Original Article

POMS

Production and Operations Management
1–16

DOI: 10.1177/10591478241234994
journals.sagepub.com/home/pao
Sage

Add-On Pricing: A Queueing Perspective

Chenguang (Allen) Wu[1], Chen Jin[2] and Ying-Ju Chen[3]

Abstract

This work is motivated by the practice of add-on services, where an add-on is not valuable unless purchased with a main service. Discrepancies in pricing have been observed in various settings such as restaurants, museums, and attractions regarding whether the add-on should be sold together with the main service, or separately from the main service at an additional charge. While there has been a vast literature on add-on pricing, its application in service-oriented businesses with congestion-prone externalities and delay-sensitive customers is less understood. We develop a queueing model and examine the optimal pricing of add-on services in such systems, and in line with practice, we focus on analyzing two pricing schemes: bundling that charges a single price to sell main and add-on services altogether, and separate selling that charges distinct prices for each service. We establish that in the absence of congestion, separate selling strictly dominates bundling across the board. When there is congestion at the main service but not the add-on, bundling can be more lucrative under a large customer demand. When congestion takes place at both services, separate selling can return to being dominant under a large customer demand. We explain these plots of reversals by illustrating an intricate interplay between pricing, market coverage, and congestion. Collectively, they reveal novel operational advantages of each pricing scheme in exploiting the fundamentals of add-on structures.

Keywords

Add-on services, congestion, queueing game, pricing

Date received 19 August 2022; accepted 1 November 2023 after two revisions

Handling Editor: Michael Pinedo

1 Introduction

In various contexts, service firms provide add-on services to complement their main service. Unlike the main service which can be purchased alone, most add-ons do not create value unless purchased and consumed together with the main service. Two competing pricing schemes have been proposed to sell the main and add-on services, one selling all services together, and the other selling add-ons separately from the main service at additional charges. The debate between these two pricing schemes has been a central theme in the add-on pricing literature. To put this into context, consider that some steakhouses (mostly low-tier chains) offer complementary sides (e.g., pasta) as free supplements to their main dishes (e.g., steak), whereas, others (often high-end brand-name ones), offer complementary sides at additional charges. In a different context, the Statue of Liberty[1] sells admission tickets that include ferry service, audio tour (i.e., main service), and access to the Ellis Island National Museum of Immigration (i.e., add-on), whereas the Hong Kong Palace Museum[2] and M+ Museum[3] sell special events and exhibitions in separate tickets and exclude them from the regular admission, and Madame Tussauds New York[4] unbundles its digital photo pass and 7D experience with its general admission to 200+ lifelike wax figures.

Despite the discrepancy in pricing observed in these examples, they share a salient feature, that is, a congestion effect, that is prevalent in service-oriented businesses and is well known to adversely impact customers' service experience. For example, preparing the main dish in a steakhouse is a queueing process that involves congestion-prone delays, and customers don't like long waiting for their orders to be delivered. Likewise, visitors incur disutility from on-site

[1]Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
[2]Department of Information Systems and Analytics, School of Computing, National University of Singapore, Singapore
[3]School of Business and Management (ISOM), Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

Corresponding author:
Chen Jin, Department of Information Systems and Analytics, School of Computing, National University of Singapore, 15 Computing Drive, 117417, Singapore.
Email: disjinc@nus.edu.sg

congestion on their visits to exhibitions or attractions.[5] Within these examples, despite a common congestion effect at the main service, discrepancies can also exist in whether congestion takes place at the add-on service too. Specifically, although main dishes in steakhouses must be cooked freshly, complementary sides such as salads can be stocked beforehand and served immediately upon a customer's order. In contrast, congestion often exists at both general and special exhibitions and causes disutility to participating visitors whoever visit them.

Our paper goes after such a unique perspective of congestion in certain service industries and revisits the classic add-on pricing for congestion-prone systems. Incorporating congestion into the add-on pricing analysis has subtle implications. Specifically, for a given price of the firm, how much congestion a customer has to experience by joining a service depends on how many other customers are present, and the level of congestion generated in the system is endogenously determined by customers' own interactions. This implies a monopoly firm aiming to maximize revenue not only faces the usual price-demand trade-off, but also must appropriately pull the pricing lever to regulate system congestion.

To fix ideas, we consider a monopoly firm selling a main service and an add-on service in a queueing context with delay-sensitive customers. All customers value the main service, but only a portion of them are interested in the add-on and the rest do not value it at all. We refer to the former customers as *high-type* and the latter as *low-type*. In line with both practice and the convention in the add-on literature, we focus on analyzing two pricing schemes: *bundling* that charges a single price to sell the main and add-on services altogether, and *separate selling* that charges distinct prices for each service. We first develop a base model without congestion and show that separate selling strictly dominates bundling across the board. We demonstrate how separate selling manages to create efficient price discrimination.

The congestion effect, if incorporated, however, can significantly alter the firm's pricing strategy. If there is congestion at the main service but not the add-on, bundling can be more lucrative under a large customer demand. When both services are congested, we consider a tandem queue model in which the main service is processed first (our formal result does not rely on this assumption though) and show that depending on the add-on capacity, separate selling may return to being dominant under a large customer demand.

The intuition behind these two plots of reversals can be explained as follows. When the main service is congested but the add-on is not, the queueing cost at the main service is a critical component of the firm's profit margin and crucially determines the firm's market coverage. Under a large customer demand, a firm practicing separate selling allows low-type customers to join the main service and occupy its limited capacity without further purchasing the add-on. Bundling, in contrast, can effectively screen out low-type customers and save limited capacity for high-type customers exclusively who become favorable targets to generate add-on revenues. We show that the effect of a limited capacity is further amplified by an endogenous queueing cost, thereby making bundling superior in a wider range of market sizes.

The story changes quickly when the add-on is congested too. We find that an extremely limited capacity of the add-on can drive away the superiority of bundling across the board, and even so under a large market size. In this case, we identify another important operational benefit of separate selling when both services are subject to congestion-prone delays: by charging a distinct price to sell the add-on, separate selling regulates the traffic flow of high-type customers and eliminates over-congestion at the add-on due to their self-interest. Bundling, in contrast, is less efficient in doing so as it charges zero nominal price for the add-on. This allows high-type customers to possibly over-join the add-on, at a rate significantly higher than the firm's optimal rate, creating over-congestion that leads to loss of revenue.

We extend our analysis in two dimensions, namely, independent valuations of main and add-on services among high-type customers, as well as endogenous capacities, and find that our main insights qualitatively extend. When the main service is congested while the add-on is not, we find that the market size required to overturn the dominance of separate selling is generally larger under independent valuations than under positively correlated valuations. When both services are congested, the effect of independent valuations is more intricate, yielding highly non-trivial comparison results between two pricing schemes (see Section 5.1 for more details). With endogenous capacities, we find that separate selling is optimal in a very broad range of cost parameters, whereas bundling is only optimal in a restricted space of cost parameters. We demonstrate the opposite effects of capacity cost and delay cost on the firm's optimal capacity level and their joint effect on the firm's optimal pricing strategy.

1.1 Related Literature

There is a vast literature on add-on pricing, and a central theme in this literature is to identify the optimal pricing scheme to sell main and add-on services depending on the context, and more specifically, whether the add-on should be sold together with the main service in a bundle, or separately from the main service at an additional charge (Ellison, 2005; Shulman and Geng, 2013; Geng et al., 2018; Cui et al., 2018). However, less is known about add-on pricing in service systems featured by congestion externalities and delay-sensitive customers. Notably, the congestion effect is known to adversely impact customers' service experience and their willingness to pay in the first place. In this work, we revisit add-on pricing in a queueing context and examine how such an effect will drive a monopoly firm's optimal pricing strategy.

Our work is also related to the research stream on product bundling that dates back to Adams and Yellen (1976). The notion of bundling has been widely studied in various fields such as economics (Fang and Norman, 2006), marketing (Ibragimov and Walden, 2010), operations (Wu et al., 2022), and information systems (Bakos and Brynjolfsson, 1999; Jin

et al., 2022). We contribute to this literature by clarifying how the congestion effect in service systems may critically overturn the normative prescriptions between separate selling and bundling in this stream of research.

Related to our work, Cao et al. (2015) considers a firm's bundling decision for two independent products, with one product having a limited capacity. The authors show that the optimal (un)bundling decision depends on both the limited capacity and the correlation between product valuations. Unlike Cao et al. (2015), we focus on add-on services, and show that the congestion effect can be a dominant driver of the firm's (un)bundling decisions regardless of correlations between service valuations (see our main model in Sections 3.2 and 4, as well as the extension in Section 5.1). In the absence of congestion, the closest paper to ours is Cui et al. (2018), which is motivated by the practice of the travel industry that charges discriminatory fees for flights based on customers' status or purchase history. Without such price discrimination, the authors show that separate selling is more likely to be adopted when customers' valuations of main and add-on services are positively correlated. In contrast, we focus on a queueing setting and demonstrate exactly the opposite to Cui et al. (2018). We show that the parameter space where separate selling is dominant can shrink under positively correlated valuations. We explain our finding by connecting it to an intricate interplay between pricing, congestion, and customers' self-interested behaviors.

Because we study add-on pricing in a queueing context, our work is also related to the queueing literature on congestion pricing; see Hassin and Haviv (2003) for a comprehensive survey of this literature. Most of this literature focuses on the optimal pricing of a single service. Notable exceptions include Wu and Yang (2018) (see also references therein on other multiservice studies), which examines the bundling problem for two independent services (i.e., each service can be consumed without the other), each serving a common pool of customers subject to its own congestion. The authors show that bundling dominates separate selling when congestion is light at both services, and separate selling dominates otherwise. Unlike Wu and Yang (2018), we focus on add-on services and we demonstrate the critical role of add-on structures in driving our key results. Specifically, the (endogenous) queueing cost in our setting is fundamentally different from an exogenous marginal cost in their revenue implications (see Section 3.2), whereas they are found to have similar directional effects on the revenue comparison between separate selling and bundling in the context of Wu and Yang (2018). See also Section 5.1 for an extended comparison between our work and Wu and Yang (2018) via a numerical study.

