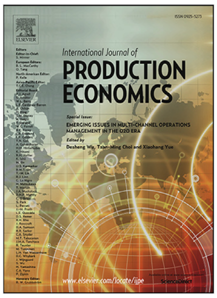


Service provision on an aggregator platform with time-sensitive customers: Pricing strategies and coordination

Myron Benioudakis, Dimitris Zissis, Apostolos Burnetas, George Ioannou

PII:	S0925-5273(22)00342-5
DOI:	https://doi.org/10.1016/j.ijpe.2022.108760
Reference:	PROECO 108760

To appear in: *International Journal of Production Economics*

Received date :	10 May 2022
Revised date :	16 November 2022
Accepted date :	27 December 2022

Service provision on an aggregator platform with time-sensitive customers: Pricing strategies and coordination

ARTICLE INFO

Keywords:
Demand aggregator platform
Strategic queueing
Time-based pricing
Double marginalization
Outsourcing
Reselling/Agency selling

ABSTRACT

The increasing tendency to fulfill customer needs via virtual platforms has led to a rapid growth of sharing economy. This practice allows non-entrepreneurs to set up a business and firms to focus on their core operations by outsourcing tasks related to attracting, finding, contracting, and invoicing customers. Hence, potential entrepreneurs and firms face the question whether to penetrate a market directly or through a platform, and to what extent. In this work, we focus on providers who offer unique services and make a choice between enlisting in a demand aggregator's platform and reaching the market directly; due to the unique services, we assume that the providers may have sufficient power to set the wholesale price that is paid to the platform. A game-theory model in a queueing framework is developed to address the questions of mode selection and pricing strategies. Such settings allow to include customers who are sensitive in delays in product/services delivery and exhibit strategic behavior. We show that under single-price contracts channel profits are adversely affected due to double marginalization. The latter effect can be mitigated by time-dependent pricing involving delay compensation or a revenue sharing contract, resulting in system coordination. We identify the equilibrium strategies and the provider's optimal policy. We also derive insights on the combined effects of key parameters such as the market size, the direct cost of customer service, and the aggregator's reservation level on the optimal pricing strategies, and quantify their impact.

1. Introduction

In this work, we study the effect of time-sensitivity and strategic customer behavior on the interaction between service providers and demand aggregator platforms in service or make-to-order production systems. We focus on cases where the provider offers unique services/products and has a choice of service mode, between direct access to the market or enlisting to an aggregator platform. We develop a game-theoretic model in a queueing framework involving the service provider, the aggregator and the customers, analyze equilibrium pricing strategies and explore how delay compensation and revenue sharing contracts can alleviate the effects of double marginalization and achieve channel coordination.

In production and distribution of physical products, a standard practice for manufacturers is to collaborate with independent retailers rather than reaching out to the market directly (Wang, Niu, Guo and Song, 2020). The benefits, as well as the ramifications of such interactions, have been extensively studied and analyzed in the literature (Legros, Jouini and Koole, 2020). Recently, there is an increasing trend for collaboration in cases of intangible products such as services. This is evidenced by the flourishing of virtual platforms (Cachon, Daniels and Lobel, 2017; Taylor, 2018); there are several examples such as: Lyft and Uber for taxi services; Deliveroo, Delivery Hero, and Wolt for food delivery services; Airbnb and Booking.com for accommodation, Fixt and Geekatoo for repair services, etc. The platform plays the role of a demand aggregator, rather than a typical retailer.

Demand aggregators can be considered as a form of outsourcing, where the enlisted companies outsource some activities to them. There are many reasons why such a collaboration may be beneficial for both parties

ORCID(s):

and thus sustainable. Among those, an important one is that the aggregators have created infrastructure and network for direct access to the end market and also have flexibility to perform customer attraction and contact processes more efficiently. As a result, firms who enlist to an aggregator platform reduce or completely eliminate the costs of direct market access. On the other hand, they may forfeit the pricing decisions to the aggregator, or share part of the revenue with them.

Another critical factor that affects demand and channel performance in general, is time. In manufacturing and distribution settings the lead time (including production time) plays a vital role. This is more intense during disruptions leading to massive shortages of not only specialized but also essential products and services (Yu, Razzaq, Rehman, Shah, Jameel and Mor, 2022). In service systems, the lead time is typically manifested as queueing delays experienced by customers. Furthermore, in most service systems customers are themselves decision makers who choose whether to join the system or not, based on their individual utility function that incorporates the value of the service and the cost of the anticipated delay. We adopt a framework in which time-sensitive customers make individual decisions whether to place an order or not, depending on the price as well as the anticipated delay. Since the decision of each customer induces externalities by affecting the delay of all others, the demand function is endogenized and derived from the equilibrium join/balk customer strategies (Hassin, 2016).

Our main objective is to explore how the strategic customer behavior impacts the interaction between service providers and demand aggregators. We focus on the equilibrium strategies and the resulting demand under various pricing settings, as well as their reflection on channel coordination. In this context, the central research questions are: i) to investigate provider's dilemma of reaching a market directly or through an aggregator; ii) to explore the effect of various parameters (such as the market size, the provider's cost of direct market access, and the aggregator's reservation level) on the profit maximizing pricing strategies of both parties; iii) to assess the impact of the resulting strategies on channel coordination.

To address these questions, we analyze a model with a service provider who seeks to penetrate a market composed of time-sensitive customers. This study focuses on services/products that are specialized; hence, the common practice is for no inventory to be stocked and service starts when a customer's order is placed. This is typical when the system offers either a pure service or a physical product on a make-to-order basis (Benioudakis, Burnetas and Ioannou, 2021). In either case, the lack of stock creates a potential for long queues and delays. Customers do not observe the number of pending orders. They base their decision on the value they receive from the service, the cost of waiting, the pricing strategy they face as well as on the expected delay. The strategic aspect of customer behavior refers to estimating the delay by considering the corresponding decisions of the other customers, thus leading to the notion of equilibrium in an appropriately defined game.

How a new entrant to the market selects the operation mode among a direct channel, a platform or a dual channel is generally more involved and includes several trade-offs. For example, selecting between a single direct channel and a dual channel brings up the trade-off of increasing the demand with customers who would not have access to the direct channel otherwise, vs cannibalizing some of the existing customer base and directing them to joining through the platform. On the other hand, the choice between a platform operation and a dual channel introduces the dilemma of making the dual channel investment in order to give existing customers the option to join directly in the future with possible savings for both the customer and the provider. In this study, we focus on the dilemma between a single direct channel and a platform mode, which is concerned with the trade-off between making the investment for direct access and forfeiting part of the market power to the aggregator. This may be the case when they are new entrants or not well-established in a market, and have not created the proper infrastructure and network for direct access. Hence, they should make a strategic decision on how to operate. Moreover, dual channel operations may not be allowed by law (e.g., clearing houses), or not be promoted due to cannibalization and special market agreements.

When the provider decides to act alone (direct service mode), he sets the retail price and shoulders the marketing and customer contact costs. Otherwise, he operates through an aggregator (indirect service mode) who undertakes all the actions related to customer contact and provides the required technology. In the latter case, a key question is who sets the retail price (Pu, Sun and Shao, 2020); it can be either the platform (reselling format) or the provider (agency selling format). Due to the specialized nature of the service/product, our primary focus is on the reselling format case, where the provider sets the wholesale price he receives, and lets the platform set the retail price. Furthermore, the provider has the option to institute a policy of delay compensation, which the aggregator passes to the customers. This is a plausible assumption when the customers are delay-sensitive and react strategically to delays. This extended type of reselling format model can be applied in settings for maintenance of specialized equipment, repair services (e.g., Fixt), etc. We also examine the agency selling format through a more general revenue sharing contract and discuss the relationships and differences between the two approaches.

In this context, we: i) formulate an outsourcing model in a service system where the demand is derived from the strategic behavior of customers; ii) show that under single-price contracts, channel profits are adversely affected by double marginalization; however, both time-dependent pricing induced by the delay compensation, and revenue sharing agreements can coordinate the system; iii) demonstrate how the provider's decision is affected by the market size, his own direct cost of marketing and customer service, as well as the aggregator's reservation level.

The rest of the paper is organized as follows: Section 2 provides the related research background. In Section 3, we present the model, the demand, the profit functions, and the centralized solution. Section 4 derives the analytical solutions for the decentralized settings under the single-price and the time-based contract, while Section 5 explores the provider's dilemma by finding appropriate thresholds and concludes with the coordination properties. In section 6, we analyze a reservation-level oriented revenue-sharing contract and discuss the connection with the single-price and the time-based contracts. Section 7 provides further insights obtained from numerical experiments. Conclusions and potential extensions are discussed in Section 8.

2. Literature Review

During recent years a large number of models have been developed regarding the creation of additional profits in decentralized settings (Vosooghidizaji, Taghipour and Canel-Depitre, 2020). A main focus is the alignment of individual and channel objectives, without harming the benefits of involved parties (Cachon and Terwiesch, 2020). In this study, we investigate how outsourcing contributes to this alignment. Specifically, we consider outsourcing practices that are related to finding customers through a demand aggregator platform. We analyze how such practices affect channel performance and promote coordination; while the impact of the aggregator is explored as well. The literature review is concentrated on the following streams: i) outsourcing and demand aggregator platforms, ii) strategic time-sensitive customers, and iii) coordination.

2.1. Outsourcing and Demand Aggregator Platforms

The practice of outsourcing has been extensively explored in the production and service management literature. In principal, outsourcing is an expensive option, but it is promoted because it provides flexibility and agility. The literature reveals that outsourcing practices have been employed extensively in several fields, highlighting also the pros and cons. A reasonable use of outsourcing can result in reduced costs and eliminate inefficiencies in many cases, since it is considered an efficient cost-cutting measure (Min, 2013). For example, Bardi and Tracey (1991) reported that 70% of US firms had chosen to outsource some of their logistics activities. However, there are barriers and significant challenges in settings that operate by employing outsourcing options (Espino-Rodríguez and Padrón-Robaina, 2006).

The logistics field is not the only one in which outsourcing practices are beneficial; queueing systems are also highly connected with service outsourcing, indicating its critical role and the potential benefits that it can offer. For instance, Abdel-Malek, Kullpattaranirun and Nanthavanij (2005) analyzed an open queueing network model of a multi-layered supply chain with a company that outsources subcomponents to several suppliers, while outsourcing strategies have been evaluated. There are also studies that consider the impact of the outsourcing practice on channel coordination, identifying service rate, service quality and the correlation between those as the critical factors for coordination (Feng, Ren and Zhang, 2019). Another popular field regarding the adoption of outsourcing practice is the call centers. Gans and Zhou (2007) analyzed and compared call-routing schemes, where the high value customers are served directly from the service provider, while the low value ones through an outsourcer. In a recent study, Legros et al. (2020) showed that postponing an outsourcing decision for customers can improve the in-house performance, but leads to severe losses in revenue.

