in valuations and capacities can strengthen the dominance of à la carte pricing in case of heavy congestion. Furthermore, when the firm endogenously determines its capacity in conjunction with the pricing strategy, we find that when the capacity cost is low (resp. high), bundle pricing often causes the firm to maintain larger (resp. smaller) capacity and generate more (resp. less) profit than à la carte pricing.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the model setup. Sections 4 and 5 formulate the à la carte pricing and bundle pricing problems, respectively. Section 6 compares these two pricing schemes. Section 7 studies several model extensions. Section 8 concludes the paper and discusses future research directions. All technical proofs are relegated to the appendix.

2 Related literature

Our paper bridges two streams of the literature, one on product bundling and the other on congestion pricing in queueing systems.

The literature on product bundling dates back to the seminal paper by Adams and Yellen [1]. Largely through illustrative examples and graphical analysis, the authors demonstrate bundling can generate more revenue than à la carte pricing even in the

absence of complementarities in consumption or economies of scales in production. Their results highlight bundling as a economic device to shape demand and practice price discrimination. Schmalensee [40] and McAfee et al. [34] find that negative dependence of customer valuations further contributes to bundling's revenue advantage. Fang and Norman [21] show that negative dependence is not necessary for bundling to work, and that even under independent customer valuations, bundling can still outperform à la carte pricing in a variety of settings. Another important message from these papers is that bundling tends to generate more (less) revenue than à la carte pricing when the marginal cost of production is low (high). In particular, information goods have negligible marginal cost and thus lend themselves well to bundle pricing [7].

The bundling literature has also explored various supply-side aspects. McCardle et al. [35] study how bundling impacts order quantities using a newsvendor model. Cao et al. [13] examine the effectiveness of bundling under a supply constraint. Banciu et al. [8] investigate bundling strategies of vertically differentiated products subject to capacity constraints. Bhargava [9] and Chakravarty et al. [15] study bundling in a distribution channel where manufacturers and retailers interact. Cui et al. [19] investigate ancillary services that are not valuable on their own but can be sold either as a separate add-on, or together with the main service. Bundling main and ancillary services can be viewed as a form of tying [10]. The reader is referred to Venkatesh and Mahajan [42] for a survey of the bundling literature. None of the papers above has looked into the question of bundling multiple congested services that involve customer waiting, which is the focus of our paper.

To that end, our work builds on the literature of congestion pricing in queueing systems. This literature originates from Naor [36] and Edelson and Hilderbrand [20]. We refer to Hassin and Haviv [25] and Hassin [24] for a comprehensive review of this literature. Most of the existing queueing research focuses on the optimal pricing of a single type of service. For instance, Chen and Frank [16] investigate the pricing problem of a monopoly service provider. Anand et al. [5] introduce the quality-speed trade-off to the congestion pricing problem. Cachon and Harker [12] and Allon and Federgruen [4] examine multiple substitutable providers competing for customers that request at most one unit of the service.

Like our paper, a scant strand of the literature does study how to price *multiple* services, yet the models and research questions are quite different. Veltman and Hassin [41] study a model of two services, one congested, and the other not; homogeneous customers must request both or neither, and therefore, à la carte pricing is ruled out by assumption. Afèche [2] considers a tandem-queue model of a service chain in which each queue is managed by a separate entity that sets its own admission fee. In that sense, that paper considers (decentralized) à la carte pricing only.

Our paper focuses on a setting in which customers can visit multiple service facilities (at most once), and therefore nicely complements the service operations literature that studies customers' repeated usage of a single service. Randhawa and Kumar [39] study the subscription pricing problem motivated by the rental business. Cachon and Feldman [11] compare the pay-per-use scheme with subscription pricing in a queueing environment.[2] Afèche et al. [3] study why priority should sometimes be offered in

[2] We contrast bundle pricing with subscription pricing in detail in Sect. 6.3.

membership programs even when customers are homogeneous in delay sensitivity. Guo et al. [23] employ a feedback-queueing model to capture patient revisits and compare fee-for-service with bundled payment as reimbursement policies.

3 Model setup

We consider a monopoly firm that operates two different service facilities, indexed by $i = 1, 2$. At each facility, the service times are independently, identically and exponentially distributed with rate μ, and services are rendered on a First-In-First-Out (FIFO) basis. We refer to μ as the capacity of each facility. There is a separate queue for each facility (as in the case of amusement parks where each ride has a line). While the base model focuses on the case of both facilities having identical and exogenous capacities, we will later take a two-pronged approach to relax these assumptions: We will consider in Sect. 7.2 an extension of asymmetric (yet exogenous) capacities and in Sect. 7.3 another extension of endogenous (yet symmetric) capacities.

Rational customers arrive (or their service needs arise) according to a Poisson process with rate Λ, where Λ is referred to as the *potential arrival rate*. Each customer is interested in requesting at most one unit of service from each facility and is delay- and price-sensitive, incurring a delay cost c per unit time spent at each facility (including service). We also refer to c as customers' delay sensitivity. Following Littlechild [32], we assume that each customer's (gross) valuation of service i, v_i, is independently drawn from a common distribution with a continuous probability density function (PDF) f over support $[\underline{v}, \bar{v}]$ such that $f(v) > 0$ for $v \in (\underline{v}, \bar{v})$ and $0 \leq \underline{v} < \bar{v} < \infty$. Let F be the corresponding cumulative distribution function (CDF) and $\bar{F} \triangleq 1 - F$ be the complementary CDF. We note that assuming customer valuations of the two services are drawn from the same distribution is common in the bundling literature [9, 21, e.g.,]; we follow this convention. In Sect. sec:asymmetricspsvaluation, we will consider an extension in which customer valuations of different services are drawn from different distributions. Table 1 provides a glossary of the main notation used in the paper.

We consider two pricing schemes:

- *À La carte pricing* The firm sells the services *separately* by charging price p_i for access to service i.
- *Bundle pricing* The firm sells the services as a *bundle* by charging price p_{B} for access to both services.

Under either pricing scheme, the firm's objective is to maximize its total revenue generated from the two services offered by setting the optimal prices. The focus on revenue (as opposed to profit) implies negligible marginal costs for serving an additional customer. This implicit assumption is applicable to many service systems, in which staff tends to be salaried and facility costs are usually fixed; the focus on revenue is also in line with the congestion pricing literature [20, 36, e.g.,]. We will extend our analysis to profit maximization in Sect. sec:endogenousspscapacity where the firm must incur higher costs to maintain larger capacity.

Table 1 Glossary of main notation

Symbol	Description
Λ	Potential arrival rate
μ	Capacity
$F, \bar{F}, f$	CDF, complementary CDF, PDF of the valuation distribution
$\underline{v}, \bar{v}$	Lower bound and upper bound for the support of the valuation distribution
c	Delay cost per unit time
k	Capacity cost per unit time
v_i	An individual customer's valuation of service $i, i = 1, 2$
p_i, p_{A}	À la carte price for service $i, i = 1, 2$, and the optimal à la carte price
p_{B}	Bundle price
$R_{\mathrm{A}}, R_{\mathrm{B}}$	Revenue of à la carte pricing and bundle pricing, respectively
Δ	Relative revenue difference $(R_{\mathrm{B}} - R_{\mathrm{A}})/R_{\mathrm{A}}$
$u_{\mathrm{A}}, u_{\mathrm{B}}$	Capacity utilization under à la carte pricing and bundle bundling, respectively
λ_i	Effective arrival rate for each facility $i, i = 1, 2$
$D(p_{\mathrm{B}}), \lambda(p_{\mathrm{B}})$	Purchase rate of the bundle and effective arrival rate for each facility, respectively, as a function of bundle price p_{B}
$W(\lambda)$	$= 1/(\mu - \lambda)$, the $M/M/1$ expected delay
W_i	Expected delay at facility $i, i = 1, 2$
$\boldsymbol{W}$	$= (W_1, W_2)$
θ, θ_i, S_i	Cutoff values on customer valuations
$G, \bar{G}, g$	CDF, complementary CDF and PDF of the average valuation distribution

Under à la carte pricing, each customer decides whether to purchase each service, and upon purchasing a service, she always visits the facility offering that service (doing so is trivially rational). Under bundle pricing, each arriving customer decides whether to purchase the bundle, and upon purchase, she further decides whether to visit each service facility. Under both schemes, if she decides to visit both facilities, then upon service completion at one facility, she joins the queue of the other one; the visit sequence can be arbitrary. Customers do not renege from the queues they join. The model primitives Λ, μ, F, c are common knowledge, but each customer is privately informed of her valuations (v_1, v_2). As customers are often physically away from the service facilities at the moment of purchase, their purchase decisions are based on the expected delay.

Under either pricing scheme, there are potentially four (Poisson) streams of joining customers differentiated by their routes through the queues: those who join queue 1 only; those who join queue 2 only; those who first join queue 1 and then queue 2; and those who first join queue 2 and then queue 1.[3] Such a queueing network (with multiple customer classes and each class having a different route) is referred to as the

[3] In an unobservable queueing system, a random customer joins each queue with a certain probability (from a system standpoint, as shown in Sects. 4 and 5). This, combined with the Poisson thinning property, implies that customers of each stream enter the queueing system according to a Poisson process.

Kelly network [18, 28–30], which is a multi-class generalization of the (single-class) Jackson network and also has a product-form steady-state distribution. Thus, each facility has the same steady-state queue-length distribution as a standalone $M/M/1$ queueing system with the same effective arrival rate and service rate. Moreover, the steady-state waiting times of the two queues are independent, making the ex post waiting time at the first queue uninformative of the waiting time at the second queue.

To avoid triviality, we make the following assumption about the model parameters.

Assumption 1 $\underline{v} < c/\mu < \bar{v}$.

Before we proceed to the formulations of the à la carte pricing problem (Sect. 4) and the bundle pricing problem (Sect. 5), we first examine a benchmark case without congestion to build intuition on the revenue comparison between the two pricing schemes.

3.1 Benchmark: pricing without congestion

Without congestion, the problem reduces to optimal pricing of goods with zero marginal costs (due to the focus on revenue), such as information goods studied in the bundling literature [7, e.g.,]. Such a non-congestion benchmark can be obtained from our model by letting capacity μ tend to infinity (which would dispel congestion).

To fix ideas, let F be a uniform distribution over $[0, 1]$. Under à la carte pricing, the optimal price for each service is $1/2$; each service is purchased by half of the customers, generating a total expected revenue $1/2$ per customer. Under bundle pricing, the optimal price for the bundle is $\sqrt{6}/3$ (≈ 0.816); 2/3 of the customers purchase the bundle, generating a total expected revenue $2\sqrt{6}/9$ (≈ 0.544) per customer. Hence, bundling generates more revenue than à la carte pricing. The key driver for bundle pricing's superiority is the reduction of customers' valuation dispersion, which allows the firm to extract more surplus. Specifically, customer valuation of each service is uniformly distributed, but their valuation distribution for the bundle is a convolution of two uniform distributions—a triangular distribution, which is more centered around and peaked at the mean, making it less dispersed than a single uniform distribution. Due to this "valuation pooling" effect, bundle pricing is more of a volume strategy as compared to à la carte pricing, with a lower total price charged ($\sqrt{6}/3 < 2 \times 1/2$) and a higher sales volume induced ($2/3 > 1/2$).

Besides the uniform distribution, bundling is also more lucrative than à la carte under many other common distributions such as Gaussian [40]. More generally, Fang and Norman [21, Proposition 4] and Ibragimov and Walden [27, Theorem 4.2] give sufficient conditions under which bundle pricing outperforms à la carte in revenue; and these conditions tend to be satisfied by a variety of valuation distributions, suggesting a widespread revenue advantage of bundle pricing.[4] In the sequel, we shall see how this advantage is affected by the presence of congestion.

[4] We acknowledge the existence of counterexamples that show bundling generates less revenue than à la carte pricing for zero-marginal-cost goods (see [21] for such an example). However, these examples typically rely on the construction of somewhat peculiar valuation distributions, which are perhaps rare pathological exceptions.

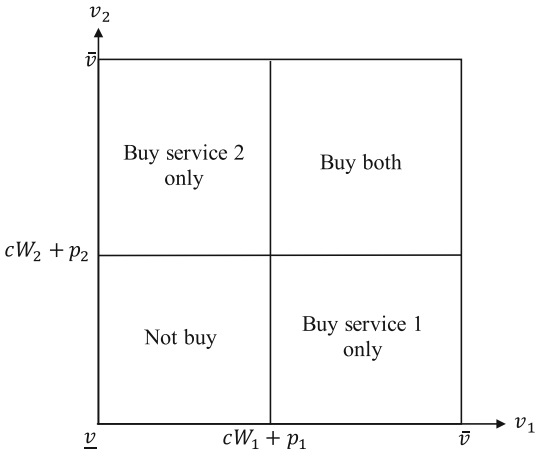


Fig. 1 Illustration of customer strategies under à la carte pricing

4 À la carte pricing

Under à la carte pricing, the firm sells each congested service separately. Given price p_i and expected delay W_i at facility i, each arriving customer with valuation v_i for service i purchases the service if and only if her expected utility from purchasing (the valuation of the service less the expected delay costs less the price) is nonnegative, i.e.,

$$v_i - cW_i - p_i \geq 0.$$

At facility i, there exists a cutoff value θ_i such that a customer purchases the service if and only if her valuation v_i (weakly) exceeds θ_i, where $\theta_i = p_i + cW_i$, i.e., a customer with valuation θ_i expects zero utility from purchasing the service. Figure 1 illustrates customer strategies (the mapping from customer valuations to their purchasing actions).

Based on the preceding argument, given (p_i, W_i), the effective (Poisson) arrival rate for facility i is $\lambda_i \triangleq \Lambda \bar{F}(\theta_i)$, which, in turn, implies that the expected delay is $W_i = W(\lambda_i)$, where $W(\lambda) \triangleq 1/(\mu - \lambda)$ is the expected delay (including time in service) of an $M/M/1$ queue with service rate μ and effective arrival rate λ.[5] In equilibrium, θ_i and W_i must be consistent such that θ_i solves the fixed-point equation $\theta_i = cW(\Lambda \bar{F}(\theta_i)) + p_i$ under p_i of interest.[6] Hence, we can recast the firm's à la carte pricing problem to maximize revenue $\sum_{i=1}^{2} \lambda_i p_i$ over prices (p_1, p_2) as an optimization problem over cutoff values (θ_1, θ_2):

[5] Recall from Sect. 3 that the underlying queueing system is a Kelly network, in which each queue operates in steady state as if it were a standalone $M/M/1$ queue.

[6] It is straightforward that θ_i must exist and is unique for any positive $p_i \in [\underline{v} - cW(\Lambda), \bar{v}]$. Any p_i outside this range cannot be optimal.

$$R_{\mathrm{A}} \triangleq \max_{\theta_1,\theta_2} \sum_{i=1}^{2} \Lambda \bar{F}(\theta_i)(\theta_i - cW(\Lambda \bar{F}(\theta_i))).$$

Thus, our à la carte pricing problem essentially boils down to the pay-per-use case in Cachon and Feldman [11] (with some minor cosmetic changes). Since the two services are symmetric, we use p_A to denote the optimal price for each service, and θ the corresponding (optimal) cutoff value. Adapted from Cachon and Feldman [11, Theorem 1], Proposition 1 characterizes the optimal à la carte price.

Proposition 1 *Assume F has a non-decreasing hazard rate,*[7] *i.e., $f(v)/\bar{F}(v)$ is non-decreasing in v. Under à la carte pricing, the optimal cutoff value θ uniquely solves*

$$\theta = cW(\Lambda \bar{F}(\theta)) + c\Lambda \bar{F}(\theta) W'(\Lambda \bar{F}(\theta)) + \frac{\bar{F}(\theta)}{f(\theta)},$$

and the optimal price p_A for each service is

$$p_A = \theta - cW(\Lambda \bar{F}(\theta)).$$

5 Bundle pricing

Under bundle pricing, the firm sells the two congested services as a bundle which grants purchasing customers access to both facilities.