2 Model

We consider a monopoly service firm (he) selling a main service and an add-on service. The main service can be purchased alone, but the add-on service does not have any value unless purchased together with the main service. We assume that customers' (she) valuations of the main service V_M are heterogeneous and uniformly distributed over $[0, \bar{v}]$. An α fraction has a need for the add-on service: a customer with valuation v for the main service values the add-on at βv $(0 < \beta < 1)$.[6] We refer to these customers with positive valuations of the add-on as *high-type* customers. The other $1 - \alpha$ fraction of customers, referred to as *low-type* customers, do not value the add-on at all. We use the generic random variable V_A to represent the entire valuations of the add-on service. Our two-segment assumption is indicative of the fact that add-on services are optional supplements to the main service and their popularity is often not so broad as the main service. For example, a diner interested in a fine steak may find complementary sides unnecessary. Visitors interested in general exhibitions (which often have a broad interest) may not be keen on special exhibitions (which often have a narrower audience). Similar assumptions on customer segments have been commonly made in the add-on literature (Shulman and Geng, 2013; Geng et al., 2018; Cui et al., 2018).

Our assumption on high-type customers' add-on valuations warrants additional discussion. First, we assume $\beta < 1$ to be reflective of the fact that most add-ons support the functionalities of a main service but are not able to fully replace it. Thus, customers typically do not value the add-on as much as the main service. Second, we assume a (perfect) positive correlation between high-type customers' valuations of main and add-on services (Dey et al., 2021). Income levels, customer loyalty, and complementarity between two services are all possible factors that contribute to such a correlation. Moreover, assuming this correlation affords a tractable analysis that exposes the critical role of queueing and congestion in the firm's optimal pricing strategy. Admittedly, this assumption drives away the well-known effect of reduced valuation dispersion established in the product bundling literature (Adams and Yellen, 1976; Fang and Norman, 2006). Incorporating this effect into our model, however, will not alter our main results. Indeed, we examine in Section 5.1 an extension in which high-type customers have independent valuations for two services, thereby allowing such effect to exist. Our main results continue to hold in that extension, suggesting some robustness of our results in the general add-on setting.

In line with practice and convention in the add-on literature, we study two pricing schemes to sell the main and add-on services, namely, *bundling* that charges a single price to sell two services altogether, and *separate selling* that charges distinct prices for each service. (Note that *mixed bundling* that works for independent services does not apply to add-on services because add-ons, by their nature, cannot be purchased alone; e.g., Dey et al. 2021.[7]) To build intuitions on how these two pricing schemes compare, we first develop a benchmark model without congestion effects.

2.1 No Congestion Effect

In the absence of congestion, we recover the classic add-on model, and for completeness, we present its main

results below. We denote the size of customers by Λ, the price under bundling by p_B, and prices of main and add-on services under separate selling by p_M and p_A, respectively.

In this formulation, bundling is a special case of separate selling by offering the add-on "for free": separate selling reduces to bundling if one sets $p_M = p_B$ and $p_A = 0$. To rule out this triviality, we impose $p_A > 0$ under separate selling throughout the paper, that is, the add-on should be purchased at a strictly positive price. Moreover, we impose $p_M < \bar{v}$, because otherwise if $p_M \geq \bar{v}$, the main service is highly costly, and all customers, irrespective of their types, receive negative utilities from solely purchasing the main service. Only high-type customers will be interested in the main service in the hope that they will receive high utilities from the add-on which are good enough to compensate for their disutilities from the main service. In other words, the add-on is effectively tied to the main service to be sold to high-type customers exclusively, making separate selling effectively equivalent to bundling (with bundle price $p_B = p_M + p_A \geq \bar{v}$). We also rule out this in our analysis to fully differentiate separate selling from bundling.

ASSUMPTION 1. *Feasible prices under separate selling are such that $p_M < \bar{v}$ and $p_A > 0$.*

Assumption 1 is imposed to ensure that a nonnegligible fraction of low-type customers will purchase the main service under separate selling (Pang and Etzion, 2012). None of them will purchase the costly add-on because they don't value it at all. Thus, in terms of market coverage, separate selling aims to sell the main service to both customer segments and sell the add-on to high-type customers only.

2.1.1 Bundling. The firm charges p_B to sell a bundle composed of main and add-on services. With zero marginal costs of serving additional customers, the firm solves

$$\max_{p_B} \Pi^B(p_B) = p_B \Lambda \mathbb{P}\{V_M + V_A \geq p_B\}. \tag{1}$$

2.1.2 Separate Selling. The firm charges p_M to sell the main service and p_A to sell the add-on. To characterize the firm's revenue, we first compute the demand D_M that purchases the main service exclusively and the demand D_{MA} that purchases both services.

$$\begin{aligned} \sup_{p_M<\bar{v}, p_A>0} \Pi^S(p_M,p_A) &= p_M(D_M + D_{MA}) + p_A D_{MA} \qquad (2)\\ \text{s.t.} \quad D_M &= \Lambda \mathbb{P}\{V_M - p_M \geq 0,\ V_M - p_M > V_M + V_A \\ &\quad - p_M - p_A\} = \Lambda \mathbb{P}\{V_M \geq p_M, V_A < p_A\}, \\ D_{MA} &= \Lambda \mathbb{P}\{V_M + V_A - p_M - p_A \geq [V_M - p_M]^+\} \\ &= \Lambda \mathbb{P}\{V_M + V_A \geq p_M + p_A, V_A \geq p_A\}. \end{aligned}$$

We characterize the optimal prices under these two pricing schemes below.

PROPOSITION 1. *Suppose there is no congestion effect.*

(i) *The optimal bundle price and the corresponding revenue are*

$$p_B^* = \frac{(1+\beta)\bar{v}}{2[1+\beta(1-\alpha)]},\ \Pi^B(p_B^*) = \frac{\Lambda(1+\beta)\bar{v}}{4[1+\beta(1-\alpha)]}.$$

(ii) *The optimal prices and revenue under separate selling are*

$$p_M^* = \frac{\bar{v}}{2},\ p_A^* = \frac{\beta\bar{v}}{2},\ \Pi^S(p_M^*,p_A^*) = \frac{\Lambda(1+\alpha\beta)\bar{v}}{4}.$$

(iii) *Separate selling strictly dominates bundling.*

Part (i) of Proposition 1 shows that the optimal bundle price is always chosen to be less than $\bar{v}$, implying that a firm experimenting with bundling should always adopt a *volume* strategy by selling the main service to both customer segments, a strategy justified by the main service's high profit margin. Thus, the firm covers both customer segments regardless of whether he adopts bundling or separate selling.

Despite a similar market coverage, part (iii) shows that separate selling strictly outperforms bundling across the board. To explain, note that separate selling can flexibly charge an affordable price to sell the main service and an extra price to sell the add-on only to those who value it highly. Through two prices, separate selling prompts customers' self-selected purchase decisions differentiated by types, using price discrimination to generate more efficient market segmentations. In contrast, bundling charges a uniform price irrespective of customers' types, allowing high-type customers to free ride and pay as much as low-type customers to receive the add-on for free. As a result, separate selling outperforms bundling across the board.

We remark that a similar comparison result continues to hold even if high-type customers have independent valuations for main and add-on services. Despite the well-known effect of reduced valuation dispersion in the product bundling literature (Adams and Yellen, 1976; Fang and Norman, 2006), one can show that *bundling is nevertheless strictly dominated by separate selling for add-on services*. This suggests a dominant effect of the add-on structure on profit comparison between the two pricing schemes.

3 Congestion at Main Service

We next develop a queueing model to incorporate the common congestion effect in service-oriented systems. We start by considering a model where the main service is capacity-constrained subject to congestion-prone delays whereas the add-on is not.[8] In the context of restaurants, complementary sides and drinks (i.e., add-ons) are often prepared beforehand. Main dishes (i.e., main service) must be freshly cooked upon receipt of orders, and preparing those dishes may involve congestion-prone delays. We model the main service as an $M/M/1$ queue to capture the general notion of congestion effect. To be aligned with the benchmark model in

Section 2.1, we assume that the demand of the main service arises in a Poisson process with rate Λ and we refer to Λ as the *market size*. Services are processed in a First-Come-First-Served manner. The processing times of each main service (with or without the add-on) are exponentially distributed with rate μ_M and we refer to μ_M as the *capacity* of the main service. Customers are delay-sensitive and incur a cost c per unit of time spent in the system (including time in service). To rule out triviality, we assume $\bar{v} > c/\mu_M$ throughout the paper.

We assume that customers are rational and can accurately estimate the *expected* wait time at each service should they join that service. Queues are unobservable when they make purchase decisions. This assumption is suited to a number of environments where customers' purchase decisions are made before visiting a service location. For example, it applies to settings where customers order meals online and pick them up offline, as well as to settings where visitors purchase admission tickets in advance before paying visits to an exhibition.

3.1 Optimal Pricing Decision

We start by analyzing the firm's bundling strategy. We first illustrate how to transform the firm's optimization problem into a tractable one that is amenable to analysis. This also lays out a pathway to solving the ensuing more complex separate selling problem.

3.1.1 Bundling. Let p_B denote the bundle price and W_M denote the expected wait time at the main service in equilibrium. A customer purchases (and joins) the main service if

$$V_M - cW_M + V_A - p_B \geq 0.$$

In equilibrium, customers' joining decisions will create a wait time consistent with their belief. This leads to the firm's bundling problem formulated as follows:

$$\begin{aligned} \max_{p_B} \ & \Pi^B(p_B) = p_B \Lambda \mathbb{P}\{V_M + V_A \geq p_B + cW_M\} \qquad (3) \\ \text{s.t.} \ & W_M = \frac{1}{\mu_M - \Lambda \mathbb{P}\{V_M + V_A \geq p_B + cW_M\}}, \\ & \mu_M - \Lambda \mathbb{P}\{V_M + V_A \geq p_B + cW_M\} > 0. \end{aligned}$$

Optimizing (3) over price p_B requires solving a fixed-point equation for the equilibrium wait time W_M (corresponding to the first constraint) under each p_B. This renders a direct analysis intractable. To overcome this challenge, we follow a standard approach in the queueing literature (Edelson and Hilderbrand, 1975; Hassin and Haviv, 2003) and convert (3) to an equivalent optimization problem that optimizes over the cut-off valuation s of the main service. Specifically, because customers pay the same price to access the bundle and incur the same waiting cost when joining, there exists a cut-off valuation s such that only customers with valuations $V_M + V_A$ higher than s will purchase the bundle. Then, using s we can rewrite the bundle price $p_B = s - cW_M$. The one-to-one correspondence between s and p_B allows us to transform (3) to the following:

PROBLEM 1.

$$\begin{aligned} \max_{s} \ & \Pi_1^B(s) = (s - cW_M)\Lambda \mathbb{P}\{V_M + V_A \geq s\} \\ s.t. \ & W_M = \frac{1}{\mu_M - \Lambda \mathbb{P}\{V_M + V_A \geq s\}}, \\ & \mu_M - \Lambda \mathbb{P}\{V_M + V_A \geq s\} > 0. \end{aligned}$$

To have a tractable characterization of the optimal bundle price, we further convert the above problem to a new equivalent one that optimizes over the demand of the main service $t := \Lambda \mathbb{P}(V_M + V_A \geq s)$. This conversion allows us to establish a close connection between pricing and market coverage, which, as we explain shortly, is the key to understanding the effect of congestion on the profitability of each pricing scheme.