A demand aggregator can be considered in part as an outsourcing agent. In general, it is a commercial entity that provides response services related to demand, facilitating all the participants (i.e., service providers, producers, customers, etc.). In this work, we focus on aggregator's activities that support the service provider with the tasks of attracting, finding, and contracting customers. This concept is common in the energy sector; Carreiro, Jorge and Antunes (2017) provided a literature review on demand aggregators in energy systems, while a case study in the Iberian energy market has been examined by using a demand aggregator (Iria, Soares and Matos, 2018). The demand aggregator concept has also attracted a lot of attention in the Operations Management; especially under the recent developments of the rapid growth of sharing economy, creating room that requires further research.

There are several works which consider outsourcing practices through demand aggregator platforms. For instance, Banerjee, Riquelme and Johari (2015) studied dynamic vs static pricing platform in a queueing network. The same dilemma has been investigated by Cachon et al. (2017), given that the demand is either low or high, emphasizing the benefits of a surge pricing strategy. A similar setting regarding the demand has been examined in Gurvich, Lariviere and Moreno (2019), where the platform adopts a compensation scheme to attract agents to provide an adequate service level. The concept becomes more realistic and complex under the consideration of time-sensitive customers; one of the main assumptions of our work. Indicative examples are the studies of Taylor (2018) and Bai, So, Tang, Chen and Wang (2019), where a platform is used to connect agents with heterogeneous time-sensitive customers. Both the customer and the wholesale prices are determined by the platform, assuming that the market size is sufficient large. In a similar concept, Choi, Guo, Liu and Shi (2020) focused on how homogeneous customers under different risk attitudes affect the pricing decisions, and the participants' profits.

The latter stream considers several service providers who operate in a large market where customer prices as well as provider wages (wholesale prices) are determined by the platform. However, in situations where the provider offers a distinctive service (such as maintenance of specialized equipment) or unique products (such as room accommodation) that are hard to be substituted, he may have power to impose his own price on the platform. Chen, Hu and Wang (2022) developed a model of a restaurant and delivery platform cooperation, where the restaurant (provider) allows the customers to select the mode of the service (either directly or through the platform) they desire. The provider sets the customer price and the platform adds a delivery fee and coordination is achieved via revenue-sharing contracts.

2.2. Strategic Time-sensitive Customers

A main aspect of this work is regarding time and the impact of it on the participants' decisions. We assume that demand is derived endogenously through the strategic behavior of customers who react to anticipated delays and pricing policies. Specifically, customers are time-sensitive in the sense that they consider their waiting time before deciding whether or not they will proceed with an order. The effect of strategic customer behavior and time-sensitivity has been mainly explored in settings where customers may decide to postpone a purchase attempting to achieve a lower price. For example, Li and Yu (2017) and Lin, Parlaktürk and Swaminathan (2018) explored the impact of strategic customer behavior on channel profitability, assuming that the customers value the product and decide whether to buy or wait. Moreover, they examined whether coordination can be achieved when the customers behave strategically.

Strategic customer behavior has also been considered in dimensions other than time; for instance when there is product differentiation in a market (Ahmadi, Iravani and Mamani, 2017) or when customers are price and quality sensitive (Lu, Chen, Tomlin and Wang, 2019). A work close to ours is Liu, Parlar and Zhu (2007) who considered pricing and lead-time compensation by incorporating a queueing model in a supplier-retailer setting with time-sensitive customers. Customers were fully compensated for late deliveries with a penalty cost per unit-time late. The importance of market and operational factors in channel performance were analyzed. However, when queueing effects are present, analyzing the effects of strategic customer behavior is generally more involved, since a customer's decision to "join" affects congestion and other customers' delays, leading to equilibrium customer strategies in appropriately defined games; which is incorporated in our study.

Such models have been extensively studied in the queueing literature in the last several years; we refer to Hassin and Haviv (2003) and Hassin (2016) for extensive reviews. In particular, many pricing schemes that include delay compensation have been used in settings where time-sensitive customers observe the length of the queue or not. Feng and Zhang (2017) considered multiple distinguishable customer types where customers are allowed to observe the length of the queue; a Markov Decision Process model has been used to derive optimal dynamic pricing and lead-time quotation policies. In the unobservable setting, Afèche, Baron and Kerner (2013) showed that when the provider offers full delay compensation and charges an entrance fee equal to the customer's service valuation, provider's profits are maximized. In a similar setting, Benioudakis et al. (2021) employed pricing compensation policies to secure a particular demand pattern over time.

2.3. Coordination

Coordination refers to the case where a decentralized setting achieves the same outcome as when all the decisions are made by a single entity (Arshinder, Kanda and Deshmukh, 2008). In principle, coordination is defined as the situation when total channel profit under the decentralized setting is equal to that under joint optimal or integrated setting (Viswanathan and Wang, 2003). Coordination can be achieved when individual entities work together by sharing information and resources seeking to capture the maximum benefits for the entire system by aligning their objectives. It is recognized as a desirable goal and several efforts have been made to reach it (Chopra and Meindl, 2019); however, in some cases coordination is a utopian situation, as there is no single entity that can enforce a globally optimal strategy. The main challenge is to propose applicable ways to optimize channel profit without violating competition rules and making unrealistic assumptions about the business partners. The inefficiency from the total welfare point of view is due to independent decision makers with different preferences, objectives, and information deciding on their actions without considering the global optimum. This individual rationality approach does not allow the system to capture the maximum level of profits that are available (Cachon and Terwiesch, 2020).

One of the most common practices to achieve coordination includes contracts between the decision makers (Cachon, 2003). Contracts, in principle, describe all the terms of an agreement, define future actions and have been considered as an important tool to align decisions and reach coordination. However, contracts are violated quite often in practice due to the dynamic character of business environments. We refer the reader to Choi and Cheng (2011) for comprehensive reviews regarding contracts and how these promote coordination. A remedy to eliminate the binding character of contracts is through mechanism design, by providing all the decision makers with appropriate incentives to decide on their actions considering as their primary objective system benefits (Myerson, 1989). That approach has been applied in Operations Management (Vohra, 2012); for example Zissis, Ioannou and Burnetas (2020) considered inventory management decisions when production and storage are controlled by different business partners. However, a challenge of the mechanism design approach is regarding who designs the mechanism and what are the benefits for him/her.

The misalignment of objectives between independent decision makers, when pricing decisions are involved, often manifests itself as a double marginalization effect. Numerous approaches have been proposed that aim to mitigate or eliminate its consequences, including vertical integration, franchise fee, resale price maintenance, sophisticated pricing schemes, etc. A work that links double marginalization and coordination is by Li, Li and Cai (2013) who studied a generalized model with uncertain supply, proposing contracts to coordinate the channel even when demand is random. In addition, Bernstein and Federgruen (2007) considered coordination approaches under price and/or service level competition in a setting with a single manufacturer and several competing retailers.

In the present paper, we develop a model of interaction between the participants, where the provider sets the payment he receives from the platform for each served customer and forfeits the decision on the final customer price. We propose a coordination mechanism that is based on delay compensation provided directly by the provider to the end customers and we make comparisons with a reservation-level oriented revenue-sharing contract. In that context, we examine the provider's decision: to be part of a platform, seeking to avoid all the operating costs related to customer contact, or to reach the market directly. We explore this dilemma of outsourcing or not under a profit maximization as well as a channel optimization point of view, leading to the consideration of coordination; one of the key objectives of this study. Another innovation of our work is including the market size and the aggregator's alternative option (in the sense that the aggregator can reject provider's offer and secure her reservation level) as model parameters. The latter allow us to explore settings and participants' decisions under the following dimensions: i) market size, ii) aggregator's reservation level, and iii) provider's customer contact cost.

3. The Model

3.1. Model Description

We consider a provider who offers a specialized service or product and plans to penetrate a market. There are two options: either to reach the market directly, or through a demand aggregator platform. In this work, we focus on cases in which dual channel operations are not considered. We denote the provider by P and the demand aggregator by A and refer to them using male and female pronouns, respectively. The aggregator has an established channel for accessing that market and undertakes all operations related to attracting customers, customer contact, payments, etc. If the provider collaborates with her (i.e., enlisted in her platform), he charges her a fixed price per served customer (wholesale price), whereas the aggregator acts as an intermediary who charges the service fee (retail price). The provider essentially outsources the customer-related operations and concentrates on providing the actual service to the customers who are allocated to him.

The provider's decision of choosing a direct or indirect service mode, as well as both parties' pricing strategies are affected by the end user demand. We endogenize the demand function by assuming that customers (end users) are price and time-sensitive and exhibit a strategic behavior. This means that the customers decide whether or not to request the service, based on the service fee and the anticipated behavior of other customers, which in turn affects the system congestion and their own delay. We assume that customers arrive according to a Poisson process with rate Λ. Every arriving customer decides to place an order or not. Order processing times are exponentially distributed with rate μ and orders are served on a First-Come First-Served basis. A completed order brings reward R to the customer; however, there is also a waiting cost c per unit of sojourn time in the system. Customers are rational; i.e., they select their decisions to maximize their individual expected net benefit (reward value minus service fee and waiting cost). We adopt an unobservable system framework where the customers are not informed about the queue length before deciding whether to place an order or balk. In case that a customer decides to balk, its net benefit is considered zero.

Regarding the provider, when he decides to reach the market directly, he incurs a customer contact cost S per unit time. The latter includes costs related to promotion, customer contact, payment processing, etc. His second option is to involve the aggregator, by offering her a deal that defines the terms of their collaboration and gives her sufficient incentives to participate. In this case, he saves the customer contact cost but he forfeits some of the (potential) profit by making the aggregator part of the process. In addition, the provider has the choice to withdraw without providing any service, direct or indirect. Thus, he penetrates that market only if he can secure a nonnegative net profit.

The aggregator is free to accept or reject the provider' offer and decides on her action seeking to maximize her own profitability. If she accepts, then she is responsible for finding the customers (marketing and sales promotion tasks) and interacting with them (arranging payments, etc.) leaving only the task of service to the provider. Incoming customers are served by the provider who charges the aggregator and receives from her a fixed price per served customer. If the aggregator rejects the provider's offer, she receives an alternative value k per unit-time. We assume that k includes any potential extra benefits that she can secure by using her resources in a different manner. The value k can be thought of as the aggregator's reservation level. In this sense, when the provider operates through the platform, a profit at least equal to k should be ensured for the aggregator. As the platform operates regardless of whether the provider joins it, we do not impose an additional cost for the aggregator if she accepts the provider, other than the opportunity cost captured by k.