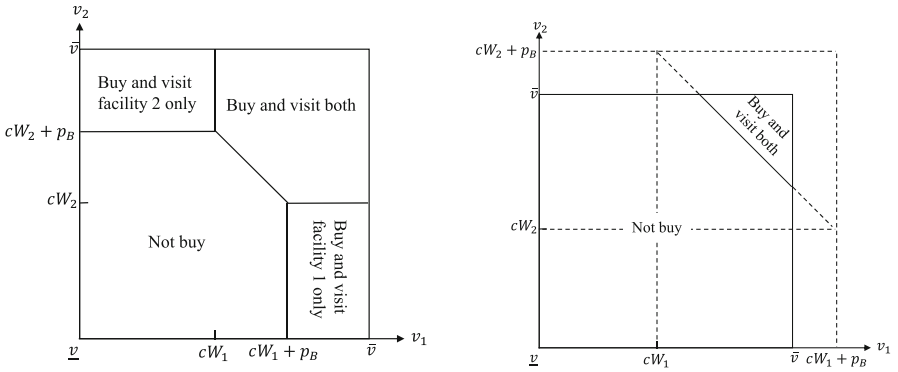
Given bundle price p_{B} and expected delays W_i for service i, a customer with valuations (v_1, v_2) decides whether to purchase the bundle in the first stage and conditional on purchase, whether to visit each facility in the second stage. She purchases the bundle in the first stage if and only if her expected utility of doing so is nonnegative, i.e.,

$$(v_1 - cW_1)^+ + (v_2 - cW_2)^+ - p_{\mathrm{B}} \geq 0. \tag{1}$$

Figure 2 illustrates two possible scenarios of customer segmentation in terms of their strategies.[8] When $cW_i + p_{\mathrm{B}} > \bar{v}$, customers either make no purchase or purchase and visit both facilities, as specified in Fig. 2b; therefore, the demand rate for the bundle coincides with the effective arrival rate of each facility. The case of $cW_i + p_{\mathrm{B}} \leq \bar{v}$ is more involved and four customer segments emerge, as specified in Fig. 2a. Specifically, it is possible for customers to purchase the bundle but use only one service and forgo the other; this happens when their valuation of the other service is too low to compensate

[7] To guarantee the first-order condition of θ to be sufficient for optimality, a non-decreasing hazard rate is required of the valuation distribution, a property that holds for a wide range of distributions (e.g., uniform, normal, exponential and Erlang). Note that if this property is not obeyed, the first-order condition would still be necessary, but may be satisfied by multiple θ's.

[8] Other scenarios of customer segmentation would lead to asymmetric effective arrival rates of the two facilities and cannot be sustained in equilibrium. We rigorously prove this result in Proposition 2.



(a) $cW_i + p_B \le \bar{v}, i = 1, 2$ (b) $cW_i + p_B > \bar{v}, i = 1, 2$

Fig. 2 Illustration of customer strategies under bundle pricing

for the delay cost that would otherwise accrue. As a result, in this scenario, the demand rate for the bundle is *strictly* larger than the effective arrival rate of each facility. This is a unique feature that arises in congested service systems.

Next, we formally characterize the effective arrival rate for each facility under a given bundle price p_{B} and expected delays $\boldsymbol{W} \triangleq (W_1, W_2)$. For service i, we use $3 - i$ to index the other service, $i = 1, 2$. Let $S_i(v_{3-i}, W, p_{\mathrm{B}})$ denote a threshold on customer valuation of service i as a function of customer valuation v_{3-i} for the other service, given expected delays $\boldsymbol{W}$ and bundle price p_{B}, such that a customer with valuation v_{3-i} for service $3 - i$ purchases the bundle and visits facility i if and only if her valuation of service i satisfies $v_i \ge S_i(v_{3-i}, \boldsymbol{W}, p_{\mathrm{B}})$. Since customers purchase the bundle if (1) holds and further visits facility i if $v_i \ge cW_i$, we have

$$S_i(v_{3-i}, \boldsymbol{W}, p_{\mathrm{B}}) = \begin{cases} cW_i + p_{\mathrm{B}} & \text{if } v_{3-i} \le cW_{3-i}, \\ c\sum_{j=1}^{2} W_j + p_{\mathrm{B}} - v_{3-i} & \text{if } v_{3-i} \in (cW_{3-i}, cW_{3-i} + p_{\mathrm{B}}], \\ cW_i & \text{otherwise.} \end{cases}$$

Given $(\boldsymbol{W}, p_{\mathrm{B}})$, the effective arrival rate for facility i, denoted by λ_i, is determined by

$$\lambda_i = \Lambda \int_{\underline{v}}^{\bar{v}} \bar{F}\left(S_i(v_{3-i}, \boldsymbol{W}, p_{\mathrm{B}})\right) \mathrm{d}F(v_{3-i}), \quad i = 1, 2. \tag{2}$$

On the other hand, given the effective arrival rate λ_i for facility i, the expected delay W_i is, in turn, determined by $W_i = W(\lambda_i) = 1/(\mu - \lambda_i)$. In equilibrium, $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2)$ and $\boldsymbol{W}$ must be consistent such that $\boldsymbol{\lambda}$ solves a system of two fixed-point equations in (2) with (W_1, W_2) replaced by $(W(\lambda_1), W(\lambda_2))$.

For a given bundle price p_{B}, it is unclear a priori whether an equilibrium characterized by the fixed-point $\boldsymbol{\lambda}$ uniquely exists. Nor is it obvious whether the two facilities always enjoy the same effective arrival rate in equilibrium. Proposition 2 provides affirmative answers to these questions.

Proposition 2 *Given any bundle price p_B, there exists a unique equilibrium (λ_1, λ_2), and the unique equilibrium induces a symmetric effective arrival rate, i.e., $\lambda_1 = \lambda_2 = \lambda$.*

Since the equilibrium is always symmetric, we can substantially simplify the formulation. We rewrite (2) (with (W_1, W_2) replaced by $(W(\lambda), W(\lambda))$) more explicitly as

$$\lambda = \Lambda \bar{F}\left(cW(\lambda) + p_{\mathrm{B}}\right) + \Lambda \int_{cW(\lambda)}^{cW(\lambda)+p_{\mathrm{B}}} \bar{F}\left(2cW(\lambda) + p_{\mathrm{B}} - v\right) \mathrm{d}F(v). \tag{3}$$

Let $\lambda(p_{\mathrm{B}})$ be the λ that solves (3) for any bundle price p_{B}, whose existence and uniqueness are justified by Proposition 2. Thus, the demand rate for the bundle (or the rate at which customers purchase the bundle) is the effective arrival rate of customers who visit one facility (and possibly visit the other) plus those who visit the other facility only:

$$D(p_{\mathrm{B}}) \triangleq \lambda(p_{\mathrm{B}}) + \Lambda F\left(cW(\lambda(p_{\mathrm{B}}))\right) \bar{F}\left(cW(\lambda(p_{\mathrm{B}})) + p_{\mathrm{B}}\right). \tag{4}$$

The firm selects bundle price p_{B} to maximize its revenue:

$$R_{\mathrm{B}} \triangleq \max_{p_{\mathrm{B}}} \; p_{\mathrm{B}} D(p_{\mathrm{B}}). \tag{5}$$

We can explicitly write $R_{\mathrm{B}} = \max\{R_{B,1}, R_{B,2}\}$, where $R_{B,1}$ and $R_{B,2}$ are defined in Problems 1 and 2, respectively.

Problem 1 ($cW + p_{\mathrm{B}} \le \bar{v}$, illustrated in Fig. 2a).

$$\begin{aligned} R_{B,1} \triangleq \max_{p_{\mathrm{B}}, W} \quad & \Lambda p_{\mathrm{B}} \left(\bar{F}(cW + p_{\mathrm{B}})(1 + F(cW)) + \int_{cW}^{p_{\mathrm{B}}+cW} \bar{F}(2cW + p_{\mathrm{B}} - v)\mathrm{d}F(v) \right), \\ \text{s.t.} \quad & W = \frac{1}{\mu - \Lambda\left[\bar{F}(cW + p_{\mathrm{B}}) + \int_{cW}^{p_{\mathrm{B}}+cW} \bar{F}(2cW + p_{\mathrm{B}} - v)\mathrm{d}F(v)\right]}, \\ & p_{\mathrm{B}} + cW \le \bar{v}. \end{aligned} \tag{6}$$

Problem 2 ($cW + p_{\mathrm{B}} > \bar{v}$, illustrated in Fig. 2b)

$$\begin{aligned} R_{B,2} \triangleq \max_{p_{\mathrm{B}}, W} \quad & \Lambda p_{\mathrm{B}} \bar{G}(cW + p_{\mathrm{B}}/2), \\ \text{s.t.} \quad & W = \frac{1}{\mu - \Lambda \bar{G}(cW + p_{\mathrm{B}}/2)}, \quad p_{\mathrm{B}} + cW \ge \bar{v}, \end{aligned}$$

where $\bar{G}$ denotes the complementary CDF ($G \equiv 1 - \bar{G}$ being the CDF and g being the density) of customers' average valuation of the two services $(v_1 + v_2)/2$:

$$\bar{G}(z) = \int_{\underline{v}}^{\bar{v}} \bar{F}(2z - v) dF(v).$$

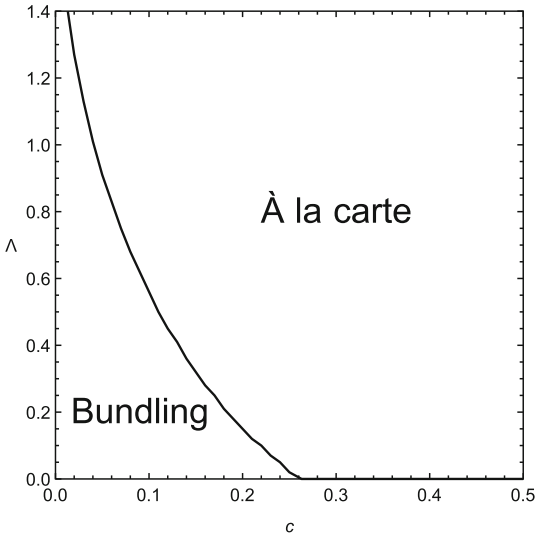


Fig. 3 Revenue comparison between à la carte and bundle pricing. Customer valuation uniformly distributed over $[0, 1]$, $\mu = 1$. The parameter region below (above) the curve is where bundle pricing generates more (less) revenue than à la carte pricing

Appendix A gives an explicit formulation of the bundle pricing problem when customer valuations are uniformly distributed.[9]

6 Comparison between À la carte and bundle pricing

This section compares bundle pricing and à la carte pricing in terms of their revenue performance, their price, operational and welfare implications.

Theorem 1 identifies conditions under which one pricing scheme outperforms the other in revenue. Following the convention of the bundling literature (see, e.g., [21]), we restrict attention to symmetric and log-concave distributions.[10]

Theorem 1 (Revenue Comparison)

1. *Suppose that the probability density function f of customer valuation is symmetric and log-concave. À la carte pricing generates more revenue than bundle pricing, i.e., $R_A > R_B$, if potential arrival rate Λ is sufficiently high.*
2. *Suppose that customer valuation is uniformly distributed over* $[0, 1]$. *Bundle pricing generates more revenue than à la carte pricing, i.e., $R_A < R_B$, if $\Lambda < 2\mu$ and delay sensitivity c is sufficiently low.*

The results of Theorem 1 are supplemented in Fig. 3. Together, they show the following. Under a low potential arrival rate relative to capacity and low delay sensitivity relative to service valuations, bundle pricing indeed outperforms à la carte pricing, echoing with the existing theory (cf. the benchmark case in Sect. 3.1). As the

[9] In the reminder of the paper, with the exception of Theorem 1-(1), all the other analytical and numerical results are established under a uniform valuation distribution unless otherwise specified.

[10] It follows from Bagnoli and Bergstrom [6] that log-concavity of probability density functions implies an increasing hazard rate as required by Proposition 1.

potential arrival rate gets high or customers get more delay-sensitive, however, bundle pricing becomes inferior to à la carte pricing, contrasting the benchmark setting without congestion. Somewhat strikingly, regardless of how favorable bundling can be in a setting without congestion, Theorem 1 shows that a sufficiently high potential arrival rate (relative to capacity) would always negate the revenue advantage of bundle pricing.[11]

Why does the presence of congestion undermine the relative appeal of bundle pricing? Recall from our discussion in Sect. 3.1 that reduced customer heterogeneity in valuations is key to the success of bundle pricing in a non-congested setting. However, when the congestion effect is considerable (either because customers are highly delay-sensitive relative to their valuation of the services or because the services are vastly popular relative to the capacity available), purchasing customers must incur sizable delay costs, which implies only customers with very high valuation of the bundle would purchase the bundle. Nevertheless, demand from these customers is likely to be too thin to retain the revenue superiority that bundling would otherwise enjoy, because most customers only have a moderate valuation of the bundle despite potential high valuation of an individual service. To that end, the driving force for the success of bundling in settings without congestion—reduced customer heterogeneity—is exactly what defeats bundling in settings with congestion.

Our results provide a potential explanation for the different pricing strategies amusement parks adopt. On the one hand, major amusement parks have a stronger brand name, potentially making its perceived value of service significantly higher than that of local parks. On the other hand, major amusement parks tend to offer waiting-area entertainment [44], which reduce the perceived cost of waiting as "occupied time feels shorter than unoccupied time" [33]. These effects combined imply that customers may be less delay-sensitive relative to their valuation of service at major amusement parks than at local ones. Therefore, major parks may prefer bundle pricing, whereas local parks may prefer à la carte pricing.[12]

If we take a broader view of our results, the delay cost incurred by customers can be interpreted as an *implicit* "marginal cost" the firm must bear in serving each customer, even in the absence of any explicit marginal cost. It is well recognized in the bundling literature [1, 21, 40] that a high marginal cost hurts the profitability of bundling. In particular, Adams and Yellen [1] argue that a high marginal cost weakens bundle pricing by forcing it to violate the *Exclusion* principle, which advises the firm not to sell a good to customers who value it less than its marginal cost. A firm that implements bundle pricing may find it difficult to follow this principle because bundling hinges on the transfer of customer surplus between goods [42]. Given the analog between

[11] We remark that the analytical results of the comparison in Theorem 1 are not exhaustive and that the analysis of the intermediate cases is numerical. Moreover, while part (1) of Theorem 1 is established under general valuation distributions, part (2) of Theorem 1 requires the assumption of a uniform valuation distribution.

[12] Note that amusement parks can differ along multiple dimensions. Even though major amusement parks may have a higher potential arrival rate (relative to capacity) than local ones, the former may still prefer bundle pricing (and the latter, à la carte pricing) if the former's cost of waiting (relative to valuation of service) is much lower than the latter's.

marginal costs of goods and delay costs in services, our findings are aligned squarely with the bundling literature.

Despite this similarity, there is one key distinction: Marginal costs of goods are exogenously specified, whereas delay costs in services are endogenously determined as an equilibrium outcome. This distinction has the following two ramifications:

Property 1 Unlike the marginal cost, the delay cost is directly borne by customers and passed on to the firm. As such, it is within the control of customers. Indeed, customers who purchase the bundle may forgo one service in equilibrium, thereby avoiding the corresponding delay cost.

Property 2 Unlike the marginal cost, the delay cost can be regulated by the firm's price. A high price would deter customers and reduce delay for those who join (see Proposition 4 for a price comparison). Thus, the amount of congestion generated (or how much the system is being utilized) varies with the pricing scheme adopted (see Proposition 5 for a utilization comparison).

6.1 Relative revenue difference

The results we have established so far speak to the direction of the revenue difference, but are silent on its magnitude. In this subsection, we explore this question and study the relative revenue difference between the two pricing schemes, Δ, defined by

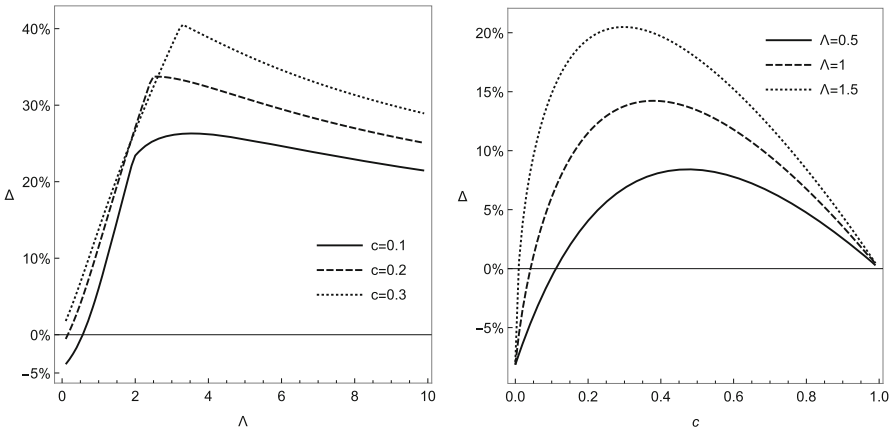
$$\Delta \triangleq \frac{R_{\mathrm{A}} - R_{\mathrm{B}}}{R_{\mathrm{B}}} \times 100\%,$$

where R_{B} and R_{A} are the optimal bundle revenue and à la carte revenue, respectively. Proposition 3 characterizes how potential arrival rate Λ impacts the relative revenue difference Δ.

Proposition 3 *Suppose that customer valuation is uniformly distributed over* $[0, 1]$. *The relative revenue difference between bundle pricing and à la carte pricing,* Δ, *is not monotone increasing in* Λ. *In particular,* $\lim_{\Lambda\to\infty} \Delta(\Lambda) = 0$.

While Theorem 1 suggests that under a sufficiently high potential arrival rate, bundle pricing falls short of à la carte pricing in revenue, Proposition 3 shows that the revenue gap between the two schemes does not always widen as the potential arrival rate increases; on the contrary, the revenue gap closes as the potential arrival rate tends to infinity. Here is the rationale. If the potential arrival rate is very high, the system becomes very congested, and only those with very high valuation will buy. Due to the capacity constraint, the volume of customers served by each facility approach a common limit and become largely invariant to the underlying pricing scheme adopted. Therefore, the revenues of the two schemes are almost identical to each other.