PROPOSITION 2. *There exists $(c/\mu_M) \leq v_1 \leq v_2$ (expressions of v_1 and v_2 can be found in Online Appendix B and are both increasing in c) such that the optimal bundle price satisfies:*

- *When $\mu_M \leq \alpha\beta\Lambda/(1+\beta)$, we have $p_B^* = \bar{v}(1+\beta)(\alpha\Lambda - t_1)/(\alpha\Lambda) - c/(\mu_M - t_1)$.*
- *When $\mu_M > \alpha\beta\Lambda/(1+\beta)$, we have*

$$p_B^* = \begin{cases} \dfrac{\bar{v}(1+\beta)(\alpha\Lambda - t_1)}{\alpha\Lambda} - \dfrac{c}{\mu_M - t_1}, & \text{if } \dfrac{c}{\mu_M} < \bar{v} \leq v_1, \\ \begin{cases} \dfrac{\bar{v}(1+\beta)(\alpha\Lambda - t_1)}{\alpha\Lambda} - \dfrac{c}{\mu_M - t_1}, \\ \dfrac{\bar{v}(1+\beta)(\Lambda - t_2)}{\Lambda[1+\beta(1-\alpha)]} - \dfrac{c}{\mu_M - t_2}, \end{cases} & \text{if } F_1(t_1) > F_2(t_2), \\ & \text{if } v_1 < \bar{v} < v_2, \\ \dfrac{\bar{v}(1+\beta)(\Lambda - t_2)}{\Lambda[1+\beta(1-\alpha)]} - \dfrac{c}{\mu_M - t_2}, & \text{otherwise}, \\ & \text{if } v \geq v_2. \end{cases}$$

where

$$F_1(t) := \left(\frac{\bar{v}(1+\beta)(\alpha\Lambda - t)}{\alpha\Lambda} - \frac{c}{\mu_M - t}\right) t$$

and

$$F_2(t) := \left(\frac{\bar{v}(1+\beta)(\Lambda - t)}{\Lambda[1+\beta(1-\alpha)]} - \frac{c}{\mu_M - t}\right) t$$

represent the firm's revenue under demands $t < \alpha\beta\Lambda/(1+\beta)$ *and* $t \geq \alpha\beta\Lambda/(1+\beta)$*, respectively, and* $t_1 \in (0, \min\{\mu_M, \alpha\beta\Lambda/(1+\beta)\})$ *and* $t_2 \in (\alpha\beta\Lambda/(1+\beta), \min\{\Lambda, \mu_M\})$ *are the unique solutions to the following equations, respectively:*

$$\begin{cases} \left.\dfrac{\mathrm{d}F_1(t)}{\mathrm{d}t}\right|_{t=t_1} = -\left[\dfrac{c\mu_M}{(\mu_M - t_1)^2} + \bar{v}(1+\beta)\left(\dfrac{2t_1}{\alpha\Lambda} - 1\right)\right] = 0, \\ \left.\dfrac{\mathrm{d}F_2(t)}{\mathrm{d}t}\right|_{t=t_2} = -\left[\dfrac{c\mu_M}{(\mu_M - t_2)^2} + \dfrac{\bar{v}(1+\beta)(2t_2 - \Lambda)}{\Lambda[1+\beta(1-\alpha)]}\right] = 0. \end{cases} \tag{4}$$

Proposition 1 is driven by a classic price–demand trade-off as it focuses on the firm's optimal prices restricted to the bundling strategy. When the main service has a high profit margin, expanding market coverage is crucial, similar to the benchmark model without congestion. In essence, by setting $\mu_M = \infty$, Proposition 2 recovers Proposition 1. Proposition 2 further generalizes Proposition 1 by showing that, when capacity μ_M is large (relative to market size Λ) and delay cost c is small (relative to service valuation $\bar{v}$), selling the main service remains highly lucrative so that the firm should continue to use a volume strategy and sell the main service to both customer segments (this corresponds to demand t_2 with $p_B + cW_M < \bar{v}$).

However, as congestion grows due to either a large market size Λ (relative to capacity μ_M) or a large delay cost c (relative to service valuation $\bar{v}$), it prompts a switch in the firm's coverage strategy. This time, because each purchasing customer has to incur a sizable queueing cost when joining the main service, this cost is passed on to the firm as an implicit marginal cost. When this cost is substantial, it heavily cuts into the main service's profit margin. As a result, using a low bundle price to reach low-type customers is no longer optimal; instead, the firm sets a high price to fully screen out low-type customers and target high-type customers exclusively (this corresponds to $p_B + cW_M \geq \bar{v}$ with demand t_1). Unlike a usual *exogenous* marginal cost, the queueing cost in our setting bears a unique characteristic: it is a result of congestion externality jointly determined by the firm's price and customers' mutual interactions. Such endogeneity has intricate pricing and profit implications different from an exogenous marginal cost.[9]

3.1.2 Separate Selling. Separate selling is more complex to analyze as it entails involved market segmentations. Let p_M and p_A denote the prices of main and add-on services, respectively, and W_M denote the equilibrium expected wait time at the main service. Similar to Assumption 1, we impose $p_M + cW_M < \bar{v}$ and $p_A > 0$ to fully differentiate separate selling from bundling.

Each customer has three options (recall that purchasing the add-on standalone is not a feasible option): purchasing the main service alone with payoff $V_M - cW_M - p_M$; purchasing both the main and add-on services with payoff $V_M - cW_M - p_M + V_A - p_A$; and forgoing purchase with zero payoff. Each customer selects the option that generates the highest payoff, leading to the firm's optimization problem formulated as follows:

$$\begin{aligned} \sup_{p_M + cW_M < \bar{v}, p_A > 0} \quad & \Pi^S(p_M, p_A) = p_M(D_M + D_A) + p_A D_{MA} \\ \text{s.t.} \quad & D_M = \Lambda\mathbb{P}\{V_M \geq p_M + cW_M,\ V_A < p_A\}, \\ & D_{MA} = \Lambda\mathbb{P}\{V_M + V_A - p_M - p_A \\ & \qquad - cW_M \geq 0,\ V_A \geq p_A\}, \\ & W_M = \frac{1}{\mu_M - D_M - D_{MA}},\ \mu_M - D_M - D_{MA} > 0. \end{aligned} \tag{5}$$

Similar to the case of bundling, (5) in itself is intractable because it requires solving a fixed-point equation for the wait time W_M under each (p_M, p_A). We thus convert (5) to an equivalent problem that optimizes over cut-off valuations. Note that for a high-type customer with valuation v of the main service, her payoff of purchasing the main service taking into account the add-on is $v - p_M - cW_M + (\beta v - p_A)^+$, which is strictly increasing in v. So, there exists a unique v_0 (if any) such that $v_0 - p_M - cW_M + (\beta v_0 - p_A)^+ = 0$. Such observation motivates us to define two cut-off valuations $s_M = p_M + cW_M < \bar{v}$ and $s_A = p_A/\beta > 0$ to be optimized in the new problems. It then follows that $v_0 + \beta(v_0 - s_A)^+ = s_M$. Depending on the relative value of s_M and s_A, there are two subcases to consider.

(1) If $s_M \leq s_A$, then high-type customers' cut-off valuation $v_0 = s_M$. In this case, high-type customers purchasing the main service will partially purchase the add-on. Demands of main and add-on services are $\Lambda[1 - s_M/\bar{v}]$ and $\alpha\Lambda[1 - s_A/\bar{v}]$, respectively, leading to

Problem 2.

$$\begin{aligned} \sup_{s_M \leq s_A < \bar{v}} \quad & \Pi_1^{S,1}(s_M, s_A) = (s_M - cW_M)\Lambda\left[1 - \frac{s_M}{\bar{v}}\right] \\ & \qquad + \beta s_A \cdot \alpha\Lambda\left[1 - \frac{s_A}{\bar{v}}\right] \\ \text{s.t.} \quad & W_M = \frac{1}{\mu_M - \Lambda\left[1 - s_M/\bar{v}\right]}, \\ & \mu_M - \Lambda\left[1 - \frac{s_M}{\bar{v}}\right] > 0. \end{aligned}$$

(2) If $s_M > s_A$, then all high-type customers purchasing the main service will continue to purchase the add-on. The cut-off valuation v_0 solves $v_0 + \beta(v_0 - s_A) = s_M$, which gives $v_0 = (s_M + \beta s_A)/(1+\beta) \in (s_A, s_M)$. Demands of

main and add-on services are $\Lambda(\alpha[1 - v_0/\bar{v}] + (1 - \alpha)[1 - s_M/\bar{v}])$ and $\alpha\Lambda[1 - v_0/\bar{v}]$, respectively, leading to

PROBLEM 3.

$$\sup_{0<s_A<s_M<\bar{v}} \Pi_1^{S,2}(s_M, s_A)$$
$$= (s_M - cW_M)\Lambda\left(\alpha\left[1 - \frac{v_0}{\bar{v}}\right] + (1-\alpha)\left[1 - \frac{s_M}{\bar{v}}\right]\right)$$
$$+ \beta s_A \cdot \alpha\Lambda\left[1 - \frac{v_0}{\bar{v}}\right]$$

s.t.

$$W_M = \frac{1}{\mu_M - \Lambda\left(\alpha\left[1 - (v_0/\bar{v})\right] + (1-\alpha)\left[1 - (s_M/\bar{v})\right]\right)},$$
$$\mu_M - \Lambda\left(\alpha\left[1 - \frac{v_0}{\bar{v}}\right] + (1-\alpha)\left[1 - \frac{s_M}{\bar{v}}\right]\right) > 0.$$

The firm's optimal separate selling scheme can be analyzed by first solving Problems 2 and 3, respectively, and then taking the solution with a higher objective value. In our analysis, we further transform Problems 2 and 3 to ones that optimize over the demands of each service denoted by t and z, respectively, and then use these quantities to characterize the optimal prices.