In this work, we consider a class of time-based pricing strategies for the customers, which includes a service fee plus a compensation per unit of time that a customer spends in the system. The compensation is determined and paid by the provider, regardless of the mode of service. Specifically, if the provider decides to penetrate the market directly, he charges a service fee p per customer, supplemented by a compensation l per unit of time in the system. If he decides to operate through the aggregator's platform, he offers her a contract which includes a wholesale charge w per customer served and a compensation l per unit time, which he commits to pay to each customer who joins. The aggregator responds to this contract by setting the service fee p herself and passing the compensation to the customers. In some business environments (such as taxi services), the aggregator may not operate under compensation schemes. In the following, we analyze the single-price policies separately, both to address situations as the above and also to assess and quantify the impact of compensation on such interactions. In the rest of this work, we use SP and TB for single-price and time-based pricing contract, respectively.

3.2. Demand and Profit Functions

The demand coincides with the rate of incoming orders under customer equilibrium, as a function of the service fee (p) and compensation rate (l). Since the compensation is accrued from the instant of joining the system, it effectively reduces the customer's waiting cost to $c - l$ per unit time. Following the standard analysis of an unobservable $M/M/1$ queue with strategic customers (c.f. Hassin and Haviv, 2003, Chapter 3), the expected benefit of a customer who places an order when the other customers place orders with rate λ is:

$$B(p, l, \lambda) = R - p - \frac{c - l}{\mu - \lambda}. \tag{1}$$

Furthermore, the equilibrium rate of incoming orders is:

$$\lambda^e(p,l) = \min\left\{\mu - \frac{c-l}{R-p}, \Lambda\right\}, \tag{2}$$

and the expected benefit of a customer who places an order in equilibrium is equal to:

$$B(p,l,\lambda^e(p,l)) = \begin{cases} 0, & \text{if } \lambda^e(p,l) < \Lambda, \\ R - p - \frac{c-l}{\mu-\Lambda}, & \text{if } \lambda^e(p,l) = \Lambda. \end{cases}$$

The expected customer benefit in equilibrium is equal to zero for any values of p, l such that $\lambda^e < \Lambda$. When customers decide strategically but individually, they tend to use the system capacity up to the point where the delay cost equals the net service profit $R - p$, resulting in zero customer benefit in equilibrium.

When p, l are such that $\lambda^e = \Lambda$; i.e., the market is captured, the expected customer benefit can be positive. However, as we show in the following sections, when the provider and the aggregator operate under profit maximizing strategies, they have the power to extract all the customer benefit even in this case. Hence, the compensation policy is mainly employed to shape the demand function in a more profitable manner, and not to improve the customer welfare.

We next consider the profit functions for the provider, the aggregator, and the entire channel. These depend on the service mode. In both cases, the provider's and the aggregator's strategies are determined by the pricing variables p, l, w, which correspond to the service fee (paid by the customers), the delay compensation rate (paid by the provider to the customers), and the wholesale price (paid by the aggregator to the provider), respectively. Depending on the service mode, p is set by the provider or the aggregator, whereas l and w are always set by the provider. In the definitions below, we indicate the dependence of the profit functions on the pricing variables p, l, w, regardless of who sets those and in which order. Based on the above, and given the equilibrium rate of incoming orders, the provider's and aggregator's expected net profit functions per unit of time are:

$$\text{Provider's Profit:} \begin{cases} G_{P,D}(p,l) = \lambda^e(p,l)\left(p - \frac{l}{\mu-\lambda^e(p,l)}\right) - S, & \text{direct service,} \\ G_{P,I}(p,l,w) = \lambda^e(p,l)\left(w - \frac{l}{\mu-\lambda^e(p,l)}\right), & \text{indirect service,} \end{cases} \tag{3}$$

$$\text{Aggregator's Profit:} \begin{cases} G_{A,D} = k, & \text{direct service,} \\ G_{A,I}(p,l,w) = \lambda^e(p,l)(p-w), & \text{indirect service.} \end{cases} \tag{4}$$

In the following analysis, we assume that $R > \frac{c}{\mu}$, so that either the provider or the aggregator attracts at least one customer when the service fee is zero and there is no delay compensation. We also assume that $0 \leq l \leq c$, since the role of compensation is to subsidize part of the waiting cost. Given this, any solution of the provider-aggregator equilibrium should satisfy $0 \leq w \leq p \leq R$.

3.3. Channel Profit and Coordination

Let consider a central planner, who decides whether service will be provided or not, as well as the mode of service and the values of λ, p, l, w. The central planner takes into account the aggregator's reservation level (k) and the provider's customer contact cost (S), with the objective of maximizing the total channel profit. In the centralized setting, we do not assume that customers determine the rate of incoming orders in equilibrium, since the central planner determines the input rate directly. In addition, the transfer payments between the provider, the aggregator and the customers are mutually canceled in the total profit expression. Therefore, the expected channel profit per unit of time under central control is:

$$G_C(\lambda) = \begin{cases} k, & \text{no service,} \\ \lambda\left(R - \frac{c}{\mu-\lambda}\right) + k - S, & \text{direct service,} \\ \lambda\left(R - \frac{c}{\mu-\lambda}\right), & \text{indirect service.} \end{cases} \tag{5}$$

Maximization of $G_C(\lambda)$ is essentially equivalent to maximizing the social profit in the unobservable $M/M/1$ queue (Edelson and Hilderbrand, 1975). If service is provided, the optimal rate of orders is equal to $\min\left\{\lambda_0, \Lambda\right\}$, where $\lambda_0 = \mu - \sqrt{\frac{c\mu}{R}}$, regardless of the mode selection. The choice between direct and indirect mode depends only on the sign of $k - S$. Therefore, the optimal channel profit is equal to:

$$G_C^* = \max\{k, G_0(\Lambda) + k - S, G_0(\Lambda)\} = \max\{k, G_0(\Lambda) + (k - S)^+\}, \tag{6}$$

where,

$$G_0(\Lambda) = \begin{cases} \left(\sqrt{R\mu} - \sqrt{c}\right)^2, & \text{if } \Lambda > \lambda_0, \\ \Lambda\left(R - \frac{c}{\mu-\Lambda}\right), & \text{if } \Lambda \leq \lambda_0. \end{cases}$$

If the optimal profit from serving customers is not sufficient to cover either the aggregator's alternative or the provider's contact cost (i.e, $G_0(\Lambda) < \min\{k, S\}$), then providing no service is socially optimal. Otherwise, the socially optimal service mode depends on the relative values of k and S and service through the aggregator is preferable if $k < S$. In practice, the channel optimal solution serves as a benchmark to assess the efficiency of the decentralized strategies, resulting from the interaction between the individual parties. In this context, coordination is defined when the channel profit under the decentralized solution is equal to the centrally optimal value G_C^*. In other words, under strategic equilibrium the selection of the service mode by the provider, the pricing policies set by the provider and the aggregator and the join/balk decisions by the customers result in the socially optimal service mode and rate of incoming orders.

4. Decentralized Pricing Strategies

In this section, we derive the decentralized solutions for the direct and indirect service mode in the reselling format framework. In each case, we consider both SP and TB strategies, corresponding to $l = 0$ and $l > 0$, respectively. We use the tilde symbol to denote prices, rate of incoming orders, and profits under the SP policy.

4.1. Direct Service Mode

Under the direct mode, the provider's profit maximization problem is:

$$G_{P,D}^* = \max_{p,l} G_{P,D}(p, l). \tag{7}$$

We express (7) as a two-stage optimization problem:

$$\max_{p,l} G_{P,D}(p, l) = \max_{\lambda} H(\lambda),$$

where,

$$H(\lambda) = \max_{p,l} \left\{ \lambda\left(p - \frac{l}{\mu - \lambda}\right) - S \,:\, \lambda^e(p, l) = \lambda \right\}.$$

In this form, the provider determines the optimal equilibrium ordering rate to induce, where $H(\lambda)$ denotes the maximum profit that can be achieved when the induced rate is equal to λ. The equilibrium condition for $\lambda^e(p,l) = \lambda$ is $p - \frac{l}{\mu-\lambda} \le R - \frac{c}{\mu-\lambda}$, with equality for $\lambda < \Lambda$. Thus, $H(\lambda) = \lambda\left(R - \frac{c}{\mu-\lambda}\right) - S$, where $0 \le \lambda \le \Lambda$.

The profit maximization problem is equivalent to maximizing the social benefit in (5) with respect to λ. Thus, the profit maximizing rate of orders is $\tilde{\lambda}^*_D = \lambda^*_D = \min\left\{\lambda_0, \Lambda\right\}$ and the maximum provider's profit $\tilde{G}^*_{P,D} = G^*_{P,D} = G_0(\Lambda) - S$, under the direct mode. In terms of the pricing strategy the provider has substantial flexibility, since any pair (p,l) such that $p - \frac{l}{\mu-\min\{\lambda_0,\Lambda\}} = R - \frac{c}{\mu-\min\{\lambda_0,\Lambda\}}$ is optimal. In particular, he can achieve the maximum profit without offering any compensation; since for $l = 0$, we obtain the profit maximizing strategy in Edelson and Hilderbrand (1975) which is:

$$p_0 = \begin{cases} R - \sqrt{\frac{cR}{\mu}}, & \text{if } \Lambda \ge \lambda_0, \\ R - \frac{c}{\mu-\Lambda}, & \text{if } \Lambda < \lambda_0. \end{cases}$$

Regarding the aggregator, she resorts to her alternative option with value k. Therefore, $\tilde{G}_{A,D} = G_{A,D} = k$.

4.2. Indirect Service Mode

In this case, the provider operates through a demand aggregator platform. We focus on two types of contracts, depending on whether a delay compensation is included or not.

4.2.1. Single-price (SP) contract

We first examine the contract where the provider charges the aggregator with a wholesale price w without offering any delay compensation. The aggregator responds optimally with a single service fee $p^*_I(w)$, as long as her optimal profit is at least k (reservation level). Therefore, the optimization problems for both parties are:

$$\tilde{G}^*_{A,I}(w) = \max_{p\ge 0} G_{A,I}(p,0,w), \tag{8}$$

$$\tilde{G}^*_{P,I} = \max_{w\ge 0}\{G_{P,I}(\tilde{p}^*_I(w),0,w) \,:\, \tilde{G}^*_{A,I}(w) \ge k\}. \tag{9}$$

Proposition 1 summarizes the aggregator's response to provider's SP strategy.