We supplement Proposition 3 with Fig. 4, from which we make the following two observations. First, when the potential arrival rate or delay sensitivity gets sufficiently high, while bundle pricing is less lucrative than à la carte (which echoes



(a) Impact of potential arrival rate Λ (b) Impact of delay sensitivity c

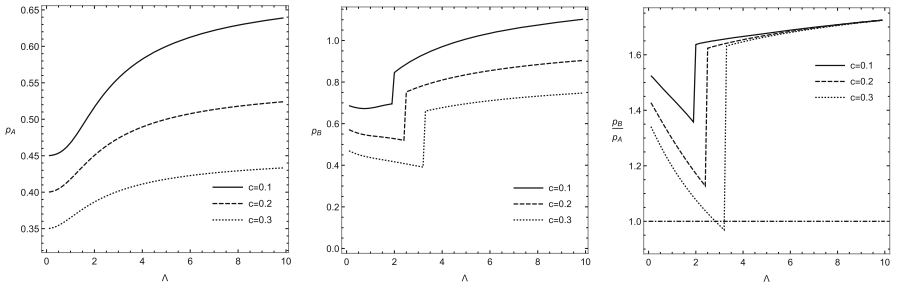
Fig. 4 Relative revenue difference between bundle pricing and à la carte pricing. Customer valuation uniformly distributed over [0, 1], $\mu = 1$; $\Delta = (R_{\mathrm{A}} - R_{\mathrm{B}})/R_{\mathrm{B}} \times 100\%$

with Theorem 1 and Fig. 3), the relative revenue gap diminishes. Second, the relative revenue difference is non-monotone in either the potential arrival rate or delay sensitivity (specifically, an inverse U-shaped relationship is observed); the revenue performance of bundle pricing relative to à la carte is the worst when the potential arrival rate or delay sensitivity is intermediate. Under such circumstances, à la carte pricing can beat bundle pricing by a sizable amount (over 40% in some instances). This non-monotonicity behavior runs counter to the prediction from the product bundling literature, which would suggest (see Appendix C) that under a uniform valuation distribution, the relative revenue difference is monotone in the (exogenous) marginal cost of production (i.e., the percentage revenue loss from suboptimally choosing bundle pricing increases with the marginal cost).

The implication from the above finding is that choosing the right pricing scheme may be most economically consequential when the potential arrival rate or delay sensitivity is intermediate. When the potential arrival rate or delay sensitivity is high, nevertheless, mis-specifying the pricing scheme—i.e., adopting bundle pricing as opposed to the optimal à la carte pricing—does not necessarily incur a huge (percentage) revenue loss. Therefore, the firm may still prefer bundle pricing under those circumstances due to its simplicity (e.g., setting up one ticket booth to sell a single ticket).

6.2 Other comparisons

In this subsection, we study the impact of bundling on price, system utilization, total visits, consumer surplus and social welfare.



(a) À la carte price (b) Bundle price (c) Price Ratio

Fig. 5 Price comparison between à la carte and bundle pricing. Customer valuation uniformly distributed over $[0, 1]$, $\mu = 1$

6.2.1 Price comparison

Proposition 4 conducts a price comparison.

Proposition 4 (Price Comparison) *Suppose that customer valuation is uniformly distributed over* $[0, 1]$. *If* Λ *is sufficiently small or sufficiently large, then* $p_B < 2p_A$. *Moreover,* $\lim_{\Lambda\to\infty} p_B(\Lambda)/p_A(\Lambda) = 2$.

Proposition 4 compares the optimal bundle price p_{B} and the optimal à la carte price p_{A}. (The total price of visiting both facilities is $2p_{\mathrm{A}}$.) Recall from Sect. 3.1 that in product bundling without congestion, bundle pricing is a volume strategy relative to à la crate pricing in the sense that it sets a lower total price (i.e., $p_{\mathrm{B}} < 2p_{\mathrm{A}}$). We analytically confirm in Proposition 4 that this relationship continues to hold in our queueing setting under either a high or a low potential arrival rate and numerically confirm it for an intermediate potential arrival rate; see Fig. 5c. This result is consistent with the anecdotal evidence that bundling offers a price discount. Further, since the two pricing schemes achieve almost identical revenue when the potential arrival rate is sufficiently high (see Proposition 3), the total prices of the two schemes are also approximately equal in this case.

Figure 5 supplements Proposition 4 and generates additional insights. We observe that while the à la carte price is increasing in the potential arrival rate (Fig. 5a), the bundle price is non-monotone (Fig. 5b), and as a result, for some intermediate potential arrival rate, the optimal bundle price can be even lower than the optimal (single) à la carte price (Fig. 5c).

When the potential arrival rate is not too high, it is best for bundle pricing to capture a large variety of customer types, including those who purchase the bundle but only visit one facility (illustrated in Fig. 2a). When the potential arrival rate increases in this regime, the service provider may have to cut the bundle price so as to retain those customers despite the increased congestion. Hence, the optimal bundle price may sometimes fall below the optimal à la carte price. However, when the potential arrival rate goes beyond a certain point, it would be optimal for bundle pricing to tame congestion and only target customers with high valuations of both services (illustrated in Fig. 2b) while pricing out those who value only one service but not so much the

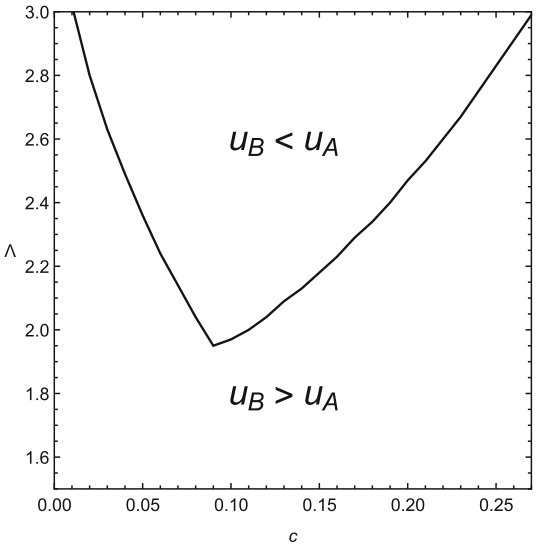


Fig. 6 Utilization comparison between à la carte and bundle pricing. Customer valuation uniformly distributed over $[0, 1]$, $\mu = 1$

other. Therefore, the firm raises the optimal bundle price. Altogether, these results collectively show the subtleties in the price implications of bundling.

6.2.2 Utilization comparison

The optimal prices have direct implications for capacity utilization, which we study next. Let λ_A and λ_B be the effective arrival rate to a facility[13] under à la carte and bundle pricing (with the prices optimally chosen), respectively. Let u_A and u_B denote the equilibrium capacity utilization under à la carte and bundle pricing, respectively. Then, $u_A = \lambda_A/\mu$, $u_B = \lambda_B/\mu$.

Proposition 5 (Utilization Comparison) *Suppose that customer valuation is uniformly distributed over* $[0, 1]$. *If* Λ *is sufficiently small,* $u_B > u_A$*; if* Λ *is sufficiently large,* $u_B < u_A$.

Capacity utilization is associated with the amount of congestion at each facility, and consequently, the implicit marginal cost incurred by the firm (as discussed earlier). Proposition 5 confirms Property 2 that the amount of congestion in the system varies with the pricing scheme adopted. As stated above, when the potential arrival rate is low, bundling works as a volume strategy that attracts more customers to the system and hence results in higher capacity utilization. By contrast, when the potential arrival rate is high, the reverse is true. In this case, the firm has an incentive to charge a relatively high price to tame congestion. Yet, as explained earlier, there is a thinner demand for the bundle as only customers who have high valuations for both services will visit. This causes capacity utilization to be lower under bundle pricing.

[13] It suffices to track one facility due to the symmetry between the two services.

Figure 6 supplements Proposition 5. Not only does it confirm the theoretical prescription from Proposition 5, it also indicates that when the potential arrival rate Λ is high, bundling leads to higher capacity utilization when delay sensitivity c is either low or high, but results in lower utilization when delay sensitivity is intermediate. Note that the utilization comparison is equivalent to the facility-level throughput comparison. Specifically, if we are interested in how bundling affects the number of customer visits per unit time (i.e., throughput) at each facility, this will follow immediately from the utilization comparison, because the throughput is equal to the utilization times capacity μ (which is fixed in our main model).

6.2.3 Total visits comparison

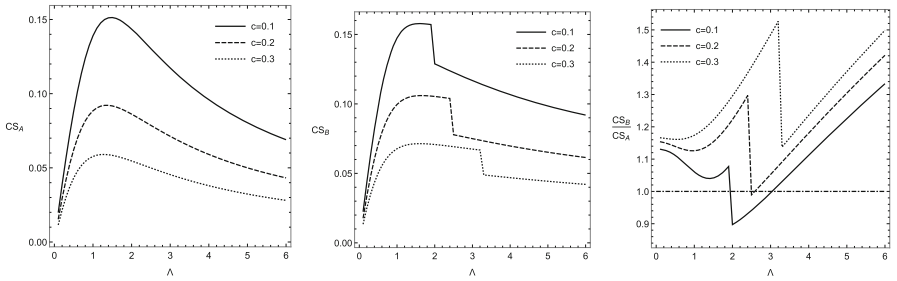
Another relevant metric is the (system-level) total visits per unit time, i.e., the total number of *unique* customers who visit the system per unit time. Note that this service-level metric differs from the aforementioned facility-level throughput, because a customer who visits the system may visit both facilities or only one of them, and in the latter case, does not contribute to the throughput of the facility that she does not visit. Formally, under à la carte pricing, the total visits per unit time are the number of unique customers who purchase at least one service per unit time, i.e.,

$$TV_A = \Lambda(1 - F^2(\theta)),$$

where the cutoff value θ follows from Proposition 1. Under bundle pricing, the total visits per unit time are the number of customers who purchase the bundle per unit time, i.e.,

$$TV_B = D(p_{\mathrm{B}}),$$

where $D(p_{\mathrm{B}})$ defined in (4) is the per-unit-time demand for the bundle under price p_{B}. We observe from our extensive numerical study that under a uniform valuation distribution, $TV_B < TV_A$ across the board, i.e., bundling always decreases the total visits per unit time. Note that this observation is in line with the utilization comparison (and equivalently, the facility-level throughput comparison) in Sect. 6.2.2 when Λ is high, but runs counter to the utilization comparison when Λ is low. In the latter case, under low Λ, while bundling increases utilization at each service facility, it decreases the number of unique customers who visit the entire service system. In such a case, as noted in Sect. 6.2.1, the bundle price is less than the sum of the à la carte prices across facilities but still higher than the standalone à la carte price at each facility. Hence, bundling can be less attractive to customers who value one service but not the other, yet more attractive to customers who value both services. The former effect can outweigh the latter effect, causing the total visits per unit time to decrease when services are sold in a bundle.



(a) À la carte pricing (b) Bundle pricing (c) Consumer surplus ratio

Fig. 7 Consumer surplus comparison between à la carte and bundle pricing. Customer valuation uniformly distributed over [0, 1], $\mu = 1$

6.2.4 Consumer surplus comparison

Next, we examine how bundling impacts consumer surplus. Let CS_{A} and CS_{B} denote the consumer surplus under the optimal à la carte and bundle pricing, respectively. Then,

$$CS_{\mathrm{A}} = 2\Lambda \int_{\theta}^{\bar{v}} [v - cW(\Lambda \bar{F}(\theta)) - p_{\mathrm{A}}] \mathrm{d}F(v),$$

where the cutoff value θ and optimal à la carte price p_{A} follow from Proposition 1, and

$$CS_{\mathrm{B}} = \Lambda \int_{\underline{v}}^{\bar{v}} \int_{\underline{v}}^{\bar{v}} \{[v_1 - cW(\lambda(p_{\mathrm{B}}))]^+ + [v_2 - cW(\lambda(p_{\mathrm{B}}))]^+ - p_{\mathrm{B}}\}^+ \mathrm{d}F(v_1)\mathrm{d}F(v_2),$$

where the optimal bundle price p_{B} and the effective arrival rate $\lambda(p_{\mathrm{B}})$ follow from (5) and (3), respectively.

Proposition 6 (Consumer Surplus Comparison) *Suppose that customer valuation is uniformly distributed over* [0, 1]. *If* Λ *is sufficiently small or sufficiently large, then* $CS_A < CS_B$. *Moreover,* $\lim_{\Lambda\to\infty} CS_B(\Lambda)/CS_A(\Lambda) = \infty$.

Proposition 6 shows that bundling results in higher consumer surplus when the potential arrival rate is either low or high. Recall from Theorem 1 that when the potential arrival rate and delay sensitivity are low, bundle pricing also achieves a higher revenue. Thus, in this case, bundling leads to a win–win outcome for both the service provider and customers. We further observe from Fig. 7c that under a broad range of conditions (beyond the ones analytically identified in Proposition 6), bundling leads to higher consumer surplus, but there are exceptions. Specifically, when delay sensitivity is relatively low and the potential arrival rate is intermediate, bundling may lead to lower consumer surplus (which is partially attributed to heavy congestion under bundling). We also observe from Fig. 7a and b that consumer surplus under either pricing scheme is non-monotone in the potential arrival rate. To see why, note that on

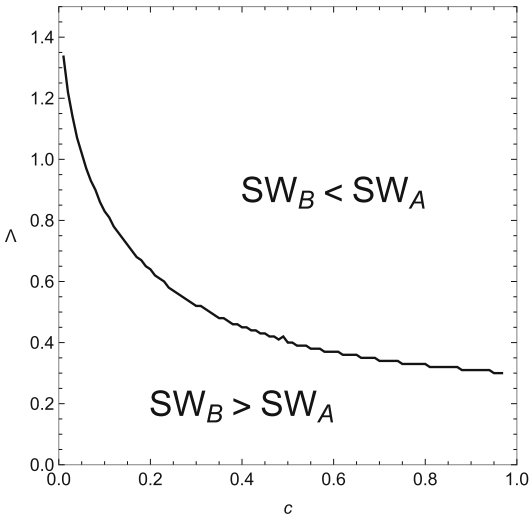


Fig. 8 Social welfare comparison between à la carte and bundle pricing. Customer valuation uniformly distributed over [0, 1], $\mu = 1$

the one hand, a high potential arrival rate implies more customers can benefit from the services; on the other hand, it also implies that the system will be more congested, potentially hurting the interest of each individual customer.

When the potential arrival rate gets sufficiently high, while the two pricing schemes resemble each other from the perspective of the service provider (both in terms of revenue and price; see Propositions 3 and 4), they (surprisingly) diverge in their impact on customers. Proposition 6 analytically shows that the consumer surplus ratio of bundle pricing to à la carte pricing tends to infinity as the potential arrival rate goes to infinity; Fig. 7c numerically shows that this ratio increases as the potential arrival rate gets sufficiently high. In this case, due to the capacity constraint, the volume of customers who use each service is similar (see the discussion after Proposition 3), but the composition of customers is markedly different. À la carte pricing sells to those with high valuation of at least one service, including those who do not value the other service much (see Fig. 1). By contrast, bundle pricing only sells to customers who have high valuation of both services (see Fig. 2b). Therefore, even though the volume of customers who obtain the services is almost the same between the two schemes, bundle pricing is better at allocating the services to those who value them the most, and hence generates more consumer surplus.

6.2.5 Social welfare comparison

Finally, we compare social welfare, defined as the sum of consumer surplus and firm revenue. Let SW_{A} and SW_{B} denote the social welfare under à la carte and bundling, respectively. Then,

$$SW_{\mathrm{A}} = CS_{\mathrm{A}} + R_{\mathrm{A}}, \quad SW_{\mathrm{B}} = CS_{\mathrm{B}} + R_{\mathrm{B}}.$$

We observe from Fig. 8 that bundling increases social welfare when the potential arrival rate Λ is low, but reduces social welfare when Λ is high. When Λ is low, bundling increases both consumer surplus (according to Proposition 6) and firm revenue (according to Theorem 1) and hence increases social welfare. However, when Λ is high, the impact of bundling on consumer surplus diverges from that on firm revenue. In such a case, while bundling continues to improve consumer surplus (according to Proposition 6), it reduces firm revenue (according to Theorem 1); thus, it is not entirely clear how bundling will impact social welfare. We show in the proof of Proposition 6 that as Λ tends to infinity, consumer surplus tends to zero under both pricing schemes due to excessive congestion. We also show in the proof of Theorem 1 that as Λ tends to infinity, firm revenue converges a positive constant under both pricing schemes. Thus, as Λ gets large, firm revenue predominates in social welfare, dwarfing the impact of consumer surplus. This suggests that the social welfare comparison largely follows from the revenue comparison in Theorem 1.

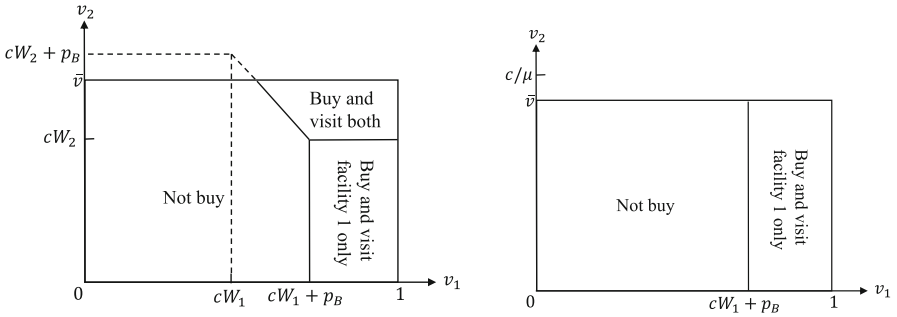
6.3 Discussion: comparison with subscription pricing

Since bundle pricing is closely related to subscription pricing, we discuss key differences in the economic and operational characteristics of these two pricing schemes. To focus our discussion, we will specifically contrast our model setup and results with those of subscription pricing as studied by the base model in Cachon and Feldman [11].