PROPOSITION 3. *The optimal prices for the main and add-on services* (p_M^*, p_A^*) *are*

$$\begin{cases} p_M^* = \bar{v}\left(\frac{z^S - t^S}{\Lambda(1-\alpha)} + 1\right) - \frac{c}{\mu_M - t^S}, \\ p_A^* = \beta\bar{v}\left(\frac{\alpha t^S - z^S}{\alpha\Lambda\beta(1-\alpha)} - \frac{z^S}{\alpha\Lambda} + 1\right). \end{cases}$$

where

$$(t^S, z^S) = \begin{cases} (t_1, t_1), & \text{if } 0 < \mu_M \le \frac{\alpha\beta\Lambda}{2(1+\beta)}, \\ (t_1, t_1), & \text{if } \frac{c}{\mu_M} < \bar{v} \le \frac{c\mu_M}{\left(\mu_M - \frac{\alpha\beta\Lambda}{2(1+\beta)}\right)^2}, \\ \left(t_2, \frac{\alpha}{2}\left(\Lambda + \frac{2t_2 - \Lambda}{1+\beta(1-\alpha)}\right)\right), & \text{if } \bar{v} > \frac{c\mu_M}{\left(\mu_M - \frac{\alpha\beta\Lambda}{2(1+\beta)}\right)^2}, \end{cases} \quad \text{if } \mu_M > \frac{\alpha\beta\Lambda}{2(1+\beta)},$$

where t_1 *and* t_2 *are given by* (4) *in Proposition* 2.

When the main service has a limited capacity (relative to the market size), or the valuation of the main service is not significantly high (relative to the delay cost), the optimal prices under separate selling induce equal demands of both main and add-on services, $t^S = z^S = t_1$. Thus, all customers purchasing the main service will purchase the add-on too, and they must all belong to the high-type. (This corresponds to a limiting case in which $p_M^* + cW_M^*$ hits $\bar{v}$ from below.) In this case, the add-on is tied to the main service to be sold, effectively ending separate selling and making it equivalent to bundling.[10] In contrast, separate selling can have its own appeal when the main service is only lightly congested. When the capacity μ_M is relatively large (relative to the market size) and the valuation of the main service is relatively high (relative to the delay cost), the firm sets a low price to sell the main service to both customer segments. Because low-type customers don't purchase the add-on, this leads to $t^S > z^S$. This observation facilitates an insightful revenue comparison between two pricing schemes which we elaborate on in the next section.

3.2 Comparison

Having derived the optimal prices under each pricing scheme, we next compare their revenues.

THEOREM 1. *Suppose there is congestion at the main service. Bundling dominates when* $\Lambda \ge ((1+\beta)(\mu_M - \sqrt{c\mu_M/\bar{v}}))/\alpha\beta$*; otherwise, separate selling strictly dominates.*

Some clarification of Theorem 1 is in order. The dominance of bundling stated in Theorem 1 when the market size is large or the delay cost is high, is not strict in the sense that the optimal revenue under bundling can also be achieved under separate selling in the limiting case ($p_M + cW_M \to \bar{v}$). However, the dominance of separate selling stated in Theorem 1 otherwise is strict in the sense that the optimal revenue under separate selling is strictly higher than that under bundling.

Theorem 1 suggests a potential pitfall in using separate selling for price discrimination in congested service systems. Recall from Proposition 1 that separate selling in general can create profitable market segmentations in the absence of congestion. Theorem 1 then imposes a condition for such benefit to extend to service systems with congestion externalities:

an ample capacity of the main service and customers' low sensitivity to delays, both collectively contributing to a low queueing cost. In this case, exploiting the main service's high margin to cover both customer types remains lucrative, preserving the dominance of separate selling. However, the value of separate selling in enforcing price discrimination dwindles and eventually vanishes as the main service becomes increasingly congested. The add-on queueing structure implies that demands of main and add-on services are both capped by the capacity of the main service adjusted by a queueing-related term (see the expression in Theorem 1). To explain more, the capacity term ensures a stable system in operations and the queueing term describes customers' endogenous interactions in the formation of an equilibrium. Theorem 1 shows that in case of considerable congestion driven by either a large market size Λ (relative to capacity μ) or a high delay cost c (relative to valuation $\bar{v}$), the firm has to abandon separate selling. Because otherwise, using separate selling to serve both customer segments only invites low-type customers to occupy the already limited capacity of the main service without further purchasing the add-on. A better strategy in this case is to save the limited capacity for high-type customers who will then become favorable targets to generate add-on revenues. Bundling manages to screen out low-type customers and restrict sales to high-type customers exclusively. This latter point can be observed by noting that demands of both main and add-on services correspond to t_1 when bundling is optimal, and so only high-type customers will purchase and join the main service.

Our result points to a unique perspective of the add-on structure in congested queueing settings. Consider a slightly different context with two *independent* services (without one being ancillary to the other). If only one service has a limited capacity but not the other (Cao et al., 2015), then the resulting dominant pricing scheme is very different from the one in Theorem 1. Specifically, a large customer demand will drive away the superiority of bundling in this new setting because demands under bundling are always constrained by the limited capacity of one service. Separate selling, however, can be significantly more profitable because the firm can sell the service with unlimited capacity to sufficiently many customers.

Note further that the boundary of market size where separate selling and bundling switch in Theorem 1 is decreasing in both the value of the add-on β and the portion of high-type customers α. A high β implies that selling the add-on brings better revenue, and so the firm has better incentives to save the capacity of the main service for high-type customers exclusively. A high α also erodes the value of separate selling and allows bundling to dominate in a wider range of market sizes. To see this, note that the optimal bundle price in Proposition 1 in the absence of congestion is increasing in α, that is, the firm can exercise a higher bundle price as the add-on becomes more popular in the market. This partially mitigates the inefficiency of bundling in price discrimination, and so separate selling is most valuable under small α. The same reasoning extends to the case with congestion, and the space where separate selling is optimal has to shrink as α increases.

The switching boundary is decreasing in the delay cost c. This implies that the effect of limited capacity is *amplified* by a queueing cost. To see why, note that a queueing cost exists and hints that the capacity of the main service cannot be fully utilized regardless of the pricing scheme adopted. A larger queueing cost further limits the capacity available to use, resulting in stronger incentives for exclusive targeting.

Our finding sheds light on the food industry that commonly bundles main dishes with complementary sides. This practice is most prevalent in low-tier restaurants but less relevant to high-end ones. In general, upscale restaurants target high-value niche markets, and cheap restaurants aim at small profit margins but quick turnovers. Such strategic positioning suggests that the former often faces low customer demands and is generally less capacity-constrained than the latter. Theorem 1 prescribes separate selling for the former and lightly congested upscale restaurants, as confirmed by a number of century-old brand-name steakhouses such as Peter Luger in New York and Jimmy's Kitchen in Hong Kong.[11] In contrast, cheap steakhouses with high volumes of demand are advised to sell sides, drinks, and puff pastry altogether with the main dish in order to fully utilize the limited kitchen capacity.

4 Congestion at Both Main and Add-on Services

In this section, we develop a queueing model in which both the main and add-on services are capacity-constrained subject to congestion-prone delays. For example, some university canteens set up different lines for collecting meals and complimentary drinks. Special exhibitions are open to visitors who have purchased a ticket to the general exhibition. Motivated by such practice, we model the queueing flow of the main and add-on services as a tandem queue. Each main and add-on service request requires a processing time that is exponentially distributed with rate μ_M and μ_A, respectively. We assume $\beta\bar{v} > c/\mu_A$ to rule out trivial cases but place no restrictions on the relative value of μ_M and μ_A. Our analysis does not rely on a specific sequence in which a customer visits two services[12], but for ease of exposition and in line with our motivating examples, we assume that the main service is always visited first.

When both services are congested, each customer has to incur a waiting cost for each service she joins. In particular, a high-type customer joining both services has to incur two separate queueing costs. In other words, bundling does not offer the add-on "for free": although it charges a zero *nominal price* for the add-on, the *full price* (the sum of nominal price and queueing cost) to purchase an add-on is necessarily positive. So, only those with high valuations of the add-on are willing to join it.

We next formulate the firm's decisions under each pricing scheme. As before, our analysis starts with bundling followed by separate selling.

4.1 Bundling

Let p_B denote the bundle price, W_M and W_A denote the expected wait time at the main and add-on services, respectively.[13] In our formulation, a customer purchases the bundle if $V_M - cW_M - p_B + (V_A - cW_A)^+ \geq 0$, where $V_M + cW_M - p_B$ is the utility from consuming the main service, and $(V_A - cW_A)^+$ is the utility taking into account the add-on. Thus, it is possible that a high-type customer purchases the bundle and ends up joining the main service only, if her valuation of the add-on is surpassed by the corresponding queueing cost.

Consider a representative high-type customer with valuation v of the main service. Her utility of purchasing the bundle, $v - cW_M - p_B + (\beta v - cW_A)^+$, is strictly increasing in v. So, there exists a unique cut-off v_0 (if any) such that $v_0 - cW_M + (\beta v_0 - cW_A)^+ - p_B = 0$. A high-type customer purchases the bundle if her valuation $V_M \geq v_0$, whereas a low-type customer purchases the bundle if her valuation is $V_M \geq p_B + cW_M$. With this observation, we formulate the firm's optimization problem as follows:

$$
\begin{aligned}
\max_{p_B} \ & p_B \Lambda\left(\alpha\left[1-\frac{v_0}{\bar{v}}\right]+(1-\alpha)\left[1-\frac{p_B+cW_M}{\bar{v}}\right]^+\right)\\
\text{s.t.} \ & v_0 - cW_M + (\beta v_0 - cW_A)^+ - p_B = 0,\\
& W_M = \frac{1}{\mu_M - \Lambda\left(\alpha\left[1-\frac{v_0}{\bar{v}}\right]+(1-\alpha)\left[1-\frac{p_B+cW_M}{\bar{v}}\right]^+\right)},\\
& W_A = \begin{cases} \dfrac{1}{\mu_A - \alpha\Lambda\left[1-\frac{v_0}{\bar{v}}\right]} & \text{if } \beta v_0 \geq \dfrac{c}{\mu_A - \alpha\Lambda\left[1-\frac{v_0}{\bar{v}}\right]},\\ \dfrac{1}{\mu_A - \alpha\Lambda\left[1-\frac{v_1}{\bar{v}}\right]} & \text{if } \beta v_0 < \dfrac{c}{\mu_A - \alpha\Lambda\left[1-\frac{v_0}{\bar{v}}\right]}, \end{cases}\\
& \qquad \text{where } v_1 > v_0 \text{ is such that } \beta v_1 = \frac{c}{\mu_A - \alpha\Lambda\left[1-\frac{v_1}{\bar{v}}\right]}.
\end{aligned}
$$

Demands of the bundle from high-type customers are $\alpha\Lambda[1 - v_0/\bar{v}]$ because the cut-off valuation of high-type customers $v_0 < \bar{v}$; otherwise, no customers will purchase the bundle. However, the cut-off valuation of low-type customers $p_B + cW_M$ can be possibly greater than $\bar{v}$. Demands of low-type customers are thus given by $(1-\alpha)\Lambda[1 - (p_B + cW_M)/\bar{v}]^+$ in a unified expression.