Proposition 1. *Given a wholesale price w,*

i.) The aggregator's optimal service fee is: $\tilde{p}^*_I(w) = w + \tilde{m}^*_I(w)$, where,

$$\tilde{m}^*_I(w) = \begin{cases} R - w - \sqrt{\frac{c(R-w)}{\mu}}, & \textit{if } w \ge \tilde{w}_{MC}, \\ R - w - \frac{c}{\mu-\Lambda}, & \textit{if } w < \tilde{w}_{MC}, \end{cases}$$

and

$$\tilde{w}_{MC} = \begin{cases} R - \frac{c\mu}{(\mu-\Lambda)^2}, & \textit{if } \Lambda \le \lambda_0, \\ 0, & \textit{if } \Lambda > \lambda_0. \end{cases}$$

ii.) The equilibrium demand is:

$$\tilde{\lambda}_I^*(w) = \begin{cases} \mu - \sqrt{\frac{c\mu}{R-w}}, & \text{if } w \geq \tilde{w}_{MC}, \\ \Lambda, & \text{if } w < \tilde{w}_{MC}. \end{cases}$$

iii.) The optimal aggregator's profit is:

$$\tilde{G}_{A,I}^*(w) = \begin{cases} \left(\sqrt{(R-w)\mu} - \sqrt{c}\right)^2, & \text{if } w \geq \tilde{w}_{MC}, \\ \Lambda\left(R - w - \frac{c}{\mu-\Lambda}\right), & \text{if } w < \tilde{w}_{MC}. \end{cases}$$

The aggregator responds to the provider's wholesale price w by adding her own profit margin $m_I^*(w)$. The margin is equal to the optimal price that a provider would set in order to maximize his profits under direct mode, if the customer service value were equal to $R - w$. In essence, the aggregator passes the wholesale price to the customers and in addition adds her own profit margin to the service fee considering that the service value to customers is equal to $R - w$. As we observe in the subsequent discussion on coordination, this double margin is generally detrimental for the channel profit. Furthermore, $\tilde{p}_I^*(w)$ is increasing with respect to w, while $\tilde{m}_I^*(w), \tilde{\lambda}_I^*(w), \tilde{G}_{A,I}^*(w)$ are decreasing. Note that, as we commented in subsection 3.2, the aggregator's best response is such that the expected customer benefit is always equal to zero, even when the market is captured.

The quantity $\tilde{w}_{MC}$ determines the wholesale price below which it is optimal for the aggregator to set her own price so that the market is captured. If the market size is large, i.e., $\Lambda > \lambda_0$, then $\tilde{w}_{MC} = 0$, i.e., the provider cannot induce market capture for any value of w. To ensure the aggregator's participation, the provider should set w so that $\tilde{G}_{A,I}^*(w) \geq k$. Since $\tilde{G}_{A,I}^*(w)$ is decreasing and continuous in w, the inequality is feasible when $k \leq \tilde{G}_{A,I}^*(0)$. By considering cases for Λ, it follows that $\tilde{G}_{A,I}^*(0) = G_0(\Lambda)$. Therefore, the provider is able to induce the aggregator's participation only if her reservation level does not exceed the optimal channel profit $G_0(\Lambda)$, by charging a wholesale price $0 \leq w \leq \tilde{w}_{max}(k)$, where $\tilde{w}_{max}(k)$ is the solution of $\tilde{G}_{A,I}^*(w) = k$. From the expression of $\tilde{G}_{A,I}^*(w)$ in Proposition 1, it follows that:

$$\tilde{w}_{max}(k) = \begin{cases} R - \frac{c}{\mu-\Lambda} - \frac{k}{\Lambda}, & \text{if } k \geq \tilde{k}_{MC}, \\ R - \frac{(\sqrt{c}+\sqrt{k})^2}{\mu}, & \text{if } k < \tilde{k}_{MC}, \end{cases} \tag{10}$$

where,

$$\tilde{k}_{MC} = \tilde{G}_{A,I}^*(\tilde{w}_{MC}) = \begin{cases} \frac{c\Lambda^2}{(\mu-\Lambda)^2}, & \text{if } \Lambda \leq \lambda_0, \\ G_0(\Lambda), & \text{if } \Lambda > \lambda_0. \end{cases}$$

When $k < \tilde{k}_{MC}$, it follows that $\tilde{w}_{max}(k) > \tilde{w}_{MC}$; thus, the allowable range of w includes wholesale prices where only a fraction of customers joins. However, when the aggregator's reservation level is high, $\tilde{k}_{MC} \leq k \leq G_0(\Lambda)$, the provider is forced to set the wholesale price so low that the market is captured. When $k > G_0(\Lambda)$, the provider and the aggregator will not collaborate; the reason is that the provider cannot set a positive wholesale price that ensures the aggregator's participation. Finally, we observe that $\tilde{\lambda}_I^*(w) \leq \lambda_D^*$ and $\tilde{p}_I^*(w) \geq p_D^*$ for all $w \geq 0$. In other words, under the SP contract, the demand is always lower compared to the corresponding demand under the direct mode. This happens because the aggregator's margin leads to a higher service fee.

We next proceed to the provider's maximization problem (9). Under the direct service mode, the optimal demand is $\lambda_D^* = \min\left\{\mu - \sqrt{\frac{c\mu}{R}}, \Lambda\right\}$. The aggregator's presence implies a double marginalization effect, which reduces the demand level from the channel optimal value to $\tilde{\lambda}_I^*(k) = \min\{\mu - Z(k), \Lambda\}$, where $Z(k)$ represents the optimal idle capacity. This quantity depends on the aggregator's reservation level k as follows:

$$Z(k) = \begin{cases} Z_0, & \text{if } k \le k_T, \\ Z_1(k), & \text{if } k > k_T, \end{cases} \tag{11}$$

where, $Z_0 = \sqrt[3]{\frac{R\sqrt{\Delta}+2c\mu^2}{2R}} - \sqrt[3]{\frac{R\sqrt{\Delta}-2c\mu^2}{2R}}$, $Z_1(k) = \frac{\mu\sqrt{c}}{\sqrt{c}+\sqrt{k}}$, $\Delta = \frac{4c^2\mu^3(27R\mu+c)}{27R^3}$ and $k_T = c\left(\frac{\mu}{Z_0} - 1\right)^2$.

Theorem 1 summarizes the optimal strategies for the provider and the aggregator.

Theorem 1. *For any reservation level $k \in [0, G_0(\lambda_0)]$, and $\Lambda \ge G_0^{-1}(k)$, it holds:*

- *For $k \le k_T$ the demand in equilibrium under the SP contract is $\tilde{\lambda}_I^*(k) = \min\{\mu - Z_0, \Lambda\}$ and we have the following cases regarding the market size:*
  - *i. When $\Lambda \in [G_0^{-1}(k), \mu - Z_1(k)]$:*
    - *(a) The optimal prices are: $\tilde{w}^* = \tilde{w}_{max}(k)$ and $\tilde{p}_I^* = R - \frac{c}{\mu-\Lambda}$.*
    - *(b) The optimal profits are: $\tilde{G}_{P,I}^* = G_0(\Lambda) - k$, and $\tilde{G}_{A,I}^* = k$.*
  - *ii. When $\Lambda \in (\mu - Z_1(k), \mu - Z_0]$:*
    - *(a) The optimal prices are: $\tilde{w}^* = \tilde{w}_{MC}$ and $\tilde{p}_I^* = R - \frac{c}{\mu-\Lambda}$.*
    - *(b) The optimal profits are: $\tilde{G}_{P,I}^* = G_0(\Lambda) - \tilde{k}_{MC}$, and $\tilde{G}_{A,I}^* = \tilde{k}_{MC}$.*
  - *iii. When $\Lambda \in (\mu - Z_0, +\infty)$:*
    - *(a) The optimal prices are: $\tilde{w}^* = R - \frac{c\mu}{Z_0^2}$ and $\tilde{p}_I^* = R - \frac{c}{Z_0}$.*
    - *(b) The optimal profits are: $\tilde{G}_{P,I}^* = (\mu - Z_0)\tilde{w}^*$ and $\tilde{G}_{A,I}^* = k_T$.*
- *For $k > k_T$ the demand in equilibrium under the SP contract is $\tilde{\lambda}_I^*(k) = \min\{\mu - Z_1(k), \Lambda\}$ and we have the following cases regarding the market size:*
  - *i. When $\Lambda \in [G_0^{-1}(k), \mu - Z_1(k)]$:*
    - *(a) The optimal prices are: $\tilde{w}^* = \tilde{w}_{max}(k)$ and $\tilde{p}_I^* = R - \frac{c}{\mu-\Lambda}$.*
    - *(b) The optimal profits are: $\tilde{G}_{P,I}^* = G_0(\Lambda) - k$, and $\tilde{G}_{A,I}^* = k$.*
  - *ii. When $\Lambda \in (\mu - Z_1(k), +\infty)$:*
    - *(a) The optimal prices are: $\tilde{w}^* = \tilde{w}_{max}(k)$ and $\tilde{p}_I^* = R - \frac{c}{Z_1(k)}$.*
    - *(b) The optimal profits are: $\tilde{G}_{P,I}^* = (\mu - Z_1(k))\tilde{w}^*$ and $\tilde{G}_{A,I}^* = k$.*

Fig. 1 summarizes the optimal provider's SP policy under the various cases of Theorem 1 and demonstrates the interaction between the market size and the aggregator's reservation level. First, service through the aggregator is feasible only when $k < G_0(\Lambda)$, or equivalently $\Lambda > G_0^{-1}(k)$, since in the opposite case the market size is not sufficient to satisfy the aggregator, even if the provider forfeits his entire profit. The threshold for sufficient market size is represented with the dotted black curve.

The quantity $\mu - Z_0$ represents the demand level that maximizes the provider's profit when the market size is sufficiently large and the reservation level k is sufficiently low such that neither of them imposes

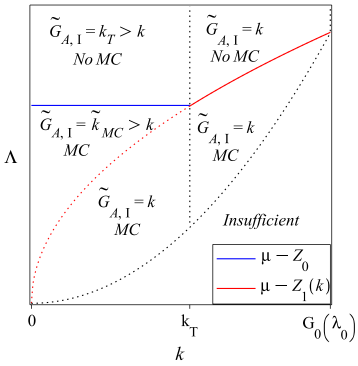


Fig. 1: Regions of Λ with respect to k

a restriction. On the other hand, the quantity $\mu - Z_1(k)$ is equal to the demand that the provider should induce, while at the same time charging at the maximum wholesale price $\tilde{w}_{max}(k)$, so that the aggregator's reservation level is attained. The solid curve in Fig. 1 corresponds to the maximum of the two quantities above and represents the optimal demand for the provider when the market size is not restrictive.

The optimal demand is generally increasing and approaches the socially optimal level as the reservation level k increases. Specifically, when the aggregator's reservation level is low ($k \leq k_T$), the optimal demand for the provider equals $\mu - Z_0$. To achieve that, he is willing to give the aggregator a price that guarantees a profit above her reservation level. This also depends on the market size. Specifically, when $\Lambda < \mu - Z_1(k)$, the provider cannot secure profit equal to k for the aggregator if he charges $w = \tilde{w}_{max}(k)$, and he is forced to lower the wholesale price below the minimum required. On the other hand, when $\mu - Z_1(k) < \Lambda < \mu - Z_0$, the provider prefers to capture the market even though he must set a wholesale price that results in aggregator profits higher than k. Finally, when $\Lambda > \mu - Z_0$, the provider prices so that the demand is equal to his profit maximizing level $\mu - Z_0$, the market is not captured, and the aggregator's profit exceeds k.