One defining feature of subscription pricing in Cachon and Feldman [11] is that customers do not know their realized valuations of future visits and are homogeneous when deciding whether to subscribe (provided that they have the same usage rate). By contrast, in our model of bundle pricing, consistent with the bundling literature, customers know their individual valuations of each service and are already heterogeneous before purchasing the bundle. As a result, our model generates notably different insights.[14]

- First, in terms of revenue comparison, if the potential arrival rate to the queue is low enough, then subscription pricing is always more lucrative than pay-per-use. By contrast, we numerically find that for any fixed potential arrival rate (that can be arbitrarily low), bundling is dominated by à la carte pricing in revenue as long as customers' delay sensitivity is sufficiently high (see Fig. 3). Hence, a high potential arrival rate is not necessary for bundling to fall short.
- Second, the relative revenue difference between subscription pricing and pay-per-use is monotone in the potential arrival rate (see Fig. 12c in Appendix D), but the relative revenue difference between bundle pricing and à la carte pricing is non-monotone. Specifically, when the potential arrival rate is sufficiently high, bundle pricing generates similar revenue to à la carte pricing, but subscription pricing severely falls short of pay-per-use. In fact, the revenue of subscription pricing diminishes to zero as the potential arrival rate gets sufficiently high.

[14] The insights mentioned below regarding the comparison between subscription pricing and pay-per-use are all based on Cachon and Feldman [11]. To be self-contained, we provide more details of their model in Appendix D.



(a) $cW_2 \le \bar{v} < cW_2 + p_B < cW_1 + p_B < 1$ (b) $\bar{v} < c/\mu < cW_1 + p_B < 1$

Fig. 9 Illustration of two additional cases of customer strategies under bundle pricing when the customer population has asymmetric valuations

- Third, subscription pricing lacks the ability to regulate congestion in the sense that the equilibrium capacity utilization under subscription pricing is always higher than that under pay-per-use. By contrast, the equilibrium capacity utilization under bundle pricing can be higher or lower than that under à la carte pricing (depending on the model primitives).
- Fourth, as for consumer surplus, subscription pricing can extract all consumer surplus ex ante, whereas pay-per-use generates positive surplus. Hence, subscription pricing always leads to lower consumer surplus than pay-per-use. By contrast, bundle pricing often generates higher consumer surplus and can create a win–win situation for the service provider and customers alike.

These differences imply that we cannot directly apply the insights from the comparison of subscription pricing and pay-per-use to the evaluation of bundle pricing vis-à-vis à la carte in our setting.

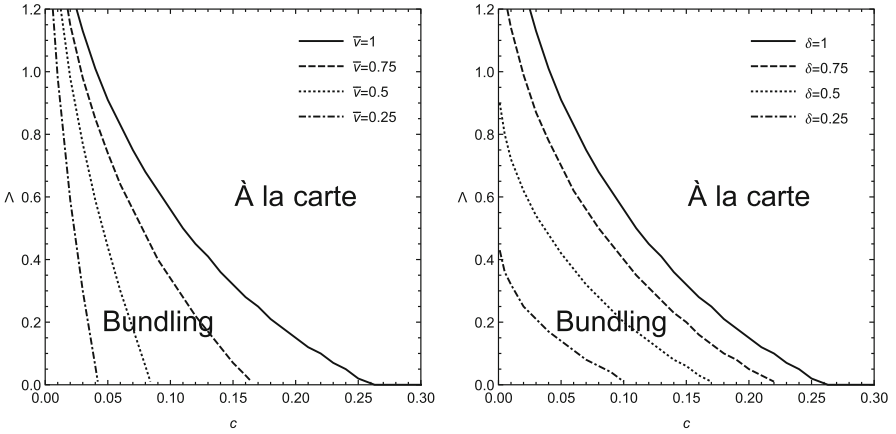
7 Extensions

In this section, we consider three extensions, namely asymmetric valuations, asymmetric capacities and endogenous capacities. Each extension changes one assumption of the base model at a time, while keeping all the other assumptions unchanged.

7.1 Asymmetric valuations

We assume that customer valuation of service 1 is uniformly distributed over [0, 1], whereas that of service 2 is uniformly distributed over $[0, \bar{v}]$ with $\bar{v} \in (0, 1)$. This captures an asymmetric scenario in which one service (service 1) is more popular than the other on average.

Our analysis starts with reexamination of customer segmentation. Recall from Fig. 2 that bundle pricing could lead to two cases of customer segmentation under symmetric valuations: a case of four segments as illustrated in Fig. 2a and a case of two segments



(a) Asymmetric valuations (b) Asymmetric capacities

Fig. 10 Revenue comparison when the two services are asymmetric. $\mu = 1$. In panel (**a**), both facilities have capacity μ; the customer valuation distributions for service 1 and service 2 are $U(0, 1)$ and $U(0, \bar{v})$, respectively. In panel (**b**), the customer valuation distribution for both services is $U(0, 1)$; the capacities of facility 1 and facility 2 are μ and $\delta\mu$, respectively

as illustrated in Fig. 2b. Under asymmetric valuations, however, two additional cases (besides the aforementioned two) might arise. Figure 9a illustrates a case in which customers are divided into three segments in terms of their strategies, i.e., those who do not buy, those who buy the bundle and visit both facilities and those who buy the bundle and visit facility 1 only. Figure 9b illustrates a case in which customers are divided into two segments in terms of their strategies: They either do not buy or buy and visit facility 1 only. This case occurs if $\bar{v} < c/\mu$, i.e., all customers have such low valuation of service 2 that nobody visits facility 2 even if it is empty. Obviously, this case reduces to à la carte pricing. Based on these cases of customer segmentation, Appendix A.2. presents an explicit formulation of the à la carte and bundle pricing problems.

We conduct a numerical study of the revenue comparison. Figure 10a encapsulates our numerical results. We observe that as the less popular service gets even less popular, i.e., as $\bar{v}$ decreases, the parameter space in which à la carte pricing dominates bundle pricing expands (the separating curve shifts left/down). On the one hand, this observation shows that an increased asymmetry makes bundling less favorable relative to à la carte. On the other hand, this observation also resonates with our earlier finding that high delay sensitivity harms the superiority of bundle pricing over à la carte pricing. With lower popularity of service 2, customers' delay cost appears higher by comparison, and the congestion effect becomes more considerable. Therefore, as explained earlier, the advantage of bundle pricing dwindles.

7.2 Asymmetric capacities

In the base model, we assume the two service facilities have equal capacity. In this subsection, we consider capacity asymmetry across the facilities. Specifically, we let the capacity of facility 1 be μ, and that of facility 2, $\delta\mu$, $\delta \in (0, 1)$. Appendix A.3 gives an explicit formulation of the à la carte and bundle pricing problems under asymmetric capacities when customer valuation is uniformly distributed over [0, 1] at both facilities. We conduct a numerical study of the revenue comparison. Figure 10b encapsulates our numerical results. We observe that as the small-capacity facility further decreases in capacity, i.e., as δ decreases, the parameter space in which à la carte pricing dominates bundle pricing expands (the separating curve shifts left/down). This observation echoes with our earlier finding that a high potential arrival rate harms the superiority of bundle pricing over à la carte pricing. With smaller capacity at one facility, the potential arrival rate seems higher, and the congestion effect becomes more considerable, which, as explained earlier, weakens the revenue advantage of bundle pricing.

7.3 Endogenous capacities

In the base model, we assume the capacities of the service facilities are exogenously given and do not vary with the underlying pricing scheme. In the long run, capacity may also be adjusted in conjunction with the pricing scheme adopted. We consider such an extension in this subsection. Specifically, we assume that the firm incurs a cost $k\mu$ per unit time for maintaining capacity μ at each facility, and unit capacity cost k measures how difficult it is to scale capacity. Thus, if both facilities operate at capacity μ, the total capacity cost per unit time is $2k\mu$. For a given pricing scheme (à la carte or bundle pricing), the firm chooses both the capacity and price(s) to maximize its total profit rate. Since the two facilities are symmetric in other dimensions, we focus on the case in which the firm allocates the same amount of capacity to each facility.

Let $\mu_{\mathrm{A}}(c)$ and $\mu_{\mathrm{B}}(c)$ be the optimal capacity under delay sensitivity c for à la carte and bundle pricing, respectively. Further, we denote the following limiting capacities: $\mu_{\mathrm{B}}(0) = \lim_{c\to 0} \mu_{\mathrm{B}}(c)$ and $\mu_{\mathrm{A}}(0) = \lim_{c\to 0} \mu_{\mathrm{A}}(c)$. Proposition 7 compares the firm's optimal capacity choice (in the limit) under the two pricing schemes.

Proposition 7 *If customer valuation is uniformly distributed over* [0, 1], *then* $\mu_B(0) > \mu_A(0)$ *when* $k < 7/16$; $\mu_B(0) \le \mu_A(0)$ *when* $k \in [7/16, 1)$.

Proposition 7 shows that when customers are delay-insensitive (which makes sharp analytical comparison possible), then bundle pricing entails a larger capacity than à la carte pricing if the capacity cost is low, but a smaller capacity if the capacity cost is relatively high; if capacity cost is too high, then neither pricing scheme is profitable. This insight continues to hold numerically when customers are mildly delay-sensitive, as demonstrated in Table 2a. The intuition is similar to that behind the utilization comparison for fixed capacity in Sect. 6.2.2. When maintaining capacity is cheap, the firm will choose a large capacity under either pricing scheme, which implies low system congestion. Therefore, as explained before, the firm practicing bundle pricing

Table 2 Profit and capacity comparison between à la carte and bundle pricing when capacity is endogenized

Capacity cost k	0.005	0.05	0.1	0.2	0.3	0.4	0.45
(a) $c = 0.005$							
À la carte profit Π_A	0.481	0.408	0.346	0.241	0.156	0.086	0.056
Bundle profit Π_B	0.521	0.429	0.349	0.213	0.106	0.034	0.010
À la carte capacity μ_{A}	1.197	0.679	0.580	0.471	0.388	0.312	0.275
Bundle capacity μ_{B}	1.469	0.870	0.754	0.617	0.436	0.281	0.211
Profit ratio Π_B/Π_A	108.4%	105.3%	101.0%	88.1%	67.9%	40.0%	17.2%
Capacity ratio $\mu_{\mathrm{B}}/\mu_{\mathrm{A}}$	122.7%	128.1%	129.9%	131.1%	112.4%	90.0%	76.6%

Capacity cost k	0.005	0.05	0.1	0.2	0.23
(b) $c = 0.05$					
À la carte profit Π_A	0.451	0.316	0.221	0.081	0.048
Bundle profit Π_B	0.487	0.326	0.210	0.040	0.000
À la carte capacity μ_{A}	2.691	1.098	0.837	0.589	0.533
Bundle capacity μ_{B}	3.173	1.328	1.020	0.709	0.632
Profit ratio Π_B/Π_A	108.0%	103.1%	95.1%	49.4%	0.0%
Capacity ratio $\mu_{\mathrm{B}}/\mu_{\mathrm{A}}$	117.9%	120.9%	121.8%	120.4%	118.6%

Note. Customer valuation uniformly distributed over [0, 1], $\Lambda = 1$

has a revenue advantage and has a stronger incentive to embrace a volume strategy by building a larger capacity to lure more customers. By contrast, when maintaining capacity is expensive, capacity level will naturally be set low under either pricing scheme, leading to heavy congestion. As explained before, this implies that the firm practicing bundle pricing faces less effective demand and consequently, maintains less capacity. However, if delay sensitivity becomes a more prominent feature, as demonstrated in Table 2-(b), the firm will make a profit under both schemes only when the capacity cost is low, and therefore, in such a case, bundle pricing is always associated with higher capacity investment.

In either case, nevertheless, we observe that bundle pricing is more profitable than à la carte pricing if the capacity cost is relatively low, but less profitable if the capacity cost is relatively high. This may provide yet another potential explanation for why firms differ in bundling strategies in practice. It is plausible that major amusement parks are more cost efficient than local amusement parks in scaling capacity. As a consequence, the former may build larger capacity and implement bundle pricing, whereas the latter may operate under smaller capacity and practice à la carte pricing instead.

8 Conclusion and discussion

This paper studies bundle pricing for firms selling multiple congestion-prone services. We find that when the potential arrival rate of customers is high relative to

service capacity, or customers' delay sensitivity is high relative to their service valuation, bundling will be less lucrative than à la carte pricing, even though customers may benefit more from bundle pricing. Our analysis highlights the need for service firms to adjust their pricing strategy as they improve branding (which increases customer valuation), develop waiting-area entertainment (which decreases customers' delay sensitivity) or expand capacity (which makes the potential arrival rate look less overwhelming by comparison).

Next, we discuss some of our modeling assumptions, acknowledge the limitations of our work and lay out future research directions. To begin with, various factors not captured in our model might affect firms' pricing strategies in practice. For example, customer demand may fluctuate at different points of the year.[15] While demand may spike during holiday seasons (at which time à la carte pricing may be more favorable according to our analysis), it might stay low for the rest of the year (at which time bundle pricing may be more advisable). The service firm may find it difficult to constantly change the pricing scheme and thus may stick to bundle pricing if it expects low demand for most of the year and high demand only on a few special occasions.

While some of our key analytical results on revenue comparison are established under fairly general valuations distributions (e.g., Theorem 1-(1)), other results assume a uniform valuation distribution for tractability. Future research can examine how to extend those results to more general distributions. Relatedly, one can consider relaxing the assumption that customer valuations of different services are independent. In addition, consistent with most of the bundling literature, our paper assumes strict additivity of customer valuations, i.e., the gross valuation of all the services combined is equal to the sum of the valuation of each service. Future research can incorporate non-strict additivity of customer valuations, although doing so would significantly complicate the model even without congestion (see, e.g., [22, 43]).

Moreover, we rule out by assumption any repeated visits to the same facility and future research can study such a feedback-queue structure.

Finally, the two pricing schemes we study in this paper (à la carte and bundle pricing) are subsumed by more complex pricing schemes such as mixed bundling [17] or two-part tariffs [37]. In mixed bundling, customers have the choice between buying individually priced services and buying them as a bundle at a discounted price. In a two-part tariff, the firm first charges an admission fee at the front entrance and then a second fee for actually using any of the services. We note that while these sophisticated schemes generate more revenue (since they pull more levers of price discrimination), firms in practice may nevertheless lean toward simple pricing schemes such as those studied in our paper. Still, future research can investigate the theoretical properties of those more complex pricing schemes.

In sum, our paper represents a first cut at investigating bundle pricing in a queueing context. We hope it will invite more future research in this area.

[15] Specifically, we envision two timescales. On a smaller timescale (e.g., within a week), customer demand is largely invariant, but on a bigger timescale (e.g., from month to month), customer demand may fluctuate. Customers may have a good idea of when demand jumps (e.g., in anticipation of a major holiday) and may be able to figure out the expected delay for each small timescale.

Appendix to "Bundle pricing of congested services"

Appendix A: Explicit formulation of pricing problems under uniform valuation distribution

A.1. Symmetric valuations and capacities

We first formulate the bundle pricing problem assuming that customer valuation is uniformly distributed over [0, 1] for both services and capacity is μ for both services.

Case (a): $cW + p_{\mathrm{B}} \le 1$. This corresponds to the case illustrated in Fig. 2a.

The effective arrival rate for each facility as a function of bundle price p_{B} is

$$\Lambda\left[\underbrace{(1-cW-p_{\mathrm{B}})cW}_{\text{the segment who visits only one facility}} + \underbrace{(1-cW)^2-\frac{p_{\mathrm{B}}^2}{2}}_{\text{the segment who visits both}}\right] = \Lambda\left(1-p_{\mathrm{B}}^2/2-cW-cWp_{\mathrm{B}}\right).$$

where W solves $W = \frac{1}{\mu-\Lambda(1-p_{\mathrm{B}}^2/2-cW-cWp_{\mathrm{B}})}$. The demand rate for the bundle as a function of bundle price p_{B} is

$$\Lambda[2(1-cW-p_{\mathrm{B}})cW+(1-cW)^2-p_{\mathrm{B}}^2/2] = \Lambda\left(1-p_{\mathrm{B}}^2/2-2cWp_{\mathrm{B}}-c^2W^2\right).$$

Hence, the revenue-maximization problem is

Problem 3

$$\begin{aligned} R_{B,1} \triangleq \max_{p_{\mathrm{B}}} \quad & \Lambda p_{\mathrm{B}}\left(1-p_{\mathrm{B}}^2/2-2cWp_{\mathrm{B}}-c^2W^2\right), \\ \text{s.t.} \quad & W = \frac{1}{\mu-\Lambda(1-p_{\mathrm{B}}^2/2-cW-cWp_{\mathrm{B}})}, \quad p_{\mathrm{B}}+cW \le 1. \end{aligned}$$

Case (b): $p_{\mathrm{B}} + cW \ge 1$. This corresponds to the case illustrated in Fig. 2b.