The last equality in the above optimization problem formulates the expected wait time at the add-on, which is endogenized by high-type customers' self-joining behaviors. Recall that high-type customers with valuations higher than v_0 will join the main service. They further self-join the add-on by comparing their add-on valuations with the corresponding queueing cost. If $\beta v_0 \geq c/(\mu_A - \alpha\Lambda[1 - v_0/\bar{v}])$, then the waiting cost at the add-on is low and all high-type customers joining the main service continue to join the add-on. In this case, v_0 solves $v_0 - cW_M + \beta v_0 - cW_A - p_B = 0$. Alternatively, if $\beta v_0 < c/(\mu_A - \alpha\Lambda[1 - v_0/\bar{v}])$, then waiting cost at the add-on is high and discourages a portion of high-type customers from joining. Only those with valuations higher than another threshold $v_1 > v_0$ will join, and the indifferent customer with valuation v_1 receives zero utility from joining.

The above optimization problem is formulated to optimize over prices. A direct analysis requires solving three nested fixed-point equations that jointly involve wait times W_M, W_A, and the cut-off valuation v_0. To overcome this challenge, we convert it to an equivalent one that optimizes over $s = p_B + cW_M \in [0, (1+\beta)\bar{v})$. Then, v_0 solves $v_0 + (\beta v_0 - cW_A)^+ = s$. For each s, there are two candidate equilibria to consider.

(1) If $\beta v_0 - cW_A < 0$, then $v_0 = s$ and $\beta v_0 < c/(\mu_A - \alpha\Lambda[1 - v_0/\bar{v}])$. Then,

PROBLEM 4.

$$
\begin{aligned}
\max_{s<\bar{v}} \ & \Pi_2^{B,1} = (s - cW_M)\Lambda\left[1-\frac{s}{\bar{v}}\right]\\
\text{s.t.} \ & W_M = \frac{1}{\mu_M - \Lambda\left[1-\frac{s}{\bar{v}}\right]}, \ \beta s < \frac{c}{\mu_A - \alpha\Lambda\left[1-\frac{s}{\bar{v}}\right]}.
\end{aligned}
$$

(2) If $\beta v_0 - cW_A \geq 0$, then v_0 solves $v_0 + \beta v_0 - c/(\mu_A - \alpha\Lambda[1 - v_0/\bar{v}]) = s$ and $\beta v_0 \geq cW_A = c/(\mu_A - \alpha\Lambda[1 - v_0/\bar{v}])$. Then,

PROBLEM 5.

$$
\begin{aligned}
\max_{s} \ & \Pi_2^{B,1} = (s - cW_M)\Lambda\left(\alpha\left[1-\frac{v_0}{\bar{v}}\right]+(1-\alpha)\left[1-\frac{s}{\bar{v}}\right]^+\right)\\
\text{s.t.} \ & W_M = \frac{1}{\mu_M - \Lambda\left(\alpha\left[1-\frac{v_0}{\bar{v}}\right]+(1-\alpha)\left[1-\frac{s}{\bar{v}}\right]^+\right)},\\
& v_0 + \beta v_0 - \frac{c}{\mu_A - \alpha\Lambda\left[1-\frac{v_0}{\bar{v}}\right]} = s,\\
& \beta v_0 \geq \frac{c}{\mu_A - \alpha\Lambda\left[1-\frac{v_0}{\bar{v}}\right]}.
\end{aligned}
$$

Here, we allow bundling to screen out low-type customers, namely, $s > \bar{v}$, and so demands from low-type customers are characterized by $(1-\alpha)[1 - s/\bar{v}]^+$.

4.2 Separate Selling

We next formulate the firm's optimization problem under separate selling. As before, we impose $p_M + cW_M < \bar{v}$ and $p_A > 0$ to differentiate separate selling from bundling. Consider a high-type customer with valuation v of the main service. Her utility of purchasing the main service taking into account the

add-on is $v - cW_M - p_M + [\beta v - cW_A - p_A]^+$, which is strictly increasing in v. So, there exists a unique v_0 (if any) such that $v_0 - cW_M + [\beta v_0 - cW_A - p_A]^+ - p_M = 0$. The firm's optimization problem then can be formulated as follows:

$$
\begin{aligned}
\sup_{p_M + cW_M < \bar{v}, p_A > 0} \quad & \Pi^S(p_M, p_A) = p_M(D_M + D_{MA}) + p_A D_{MA} \\
\text{s.t.} \quad & D_M + D_{MA} = \Lambda \mathbb{P}\{V_M - cW_M + [V_A - cW_A - p_A]^+ - p_M \geq 0\} \\
& = \Lambda \left(\alpha \left[1 - \frac{v_0}{\bar{v}} \right] + (1-\alpha) \left[1 - \frac{p_M + cW_M}{\bar{v}} \right]^+ \right), \\
& D_{MA} = \Lambda \mathbb{P}\{V_M + V_A - p_M - p_A - cW_M - cW_A \geq 0,\ V_A \geq p_A + cW_A\} \\
& = \begin{cases}
\alpha \Lambda \left[1 - \frac{v_0}{\bar{v}} \right] & \text{if } \beta v_0 - \dfrac{c}{\mu_A - \alpha \Lambda \left[1 - \frac{v_0}{\bar{v}} \right]} - p_A \geq 0, \\
\alpha \Lambda \left[1 - \frac{v_1}{\bar{v}} \right] & \text{if } \beta v_0 - \dfrac{c}{\mu_A - \alpha \Lambda \left[1 - \frac{v_0}{\bar{v}} \right]} - p_A < 0, \quad \text{where } v_1 \text{ is such that} \\
 & \qquad \beta v_1 - \dfrac{c}{\mu_A - \alpha \Lambda \left[1 - \frac{v_1}{\bar{v}} \right]} - p_A = 0.
\end{cases} \\
& v_0 - cW_M + [\beta v_0 - cW_A - p_A]^+ - p_M = 0, \\
& W_M = \frac{1}{\mu_M - D_M - D_{MA}}, \ W_A = \frac{1}{\mu_A - D_{MA}}.
\end{aligned}
$$

The above optimization problem again requires solving three nested fixed-point equations that jointly involve W_M, W_A, and v_0. As before, we convert it to an equivalent one that optimizes over cut-off valuations. Let $s_M = p_M + cW_M < \bar{v}$ and $s_A = (p_A + cW_A)/\beta$. Then $v_0 + \beta(v_0 - s_A)^+ = s_M$, and depending on the relative value of s_M and s_A, there are two possible equilibria to consider.

(1) If $s_M \leq s_A$, then $v_0 = s_M$. Then,

PROBLEM 6.

$$
\begin{aligned}
\sup_{s_M \leq s_A < \bar{v}} \quad & \Pi_2^{S,1} = (s_M - cW_M)\Lambda \left[1 - \frac{s_M}{\bar{v}} \right] \\
& + (\beta s_A - cW_A)\alpha \Lambda \left[1 - \frac{s_A}{\bar{v}} \right] \\
\text{s.t.} \quad & W_M = \frac{1}{\mu_M - \Lambda \left[1 - \frac{s_M}{\bar{v}} \right]}, \\
& W_A = \frac{1}{\mu_A - \alpha \Lambda \left[1 - \frac{s_A}{\bar{v}} \right]}.
\end{aligned}
$$

(2) If $s_M > s_A$, then v_0 solves $v_0 + \beta(v_0 - s_A) = s_M$. Then,

PROBLEM 7.

$$
\begin{aligned}
\sup_{s_A < s_M < \bar{v}} \quad & \Pi_2^{S,2} = (s_M - cW_M)\Lambda \left(\alpha \left[1 - \frac{v_0}{\bar{v}} \right] + (1-\alpha) \left[1 - \frac{s_M}{\bar{v}} \right] \right) \\
& + (\beta s_A - cW_A)\alpha \Lambda \left[1 - \frac{v_0}{\bar{v}} \right] \\
\text{s.t.} \quad & v_0 + \beta(v_0 - s_A) = s_M, \\
& W_M = \frac{1}{\mu_M - \Lambda \left(\alpha \left[1 - \frac{v_0}{\bar{v}} \right] + (1-\alpha) \left[1 - \frac{s_M}{\bar{v}} \right] \right)}, \\
& W_A = \frac{1}{\mu_A - \alpha \Lambda \left[1 - \frac{v_0}{\bar{v}} \right]}.
\end{aligned}
$$

Here, because we impose $s_M < \bar{v}$, demands from low-type customers are simply $(1-\alpha)\Lambda[1 - s_M/\bar{v}]$.

The optimization problems after transformations, nevertheless, cannot be solved in closed forms under either pricing scheme due to intertwined waiting costs. However, in terms of revenue, the above transformations allow us to derive sufficient conditions that prompt an easy comparison between the two pricing schemes. We demonstrate these conditions below.

THEOREM 2. *Suppose there is congestion at both the main and add-on services.*

(i) *If* $\mu_A \geq \mu_M/\beta$ *and* $\bar{v} \geq c\mu_A/\beta(\mu_A - \mu_M)^2$, *then bundling (weakly) dominates for* $\Lambda > ((2\mu_M(1+\beta)\bar{v})/\alpha)/(\beta\bar{v} - (c\mu_A/(\mu_A - \mu_M)^2))$.