When the reservation level is high ($k > k_T$), the provider's flexibility is reduced, since now $\mu - Z_1(k) > \mu - Z_0$. Thus, he is forced to set the wholesale price at $\tilde{w}_{max}(k)$ to ensure the aggregator's participation with a profit exactly to her reservation level. If the market size is small, i.e., $\Lambda \leq \mu - Z_1(k)$, then $\tilde{w}_{max}(k) < \tilde{w}_{MC}$; thus, the wholesale price is set below the maximum level that would allow market capture. The price difference between $\tilde{w}_{max}(k)$ and $\tilde{w}_{MC}$ is sacrificed by the provider in order to increase the aggregator's profit to the minimum level required for participation. When $\Lambda > \mu - Z_1(k)$, the market is not captured, since $\tilde{w}_{max}(k)$ results in demand equal to $\mu - Z_1(k)$.

Another interesting insight arises from the behavior of the optimal wholesale price in this range of Λ. It is straightforward to show that for $\Lambda \leq \mu - Z_1(k)$, the optimal wholesale price $\tilde{w}_{max}(k)$ is increasing in Λ. This means that as the market size increases, the market is captured with a higher wholesale price. Although this seems contrary to the intuitive property that a larger market is captured with a lower price, it is explained by the aggregator's reservation level. Since the optimal price in this range is equal to $\tilde{w}_{max}(k) < \tilde{w}_{MC}$, as Λ increases the provider can increase the price and still induce market capture, thus it is optimal for him to do so. As Λ increases above $\mu - Z_1(k)$, the market capture price decreases below $\tilde{w}_{max}(k)$ and the optimal wholesale price is such that a fraction of the customers are served.

4.2.2. Time-based (TB) contract

We examine the case under which the contract includes compensation with along the service charge (w). Specifically, the provider offers customers a delay compensation l per unit of time they spend in the system. The aggregator responds optimally with a service fee $p_I^*(w,l)$, as long as her profit is at least equal to her reservation level k. Hence, the optimization problems for aggregator and provider are defined as follows:

$$G_{A,I}^*(w,l) = \max_{p\geq w} G_{A,I}(p,w,l), \tag{12}$$

$$G_{P,I}^* = \max_{w,l\geq 0}\{G_{P,I}(p_I^*(w,l),l,w) \,:\, G_{A,I}^*(w,l) \geq k\}. \tag{13}$$

For a fixed value of the delay compensation (l), the aggregator's response is equivalent to that under a SP strategy and customer waiting cost $c-l$. We thus obtain the following generalization of Proposition 1.

Proposition 2. *Given a wholesale price w and a delay compensation l,*

i.) The aggregator's optimal service fee is: $p_I^*(w,l) = w + m_I^*(w,l)$, where,

$$m_I^*(w,l) = \begin{cases} R - w - \sqrt{\frac{(c-l)(R-w)}{\mu}}, & \text{if } w \geq w_{MC}(l), \\ R - w - \frac{c-l}{\mu-\Lambda}, & \text{if } w < w_{MC}(l), \end{cases}$$

and

$$w_{MC}(l) = \begin{cases} R - \frac{(c-l)\mu}{(\mu-\Lambda)^2}, & \text{if } \Lambda \leq \mu - \sqrt{\frac{(c-l)\mu}{R}}, \\ 0, & \text{if } \Lambda > \mu - \sqrt{\frac{(c-l)\mu}{R}}. \end{cases}$$

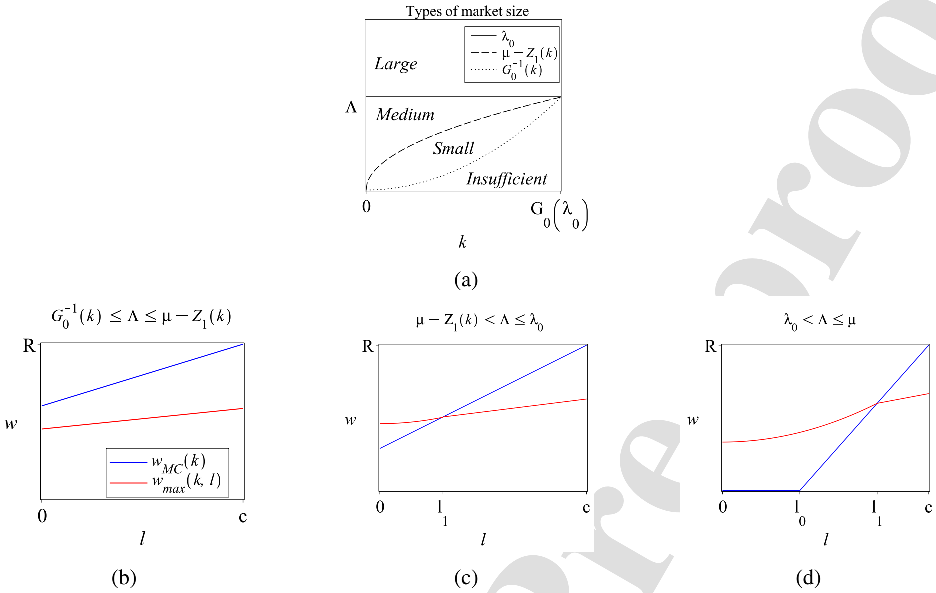
ii.) The equilibrium demand is:

$$\lambda_I^*(w,l) = \begin{cases} \mu - \sqrt{\frac{(c-l)\mu}{R-w}}, & \text{if } w \geq w_{MC}(l), \\ \Lambda, & \text{if } w < w_{MC}(l). \end{cases}$$

iii.) The optimal aggregator's profit is:

$$G_{A,I}^*(w,l) = \begin{cases} \left(\sqrt{(R-w)\mu} - \sqrt{c-l}\right)^2, & \text{if } w \geq w_{MC}(l), \\ \Lambda\left(R - w - \frac{c-l}{\mu-\Lambda}\right), & \text{if } w < w_{MC}(l). \end{cases}$$

The aggregator responds to the provider's strategy (w,l) by adding a profit margin $m_I^*(w,l)$ for herself. This margin is equal to the optimal price that a provider would set under direct service mode, customer service value $R-w$ and waiting cost $c-l$. The delay compensation allows the aggregator to increase her service fee, without hurting the demand. Indeed, in Section 5 we show that the optimal level of compensation leads the system to coordination. We also observe that $p_I^*(w,l)$ is increasing in l and w, while $m_I^*(w,l)$, $\lambda_I^*(w,l)$, $G_{A,I}^*(w,l)$ are decreasing in w and increasing in l. By incorporating a delay compensation in his pricing strategy, the provider increases the range of wholesale prices and market sizes Λ, under which market is captured. Indeed, both the maximum market size for market capture, which is equal to $\mu - \sqrt{\frac{(c-l)\mu}{R}}$,



(a)

(b) (c) (d)

Fig. 2: Types of market size and pairs (w, l) of aggregator's participation

as well as the wholesale price threshold $w_{MC}(l)$ are increasing in l. This is intuitive since as l increases, the option of joining becomes more attractive to customers.

To ensure the aggregator's participation, the provider should set pairs of w, l such that $G^*_{A,I}(w, l) \ge k$. In a straightforward generalization of the SP strategy, by substituting c with $c - l$, we obtain the maximum feasible wholesale price as a function of l:

$$w_{max}(k, l) = \begin{cases} R - \frac{c-l}{\mu-\Lambda} - \frac{k}{\Lambda}, & \text{if } k \ge k_{MC}(l), \\ R - \frac{(\sqrt{c-l}+\sqrt{k})^2}{\mu}, & \text{if } k \le k_{MC}(l), \end{cases}$$

where,

$$k_{MC}(l) = G^*_{A,I}(w_{MC}(l), l) = \begin{cases} \frac{(c-l)\Lambda^2}{(\mu-\Lambda)^2}, & \text{if } \Lambda \le \mu - \sqrt{\frac{(c-l)\mu}{R}}, \\ \left(\sqrt{(R-w)\mu} - \sqrt{c-l}\right)^2, & \text{if } \Lambda > \mu - \sqrt{\frac{(c-l)\mu}{R}}. \end{cases}$$

Given any compensation l, if the aggregator's reservation level is below the threshold $k_{MC}(l)$, the provider has flexibility to capture the market or not, by setting an appropriate wholesale price. In the opposite case, his feasible range consists only of wholesale prices such that the market is captured. The maximum feasible price $w_{max}(k, l)$ is increasing in l, consistently with the fact that a higher delay compensation increases the customers' willingness to join, allowing the aggregator's participation even for higher values of the wholesale price.

Fig. 2 illustrates the relationship between the two threshold prices ($w_{MC}(l)$ for market capture and $w_{max}(l)$ for aggregator's participation). The feasible pairs (w, l) that secure the aggregator's participation correspond to the points below the $w_{max}(k,l)$ curve. Therefore, if $w_{MC}(l) > w_{max}(k,l)$, the provider is always forced to capture the market. If $0 < w_{MC}(l) < w_{max}(k,l)$, he has a choice to capture the market or not by setting an appropriate w, and if $w_{MC}(l) = 0$, market capture is not possible for the given value of l.

By comparing $w_{MC}(l)$ and $w_{max}(k,l)$ for different values of k and Λ, we may characterize the market size as insufficient, small, medium or large. Specifically, the insufficient market corresponds to the case where the provider cannot ensure a positive profit from the aggregator's participation, i.e., $\Lambda < G_0^{-1}(k)$. When $G_0^{-1}(k) \le \Lambda \le \mu - Z_1(k)$, which corresponds to Fig. 2b, $w_{MC}(l) > w_{max}(k,l)$ for all $l \in [0,c]$. The market size is so small that, in order to ensure the aggregator's participation, the provider must price so that the market is captured regardless of his choice of compensation level. We refer to this case as small market size. The case of medium market size, in Fig. 2c, corresponds to values of k,Λ such that $\mu - Z_1(k) < \Lambda \le \lambda_0$. If the delay compensation is low, $l \le l_1(k,\Lambda) = c - \frac{k(\mu-\Lambda)^2}{\Lambda^2}$, the market may or may not be captured depending on the value of w. However for larger values of l, any feasible wholesale price induces market capture. Finally, when $\Lambda > \lambda_0$, which is displayed in Fig. 2d, the market size is large, in the sense that for low compensation levels $l \le l_0 = c - \frac{R(\mu-\Lambda)^2}{\mu}$ capturing the market is never feasible. For $l_0 < l \le l_1$ it depends on the choice of w, and for $l > l_1$ is always forced. Note that when $\Lambda > \mu$, $l_0 = c$, as expected, since market capture is never possible.

The ranges of (k,Λ) corresponding to the three cases above are displayed in Fig. 2a. The graph shows that the market type depends not only on Λ, but also on the aggregator's reservation level. In particular, as k increases, the range of Λ where the market is considered small increases, which reflects the fact that the provider is under stronger pressure to capture the market in order to ensure the aggregator's participation.