The effective arrival rate for each facility is equal to the demand rate of the bundle and is equal to

$$\Lambda\frac{(1-cW-(cW+p_{\mathrm{B}}-1))^2}{2} = \Lambda\frac{(2-2cW-p_{\mathrm{B}})^2}{2}.$$

where W solves $W = \frac{1}{\mu-\Lambda(2-2cW-p_{\mathrm{B}})^2/2}$. Hence, the revenue-maximization problem is

Problem 4

$$\begin{aligned} R_{B,2} \triangleq \max_{p_{\mathrm{B}}} \quad & \Lambda p_{\mathrm{B}}\,(2-2cW-p_{\mathrm{B}})^2/2, \\ \text{s.t.} \quad & W = \frac{1}{\mu-\Lambda(2-2cW-p_{\mathrm{B}})^2/2}, \quad 2-2cW-p_{\mathrm{B}} \ge 0, \quad p_{\mathrm{B}}+cW \ge 1. \end{aligned}$$

The optimal bundle revenue is $R_{\mathrm{B}} = \max\{R_{B,1}, R_{B,2}\}$, where $R_{B,1}$ and $R_{B,2}$ are obtained from solving Problems 3 and 4, respectively.

A.2. Asymmetric valuations

A.2.1. À La carte pricing

The optimal à la carte revenue R_{A} is:

$$R_{\mathrm{A}} = \max_{\theta_1,\theta_2} \quad \Lambda(1-\theta_1)\left[\theta_1 - \frac{c}{\mu - \Lambda(1-\theta_1)}\right] + \Lambda(1-\theta_2/\bar{v})\left[\theta_2 - \frac{c}{\mu - \Lambda(1-\theta_2/\bar{v})}\right]. \quad \text{(A.1)}$$

A.2.2. Bundling pricing

The optimal bundle revenue is $R_{\mathrm{B}} = \max\{R_{B,1}, R_{B,2}, R_{B,3}, R_{B,4}\}$, where $R_{B,1}$, $R_{B,2}$, $R_{B,3}$, $R_{B,4}$ are defined in Problems 5 through 8, respectively.

Problem 5 (Four segments: visit both, visit 1 only, visit 2 only, not buy, as illustrated in Fig. 2a)

$$\begin{aligned}
R_{B,1} = \max_{p_{\mathrm{B}}, W_1, W_2} \quad & \Lambda p_{\mathrm{B}}\left[\bar{v} - (cW_1 + p_{\mathrm{B}})(cW_2 + p_{\mathrm{B}}) + \frac{1}{2}p_{\mathrm{B}}^2\right]/\bar{v}, \\
\text{s.t.} \quad W_1 = & \frac{1}{\mu - \Lambda\left[\bar{v} - (cW_1 + p_{\mathrm{B}})(cW_2 + p_{\mathrm{B}}) + \frac{1}{2}p_{\mathrm{B}}^2 - (\bar{v} - p_{\mathrm{B}} - cW_2)cW_1\right]/\bar{v}}, \\
W_2 = & \frac{1}{\mu - \Lambda\left[\bar{v} - (cW_1 + p_{\mathrm{B}})(cW_2 + p_{\mathrm{B}}) + \frac{1}{2}p_{\mathrm{B}}^2 - (1 - p_{\mathrm{B}} - cW_1)cW_2\right]/\bar{v}}, \\
& cW_1 + p_{\mathrm{B}} \le 1, \quad cW_2 + p_{\mathrm{B}} \le \bar{v}.
\end{aligned}$$

Problem 6 (Three segments: buy and visit both, buy and visit 1 only, not buy, as illustrated in Fig. 9a)

$$\begin{aligned}
R_{B,2} = \max_{p_{\mathrm{B}}, W_1, W_2} \quad & \Lambda p_{\mathrm{B}}\left[\bar{v}(1 - cW_1 - p_{\mathrm{B}}) + (\bar{v} - cW_2)^2/2\right]/\bar{v}, \\
\text{s.t.} \quad W_1 = & \frac{1}{\mu - \Lambda\left[\bar{v}(1 - cW_1 - p_{\mathrm{B}}) + (\bar{v} - cW_2)^2/2\right]/\bar{v}}, \\
W_2 = & \frac{1}{\mu - \Lambda\left[(1 - cW_1 - (cW_2 + p_{\mathrm{B}} - \bar{v}) + 1 - cW_1 - p_{\mathrm{B}})(\bar{v} - cW_2)/2\right]/\bar{v}}, \\
cW_1 + p_{\mathrm{B}} \le 1, \quad & cW_2 + p_{\mathrm{B}} \ge \bar{v}, \quad \bar{v} \ge cW_2.
\end{aligned}$$

Problem 7 (Two segments: buy and visit both, not buy, as illustrated in Fig. 2b)

$$\begin{aligned} R_{B,3} &= \max_{p_{\mathrm{B}},W} \quad \Lambda p_{\mathrm{B}}(1+\bar{v}-2cW-p_{\mathrm{B}})^2/(2\bar{v}), \\ \text{s.t.} \quad W &= \frac{1}{\mu-\Lambda(1+\bar{v}-2cW-p_{\mathrm{B}})^2/(2\bar{v})}, \\ cW+p_{\mathrm{B}} &\geq 1, \quad cW \leq \bar{v}. \end{aligned}$$

Problem 8 (Two segments: buy and visit 1 only, not buy, as illustrated in Fig. 9b)

$$\begin{aligned} R_{B,4} &= \max_{p_{\mathrm{B}},W} \quad \Lambda p_{\mathrm{B}}(1-cW-p_{\mathrm{B}}), \quad \text{s.t.} \\ W &= \frac{1}{\mu-\Lambda(1-cW-p_{\mathrm{B}})}, \quad cW+p_{\mathrm{B}} \leq 1, \quad c/\mu \geq \bar{v}. \end{aligned}$$

A.3. Asymmetric capacities

A.3.1. À La carte pricing

The optimal à la carte revenue R_{A} is:

$$R_{\mathrm{A}} = \max_{\theta_1,\theta_2} \quad \Lambda(1-\theta_1)\left[\theta_1 - \frac{c}{\mu-\Lambda(1-\theta_1)}\right] + \Lambda(1-\theta_2)\left[\theta_2 - \frac{c}{\delta\mu-\Lambda(1-\theta_2)}\right].$$

A.3.2. Bundling pricing

The optimal bundle revenue is $R_{\mathrm{B}} = \max\{R_{B,1}, R_{B,2}, R_{B,3}, R_{B,4}\}$, where $R_{B,1}$, $R_{B,2}$, $R_{B,3}$, $R_{B,4}$ are defined in Problems 9 through 12, respectively.

Problem 9 (Four segments: visit both, visit 1 only, visit 2 only, not buy)

$$\begin{aligned} R_{B,1} &= \max_{p_{\mathrm{B}},W_1,W_2} \quad \Lambda p_{\mathrm{B}}\left[1-(cW_1+p_{\mathrm{B}})(cW_2+p_{\mathrm{B}})+\frac{1}{2}p_{\mathrm{B}}^2\right], \\ \text{s.t.} \quad W_1 &= \frac{1}{\mu-\Lambda\left[1-(cW_1+p_{\mathrm{B}})(cW_2+p_{\mathrm{B}})+\frac{1}{2}p_{\mathrm{B}}^2-(1-p_{\mathrm{B}}-cW_2)cW_1\right]}, \\ W_2 &= \frac{1}{\delta\mu-\Lambda\left[1-(cW_1+p_{\mathrm{B}})(cW_2+p_{\mathrm{B}})+\frac{1}{2}p_{\mathrm{B}}^2-(1-p_{\mathrm{B}}-cW_1)cW_2\right]}, \\ cW_1+p_{\mathrm{B}} &\leq 1, \quad cW_2+p_{\mathrm{B}} \leq 1. \end{aligned}$$

Problem 10 (Three segments: buy and visit both, buy and visit 1 only, not buy)

$$\begin{aligned} R_{B,2} &= \max_{p_{\mathrm{B}}, W_1, W_2} \quad \Lambda p_{\mathrm{B}} \left[(1 - cW_1 - p_{\mathrm{B}}) + (1 - cW_2)^2/2 \right], \\ \text{s.t.} \quad W_1 &= \frac{1}{\mu - \Lambda \left[(1 - cW_1 - p_{\mathrm{B}}) + (1 - cW_2)^2/2 \right]}, \\ W_2 &= \frac{1}{\delta\mu - \Lambda \left[(1 - cW_1 - (cW_2 + p_{\mathrm{B}} - 1) + 1 - cW_1 - p_{\mathrm{B}})(1 - cW_2)/2 \right]}, \\ & cW_1 + p_{\mathrm{B}} \le 1, \quad cW_2 + p_{\mathrm{B}} \ge 1, \quad cW_2 \le 1. \end{aligned}$$

Problem 11 (Two segments: buy and visit both, not buy)

$$\begin{aligned} R_{B,3} &= \max_{p_{\mathrm{B}}, W_1, W_2} \quad \Lambda p_{\mathrm{B}} (2 - cW_1 - cW_2 - p_{\mathrm{B}})^2/2, \\ \text{s.t.} \quad W_1 &= \frac{1}{\mu - \Lambda(2 - cW_1 - cW_2 - p_{\mathrm{B}})^2/2}, \quad W_2 = \frac{1}{\delta\mu - \Lambda(2 - cW_1 - cW_2 - p_{\mathrm{B}})^2/2}, \\ & cW_1 + p_{\mathrm{B}} \ge 1, \quad cW_2 + p_{\mathrm{B}} \ge 1, \quad cW_1 \le 1, \quad cW_2 \le 1. \end{aligned}$$

Problem 12 (Two segments: buy and visit 1 only, not buy)

$$\begin{aligned} R_{B,4} &= \max_{p_{\mathrm{B}}, W} \quad \Lambda p_{\mathrm{B}} (1 - cW - p_{\mathrm{B}}), \quad \text{s.t.} \\ W &= \frac{1}{\mu - \Lambda(1 - cW - p_{\mathrm{B}})}, \quad cW + p_{\mathrm{B}} \le 1, \quad c/(\delta\mu) \ge 1. \end{aligned}$$

Appendix B: Proofs

Proof of Proposition 2 We first show there always exists a symmetric fixed point (λ_1, λ_2) with $\lambda_1 = \lambda_2 = \lambda$ to the set of equations in (2). We next show that the fixed point is unique. To show a symmetric fixed point exists, we need to show that there exists λ that satisfies $\lambda = \Lambda \int_{\underline{v}}^{\bar{v}} \bar{F}(S(v, W(\lambda), p_{\mathrm{B}}) \mathrm{d}F(v)$, where

$$S(v, W(\lambda), p_B) = \begin{cases} cW(\lambda) & \text{if } p_B - (v - cW(\lambda))^+ < 0, \\ p_B - (v - cW(\lambda))^+ + cW(\lambda) & \text{if } p_B - (v - cW(\lambda))^+ \ge 0. \end{cases}$$

Define $\zeta(\lambda) \triangleq \Lambda \int_{\underline{v}}^{\bar{v}} \bar{F}(S(v, W(\lambda), p_B) \mathrm{d}F(v) - \lambda$. One can easily verify that $\zeta(\lambda)$ is decreasing in λ. Further, $\zeta(0) > 0$ and $\zeta(\Lambda) < 0$. Hence, there exists a unique λ_0 such that $\zeta(\lambda_0) = 0$. Define $\boldsymbol{\lambda} \triangleq (\lambda_0, \lambda_0)$, which is the unique symmetric fixed point. We next show the symmetric fixed point is the only fixed point. Otherwise, suppose there exists an asymmetric fixed point $\boldsymbol{\lambda} = (\lambda, \lambda + \delta)$. By the symmetry of the two

services, we can assume $\delta > 0$ without loss of generality. Thus,

$$\Lambda \int_{\underline{v}}^{\bar{v}} \bar{F}(S(v_1, W(\lambda), W(\lambda+\delta), p_B) dF(v_1) = \lambda + \delta,$$
$$\Lambda \int_{\underline{v}}^{\bar{v}} \bar{F}(S(v_2, W(\lambda+\delta)), W(\lambda), p_B) dF(v_2) = \lambda.$$

Subtracting the second equation from the first gives

$$\Lambda \int_{\underline{v}}^{\bar{v}} \left[\bar{F}(S(v, W(\lambda), W(\lambda+\delta), p_B)) - \bar{F}(S(v, W(\lambda+\delta), W(\lambda), p_B)\right] \mathrm{d}F(v) = \delta.$$

We compare $S(v_1, W(\lambda), W(\lambda+\delta), p_B)$ and $S(v_2, W(\lambda+\delta)), W(\lambda), p_B)$ in three regions $\mathcal{I}_1 = \{v : v \le p_B + cW(\lambda)\}, \mathcal{I}_2 = \{v : p_B + cW(\lambda) \le v < p_B + cW(\lambda+\delta)\}$, $\mathcal{I}_3 = \{v : v > p_B + cW(\lambda+\delta)\}$.

We rewrite $S(v_1, W(\lambda), W(\lambda+\delta), p_B)$ and $S(v_2, W(\lambda+\delta)), W(\lambda), p_B)$ in the three regions,

$$\mathcal{I}_1 : \begin{cases} S(v, W(\lambda), W(\lambda+\delta), p_B) = \min\{p_B - v + cW(\lambda) + cW(\lambda+\delta), p_B + cW(\lambda+\delta)\} \\ S(v, W(\lambda+\delta), W(\lambda), p_B) = \min\{p_B - v + cW(\lambda) + cW(\lambda+\delta), p_B + cW(\lambda)\} \end{cases}$$
$$\mathcal{I}_2 : \begin{cases} S(v, W(\lambda), W(\lambda+\delta), p_B) = cW(\lambda+\delta) \\ S(v, W(\lambda+\delta), W(\lambda), p_B) = \min\{p_B - v + cW(\lambda) + cW(\lambda+\delta), p_B + cW(\lambda, p_B)\} \end{cases}$$
$$\mathcal{I}_3 : \begin{cases} S(v, W(\lambda), W(\lambda+\delta), p_B) = cW(\lambda+\delta) \\ S(v, W(\lambda+\delta), W(\lambda), p_B) = cW(\lambda) \end{cases}$$

$S(v, W(\lambda), W(\lambda+\delta), p_B) > S(v, W(\lambda+\delta)), W(\lambda), p_B)$ in $\mathcal{I}_1$ and $\mathcal{I}_3$. Also, notice that the definition of $\mathcal{I}_2$ implies $p - v + cW(\lambda) \le 0$, or equivalently, $cW(\lambda+\delta) \ge p - v + cW(\lambda) + cW(\lambda+\delta)$. It follows that

$$\begin{aligned} S(v, W(\lambda), W(\lambda+\delta), p_B) = cW(\lambda+\delta) &\ge p - v + cW(\lambda) + cW(\lambda+\delta) \\ &\ge \min\{p - v + cW(\lambda) + cW(\lambda+\delta), p + cW(\lambda)\} = S(v, W(\lambda+\delta), W(\lambda), p_B). \end{aligned}$$

Hence, $S(v, W(\lambda), W(\lambda+\delta), p_B) > S(v, W(\lambda+\delta), W(\lambda), p_B)$ for all v, which leads to the following contradiction, $\delta = \Lambda \int_{\underline{v}}^{\bar{v}} \Big[\bar{F}(S(v, W(\lambda), W(\lambda+\delta), p_B)) - \bar{F}(S(v, W(\lambda+\delta)), W(\lambda), p_B)\Big] \mathrm{d}F(v) \le 0$. Thus, no asymmetric fixed point can exist, i.e., the symmetric fixed point is the only fixed point. □

Proof of Theorem 1 Part (1):

Step 1: We first show that in Problem 1, $R_{B,1} \to 0$ as $\Lambda \to \infty$. In case (a), for any feasible p_B such that $cW + p_B \le \bar{v}$, the effective arrival rate for each facility

is $\lambda = \Lambda\left[\bar{F}(cW + p_{\mathrm{B}}) + \int_{cW}^{p_{\mathrm{B}}+cW} \bar{F}(2cW + p_{\mathrm{B}} - v)\mathrm{d}F(v)\right]$. The stability of the system requires $\lambda < \mu$.