(ii) *If* $\mu_A \leq \alpha\mu_M$, *then separate selling strictly dominates for all* $\Lambda > 0$.

(iii) *If* $\mu_M < \mu_A < \mu_M/\beta$, *then there exists* $\lambda_0 < \mu_M$ *that depends on* $\mu_A, \mu_M, \alpha, \beta$. *For* $\Lambda > 2\lambda_0$, *define* $\hat{v} := c\mu_M/((\mu_M - \lambda_0)^2(1 - 2\lambda_0/\Lambda))$.

  (1) *If* $\bar{v} \geq \max\{\hat{v}, c\mu_A/(\beta(\mu_A - \mu_M)^2)\}$, *then bundling dominates for* $\Lambda > \max\{2\lambda_0, ((2\mu_M(1+\beta)\bar{v})/\alpha)/(\beta\bar{v} - (c\mu_A/((\mu_A - \mu_M)^2))\}$.

  (2) *If* $\bar{v} < \hat{v}$, *then separate selling strictly dominates for* $\Lambda > 2\lambda_0$.

Theorem 2 compares the revenues between two pricing schemes when both services are congested. Part (i) partially recovers Theorem 1, if one sets $\mu_A = \infty$, namely, when the add-on can be processed infinitely fast (or prepared in advance and does not require real-time processing). In general, for a finite but sufficiently large μ_A, as long as the valuation of the main service is sufficiently high, the comparison result established in Theorem 1 under a large market size will carry through. An ample capacity of the add-on implies that congestion at the add-on is negligible. The sales of each service then are capped by the capacity of the main service, and to exploit this capacity, the firm adopts bundling to screen out low-type customers.

The story changes quickly when the add-on is severely capacity-constrained, $\mu_A \leq \alpha\mu_M$. Part (ii) considers such a case and shows that separate selling strictly dominates bundling across the board, and does so even under a large market size. This is in sharp contrast to Theorem 1, which illustrates the dominance of bundling under a large market size when congestion only exists at the main service. To explain this contrast, we identify another operational benefit of separate selling that works best when both services involve congestion-prone delays: charging an additional price to sell the add-on can regulate congestion at its own location. Bundling, in contrast, loses admission control at the add-on due to a zero nominal price. In this case, it is likely that high-type customers will act on self-interest to over-join the add-on, at a rate significantly higher than the firm's optimal rate, creating over-congestion and revenue losses. Naturally, the strength of separate selling in regulating add-on traffic is most pronounced when the add-on has a very limited capacity, in which case, the firm selectively sells the add-on to customers with the highest valuations.

The finding in Part (ii) shares a similar spirit with Wu and Yang (2018). Focusing on two independent services (i.e., each service can be consumed without the other; e.g., an amusement park operating two different roller coasters), Wu and Yang (2018) studies the revenue comparison between bundling and separate selling and shows that separate selling dominates bundling in systems with significant congestion due to either a large market size or high delay sensitivity. The intuition is similar. Separate selling regulates congestion at each service location through distinct prices, whereas bundling is less effective in doing so as it cannot eliminate the overjoining of purchasing customers. Nevertheless, there is a fundamental difference between our Theorem 2 and the finding in Wu and Yang (2018). This difference stems from the underlying queueing structures. Wu and Yang (2018) shows that any capacity asymmetry between two services will strengthen the dominance of separate selling over bundling under a large market size, whereas in our setting, as shown in Parts (i) and (ii) of Theorem 2, the optimal pricing strategy depends intimately on how the capacity asymmetry is realized in its form. Specifically, bundling is a dominant strategy when the main service is extremely capacity-constrained while the add-on is not, and separate selling is a better strategy when the opposite is true. A more straightforward comparison between our setting and Wu and Yang (2018) is presented in Section 5.1 via a numerical study, where we demonstrate contrasting pricing and revenue implications of congestion under different queueing structures.

Part (iii) of Theorem 2 considers an add-on service with an intermediate capacity such that $\mu_M < \mu_A < \mu_M/\beta$. Recall that a large market size allows bundling to be optimal when congestion only exists at the main service, and this continues to hold when the valuation of the main service is sufficiently high despite congestion at the add-on. This is because a high valuation of the main service translates to high margins of both main and add-on services, and this narrows down the queueing effect at the add-on. In this case, the limited capacity of the main service continues to play a dominant role and the firm adopts bundling to serve high-type customers exclusively. Otherwise, when the valuation of the main service is relatively low, queueing cost becomes a critical component of the profit margins of selling each service. Separate selling then emerges as a better strategy to enforce necessary congestion regulation.

Theorem 2 provides a tentative explanation for the practice of selling add-ons separately from a main service in some congested settings. For example, museums may sell special exhibitions in separate tickets and exclude them from admission to general exhibitions. Amusement parks may unbundle meals from their popular restaurants (which often requires waiting in a check-out line or queueing for a table) from the general admission tickets. While we do not claim congestion effects as the exclusive driver of these unbundling tactics, pulling pricing levers to regulate congestion is often necessary in some of these settings. When this becomes a first-order factor (e.g., featured by a limited μ_M, μ_A, and nonnegligible c), our results provide normative guidance for the applicability of each pricing scheme.

Theorem 2 only gives sufficient conditions to compare the revenues of two pricing schemes. A precise comparison can be established by numerically solving Problems 4 to 7. We provide an example below, fixing parameters $\mu_M = 1, \bar{v} = 1, \alpha = 0.9$, and $\beta = 0.8$, and varying Λ and c to produce different congestion levels. Figure 1 presents the results and validates

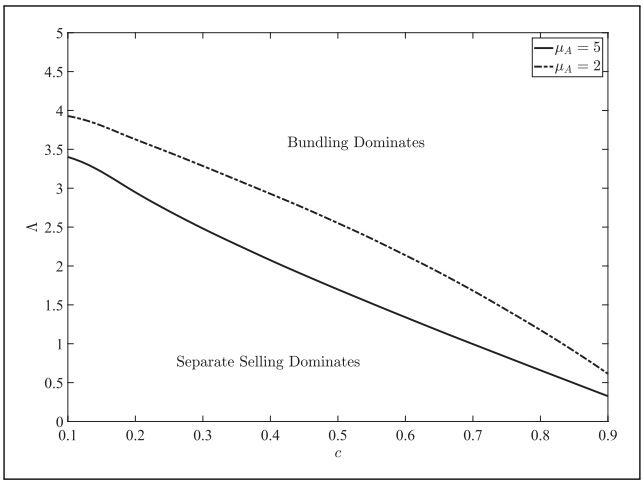

Figure 1. Revenue comparison between bundling and separate selling: congestion at both main and add-on services. $\bar{v} = 1,\ \mu_M = 1,\ \alpha = 0.9,\ \beta = 0.8$.

our prescriptions in Theorem 2. When the add-on has an ample capacity ($\mu_A = 5$), a large market size favors bundling whereas a small market size disfavors bundling. A limited capacity of the add-on ($\mu_A = 2$) exacerbates congestion and expands the parameter space where separate selling is optimal.

5 Extensions

In this section, we examine two model extensions, namely, high-type customers' independent valuations of main and add-on services, as well as endogenous capacities, each changing one assumption of our main model at a time, while keeping all others the same.

5.1 Independent Valuations

We first examine an extension by relaxing the assumption in the main model that high-type customers' valuations of two services are positively correlated. Specifically, in this extension, we assume that high-type customers have *independent* valuations for the main and add-on services, which are uniformly distributed over $[0, \bar{v}]$ and $[0, \beta\bar{v}]$, respectively. So, the *marginal* valuation distributions of each service considered in this extension are the same as those in the main model. This extension captures the well-known reduced valuation dispersion under bundling and examines its effect in the add-on queueing context. Following Sections 3 and 4, we consider two configurations of the add-on, one with an unlimited capacity and one with a constrained capacity. A detailed analysis of this extension can be found in Online Appendix A, where we formulate five separate optimization problems (three for bundling and two for separate selling) that involve nested fixed-point equations. The resulting analysis is intractable and we numerically compute the optimal solutions under each configuration.

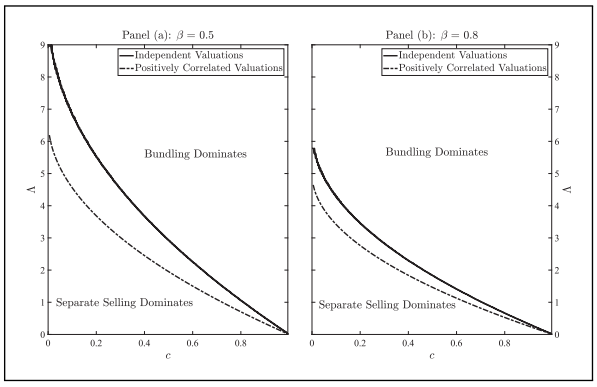

Figure 2. Revenue comparison between bundling and separate selling: congestion at main service only. Independent valuations versus positively correlated valuations. $\bar{v} = 1,\ \mu_M = 1,\ \alpha = 0.9$.

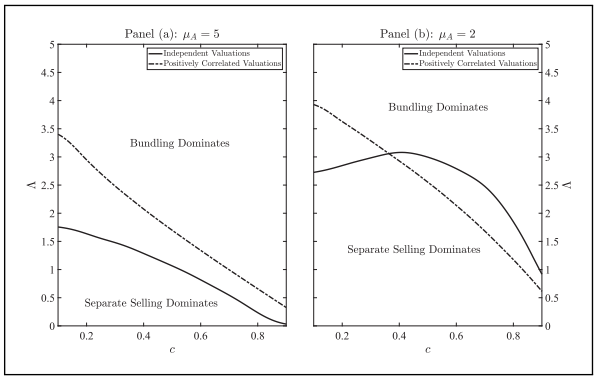

Figure 3. Revenue comparison between bundling and separate selling: congestion at both main and add-on services. Independent valuations versus positively correlated valuations. $\bar{v} = 1,\ \mu_M = 1,\ \alpha = 0.9,\ \beta = 0.8$.