We next proceed to the solution of the provider's optimization problem (13). Theorem 2 summarizes the optimal pricing strategies for both parties. The key insight of this result is that the extra flexibility offered by the TB contract allows the service provider to bring the demand back to the channel optimal level. The main idea of the proof is to seek pricing strategies of the form $(w_{max}(k,l), l)$, so that: i) the aggregator's profit is restricted to the reservation level k; ii) the optimal value of l results in provider's profit equal to the upper bound $G_0(\Lambda) - k$ in all market size cases.

Theorem 2. *For any reservation level $k \in [0, G_0(\lambda_0)]$, and $\Lambda \ge G_0^{-1}(k)$, the demand in equilibrium under the TB contract is $\lambda_I^* = \min\{\lambda_0, \Lambda\}$ and the optimal profits of the two parties are $G_{P,I}^* = G_0(\Lambda) - k$ and $G_{A,I}^* = k$. The optimal pricing strategies for provider and aggregator are any (w^*, l_I^*) and p_I^* respectively that satisfy the following:*

i. When $\Lambda \in [G_0^{-1}(k), \mu - Z_1(k)]$:

$$l_I^* \in [0,c],\ w^* = w_{max}(k, l_I^*) = R - \frac{c - l_I^*}{\mu - \Lambda} - \frac{k}{\Lambda},\ p_I^* = R - \frac{c - l_I^*}{\mu - \Lambda}.$$

ii. When $\Lambda \in (\mu - Z_1(k), \lambda_0]$:

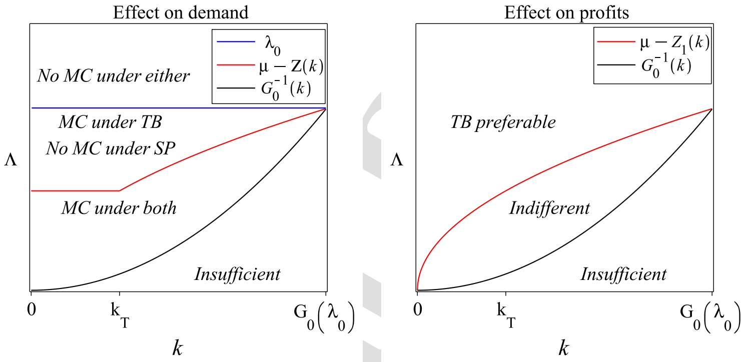
$$l_I^* \in [l_1(k,\Lambda), c],\ w^* = w_{max}(k, l_I^*) = R - \frac{c - l_I^*}{\mu - \Lambda} - \frac{k}{\Lambda},\ p_I^* = R - \frac{c - l_I^*}{\mu - \Lambda}.$$

iii. When $\Lambda \in (\lambda_0, +\infty)$:

$$l_I^* = c - \frac{k(\mu - \lambda_0)^2}{\lambda_0^2} = l_1(k, \lambda_0),\ w^* = w_{max}(k, l_I^*) = R - \frac{k\mu}{\lambda_0^2},\ p_I^* = R - \frac{k(\mu - \lambda_0)}{\lambda_0^2}.$$

Theorem 2 shows that the provider's optimal strategy is to capture the market in the small and medium market cases. In both cases, the optimal strategy is not unique, since any pair $(w_{max}(k,l),l)$ such that $w_{max}(l) \leq w_{MC}(l)$ is optimal. In particular under the small market case, the value $l = 0$ is in the range of optimal compensation levels, which is consistent with Theorem 1 and implies that the provider is indifferent between SP and TB. However in the medium market case, the optimal strategies are those with $l \geq l_1(k,\lambda_0)$; therefore, the SP strategy is not optimal. Finally when the market is large, there is a unique optimal strategy such that the demand is equal to the optimal value under direct mode (λ_0), requiring a positive compensation level. Thus, although under the direct mode structure the provider may induce the profit maximizing demand without the need of delay compensation, under the indirect mode this can be achieved only with TB pricing.

Fig. 3 summarizes the comparison between SP and TB with respect to the final demand and the provider's profit, on Fig. 3a and Fig. 3b, respectively. The provider is indifferent between SP and TB under small market, while he strictly prefers TB for medium and large markets. Regarding the demand, the comparison is somewhat more involved. Specifically, both pricing strategies induce market capture under small market and in part of the medium market range, they have different effects in the remaining part of the medium range, and none of them induces market capture for large market.



(a) (b)

Fig. 3: The effect of TB on market capture and provider's profit

5. Selection of Service Mode and Coordination

In this section, we consider the provider's choice of service mode under the reselling format framework, and establish appropriate thresholds on the contact cost S for selecting indirect service, depending on the pricing scheme in use. We then identify conditions so that channel coordination is attained.

5.1. Selection of service mode

The provider's choice of service mode depends on the customer contact cost under direct service, the aggregator's reservation level, the market size and the pricing scheme, SP or TB, used in indirect service. First, when $G_0(\Lambda) < \min\{S,k\}$, then the market size is insufficient for both direct and indirect service and the provider will not participate under either service mode. When $S \leq G_0(\Lambda) \leq k$ he prefers direct service. Similarly, when $k \leq G_0(\Lambda) \leq S$ he prefers indirect service, under any pricing scheme.

When the market size is sufficient for both service modes, i.e., $G_0(\Lambda) \geq \max\{S,k\}$, the provider's choice between them depends on the pricing scheme. Under direct service his optimal profit is $G_0(\Lambda) - S$ and can be achieved either with a SP or by selecting from an infinite collection of TB pricing strategies. If he selects indirect service under SP, then his profit is $\tilde{G}^*_{P,I}$. Since $\tilde{G}^*_{P,I} \leq G_0(\Lambda)$, by selecting the indirect mode the provider sacrifices part of his profit from providing service in order to save the customer contact cost. Thus, it is optimal to select indirect service if $S \geq \tilde{S}_0 = G_0(\Lambda) - \tilde{G}^*_{P,I}$. By considering the various cases of Theorem 1, it follows that $\tilde{S}_0$ is increasing in Λ and is stabilized when $\Lambda \geq \lambda_0$. Therefore, for a fixed value of $S < \tilde{S}_0(\lambda_0)$, the provider prefers the indirect service if the market size is lower than a threshold (that also depends on k) and remains in direct service otherwise. The intuition is that if the market is not sufficiently large, the provider cannot make enough profit from directly servicing the customers to compensate for the contact costs and prefers to enlist to the aggregator's platform. For larger values of S direct service is never preferred.

On the other hand, under TB pricing the provider's profit is always equal to $G^*_{P,I} = G_0(\Lambda) - k$ and it is optimal to select the indirect service if $S \geq S_0 = k$. Since $\tilde{G}^*_{P,I} \leq G^*_{P,I}$, the two cost thresholds satisfy $S_0 \leq \tilde{S}_0$, thus for a given market size the provider is generally more willing to select indirect service if he can employ a TB pricing strategy than if he is restricted to a SP scheme. In summary, when $k > G_0(\Lambda)$, the provider will offer direct service if $S \leq G_0(\Lambda)$ and no service otherwise. When $k \leq G_0(\Lambda)$:

i. If $S < S_0$, then the provider will always select direct service.
ii. If $S_0 \leq S < \tilde{S}_0$, the provider will select indirect service mode only if TB pricing is allowed.
iii. If $S \geq \tilde{S}_0$, the provider will always select indirect service.

Regarding the centrally optimal decision about the service mode, it follows from the discussion in subsection 3.3 and equation (6) that when $G_0(\Lambda) < \min\{S, k\}$ the channel cannot afford to provide service under either mode. If $S < \min\{k, G_0(\Lambda)\}$, the preferable mode is the direct, and otherwise the indirect. Fig. 4a shows the different ranges of (k, S) in which direct or indirect service mode is optimal for the Provider (P) and the channel (C). The blue and the red lines correspond to the thresholds $\tilde{S}_0$ and S_0 respectively.

5.2. Channel Coordination

Coordination is attained when the provider's choice of the service mode coincides with the channel optimal one, and at the same time the demand induced by the provider's decentralized strategy is equal to the socially optimal level λ_0.

When $G_0(\Lambda) < \min\{S, k\}$ coordination is trivially attained since providing no service is optimal for all parties. The following two theorems summarize the conditions for coordination under SP and TB pricing, when $G_0(\Lambda) \geq \min\{S, k\}$. For both results the proof is based on comparing the service mode selections and the channel profits between the decentralized and centralized cases, for the various ranges of S and k, derived in Sections 4 and 5.1.

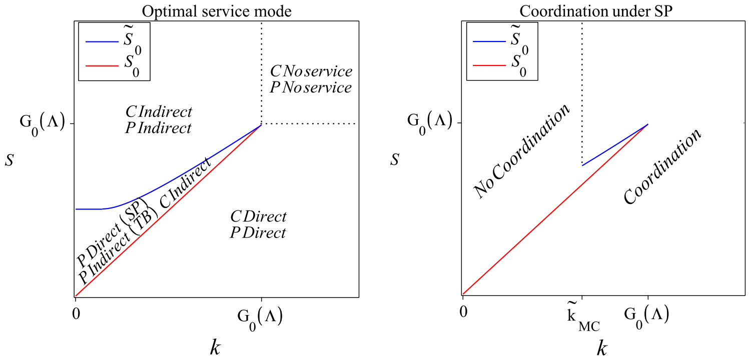
Theorem 3. *Under SP:*

i. When $S < k$ coordination is always attained.

ii. When $S \geq k$ coordination is attained if and only if the optimal provider's policy is to select indirect service and capture the market.

Theorem 4. *Under TB, coordination is always attained.*

In Fig. 4 we illustrate the results of this Section with respect to the cost S and reservation level k, for a given value of the market size Λ. Fig. 4a presents the comparison between provider and channel selection of service mode. The decisions of the channel and the provider about the optimal service mode always coincide, with the only exception when $S \in [S_0, \tilde{S}_0]$ and $k \leq G_0(\Lambda)$ under SP.



(a) (b)

Fig. 4: Optimal service modes and coordination under SP

Fig. 4b shows the coordination region under SP. Coordination is not attained in two cases: (i) when the provider and channel optimal service modes do not coincide, and (ii) when indirect service is provider and channel optimal and in addition $k < \tilde{k}_{MC}$. In the last case the market is sufficiently large so that the provider prefers to set a high price that does not capture it and at the same time the aggregator secures a profit higher than her reservation level. The lack of coordination is due to double marginalization. As the market size Λ increases, $\tilde{k}_{MC}$ approaches $G_0(\Lambda)$ and the no-coordination region expands. Under TB, the optimal choice regarding the service mode for both the provider and the centralized system coincides, since it only depends on the difference between the aggregator's reservation level and the provider's contact cost. In that case, coordination is always attained.

6. Revenue-Sharing (RS) Contract

A prevalent contracting mechanism in several demand aggregator platforms is Revenue Sharing (RS) (Cachon and Lariviere, 2005), under which one of the two parties sets the customer price and the revenue is shared between the provider and the aggregator in specified proportions. This framework includes the agency-selling format, where the aggregator sets the sharing proportion and the provider sets the customer price. In this section we explore the role that RS agreements may play, in connection with the single-price and time-based contracts we analyzed above.