Further,

$$\begin{aligned}
\lambda =& \Lambda\left[\bar{F}(cW + p_{\mathrm{B}}) + \int_{cW}^{p_{\mathrm{B}}+cW} \bar{F}(2cW + p_{\mathrm{B}} - v)\mathrm{d}F(v)\right] \\
\geq& \Lambda \int_{cW}^{p_{\mathrm{B}}+cW} \bar{F}(2cW + p_{\mathrm{B}} - v)\mathrm{d}F(v) \\
=& \Lambda\left[\int_{cW}^{p_{\mathrm{B}}+cW} [F(p_{\mathrm{B}} + cW) - F(p_{\mathrm{B}} + 2cW - v)]\mathrm{d}F(v)\right] \\
\geq& \Lambda\left[\int_{p_{\mathrm{B}}/2+cW}^{p_{\mathrm{B}}+cW} [F(p_{\mathrm{B}} + cW) - F(p_{\mathrm{B}} + 2cW - v)]\mathrm{d}F(v)\right] \\
\geq& \Lambda\left[\int_{p_{\mathrm{B}}/2+cW}^{p_{\mathrm{B}}+cW} [F(p_{\mathrm{B}} + cW) - F(p_{\mathrm{B}}/2 + cW)]\mathrm{d}F(v)\right] \\
=& \Lambda(F(p_{\mathrm{B}} + cW) - F(p_{\mathrm{B}}/2 + cW))^2.
\end{aligned}$$

The stability condition requires $F(p_{\mathrm{B}} + cW) - F(p_{\mathrm{B}}/2 + cW) \to 0$ as $\Lambda \to \infty$, which implies $p_{\mathrm{B}} \to 0$. The bundle revenue is $R_{B,1} = p_{\mathrm{B}}(\lambda + \Lambda\bar{F}(cW + p_{\mathrm{B}})F(cW))$. Since $\lambda < \mu$ and $\Lambda\bar{F}(cW + p_{\mathrm{B}})F(cW) \leq \lambda < \mu$, it follows that $R_{B,1} < 2p_{\mathrm{B}}\mu$. As $\Lambda \to \infty$, $p_{\mathrm{B}} \to 0$, thus $R_{B,1} \to 0$.

Haviv and Randhawa [26] (Lemma 1) show that under à la carte pricing, the maximum revenue does not diminish as $\Lambda \to \infty$, i.e., $\lim_{\Lambda\to\infty} R_{\mathrm{A}} > 0$. Hence, there exists a finite threshold K_1 such that $R_{\mathrm{A}} > R_{B,1}$ for $\Lambda > K_1$.

Step 2: We next consider Problem B.1 as a relaxation problem of Problem 2.

Problem B.1

$$\hat{R}_{B,2} \triangleq \max_{p_{\mathrm{B}},W} \quad \Lambda p_{\mathrm{B}}\bar{G}(cW + p_{\mathrm{B}}/2), \quad \text{s.t.} \quad W = \frac{1}{\mu - \Lambda\bar{G}(cW + p_{\mathrm{B}}/2)}.$$

Notice that Problem 2 has one more constraint than Problem B.1, and therefore, $R_{B,2} \leq \hat{R}_{B,2}$.

We shall show that there exists a threshold K_2 such that $R_{\mathrm{A}} > \hat{R}_{B,2}$ if $\Lambda > K_2$.

We first state an auxiliary result in Lemma B.1, which follows from Ibragimov and Walden [27, Proposition B.1].

Lemma B.1 *If X_1 and X_2 are two i.i.d. random variables with a symmetric and log-concave probability density, then $\mathbb{P}\left(\frac{X_1+X_2}{2} - \alpha > x\right) < \mathbb{P}(X_1 - \alpha > x)$ for all $x > 0$ and $\mathbb{P}\left(\frac{X_1+X_2}{2} - \alpha > x\right) > \mathbb{P}(X_1 - \alpha > x)$ for all $x < 0$, where $\alpha = \mathbb{E}[X_1]$ is the mean of X_1.*

Recall that the firm's maximum à la carte revenue can be expressed as $R_{\mathrm{A}} = \max_\theta\ 2\Lambda\mathbb{P}(V_1 > \theta)(v - cW(\Lambda\mathbb{P}(V_1 > \theta)))$.

Let $\theta_B = cW + p_B/2$. Problem B.1 can be equivalently formulated as $\hat{R}_{B,2} = \max_{\theta_B} 2\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\left[\theta_B - cW\left(\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\right)\right]$, where V_1, V_2 are i.i.d. random variables following distribution F.

Any feasible θ_B must satisfy $\Lambda\mathbb{P}(\frac{V_1+V_2}{2} > \theta_B) < \mu$. In particular, if $\Lambda > K_2 \triangleq \mu/\mathbb{P}\left(\frac{V_1+V_2}{2} > \mathbb{E}[V_1]\right)$, then any feasible θ_B must satisfy $\theta_B > \mathbb{E}[V_1]$. Given θ_B, define $v_A(\theta_B) \triangleq \{v : \Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right) = \Lambda\mathbb{P}(V_1 > v)\}$. Since $\theta_B > \mathbb{E}[V_1]$ when $\Lambda > K_2$, Lemma B.1 implies $v_A(\theta_B) > \theta_B$ when $\Lambda > K_2$. Hence, when $\Lambda > K_2$,

$$\begin{aligned}
&2\Lambda\mathbb{P}(V_1 > v_A(\theta_B))\left[v_A(\theta_B) - cW(\Lambda\mathbb{P}(V_1 > v_A(\theta_B)))\right] \\
&= 2\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\left(v_A(\theta_B) - cW\left(\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\right)\right) \\
&> 2\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\left(\theta_B - cW\left(\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\right)\right).
\end{aligned}$$

It follows that when $\Lambda > K_2$,

$$\begin{aligned}
&\max_{\theta_B} 2\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\left(\theta_B - cW\left(\Lambda\mathbb{P}\left(\frac{V_1+V_2}{2} > \theta_B\right)\right)\right) \\
&< \max_{v_A} 2\Lambda\mathbb{P}(V_1 > v_A)(v_A - cW(\Lambda\mathbb{P}(V_1 > v_A))).
\end{aligned}$$

In other words, $R_A > \hat{R}_{B,2}$ when $\Lambda > K_2$. Since $\hat{R}_{B,2} \geq R_{B,2}$, it implies that $R_A > R_{B,2}$ if $\Lambda > K_2$. Since we have established in Step 1 that $R_A > R_{B,1}$ when $\Lambda > K_1$, it further follows that $R_A > R_B = \max\{R_{B,1}, R_{B,2}\}$ if $\Lambda > K_3 \triangleq \max\{K_1, K_2\}$, which completes the proof of Part (1).

Part (2): We express the maximum à la carte revenue as a function of delay sensitivity c:

$$R_A(c) = \sup_{v_A} R_A^1(v_A, c) \triangleq 2\Lambda(1 - v_A)\left(v_A - \frac{c}{\mu - \Lambda(1 - v_A)}\right), \quad \text{s.t.} \quad \mu > \Lambda(1 - v_A).$$

We shall show $\lim_{c\to 0} R_A(c) = R_A^0$, where $R_A^0 = \sup_{v_A} R_A^2(v_A) \triangleq 2\Lambda(1 - v_A)v_A$, s.t. $\mu > \Lambda(1 - v_A)$. First, note that $R_A(c) \leq R_A^0$ for all c. Hence, $\limsup_{c\to 0} R_A(c) \leq R_A^0$. Second, by definition of R_A^0, for any $\epsilon > 0$, there exists $\hat{v}_A$ satisfying $\mu > \Lambda(1 - \hat{v}_A)$ such that $R_A^2(\hat{v}_A) > R_A^0 - \epsilon/2$. Fix $\hat{v}_A$; consider $R_A^3(c) \triangleq R_A^1(\hat{v}_A, c) = 2\Lambda(1 - \hat{v}_A)\left(\hat{v}_A - \frac{c}{\mu - \Lambda(1-\hat{v}_A)}\right)$. Let $c \to 0$, it holds that $R_A^3(c) > R_A^2(\hat{v}_A) - \epsilon/2$ for sufficiently small c. Hence, $R_A^1(\hat{v}_A, c) = R_A^3(c) \geq R_A^0 - \epsilon$, implying $R_A(c) = \sup_{v_A} R_A^1(v_A, c) \geq R_A^0 - \epsilon$. Since ϵ is arbitrary, it follows that $\liminf_{c\to 0} R_A(c) \geq R_A$. We conclude that $\lim_{c\to 0} R_A(c) = R_A^0$.

We next solve R_{A}^0. If $\Lambda < 2\mu$, then $R_{\mathrm{A}}^0 = \Lambda/2$. If $\Lambda \geq 2\mu$, then the supremum can be achieved by letting $\Lambda(1 - v_A) = \mu$; $R_{\mathrm{A}}^0 = 2\mu[1 - \mu/\Lambda]$. Hence,

$$R_{\mathrm{A}}^0 = \begin{cases} \frac{\Lambda}{2}, & \Lambda < 2\mu; \\ 2\mu(1 - \mu/\Lambda), & \Lambda \geq 2\mu. \end{cases} \tag{B.1}$$

Similarly, the maximum bundle revenue $R_{\mathrm{B}}(c)$ as $c \to 0$ is $\lim_{c\to 0} R_{\mathrm{B}}(c) = R_{\mathrm{B}}^0 = \max\{R_{B,1}^0, R_{B,2}^0\}$, where $R_{B,1}^0$ and $R_{B,2}^0$ are obtained from the following problems: $R_{B,1}^0 = \sup_{p_{\mathrm{B}}} \Lambda p_{\mathrm{B}}(1 - p_{\mathrm{B}}^2/2)$, s.t. $p_{\mathrm{B}} \leq 1$, $\mu > \Lambda(1 - p_{\mathrm{B}}^2/2)$.; $R_{B,2}^0 = \sup_{p_{\mathrm{B}}} \Lambda p_{\mathrm{B}}(2 - p_{\mathrm{B}})^2/2$, s.t. $1 \leq p_{\mathrm{B}} \leq 2$, $\mu > \Lambda(2 - p_{\mathrm{B}})^2/2$. Solving these two problems gives

$$R_{B,1}^0 = \begin{cases} \frac{2\sqrt{6}}{9}\Lambda, & \Lambda < 3\mu/2; \\ \mu\sqrt{2(1 - \mu/\Lambda)}, & \Lambda \in [3\mu/2, 2\mu) \\ \text{infeasible}, & \Lambda \geq 2\mu; \end{cases} \quad R_{B,2}^0 = \begin{cases} \Lambda/2, & \Lambda < 2\mu; \\ \mu(2 - \sqrt{2\mu/\Lambda}), & \Lambda \geq 2\mu. \end{cases}$$

Since $R_{\mathrm{B}}^0 = \max\{R_{B,1}^0, R_{B,2}^0\}$, we have

$$R_B^0 = \begin{cases} \frac{2\sqrt{6}}{9}\Lambda, & \Lambda < 3\mu/2; \\ \mu\sqrt{2(1 - \mu/\Lambda)}, & \Lambda \in [3\mu/2, 2\mu); \\ \mu(2 - \sqrt{2\mu/\Lambda}), & \Lambda \geq 2\mu. \end{cases} \tag{B.2}$$

Combining (B.1) and (B.2), we have $R_{\mathrm{B}}^0 > R_{\mathrm{A}}^0$ if $\Lambda < 2\mu$. Since $\lim_{c\to 0} R_{\mathrm{A}}(c) = R_{\mathrm{A}}^0$ and $\lim_{c\to 0} R_{\mathrm{B}}(c) = R_{\mathrm{B}}^0$, it follows that if $\Lambda < 2\mu$, there exists δ such that for $c < \delta$, $R_{\mathrm{B}}(c) > R_{\mathrm{A}}(c)$. □

Proof of Proposition 3 Let $\Lambda \to \infty$. Under à la carte pricing,

$$R_{\mathrm{A}} = \sup_{v_A} \; 2\Lambda(1 - v_A)\left(v_A - \frac{c}{\mu - \Lambda(1 - v_A)}\right), \quad \text{s.t.} \quad \mu > \Lambda(1 - v_A).$$

In order for $\mu > \Lambda(1 - v_A)$ to hold, we must have $v_A \to 1$ as $\Lambda \to \infty$. The FOC gives

$$v_A = \frac{1}{2}\left[1 + \frac{c\mu}{(\mu - \Lambda(1 - v_A))^2}\right]. \tag{B.3}$$

Hence, $\Lambda(1 - v_A) = \mu - \sqrt{\frac{c\mu}{2v_A - 1}}$. Since $v_A \to 1$, $\Lambda(1 - v_A) \to \mu - \sqrt{c\mu}$.

$$\lim_{\Lambda\to\infty} R_{\mathrm{A}} = \lim_{\Lambda\to\infty} 2\Lambda(1 - v_A)\left[v_A - \frac{c}{\mu - \Lambda(1 - v_A)}\right] = 2\mu(1 - \sqrt{c/\mu})^2.$$

Under bundle pricing, we have shown in the proof of Theorem 1 that $R_{\mathrm{B}} = R_{B,2}$ since $R_{B,2} > R_{B,1}$ for sufficiently large Λ. $R_{B,2}$ is obtained from Problem 4. We consider a relaxation problem:

Problem B.2

$$\hat{R}_{B,2} = \max_{p_{\mathrm{B}}} \quad \Lambda p_{\mathrm{B}} \left(2 - 2cW - p_{\mathrm{B}}\right)^2/2,$$
$$\text{s.t.} \quad W = \frac{1}{\mu - \Lambda(2 - 2cW - p_{\mathrm{B}})^2/2}, \quad 2 - 2cW - p_{\mathrm{B}} \geq 0, \quad p_{\mathrm{B}} + 2cW \geq 1.$$

The only difference between Problems 2 and B.2 is that Problem 2 requires $p_{\mathrm{B}} + cW \geq 1$, whereas Problem B.2 only requires $p_{\mathrm{B}} + 2cW \geq 1$. Hence, $\hat{R}_{B,2} \geq R_{B,2}$.

Let $1-y = (2-2cW-p_{\mathrm{B}})^2/2 \in [0, 1/2]$. Then, $p_{\mathrm{B}}=2\left(1 - \sqrt{\frac{1-y}{2}} - \frac{c}{\mu-\Lambda(1-y)}\right)$.

$$\hat{R}_{B,2} = \max_{y\in[1/2,1]} 2\Lambda(1-y)\left(1 - \sqrt{\frac{1-y}{2}} - \frac{c}{\mu - \Lambda(1-y)}\right), \quad \text{s.t.} \quad \mu > \Lambda(1-y).$$

The unconstrained FOC gives $\frac{c\mu}{[\mu-\Lambda(1-y)]^2} = 1 - \frac{3\sqrt{2}}{4}\sqrt{1-y}$. For $\mu > \Lambda(1-y)$ to hold, we must have $y \to 1$ as $\Lambda \to \infty$. Hence, $1 - \frac{3\sqrt{2}}{4}\sqrt{1-y} \to 1$ as $\Lambda \to \infty$. From the FOC, $\lim_{\Lambda\to\infty} \frac{c\mu}{[\mu-\Lambda(1-y)]^2} = 1$. This implies $\lim_{\Lambda\to\infty} \Lambda(1-y) = \mu - \sqrt{c\mu}$. Plugging this into Problem B.2 gives $\lim_{\Lambda\to\infty} \hat{R}_{B,2} = 2\mu(1 - \sqrt{c/\mu})^2$. Further, $p_{\mathrm{B}} + cW = 2\left(1 - \sqrt{\frac{1-y}{2}}\right) - \frac{c}{\mu-\Lambda(1-y)}$. As $\Lambda \to \infty$, $y \to 1$, and $p_{\mathrm{B}} + cW \to 2 - \sqrt{c/\mu} \geq 1$. Hence, the relaxation Problem B.2 recovers the optimal solution to the original Problem 2. Therefore, $\lim_{\Lambda\to\infty} R_{B,2} = \lim_{\Lambda\to\infty} \hat{R}_{B,2} = 2\mu(1 - \sqrt{c/\mu})^2$. Hence, $\lim_{\Lambda\to\infty} \Delta(\Lambda) = 0$.

Also, Theorem 1 shows $\Delta(\Lambda) > 0$ for sufficiently large (but finite) Λ, which implies that $\Delta(\Lambda)$ cannot be monotone increasing in Λ. □

Proof of Proposition 4 *Step 1*: We consider sufficiently small Λ. Let $\Lambda \to 0$. Thus, $W \to c/\mu$ under either à la carte or bundle pricing. Under à la carte pricing, similar to the proof of Theorem 1, the scaled revenue R_{A}/Λ is $R_{\mathrm{A}}/\Lambda = \max_{v_A} 2(1 - v_A)(v_A - c/\mu)$. The first-order condition (FOC) gives $v_A = \frac{1}{2}\left(1 + \frac{c}{\mu}\right)$. Plugging it into the revenue function gives the scaled revenue R_{A}/Λ and price p_{A}: $R_{\mathrm{A}}/\Lambda = \frac{1}{2}\left(1 - \frac{c}{\mu}\right)^2$, $p_{\mathrm{A}} = v_A - c/\mu = \frac{1}{2}\left(1 - \frac{c}{\mu}\right)$. The equilibrium utilization satisfies

$$\frac{u_A}{\Lambda/\mu} = 1 - v_A = \frac{1}{2}\left(1 - \frac{c}{\mu}\right). \tag{B.4}$$

Under bundle pricing, the scaled revenue $R_{\mathrm{B}}/\Lambda = \max\{R_{B,1}/\Lambda, R_{B,2}/\Lambda\}$, where $R_{B,1}$ and $R_{B,2}$ are obtained from the following two subproblems, respectively.