Figures 2 and 3 present the results, showing that the insights of our main model qualitatively extend. When the main service is capacity-constrained while the add-on is not (Figure 2), separate selling outperforms bundling under a small market size and bundling outperforms otherwise. The market size required to reverse the dominance of separate selling, however, is found to be larger under independent valuations than under correlated valuations. This is because, under a positive correlation, high-type customers have high valuations for both main and add-on services. Bundling can exploit the high margins of both services by selling the main service exclusively to high-type customers. Thus, a positive correlation can enhance the appeal of bundling, making it a dominant strategy in a wider range of parameter space.

When both services are congested (Figure 3), the effect of correlation in valuations is more intricate. On the one hand, as said, high-type customers tend to value both services highly under a positive correlation, a fact that favors bundling in case of exclusive targeting. On the other hand, because these customers value the add-on highly, their chance of over-joining

the add-on and creating over-congestion is higher too. So, the overall effect of a positive correlation is mixed and how it materializes depends on the underlying parameters. Figure 3 shows that when customers are less sensitive to delays, they will likely over-join the add-on under a positive correlation, making separate selling a dominant strategy in a wider range of market sizes. However, as c increases, a limited capacity of the add-on ($\mu_A = 2$) leads to a considerable queueing cost at the add-on location and this effectively precludes over-joining. As a result, bundling can benefit from a positive correlation in a wider range of market sizes.

5.1.1 Comparison with Wu and Yang (2018). The above revenue patterns are instructive as they are fundamentally different from those in Wu and Yang (2018). To see this better, we replicate our previous analysis in the context of Wu and Yang (2018). This will elucidate the critical role of add-on structures in driving our key results. To mimic Wu and Yang (2018), which focuses on two *independent* services (i.e., each service can be consumed without the other) and two independent valuations (i.e., each customer has independent valuations for two services), we consider the following two-segment market: a segment (accounting for fraction α) with *independent* valuations $V_1 \sim U(0, \bar{v})$ and $V_2 \sim U(0, \beta\bar{v})$ for service 1 and 2, respectively, and another segment (accounting for fraction $1-\alpha$) with valuations $V_1 \sim U(0, \bar{v})$ and $V_2 = 0$ for service 1 and 2, respectively.[14] This, on the one hand, allows us to compare the results in the context of Wu and Yang (2018) to those in the current extension by fixing the *joint* valuation distribution of two services, and on the other hand, does not impose a strong correlation between services valuation as required by the (independent) service nature in Wu and Yang (2018). Formulations of the new problem follow a similar spirit of Wu and Yang (2018), and we omit the details for brevity.

Figure 4 presents the revenue comparison between separate selling and bundling in the new setting of two *nearly* independent services, which is largely in line with Wu and Yang (2018). In our study, we fix the capacity of the superior service 1 to one and vary the capacity of the inferior service 2 such that $\mu_2 \in \{0.7, 2, 5\}$. We find that separate selling dominates bundling in the upper right region where Λ and c are both large and bundling dominates otherwise in the lower left region, both echoing Wu and Yang (2018). Moreover, the switching boundary that separates these regions has a similar curvature as that in Wu and Yang (2018). In the case of $\mu_2 \in \{2, 5\}$ (i.e., ample capacity at service 2), this fully reverses our prescription in Figure 3 corresponding to the add-on configuration. Furthermore, in the case of $\mu_2 = 0.7$ (i.e., highly constrained capacity at service 2), this also sharply contrasts with our finding in the add-on setting that separate selling can dominate bundling across the board (under all market sizes and delay sensitivities) in the spirit of Theorem 2, Part (ii). Therefore, despite similar mechanisms of using prices to regulate congestion when necessary, the pricing and revenue implications of congestion are fundamentally different under these two

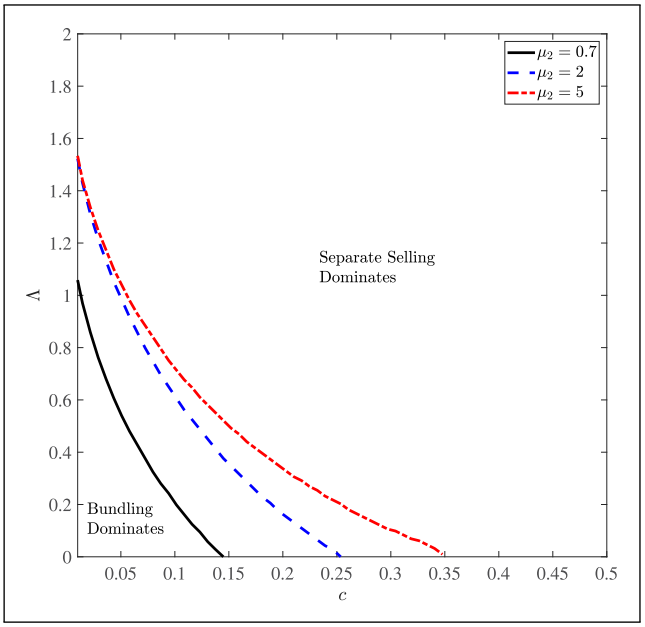

Figure 4. Revenue comparison between bundling and separate selling: two nearly independent services. $\bar{v} = 1$, $\mu_1 = 1$, $\alpha = 0.9$, $\beta = 0.8$.

queueing structures. Our analysis also suggests that pricing for add-on queues is more sensitive to the capacity asymmetry between different services.

5.2 Endogenous Capacity

Inthe main model, we assumed that the capacity of each service is exogenously given and does not vary with the pricing scheme adopted. However, in the long run, capacity may also be adjusted in conjunction with prices. This section studies such an extension. Specifically, we assume that the firm incurs a cost k_M and k_A to maintain a unit capacity of main and add-on services per unit of time, respectively. A central question in this extension is how the optimal capacity varies with the pricing scheme adopted, and vice versa. Our analysis considers both $k_A = 0$ (so that the firm sets $\mu_A = \infty$ and processes the add-on infinitely fast, allowing congestion only to be present at the main service) and $k_A > 0$ (so that the add-on is also capacity-constrained). Our two-step optimization first characterizes the optimal pricing scheme under each (μ_M, μ_A) and then optimizes over (μ_M, μ_A) to select the best capacity configuration.

We present an example of $k_A = 0$ in Table 1 (and Figure 5). We find that separate selling is a dominant strategy in a broad range of cost parameters, whereas bundling is only more profitable when the capacity cost is sufficiently large and the delay sensitivity is sufficiently small. In the latter case, because maintaining capacity is very costly, the firm can only afford a very limited capacity of the main service, and this favors bundling as we demonstrated in Section 3.2. The delay sensitivity, however, must be sufficiently small to guarantee a

Table 1. Optimal capacity and pricing strategy when capacity is endogenized: congestion at main service only. $\Lambda = 1, \bar{v} = 1, \alpha = 0.9$, $\beta = 0.5, k_A = 0$.

Cost	$c = 0.002$		$c = 0.007$		$c = 0.02$	
k_M	μ_M^*	Strategy	μ_M^*	Separate	μ_M^*	Strategy
0.1	0.55	Separate	0.63	Separate	0.74	Separate
0.4	0.38	Separate	0.40	Separate	0.43	Separate
0.85	0.18	Bundle	0.16	Bundle	/	/
0.9	0.16	Bundle	/	/	/	/

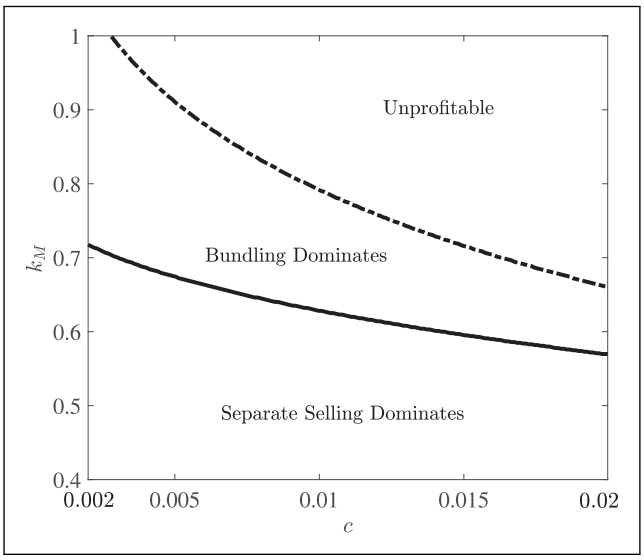

Figure 5. Graphical illustration of Table 1.

positive margin to sell the main service. Indeed, when both capacity and delay costs are high, neither pricing scheme can generate a positive profit and we use a slash '/' to denote this unprofitable case. Although both of these costs will reduce the margin of selling the main service,they can have opposite effects on the firm's optimal capacity decision. A high capacity cost brings down the firm's capacity level, whereas a high delay cost strengthens the firm's incentive of increasing capacity to alleviate congestion and wait times.

When the add-on requires costly investment, we present such a case in Table 2 (and Figure 6) by fixing $k_A = 0.01$. Because the queueing cost is likely high at the add-on, we set $\beta = 0.8$ to ensure that a significant portion of high-type customers will join the add-on. As before, we find that a large capacity cost will cut the firm's capacity levels of both services, whereas a high delay cost tends to do the opposite. Separate selling is optimal under most cost parameters, but bundling can be a dominant strategy when the firm maintains a limited capacity for the main service and ample capacity for the add-on. This happens when the capacity cost is high and the delay sensitivity is moderate. To explain why moderate delay sensitivities are necessary, note first that a low delay sensitivity will lead to underinvestment of capacity, which generates considerable congestion and favors separate selling for its strength in regulating congestion. A high delay sensitivity, in contrast, will drag the profit margin below zero to render service offerings unprofitable.

6 Conclusion

In this work, we revisited the classic add-on pricing in a queueing context with delay-sensitive customers and explored how congestion and queueing in service systems impact a monopoly firm's pricing strategy. We examined two commonly used pricing schemes, bundling and separate selling, to sell the main and add-on services and compared their profitability in various scenarios. We demonstrated the superiority of separate selling in environments without congestion effects. When congestion occurs at the main service only, we showed that separate selling may fall short of bundling under a large customer demand. When both services are congested, we showed that separate selling can again dominate under a large customer demand. We extended our analysis to consider independent valuations of two services and endogenous capacities and found that our main insights qualitatively carried over.

These plots of reversals reveal novel operational benefits of each pricing scheme. Separate selling entails efficient price discrimination in environments without congestion, but such strength is undermined by a congested main service. It regains

Table 2. Optimal capacity and pricing strategy when capacity is endogenized: congestion at both main and add-on services. $\Lambda = 1, \bar{v} = 1$, $\alpha = 0.9, \beta = 0.8, k_A = 0.01$.