In our model, under a RS contract with customer price p and aggregator's proportion γ, the provider's and aggregator's revenue functions are $G_A(p,\gamma) = \gamma p \lambda^e(p)$ and $G_P(p,\gamma) = (1-\gamma)\lambda^e(p)p$, respectively. In the presence of reservation levels, the necessary conditions for participation are: $G_A(p,\gamma) \geq k$ and $G_P(p,\gamma) \geq G_0(\Lambda) - S$, which imply that $\gamma \geq \frac{k}{p\lambda^e(p)}$ and $\gamma \leq 1 - \frac{G_0(\Lambda)-S}{p\lambda^e(p)}$ respectively. In this case, the feasible range of γ is:

$$\underline{\gamma}(p) \leq \gamma \leq \overline{\gamma}(p), \tag{14}$$

where $\underline{\gamma}(p) = \frac{k}{p\lambda^e(p)}$ and $\overline{\gamma}(p) = 1 - \frac{G_0(\Lambda)-S}{p\lambda^e(p)}$.

The collaboration between the provider and the aggregator is feasible if the customer price p is such that $\underline{\gamma}(p) \leq \overline{\gamma}(p)$. From the above expressions this condition is equivalent to:

$$p\lambda^e(p) \geq G_0(\Lambda) - S + k, \tag{15}$$

which is the intuitively expected requirement that the revenue generated must be sufficient in order to cover at least both parties' reservation levels.

Such prices exist if and only if $\max_{0\le p<R}(p\lambda^e(p)) \ge G_0(\Lambda)-S+k$. However as we have seen in Section 4, $p\lambda^e(p)$ is maximized for $p = p_0$ and the maximum value is equal to $G_0(\Lambda)$. Combining the above, a necessary and sufficient condition for the existence of a feasible RS contract is that $k \le S \le G_0(\Lambda)$.

Based on this discussion we first observe that if $k \le S \le G_0(\Lambda)$, then the RS contract can coordinate the channel, since the optimal price p_0 is feasible for collaboration. In this case, the range of the aggregator's revenue share from 14 becomes:

$$\underline{\gamma}(p_0) = \frac{k}{G_0(\Lambda)} \le \gamma \le \frac{S}{G_0(\Lambda)} = \overline{\gamma}(p_0).$$

The condition states that, in order for the collaboration to be profitable for both parties, the aggregator must achieve her reservation level k, however she cannot demand more than the provider's savings from moving to the indirect channel. In connection with Theorem 3, when $k \le S$ the RS contract always allows coordination under collaboration, whereas the SP contract does so only under market capture. Fig. 5 shows the provider's and aggregator's revenues as a function of γ under coordination, as well as the range of γ where collaboration is preferred.

Furthermore, although p_0 is the optimal price for both the provider and the aggregator regardless of which party determines it, there may be situations where this price is not feasible for other reasons such as price regulation, etc. (see, e.g. Benioudakis et al. (2021)). If $k \le S \le G_0(\Lambda)$ holds, then from (15) and the fact that $p\lambda^e(p)$ is concave in p it follows that there exists a price range $p_L \le p \le p_H$ in which collaboration is profitable for both parties under the RS contract. Clearly this range includes p_0. We summarize the above in the following Theorem:

Theorem 5. *Under the RS contract, when $k \le S \le G_0(\Lambda)$,*

i. There exists a price range $p \in [p_L, p_H]$ in which collaboration is profitable for both parties, for values of the aggregator's portion in the range of $\gamma \in [\underline{\gamma}(p), \overline{\gamma}(p)]$.

ii. The optimal price for both the provider and the aggregator is p_0, which is inside the feasible price range $[p_L, p_H]$. Price p_0 leads to coordination for the aggregator's portion in the range $\gamma \in [\underline{\gamma}(p_0), \overline{\gamma}(p_0)]$.

We next consider the effect of the market size. We observe that $\underline{\gamma}(p_0)$ and $\overline{\gamma}(p_0)$ are decreasing in Λ for $\Lambda \le \lambda_0$, and become constant when the market is large ($\Lambda > \lambda_0$) and it is not optimal to be captured. Thus, under small and medium market the provider must generally give a larger portion of the revenue to the aggregator. Although this may seem counter-intuitive, the reason is that the aggregator's reservation level is independent of the market size, whereas for the provider it is increasing. Therefore, when the market and consequently the total revenue is not large, the provider's reservation level is reduced. On the other hand he must yield a higher portion of the revenue to the aggregator in order for her to participate.

Theorem 5 provides ranges for the aggregator's portion so that cooperation is profitable for both parties under any price, while coordination is achieved under price p_0. Which value of γ is actually selected depends on the market power of each party. It may be arbitrarily set by one of the two or determined through bargaining. For example, in prevalent platforms such as Airbnb and Booking, the common practice is that the aggregator has sufficient power to set the value of γ. In such platforms there is not a real need for delay compensation, and RS contracts are appropriate instruments for coordination.

On the other hand, the main application framework of our model is on platforms where the provider offers services that are tailored to the needs of individual customers such as maintenance of specialized equipment, repair services, etc., where customers are delay sensitive and react to delays in a strategic manner. In such

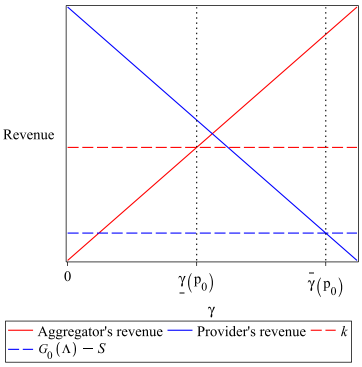


Fig. 5: Revenue-sharing with respect to γ under the optimal price p_0

situations pricing policies such as the TB contract are meaningful, since they offer customers a sense of fairness (although the customer expected net benefit is zero in equilibrium). In terms of profit allocation, the TB contract is equivalent to a coordinating RS contract where the provider determines the value of γ, since under both the aggregator receives her reservation level k. However, to the extent that TB pricing offers desirable side benefits such as the above, the negotiations between the provider and the aggregator may result in a combination of TB and RS contracts that coordinate the channel and are satisfactory for both parties.

7. Computational Experiments

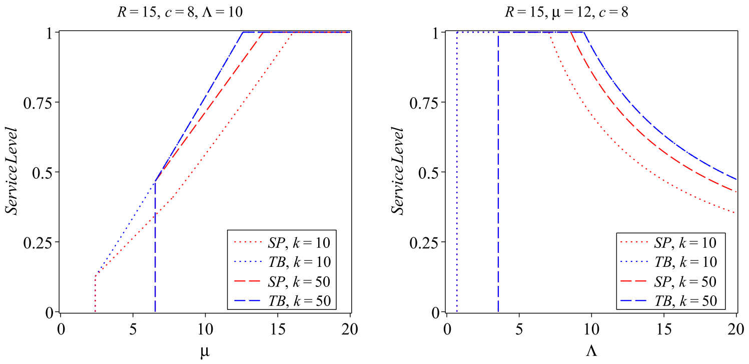
In this section, we conduct several numerical experiments that allow us to investigate the sensitivity of the service level and the degree of inefficiency due to the selfish behavior of the participants. Specifically, in subsection 7.1, we examine how model parameters such as the service rate (μ), the aggregator's reservation level (k), and the market size (Λ) affect the service level. In subsection 7.2, we explore the Price of Anarchy (PoA), which is a common measure of inefficiency (Ghosh and Hassin, 2021). In the following computational experiments, we use a base case of parameter values $R = 15, c = 8, \mu = 12, \Lambda = 10$ and in each case let one or more of the parameters to vary.

7.1. Service level

The service level is defined as the fraction of the market that is actually served, i.e., $\frac{\lambda^*}{\Lambda}$, where λ^* denotes the demand under equilibrium. We compare the service level between SP and TB contracts under indirect service. We focus on its sensitivity with respect to the service rate and the market size under both pricing strategies, for a low ($k = 10 < k_T$) and a high ($k = 50 < k_T$) value of the aggregator's reservation level k. The results are presented in Fig. 6, where the red and the blue lines represent the SP and the TB contract, respectively, while the dashed and the dotted lines correspond to the high and the low values of k.

In contrast to the direct service case, the service level has a discontinuity jump. In order for the indirect service to be feasible for the provider, the rate of incoming orders must be sufficiently high so that $G_0(\Lambda) \geq k$. The point of discontinuity in Fig. 6a corresponds to to the minimum capacity under the given values of Λ and k, such that this condition is satisfied. This minimum is the same for both contracts (SP and TB). Similarly, the discontinuity in Fig. 6b corresponds to the minimum market size that makes indirect service feasible for the provider under the given μ and k.

In general, TB results in higher service level than SP as expected. Moreover, under SP, the service level increases in k, although the provider's profit is decreasing. This is so because the provider sets his price so



(a)

(b)

Fig. 6: Service level with respect to μ and Λ

that customer demand is higher, in order to keep the aggregator in the deal. Furthermore, as k increases, the minimum required capacity for collaboration increases. The behavior with respect to k is different under TB. Here the profit maximizing demand is $\lambda^* = \lambda_0$, independent of k. The only effect of the reservation level is that it increases the minimum required capacity so that λ^* is attainable.

Regarding the sensitivity with respect to the market size in Fig. 6b, we obtain similar insights for the minimum required market size for collaboration. For completeness we provide the corresponding figures regarding the waiting cost c and the customer's service value R in the appendix, from which analogous insights can be derived.

7.2. Price of Anarchy

PoA is a measure that quantifies the loss in system profits due to the selfish behavior of individual parties, which in this model is manifested through double marginalization under the SP contract. PoA is defined as the ratio of the optimal profit between the centralized and the decentralized setting. Since the system is coordinated under TB, we explore the behavior of PoA under SP, with respect to the market size Λ. Thus,

$$PoA(\Lambda) = \frac{G^*_C(\Lambda)}{\tilde{G}^*_{P,I}(\Lambda) + \tilde{G}^*_{A,I}(\Lambda)}. \tag{16}$$

The results are presented in Fig. 7, for a low and a high value of k ($k = 10$ and $k = 18$). In both figures the blue curves correspond to a low value of the direct service cost $S = 20$, which makes the provider's choice of channel dependent on the market size. The red curves correspond to any S above the corresponding value of $\tilde{S}_0(\lambda_0)$ (approximately equal to 33 and 35 in the left and right panel, respectively), under which the provider prefers indirect service for any market size. This allows us to compare the case where the provider's choice about the service mode may deviate from the centrally optimally one, with the case that the choice of service mode is always centrally optimal. In all cases $S > k$, since otherwise coordination would occur and $PoA = 1$ would hold trivially. The dotted vertical lines correspond to the boundaries between insufficient, small, medium and large market sizes, as discussed in Section 4.