Subproblem 1:

$$R_{B,1}/\Lambda = \max_{p_{\mathrm{B}}} \quad p_{\mathrm{B}}\left(1 - p_{\mathrm{B}}^2/2 - 2p_{\mathrm{B}}c/\mu - c^2/\mu^2\right), \quad \text{s.t.} \quad p_{\mathrm{B}} + c/\mu \le 1.$$

The unconstrained FOC gives $1 - (c/\mu)^2 - -4(c/\mu)p_{\mathrm{B}} - 3p_{\mathrm{B}}^2/2 = 0$. Solving the FOC gives $p_{\mathrm{B}} = \frac{1}{3}\left(-4c/\mu + \sqrt{6 + 10(c/\mu)^2}\right)$. Note that the other root of the FOC is negative and thus omitted. One can verify p_{B} above satisfies $p_{\mathrm{B}} + c/\mu \le 1$ and thus is the optimizer to the constrained problem. Plugging p_{B} into the revenue function gives the scaled revenue is $R_{B,1}/\Lambda = \frac{2}{27}\left(\sqrt{10(c/\mu)^2 + 6} - 4(c/\mu)\right)\left(3 + (c/\mu)^2 - (c/\mu)\sqrt{10(c/\mu)^2 + 6}\right)$. The equilibrium utilization u_B satisfies

$$\frac{u_B}{\Lambda/\mu} = 1 - p_{\mathrm{B}}^2/2 - c/\mu - p_{\mathrm{B}}c/\mu = \frac{1}{9}\left[6 - (c/\mu)^2 + \frac{c}{\mu}\left(\sqrt{10(c/\mu)^2 + 6} - 9\right)\right]. \tag{B.5}$$

Subproblem 2:

$$R_{B,2}/\Lambda = \max_{p_{\mathrm{B}}} \quad p_{\mathrm{B}}\left(2 - 2c/\mu - p_{\mathrm{B}}\right)^2/2, \quad \text{s.t.} \quad p_{\mathrm{B}} + c/\mu \ge 1,\ p_{\mathrm{B}} + 2c/\mu \le 2.$$

The unconstrained first derivative of the objective function with respect to p_{B} is $(2 - -2c/\mu - p_{\mathrm{B}})^2/2 - -2p_{\mathrm{B}}(2 - -2c/\mu - p_{\mathrm{B}}) = (2 - -2c/\mu - p_{\mathrm{B}})[(2 - -2c/\mu - p_{\mathrm{B}})^2 - -4p_{\mathrm{B}}]$. The first term $(2 - -2c/\mu - p_{\mathrm{B}})$ is nonnegative, so the sign of the derivative is the same as the sign of the second term $[(2 - -2c/\mu - p_{\mathrm{B}})^2 - -4p_{\mathrm{B}}]$. The second term is decreasing in p_{B}. When $p_{\mathrm{B}} = 1 - c/\mu$, the second term is $(2 - -2c/\mu - (1 - c/\mu))^2 = 4(1 - c/\mu) = (1 - c/\mu)(1 - c/\mu - 4) < 0$.

Therefore, the first derivative is negative for $p_{\mathrm{B}} \in [1 - c/\mu, 2 - -2c/\mu]$. The revenue function is decreasing in $p_{\mathrm{B}} \in [1 - c/\mu, 2 - -2c/\mu]$. Thus, the maximum of the revenue function is attained at $p_{\mathrm{B}} = 1 - c/\mu$. Hence, $R_{B,1}/\Lambda \ge R_{B,2}/\Lambda$; $R_{\mathrm{B}}/\Lambda = R_{B,1}/\Lambda$.

We have shown that $2p_{\mathrm{A}} = 1 - c/\mu$ and that $p_{\mathrm{B}} = \frac{1}{3}\left(-4c/\mu + \sqrt{6 + 10(c/\mu)^2}\right)$. One can verify that $2p_{\mathrm{A}} > p_{\mathrm{B}}$ always holds for all $c/\mu < 1$.

Step 2: We consider sufficiently large Λ. The FOC for à la carte pricing (B.3) gives

$$\frac{c\mu}{(\mu - \Lambda(1 - v_A))^2} = 1 - 2(1 - v_A), \tag{B.6}$$

Letting $\lambda_A = \Lambda(1 - v_A)$, we rewrite (B.6) as $c\mu/(\mu - \lambda_A)^2 = 1 - 2\lambda_A/\Lambda$. The optimal à la carte price is

$$p_{\mathrm{A}} = v_A - \frac{c}{\mu - \Lambda(1 - v_A)}$$

$$= 1 - \frac{\lambda_A}{\Lambda} - \frac{c}{\mu - \lambda_A} \overset{(a)}{=} 1 - \frac{1 - \frac{c\mu}{(\mu-\lambda_A)^2}}{2} - \frac{c}{\mu - \lambda_A}$$
$$= \frac{1}{2} + \frac{c(2\lambda_A - \mu)}{2(\mu - \lambda_A)^2},$$

where equality (a) follows from $c\mu/(\mu - \lambda_A)^2 = 1 - 2\lambda_A/\Lambda$. Define $Y_A(\lambda) = \frac{1}{2} + \frac{c(2\lambda-\mu)}{2(\mu-\lambda)^2}$ which is increasing in λ since $Y_A'(\lambda) = \frac{c\lambda}{(\mu-\lambda)^3} > 0$.

The FOC for bundle pricing gives

$$\frac{c\mu}{(\mu - \Lambda(1-y))^2} = 1 - \frac{3\sqrt{2}}{4}\sqrt{1-y}. \tag{B.7}$$

Note that as we argued in the proof of Proposition 3, when Λ is sufficiently large, (B.7) gives the optimal solution to the original constrained bundle pricing problem. Letting $\lambda_B = \Lambda(1-y)$, we rewrite (B.7) as $c\mu/(\mu - \lambda_B)^2 = 1 - \frac{3\sqrt{2}}{4}\sqrt{\lambda_B/\Lambda}$. The optimal bundle price satisfies

$$\frac{p_{\mathrm{B}}}{2} = 1 - \sqrt{\frac{1-y}{2}} - \frac{c}{\mu - \Lambda(1-y)} = 1 - \sqrt{\frac{\lambda_B}{2\Lambda}} - \frac{c}{\mu - \lambda_B} \overset{(b)}{=} 1 - \frac{2}{3}\left[1 - \frac{c\mu}{(\mu - \lambda_B)^2}\right]$$
$$- \frac{c}{\mu - \lambda_B} = \frac{1}{3} + \frac{c(3\lambda_B - \mu)}{3(\mu - \lambda_B)^2},$$

where (b) follows from $c\mu/(\mu - \lambda_B)^2 = 1 - \frac{3\sqrt{2}}{4}\sqrt{\lambda_B/\Lambda}$.

Define $Y_B(\lambda) = \frac{1}{3} + \frac{c(3\lambda-\mu)}{3(\mu-\lambda)^2}$ which is increasing in λ since $Y_B'(\lambda) = \frac{c(\mu+3\lambda)}{3(\mu-\lambda)^3} > 0$. We next show $Y_A(\lambda_A) \geq Y_B(\lambda_A) > Y_B(\lambda_B)$ which implies $2p_{\mathrm{A}} > p_{\mathrm{B}}$. One can show λ_A is increasing in Λ and has a limit $\mu - \sqrt{c\mu}$ as $\Lambda \to \infty$, and so $\lambda_A \leq \mu - \sqrt{c\mu}$. Comparing $Y_A(\lambda_A)$ with $Y_B(\lambda_B)$ gives

$$Y_A(\lambda_A) - Y_B(\lambda_A) = \frac{1}{6} - \frac{1}{6}\frac{c\mu}{(\mu - \lambda_A)^2} \geq \frac{1}{6} - \frac{1}{6}\frac{c\mu}{(\mu - (\mu - \sqrt{c\mu}))^2} = 0.$$

In Proposition 5, we show $\lambda_A > \lambda_B$ for sufficiently large Λ. Since $Y_B(\lambda)$ is increasing in λ, it follows that $Y_B(\lambda_A) > Y_B(\lambda_B)$.

Finally, we consider the limit of p_{A} and p_{B} as $\Lambda \to \infty$. Recall that $p_{\mathrm{A}} = v_A - c/(\mu - \lambda_A)$ and $\frac{p_{\mathrm{B}}}{2} = 1 - \sqrt{\frac{1-y}{2}} - \frac{c}{\mu - \Lambda(1-y)}$. Since $\lambda_A \to \mu - \sqrt{c\mu}$ and $v_A \to 1$ as $\Lambda \to \infty$, we have $\lim_{\Lambda\to\infty} p_{\mathrm{A}} = 1 - \sqrt{c/\mu}$. Since $y \to 1$ and $\Lambda(1-y) \to \mu - \sqrt{c\mu}$ as $\Lambda \to \infty$, we have $\lim_{\Lambda\to\infty} p_{\mathrm{B}}/2 = 1 - \sqrt{c/\mu}$. □

Proof of Proposition 5 *Step 1*: We consider sufficiently small Λ. From (B.4), the equilibrium utilization u_A under à la carte pricing satisfies $\frac{u_A}{\Lambda/\mu} = \frac{1}{2}\left(1 - \frac{c}{\mu}\right)$. From (B.5), the equilibrium utilization u_B under bundle pricing satisfies $\frac{u_B}{\Lambda/\mu} = \frac{1}{9}\left[6 - (c/\mu)^2 + \frac{c}{\mu}\left(\sqrt{10(c/\mu)^2 + 6} - 9\right)\right]$. Now, we shall show $u_B/(\Lambda/\mu) > u_A/(\Lambda/\mu)$. Let $x \triangleq c/\mu \in (0,1)$. We show, for any $x \in (0,1)$,

$\frac{1}{9}\left[6-x^2+x\left(\sqrt{10x^2+6}-9\right)\right] > \frac{1}{2}(1-x)$. We first show that for $x \in (0,1)$, $\sqrt{10x^2+6} > \frac{5x+3}{2}$. This is equivalent to showing $10x^2+6 > \frac{(5x+3)^2}{4}$, which is further equivalent to showing $40x^2+24 > (5x+3)^2$. Collecting terms gives $15(1-x)^2 > 0$, which holds for $x \in (0,1)$. Therefore, $\sqrt{10x^2+6} > \frac{5+3x}{2}$. We have $\frac{1}{9}\left[6-x^2+x\left(\sqrt{10x^2+6}-9\right)\right] > \frac{1}{9}\left[6-x^2+x\left(\frac{5+3x}{2}-9\right)\right] = \frac{1}{6}(4--5x+x^2)$. Next, we show $(4-5x+x^2)/6 > (1-x)/2$, or equivalently, $4--5x+x^2 > 3(1-x)$. For $x \in (0,1)$, $(4--5x+x^2)-3(1-x) = 1--2x+x^2 = (x-1)^2 > 0$. Hence, by continuity, $u_B > u_A$ for sufficiently small Λ.

Step 2: We consider sufficiently large Λ. In the FOC (B.6) and (B.7), the left-hand sides of (B.6) and (B.7) are the same and decreasing in their corresponding argument (v_A and y, respectively), and the right-hand sides of (B.6) and (B.7) are increasing in their corresponding argument. As Λ gets sufficiently large, both v_A and y will be sufficiently close to 1. Hence, $v_A < y$ if and only if $1-2(1-x) > 1-\frac{3\sqrt{2}}{4}\sqrt{1-x}$ for x close to 1. To this end, let $f_1(x) = 1--2(1-x)$ and $f_2(x) = 1-\frac{3\sqrt{2}}{4}\sqrt{1-x}$. First note that $f_1(1) = f_2(1) = 1$. Next, $\frac{df_1(x)}{dx} = 2$, $\frac{df_2(x)}{dx} = \frac{3}{3\sqrt{2}\sqrt{1-x}}$. $\lim_{x\to 1^-}\frac{df_2(x)}{dx} = \infty > 2$. Hence, there exists $\bar{x} \in (0,1)$ such that $f_1(x) > f_2(x)$ for $x \in (\bar{x},1)$. This further implies that for sufficiently large Λ, $v_A < y$. Since $u_A = \Lambda(1-v_A)/\mu$ and $u_B = \Lambda(1-y)/\mu$, it follows that $u_A > u_B$ for sufficiently large Λ. □

Proof of Proposition 6 *Step 1*: We consider sufficiently small Λ. Define $x = c/\mu \in (0,1)$. Then,

$$CS_{\mathrm{A}}/\Lambda = 2\int_{v_A}^{1}(v-v_A)dv = \frac{1}{4}(1-x)^2.$$
$$CS_{\mathrm{B}}/\Lambda = \int_{\underline{v}}^{\bar{v}}\int_{\underline{v}}^{\bar{v}}\{[v_1-x]^+ + [v_2-x]^+ - p_B\}^+ dF(v_1)dF(v_2),$$

where $v_A = (1+c/\mu)/2$ and $p_{\mathrm{B}} = \frac{1}{3}\left(-4x+\sqrt{6+10x^2}\right)$ by the proof of Proposition 4. Simple computation yields

$$CS_{\mathrm{B}}/\Lambda = \frac{1}{81}\left(34x^3 - 16\sqrt{2}\sqrt{5x^2+3}x^2 + 81x^2 - 24\sqrt{2}\sqrt{5x^2+3} - 36x + 81\right).$$

Thus,

$$CS_{\mathrm{B}}/\Lambda - CS_{\mathrm{A}}/\Lambda = \frac{1}{324}\left(136x^3+243x^2+18x+243-(96+64x^2)\sqrt{10x^2+6}\right).$$

Since

$$\begin{aligned}&(136x^3+243x^2+18x+243)^2 - [(96+64x^2)\sqrt{10x^2+6}]^2\\ &= 27(1-x)^3\left(832x^3+48x^2+741x+139\right) > 0,\end{aligned}$$

it follows that $CS_{\mathrm{B}}/\Lambda > CS_{\mathrm{A}}/\Lambda$ as $\Lambda \to 0$.

Step 2: We consider sufficiently large Λ. Let λ_A and λ_B be the optimal joining rate for each service under à la carte and bundle pricing, respectively. Under à la carte, the cutoff valuation for each service is $v_A = 1 - \lambda_A/\Lambda$, and so the consumer surplus is $CS_{\mathrm{A}} = 2\Lambda \int_{v_A}^{1} (v - v_A)dv = \lambda_A^2/\Lambda$. Under bundle pricing, the cutoff valuation for the bundle is $v_B = 2 - \sqrt{2\lambda_B/\Lambda}$, and so the consumer surplus is $CS_{\mathrm{B}} = \Lambda \int_{v_B}^{2} (v - v_B)(2 - v)dv = \frac{\Lambda}{6}(2\lambda_B/\Lambda)^{3/2}$, where $(2 - v)$ is the PDF of bundle valuation for $v > 1$. Now,

$$\frac{CS_{\mathrm{A}}}{CS_{\mathrm{B}}} = \frac{3}{\sqrt{2}} \frac{\lambda_A^2}{\lambda_B^{3/2}} \Big/ \sqrt{\Lambda} < 1 \text{ for sufficiently large } \Lambda.$$

Since $\lambda_A \to \mu - \sqrt{c\mu} > 0$ and $\lambda_B \to \mu - \sqrt{c\mu} > 0$ as $\Lambda \to \infty$, it follows that $CS_{\mathrm{A}}/CS_{\mathrm{B}} \to 0$ as $\Lambda \to \infty$, or equivalently, $\lim_{\Lambda\to\infty} CS_{\mathrm{B}}/CS_{\mathrm{A}} = \infty$. □

Proof of Proposition 7 From (B.1), the à la carte profit as a function of μ is

$$\Pi_A^0(\mu) = \begin{cases} \frac{\Lambda}{2} - 2k\mu, & \Lambda < 2\mu; \\ 2\mu(1 - \mu/\Lambda) - 2k\mu, & \Lambda \geq 2\mu. \end{cases}$$

For $\mu > \Lambda/2$, $\Pi_A^0(\mu)$ is decreasing in μ, which implies the optimal capacity should be no greater than $\Lambda/2$. For $\mu \leq \Lambda/2$, the FOC with respect to μ gives $2 - -4\mu/\Lambda - 2k = 0$. This gives $\mu = (1-k)\Lambda/2 \in (0, \Lambda/2)$ for $k < 1$. Plugging it into the profit function gives $\Pi_A^0 = \Lambda(1 - k)^2/2 > 0$. Hence, for $k < 1$, the optimal à la carte capacity is $\mu_{\mathrm{A}} = (1 - k)\Lambda/2$; for $k \geq 1$, the firm cannot generate a positive profit.

From (B.2), the bundling profit as a function of μ is

$$\Pi_B^0(\mu) = \begin{cases} \frac{2\sqrt{6}}{9}\Lambda - 2k\mu, & \Lambda < 3\mu/2; \\ \mu\sqrt{2(1 - \mu/\Lambda)} - 2k\mu, & \Lambda \in [3\mu/2, 2\mu); \\ \mu(2 - \sqrt{2\mu/\Lambda}) - 2k\mu, & \Lambda \geq 2\mu. \end{cases}$$

For $\mu > 2\Lambda/3$, $\Pi_B^0(\mu)$ is decreasing in μ, which implies the optimal capacity should be no greater than $2\Lambda/3$. We thus focus on $\mu \in (0, 2\Lambda/3)$ and discuss two cases depending on the relative value of k and $1/4$.