Cost	$c = 0.002$		$c = 0.007$		$c = 0.02$	
k_M	(μ_M^*, μ_A^*)	Strategy	(μ_M^*, μ_A^*)	Strategy	(μ_M^*, μ_A^*)	Strategy
0.1	(0.56, 0.71)	Separate	(0.63, 0.94)	Separate	(0.74, 1.29)	Separate
0.4	(0.40, 0.59)	Separate	(0.42, 0.79)	Separate	(0.45, 1.08)	Separate
0.85	(0.24, 0.42)	Separate	(0.22, 0.54)	Bundle	/	/
0.9	(0.22, 0.40)	Bundle	(0.21, 0.52)	Bundle	/	/

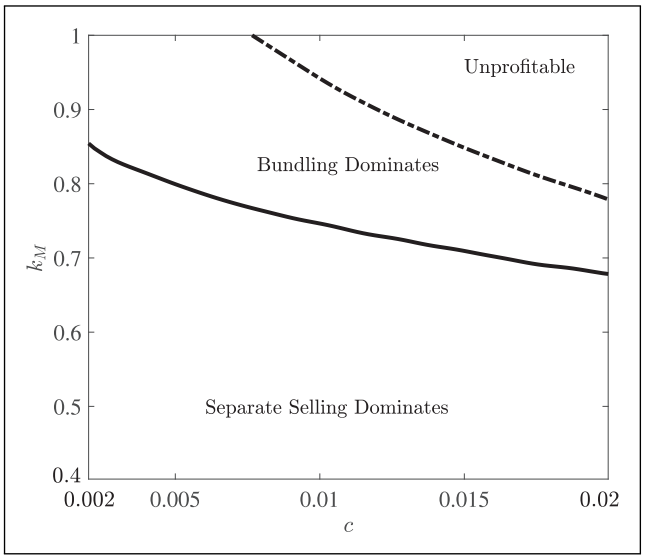

Figure 6. Graphical illustration of Table 2.

appeal when the add-on service is congested too, and utilizing it to regulate congestion can play a vital role in this latter context.

Our model has its limitations. There are other realistic factors not captured in our model such as travel time between different services (Hassin and Roet-Green, 2020), heterogeneity in customers' delay sensitivities (Mendelson and Whang, 1990; Afeche, 2013), customers' bounded rationality when purchasing multiple services (Ellison 2005; Cui et al. 2018; Dey et al. 2021), and uncertain valuations (Wu et al. 2022). Moreover, the practice of bundling can also be influenced by more nuanced features (e.g., menu costs) and historical reasons (e.g., business norms) which we don't model in our work. Extending our framework to incorporate these realistic factors will be promising endeavors for future research. We hope our paper will invite more investigations into this exciting strand of research.

Acknowledgments

The authors thank the department editor Michael Pinedo, a senior editor, and three anonymous reviewers for constructive feedback in the review process. The authors also thank Yihao Ang for research assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Chenguang (Allen) Wu received support from the Hong Kong General Research Fund (Grant Number: 16506122, 16501023). Chen Jin received support from the Singapore Ministry of Education Academic Research Fund (Tier 1 Grant 251RES2101). Ying-Ju Chen received support from the Hong Kong General Research Fund (Grant Number: 16500821, 16212821, and HKUST C6020-21GF).

ORCID iDs

Chenguang (Allen) Wu https://orcid.org/0000-0002-2528-0286
Ying-Ju Chen https://orcid.org/0000-0002-5712-1829

Supplemental Material

Supplemental material for this article is available online (doi: 10.1177/10591478241234994).

Notes

1. See https://www.cityexperiences.com/new-york/city-cruises/statue/.
2. See https://www.hkpm.org.hk/en/home.
3. See https://www.mplus.org.hk/en/.
4. See https://www.madametussauds.com/new-york/.
5. Visitors to Musee du Louvre write comments on Tripadvisor: "When we eventually found the Mona Lisa, the people crowding around were 50+ and it was impossible to see anything."
6. We consider in Section 5.1 an extension in which these customers have independent valuations for the main and add-on services.
7. Specifically, if there are three prices, p_M, p_A, and p_B for the main service, add-on service, and service bundle, respectively, then any rational customer who purchases the add-on will first purchase the main service and then end up paying a total of $\min\{p_M + p_A, p_B\}$, effectively making one of p_A and p_B irrelevant: if $p_M + p_A \leq p_B$, then p_B is irrelevant and all customers who purchase the main and add-on services pay $p_M + p_A$; if $p_M + p_A > p_B$, then p_A is irrelevant and the effective price of the add-on is $p_B - p_M$.
8. We consider in Section 4 a variation of this model that allows the add-on service to be capacity-constrained subject to congestion too.
9. We show in Proposition 4, Online Appendix B that an exogenous marginal cost does not alter the qualitative profit comparison between bundling and separate selling in Proposition 1, whereas an endogenous queueing cost, as we show in Section 3.2, does. This contrasts Wu and Yang (2018) in which the queueing cost is found to have a similar directional effect as an exogenous marginal cost on the profit comparison between bundling and separate selling.
10. The optimal price of the main service p_M^* described by Proposition 2 can be possibly negative in extreme cases in which the firm's profits are primarily driven by the add-on. A sufficient condition to rule out this uninteresting case is $\bar{v} > (1 + \beta)c/\mu_M$, which does not involve Λ. Under this condition, the queueing cost at the main service will never exceed $\bar{v}$. In other words, separate selling can be suboptimal under a large customer demand (see Theorem 1) even when there are positive values created from the main service.
11. Peter Luger in New York charges separate prices for desserts, fries, and salads. Jimmy's Kitchen in Hong Kong charges sparkling water at around $50 HKD per bottle.
12. Without a prespecified route of visiting two services, our analysis only stipulates that customers have the same probability of visiting each service first. In this case, there are potentially three (Poisson) streams of joining customers differentiated by their respective routes through two services: those who join the main service only; those who first join the main service and

then the add-on service; and those who first join the add-on service and then the main service. Such a queueing network (with multiple customer classes, each class having a different route) is known as the Kelly network (Kelly, 1975, 1976; Chen and Yao, 2001), which is a multiclass generalization of the (single-class) Jackson network and is known to have a product-form steady-state queue-length distribution. Hence, each service station has the same steady-state queue-length distribution as an $M/M/1$ queue with the same effective arrival rate and service capacity. However, we also note that a departure from our homogeneous-visiting-probability assumption may lead to a possibly different analysis and set of results (e.g., Parlaktürk and Kumar 2004; Arlotto et al. 2019), and we leave this to future research.

13. We use the expected wait time at the add-on in steady state as a proxy to compute a customer's waiting cost of joining the add-on. In general, a customer's wait time at the add-on depends on the queueing dynamics after she first joins the main service. We leave a more refined analysis (e.g., Ji et al., 2020) to future research.
14. Wu and Yang (2018) assume that each customer has independent valuations for two services. A slight departure from this assumption does not affect the qualitative comparison between bundling and separate selling, as shown in Figure 4.

References

Adams WJ and Yellen JL (1976) Commodity bundling and the burden of monopoly. *The Quarterly Journal of Economics* 90(3): 475–498.

Afeche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management* 15(3): 423–443.

Arlotto A, Frazelle AE and Wei Y (2019) Strategic open routing in service networks. *Management Science* 65(2): 735–750.

Bakos Y and Brynjolfsson E (1999) Bundling information goods: Pricing, profits, and efficiency. *Management Science* 45(12): 1613–1630.

Cao Q, Stecke KE and Zhang J (2015) The impact of limited supply on a firm's bundling strategy. *Production and Operations Management* 24(12): 1931–1944.

Chen H and Yao DD (2001) *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Vol. 4. New York: Springer.

Cui Y, Duenyas I and Sahin O (2018) Unbundling of ancillary service: How does price discrimination of main service matter? *Manufacturing & Service Operations Management* 20(3): 455–466.

Dey D, Ghoshal A and Lahiri A (2021) Support forums and software vendor's pricing strategy. *Information Systems Research* 32(2): 653–669.

Edelson NM and Hilderbrand DK (1975) Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society* 43(1): 81–92.

Ellison G (2005) A model of add-on pricing. *The Quarterly Journal of Economics* 120(2): 585–637.

Fang H and Norman P (2006) To bundle or not to bundle. *The RAND Journal of Economics* 37(4): 946–963.

Geng X, Tan Y and Wei L (2018) How add-on pricing interacts with distribution contracts. *Production and Operations Management* 27(4): 605–623.

Hassin R and Haviv M (2003) *To Queue Or Not to Queue: Equilibrium Behavior in Queueing Systems*. Vol. 59. Boston: Springer Science & Business Media.

Hassin R and Roet-Green R (2020) On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management* 23(4): 989–1004.

Ibragimov R and Walden J (2010) Optimal bundling strategies under heavy-tailed valuations. *Management Science* 56(11): 1963–1976.

Ji J, Roet-Green R and Snitkovsky RI (2020) Foresee the next line: On information disclosure in tandem queues. *Working paper*.

Jin C, Wu CA and Lahiri A (2022) Piracy and bundling of information goods. *Journal of Management Information Systems* 39(3): 906–933.

Kelly FP (1975) Networks of queues with customers of different types. *Journal of Applied Probability* 12(3): 542–554.

Kelly FP (1976) Networks of queues. *Advances in Applied Probability* 8(2): 416–432.

Mendelson H and Whang S (1990) Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations research* 38(5): 870–883.

Pang M-S and Etzion H (2012) Research note—analyzing pricing strategies for online services with network effects. *Information Systems Research* 23(4): 1364–1377.

Parlaktürk AK and Kumar S (2004) Self-interested routing in queueing networks. *Management Science* 50(7): 949–966.

Shulman JD and Geng X (2013) Add-on pricing by asymmetric firms. *Management Science* 59(4): 899–917.

Wu CA, Jin C and Chen Y-J (2022) Managing customer search via bundling. *Manufacturing & Service Operations Management* 24(4): 1906-–1925.

Wu CA and Yang L (2018) Bundle pricing of congested services. *Working paper*.

How to cite this article

Wu C (A), Jin C and Chen Y-J (2024) Add-On Pricing: A Queueing Perspective. *Production and Operations Management* xx(x): 1–17.