Case $k \leq 1/4$. For $\mu \leq \Lambda/2$, $\Pi_B^0(\mu) = \mu(2 - \sqrt{2\mu/\Lambda}) - 2k\mu$ and the unconstrained FOC gives $\mu = 8(1 - k)^2\Lambda/9 \geq \Lambda/2$. Hence, $\Pi_B^0(\mu)$ is increasing in μ for $\mu \in (0, \Lambda/2]$. The optimal capacity is attained for $\mu \in [\Lambda/2, 2\Lambda/3]$, where $\Pi_B^0(\mu) = \mu\sqrt{2(1 - \mu/\Lambda)} - 2k\mu$. The first derivative gives $\frac{d\Pi_B^0(\mu)}{d\mu} = \frac{(2 - -3\mu/\Lambda)}{\sqrt{2}\sqrt{1-\mu/\Lambda}} - 2k = \left[3\sqrt{1 - \mu/\Lambda} - \frac{1}{\sqrt{1-\mu/\Lambda}}\right]/\sqrt{2} - 2k$. By inspection, $d\Pi_B^0(\mu)/d\mu$ is decreasing in μ/Λ. The maximum is obtained at $\mu = \Lambda/2$, at which point $d\Pi_B^0(\mu)/d\mu \mid_{\mu=\Lambda/2} = 1/2 - 2k \geq 0$ since $k \leq 1/4$. The minimum $d\Pi_B^0(\mu)/d\mu \mid_{\mu=2\Lambda/3} = -2k < 0$. Hence, the optimal capacity μ_{B} solving the FOC must satisfy $\mu_{\mathrm{B}} \geq \Lambda/2$. Since

$\mu_A = (1-k)\Lambda/2$, it follows that $\mu_B > \mu_A$. Bundling is profitable in this case because $\Pi_B^0(\mu_B) \geq \Pi_B^0(\Lambda/2) = \Lambda(1/2-k) > 0$.

Case $k > 1/4$. For $\mu \leq \Lambda/2$, $\Pi_B^0(\mu) = \mu(2-\sqrt{2\mu/\Lambda}) - 2k\mu$ and the unconstrained FOC gives $\mu = 8(1-k)^2\Lambda/9 < \Lambda/2$. Hence, the optimal capacity in this region is $8(1-k)^2\Lambda/9$. Later we show this capacity is also globally optimal. Since $\Pi_B^0(8(1-k)^2\Lambda/9) = 2(1-k)/3$, bundling generates positive profits if $k < 1$. Recall that the optimal à la carte capacity is $\mu_A = (1-k)\Lambda/2$. Hence, $\mu_B > \mu_A$ if and only if $8(1-k)^2\Lambda/9 > (1-k)\Lambda/2$, which is equivalent to $k < 7/16$. Now, we show $\mu_B = 8(1-k)^2\Lambda/9$ is globally optimal. It suffices to show the optimal capacity is achieved for $\mu \in (0, \Lambda/2]$. For $\mu \in (\Lambda/2, 2\Lambda/3]$, the first derivative $d\Pi_B^0(\mu)/d\mu$ is decreasing in μ/Λ. The maximum is obtained at $\mu = \Lambda/2$, at which point $d\Pi_B^0(\mu)/d\mu \mid_{\mu=\Lambda/2} = 1/2 - 2k < 0$ since $k > 1/4$. Thus, $\frac{d\Pi_B^0(\mu)}{d\mu} < 0$ for all $\mu \in (\Lambda/2, 2\Lambda/3]$; $\Pi_B^0(\mu)$ is decreasing in μ for $\mu \in (\Lambda/2, 2\Lambda/3]$. Therefore, the optimal μ_B is in region $(0, \Lambda/2]$. □

Appendix C: Relative revenue difference in product bundling

We consider the classical setting of product bundling (without congestion) in which the marginal cost of production is κ. Customer valuation is uniformly distributed over $[0, 1]$.

À la carte pricing $R_A = \max_{p_A} \quad 2(1-p_A)(p_A - \kappa)$. The FOC gives $p_A = \frac{1}{2}(1+\kappa), \quad R_A = \frac{1}{2}(1-\kappa)^2$.

Bundle pricing

Subproblem 1: $R_{B,1} = \max_{p_B} \quad (1 - p_B^2/2)(p_B - 2\kappa), \quad \text{s.t.} \quad p_B \leq 1.$

Subproblem 2: $R_{B,2} = \max_{p_B} \quad \frac{1}{2}(2-p_B)^2(p_B - 2\kappa), \quad \text{s.t.} \quad p_B \geq 1.$

The unconstrained FOC of subproblem 1 gives $1 + 2\kappa p_B - 3p_B^2/2 = 0$. When $\kappa > 1/4$, the revenue function is always increasing in p_B. Otherwise, $p_B = \frac{1}{3}(2\kappa + \sqrt{6+4\kappa^2})$. and $R_{B,1} = \frac{2}{27}\left(\sqrt{4\kappa^2+6} - 4\kappa\right)\left(3 - -2\kappa^2 - \kappa\sqrt{4\kappa^2+6}\right), \ \kappa < 1/4$. The unconstrained FOC of subproblem 2 gives $p_B = \frac{2+4\kappa}{3}$. For $\kappa < 1/4$, $R_{B,2}$ is no better than $R_{B,1}$. Otherwise, $R_{B,2} = \frac{16}{27}(1-\kappa)^3, \ \kappa > 1/4$. In sum,

$$R_B = \begin{cases} \frac{2}{27}\left(\sqrt{4\kappa^2+6} - 4\kappa\right)\left(3 - 2\kappa^2 - \kappa\sqrt{4\kappa^2+6}\right), & \kappa < 1/4; \\ \frac{16}{27}(1-\kappa)^3, & \kappa > 1/4. \end{cases}$$

$$p_B = \begin{cases} \frac{1}{3}(2\kappa + \sqrt{6+4\kappa^2}), & \kappa < 1/4; \\ \frac{2+4\kappa}{3}, & \kappa > 1/4. \end{cases}$$

Figure 11 plots the relative revenue difference $(R_A - R_B)/R_B \times 100\%$ as a function of the marginal cost and shows that it is monotone increasing.

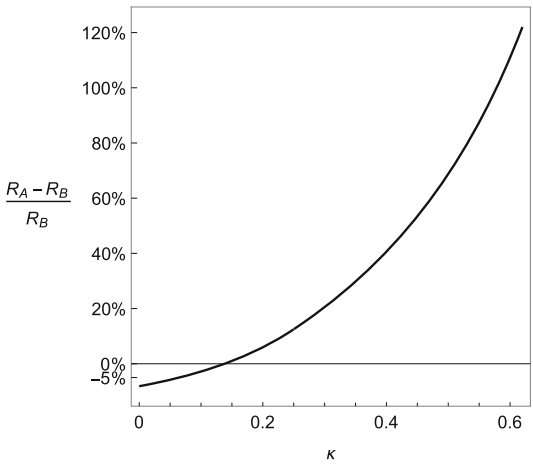
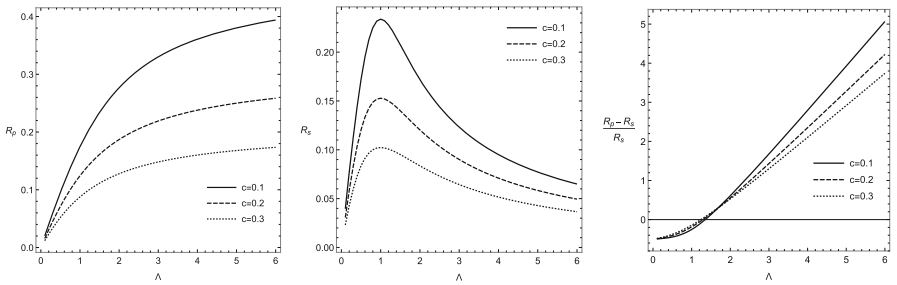


Fig. 11 The relative revenue difference in classical product bundling as a function of the marginal cost



(a) Pay-per-use (b) Subscription pricing (c) Relative revenue difference

Fig. 12 Revenue comparison of subscription pricing and pay-per-use. Customer valuation uniformly distributed over [0, 1], $\mu = 1$

Appendix D: Relative revenue difference between subscription pricing and pay-per-use

This section compares subscription pricing with the pay-per-use scheme, following Cachon and Feldman [11]. We first recap the revenue functions of each pricing scheme. According to Cachon and Feldman [11], the maximum revenue of pay-per-use is

$$R_p \triangleq \max_v \Lambda \bar{F}(v)[v - cW(\Lambda \bar{F}(v))].$$

The maximum revenue of subscription pricing is

$$R_s \triangleq \max_{v_s} \Lambda \bar{F}(v_s)(\mathbb{E}[V|V \geq v_s] - v_s),$$

where $v_s = cW(\Lambda \bar{F}(v_s)))$.

Figure 12 plots the maximum revenue of each pricing scheme and the relative revenue difference as a function of potential arrival rate Λ (under a uniform valuation distribution). We make the following three observations.

1. We observe from Fig. 12a that the revenue of pay-per-use is increasing in Λ, as expected.
2. We observe from Fig. 12b that the revenue of subscription pricing is non-monotone in Λ with an inverse U-shaped relationship. In fact, one can analytically show that the revenue of subscription pricing tends to zero as Λ goes to infinity. This non-monotone relationship is driven by the inability of subscription pricing to control congestion (an effect highlighted by [11]).
3. We observe from Fig. 12c that the relative revenue difference between pay-per-use and subscription pricing is increasing in Λ (i.e., the percentage revenue loss from suboptimally choosing subscription pricing grows with Λ). Note that this observation contrasts the comparison between bundle pricing and à la crate pricing, where the relative revenue difference is non-monotone in Λ, and bundle pricing becomes largely identical to à la carte pricing in revenue performance when Λ is large.

Funding Open access funding provided by Hong Kong University of Science and Technology Chenguang (Allen) Wu acknowledges support from Hong Kong General Research Fund (Grant 16506122).

References

1. Adams, W.J., Yellen, J.L.: Commodity bundling and the burden of monopoly. Q. J. Econ. **90**, 475–498 (1976)
2. Afèche, P.: Decentralized service supply chains with multiple time-sensitive customer segments: pricing, capacity, and coordination. Working paper, University of Toronto (2013)
3. Afèche, P., Baron, O., Milner, J., Roet-Green, R.: How to charge and prioritize time-sensitive customers with heterogeneous demands. Oper. Res. **67**(4), 1184–1208 (2019)
4. Allon, G., Federgruen, A.: Competition in service industries. Oper. Res. **55**(1), 37–55 (2007)
5. Anand, K.S., Paç, M.F., Veeraraghavan, S.: Quality-speed conundrum: trade-offs in customer-intensive services. Manag. Sci. **57**(1), 40–56 (2011)
6. Bagnoli, M., Bergstrom, T.: Log-concave probability and its applications. Econ. Theor. **26**(2), 445–469 (2005)
7. Bakos, Y., Brynjolfsson, E.: Bundling information goods: pricing, profits, and efficiency. Manag. Sci. **45**(12), 1613–1630 (1999)
8. Banciu, M., Gal-Or, E., Mirchandani, P.: Bundling strategies when products are vertically differentiated and capacities are limited. Manag. Sci. **56**(12), 2207–2223 (2010)
9. Bhargava, H.K.: Retailer-driven product bundling in a distribution channel. Mark. Sci. **31**(6), 1014–1021 (2012)
10. Burstein, M.L.: The economics of tie-in sales. Rev. Econ. Stat. **42**(1), 68–73 (1960)

11. Cachon, G.P., Feldman, P.: Pricing services subject to congestion: charge per-use fees or sell subscriptions? Manuf. Serv. Oper. Manag. **13**(2), 244–260 (2011)
12. Cachon, G.P., Harker, P.T.: Competition and outsourcing with scale economies. Manag. Sci. **48**(10), 1314–1333 (2002)
13. Cao, Q., Stecke, K.E., Zhang, J.: The impact of limited supply on a firm's bundling strategy. Prod. Oper. Manag. **24**(12), 1931–1944 (2015)
14. Catto, S.: Travel advisory; Toronto's top draws in a single package. New York Times (July 11, 2004), https://www.nytimes.com/2004/07/11/travel/travel-advisory-toronto-s-top-draws-in-a-single-package.html (2004)
15. Chakravarty, A., Mild, A., Taudes, A.: Bundling decisions in supply chains. Eur. J. Oper. Res. **231**(3), 617–630 (2013)
16. Chen, H., Frank, M.: Monopoly pricing when customers queue. IIE Trans. **36**(6), 569–581 (2004)
17. Chen, Y., Riordan, M.H.: Profitability of product bundling. Int. Econ. Rev. **54**(1), 35–57 (2013)
18. Chen, H., Yao, D.D.: Kelly networks. In: Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization, pp. 69–96. Springer, New York, New York, NY (2001)
19. Cui, Y., Duenyas, I., Sahin, O.: Unbundling of ancillary service: how does price discrimination of main service matter? Manuf. Serv. Oper. Manag. **20**(3), 455–466 (2018)
20. Edelson, N.M., Hilderbrand, D.K.: Congestion tolls for Poisson queuing processes. Econ. J. Econ. Soc. **43**, 81–92 (1975)
21. Fang, H., Norman, P.: To bundle or not to bundle. Rand J. Econ. **37**(4), 946–963 (2006)
22. Guiltinan, J.P.: The price bundling of services: a normative framework. J. Mark. **51**(2), 74–85 (1987)
23. Guo, P., Tang, C.S., Wang, Y., Zhao, M.: The impact of reimbursement policy on social welfare, revisit rate, and waiting time in a public healthcare system: fee-for-service versus bundled payment. Manuf. Serv. Oper. Manag. **21**(1), 154–170 (2019)
24. Hassin, R.: Rational Queueing. CRC Press, Boca Raton, FL (2016)
25. Hassin, R., Haviv, M.: To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, vol. 59. Springer Science & Business Media, New York, NY (2003)
26. Haviv, M., Randhawa, R.S.: Pricing in queues without demand information. Manuf. Serv. Oper. Manag. **16**(3), 401–411 (2014)
27. Ibragimov, R., Walden, J.: Optimal bundling strategies under heavy-tailed valuations. Manag. Sci. **56**(11), 1963–1976 (2010)
28. Kelly, F.P.: Networks of queues with customers of different types. J. Appl. Probab. **12**(3), 542–554 (1975)
29. Kelly, F.P.: Networks of queues. Adv. Appl. Probab. **8**(2), 416–432 (1976)
30. Kelly, F.P.: Reversibility and Stochastic Networks. Cambridge University Press, New York, NY (2011)
31. Levine, A.: What is an e-ticket ride? TripSavvy (April 12, 2018), https://www.tripsavvy.com/what-is-an-e-ticket-ride-3225791 (2018)
32. Littlechild, S.C.: Optimal arrival rate in a simple queueing system. Int. J. Prod. Res. **12**(3), 391–397 (1974)
33. Maister, D.H.: The psychology of waiting lines. In: Czepiel, J.A., Solomon, M.R., Surprenant, C.F. (eds.) The Service Encounter, pp. 113–123. Lexington Books, Lexington, MA (1985)
34. McAfee, R.P., McMillan, J., Whinston, M.D.: Multiproduct monopoly, commodity bundling, and correlation of values. Q. J. Econ. **104**(2), 371–383 (1989)
35. McCardle, K.F., Rajaram, K., Tang, C.S.: Bundling retail products: models and analysis. Eur. J. Oper. Res. **177**(2), 1197–1217 (2007)
36. Naor, P.: The regulation of queue size by levying tolls. Econ. J. Econ. Soc. **37**, 15–24 (1969)
37. Oi, W.Y.: A Disneyland dilemma: two-part tariffs for a mickey mouse monopoly. Q. J. Econ. **85**(1), 77–96 (1971)
38. Quezada, Z.: Buffet of buffets with caesars entertainment in las vegas. TripSavvy (May 15, 2017), https://www.tripsavvy.com/buffet-of-buffets-caesars-palace-4135976 (2017)
39. Randhawa, R.S., Kumar, S.: Usage restriction and subscription services: operational benefits with rational users. Manuf. Serv. Oper. Manag. **10**(3), 429–447 (2008)
40. Schmalensee, R.: Gaussian demand and commodity bundling. J. Bus. **57**, S211–S230 (1984)
41. Veltman, A., Hassin, R.: Equilibrium in queueing systems with complementary products. Queueing Syst. **50**(2), 325–342 (2005)

42. Venkatesh, R., Mahajan, V.: The design and pricing of bundles: a review of normative guidelines and practical approaches. In: Rao, V,R,, (ed.) *Handbook of Pricing Research in Marketing*, chapter 11, 232–257. Edward Elgar Publishing Inc., MA (2009)
43. Venkatesh, R., Kamakura, W.: Optimal bundling and pricing under a monopoly: contrasting complements and substitutes from independently valued products. J. Bus. **76**(2), 211–231 (2003)
44. Ziese, A.: 8 incredible queue lines you mustn't skip at Walt Disney World. Theme Park Tourist (June 19, 2014), https://www.themeparktourist.com/features/20140616/18602/8-incredible-queues-you-mustnt-skip-walt-disney-world (2014)