A first observation is that PoA is always bounded for each market size. Specifically, for small market, there is no inefficiency, which is reasonable due to coordination, thus $PoA = 1$. When the market becomes medium, it starts depending on the requirements of the aggregator and the cost S. For low S, the provider

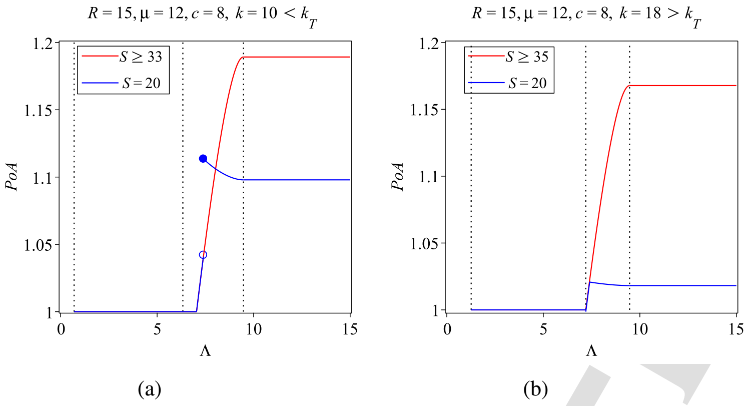


Fig. 7: Price of anarchy

will deviate from the centrally optimal service mode. At this point the inefficiency reaches its maximum level and afterwards decreases. On the other hand, for higher values of k, the level of inefficiency increases up to the beginning of the large market region where it is stabilized and takes its maximum value.

Another interesting insight is that when k is low, there is a point of discontinuity when the provider deviates from the centrally optimal service mode. This happens because at the point of switch, the aggregator loses the opportunity to earn $k_T > k$, while the provider is indifferent between the two service modes. This gap does not exist when k is high, because the aggregator receives the same profit k regardless of the service mode. In this case the maximum inefficiency of the system is significantly lower. Finally, the aggregator's reservation level and the provider's contact cost have opposite effects on the inefficiency. As we have seen in Theorem 1 and Fig. 1, when k increases, the number of incoming orders approaches the socially optimal level and PoA is reduced. On the contrary, a high value of S reduces the provider's flexibility to change the service mode and this generally increases the inefficiency.

8. Conclusion

This paper is motivated by a practical dilemma that several manufacturers and/or service providers face on how to penetrate a market, either directly or through a (demand aggregator) platform. We focused on the reselling format case, where providers set the wholesale price and let the platform determine the retail price. On the customer side, we assumed that they are time-sensitive and react strategically to delays. We investigated three contracting and pricing mechanisms; two for the reselling model (single-price and time-based pricing) and one for the agency selling model (revenue-sharing).

The key managerial insights derived from our work are the following. Under the direct service mode, the delay compensation does not make any difference in the provider's profitability; however, it offers flexibility to affect the demand pattern, by selecting among appropriately defined combinations of service price and delay compensation. Under the indirect service, we considered both a reselling and a agency selling model. In the former model, offering delay compensation is beneficial when the market size is large, since it allows to increase the demand significantly. We showed that coordination is always attained under time-based pricing, while under a single price it depends on the interaction among the provider's contact cost, the aggregator's reservation level, and the market size. Under the agency selling model, a revenue-sharing contract has been analyzed. We identified appropriate ranges of price and aggregator's share that lead to coordination. Through extensive numerical experiments, we explored the behavior of the service level under different contracts and its sensitivity to model parameters. We also showed that: i) the price of anarchy is always bounded and ii) the aggregator's reservation level and the provider's contact cost have opposite effects on inefficiency.

Many directions seem promising for future research. First, the model structure can be generalized, to allow for more than one providers and/or more than one platforms. The providers may have flexibility in the level of cooperation with each platform, for instance to allocate only customer contact activities, part of the service, product delivery, etc. In this way a multiple service mode can be incorporated into a provider's choices; in addition to direct service, he may also cooperate with one or more platforms. Furthermore, all customers may not have access to all platforms, which effectively separates the market into multiple segments, and different pricing mechanisms may be allowed for each one of those. For example, a provider may decide to keep his own customer base through direct service and offer a group pricing option to the aggregator who brings to him a new market. Other directions of interest include exploring the effect of platform enrollment on the provider's capacity decisions, and on inventory control policies for cases where stocking a physical product is also involved, as well as incorporating information asymmetry between the provider and the aggregator when the latter has private information about the market.

Appendices

A. Notation

We have used the following notation throughout the paper. The symbols "$\sim$" and "$*$" represent the case in which the delay compensation is not allowed and the optimal value/profits respectively.

- D, I: Direct mode/Indirect mode
- P, A: Provider/Aggregator
- SP, TB: Single-price/Time-based pricing strategy
- MC: Market capture
- μ: Service rate
- R: Customer's service value
- c: Cost of waiting per unit time
- k: Aggregator's reservation level
- S: Provider's customer contact cost rate under direct mode
- Λ: Market size
- G_0: Socially optimal total benefit per unit time
- λ_0: Socially optimal rate of incoming orders
- $S_0, \tilde{S}_0$: Contact cost thresholds
- λ: Rate of incoming orders
- p: Service fee
- l: Delay compensation per unit time
- w: Wholesale price
- w_{MC}: The wholesale price below which it is optimal for the aggregator to set her own price so that the market is captured
- k_{MC}: Aggregator's profit for w_{MC}
- λ^e: Equilibrium rate of incoming orders

- γ: Aggregator's revenue share
- $G_{A,D}, G_{A,I}$: Aggregator's profit under direct and indirect mode
- $G_{P,D}, G_{P,I}$: Provider's profit under direct and indirect mode
- G_C: Centralized profit

B. Proofs

Proof of Proposition 1. If we set $m = p - w$ we have:

$$\tilde{G}^*_{A,I}(p) = \lambda^e(m + w, 0)m. \tag{B.1}$$

Moreover, from (2) with $p = m + w$ and $l = 0$:

$$B(p, 0, \lambda; R) = R - w - m - \frac{c}{\mu - \lambda} = B(m, 0, \lambda; R - w), \tag{B.2}$$

which implies that $\lambda^e(m + w, 0; R) = \lambda^e(m, 0; R - w)$

Therefore, the problem is equivalent with the profit maximization problem in Edelson and Hilderbrand (1975) with customer service value equal to $R - w$. Thus the profit maximizing value $\tilde{m}^*_I$ is such that the demand is equal to:

$$\tilde{\lambda}^*_I(w) = \min\left\{\mu - \sqrt{\frac{c\mu}{R - w}}, \Lambda\right\}.$$

Under the optimal policy the market is captured when $\Lambda \leq \mu - \sqrt{\frac{c\mu}{R-w}}$, i.e., when $\Lambda \leq \mu - \sqrt{\frac{c\mu}{R}} = \lambda_0$ and $w \leq R - \frac{c\mu}{(\mu-\Lambda)^2}$. When $\Lambda > \lambda_0$, market capture is not possible for any nonnegative value of w. We thus define,

$$\tilde{w}_{MC} = \begin{cases} R - \frac{c\mu}{(\mu-\Lambda)^2}, & \text{if } \Lambda \leq \lambda_0, \\ 0, & \text{if } \Lambda > \lambda_0, \end{cases}$$

and we have that the optimal service fee is equal to $\tilde{p}^*_I(w) = w + \tilde{m}^*_I(w)$, where:

$$\tilde{m}^*_I(w) = \begin{cases} R - w - \sqrt{\frac{c(R-w)}{\mu}}, & \text{if } w \geq \tilde{w}_{MC}, \\ R - w - \frac{c}{\mu-\Lambda}, & \text{if } w < \tilde{w}_{MC}. \end{cases}$$

The optimal aggregator's profit is:

$$\tilde{G}^*_{A,I}(w) = \begin{cases} \left(\sqrt{(R - w)\mu} - \sqrt{c}\right)^2, & \text{if } w \geq \tilde{w}_{MC} \\ \Lambda\left(R - w - \frac{c}{\mu-\Lambda}\right), & \text{if } w < \tilde{w}_{MC} \end{cases}.$$

□

Proof of Theorem 1. We first consider the case where the provider is not constrained by the market size and the aggregator's reservation level. Then from Proposition 1, the range of λ he can achieve with a nonnegative

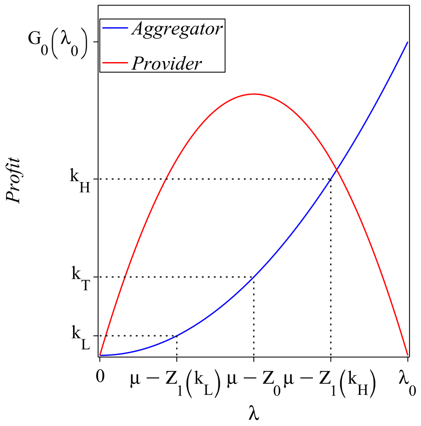


Fig. 8: Profits with respect to λ

price $w \geq 0$ is $0 \leq \lambda \leq \lambda_0$. The inverse form of the demand is:

$$w(\lambda) = \tilde{w}_{MC}(\lambda) = R - \frac{c\mu}{(\mu - \lambda)^2}. \tag{B.3}$$

Therefore, the provider's optimization problem becomes:

$$\max_{0 \leq \lambda \leq \lambda_0} \lambda \left(R - \frac{c\mu}{(\mu - \lambda)^2} \right) = \max_{0 \leq \lambda \leq \lambda_0} f(\lambda).$$

The first and the second derivative of this function are: $f^{'}(\lambda) = R - c\mu\frac{\mu+\lambda}{(\mu-\lambda)^3}$ and $f^{''}(\lambda) = -c\mu\frac{2(2\mu+\lambda)}{(\mu-\lambda)^4} < 0$. Thus, $f(\lambda)$ is concave in $[0, \lambda_0]$, where $f(0) = f(\lambda_0) = 0$. Furthermore, $f^{'}(0) = R - \frac{c}{\mu} > 0$ and $f^{'}(\lambda_0) < 0$. $f(\lambda)$ is illustrated by the red line in Fig. 8.

Therefore, there is a unique maximizing value of $\lambda \in [0, \lambda_0]$ such that $f^{'}(\lambda) = 0$. This is equivalent to $y^3 + \frac{c\mu}{R}y = \frac{2c\mu^2}{R}$, where $y = \mu - \lambda$ and $y \in [\sqrt{\frac{c\mu}{R}}, \mu]$. We can solve this equation with respect to y using the Cardano Method. Specifically let: $3st = \frac{c\mu}{R}$ and $s^3 - t^3 = \frac{2c\mu^2}{R}$, where the solution for the latter pair of equations is $y = s - t$. By solving the above system we derive that:

$$t = \sqrt[3]{\frac{\frac{-2c\mu^2}{R} \pm \sqrt{\Delta}}{2}}.$$

Therefore we have two solutions. From the properties of the cube root and since $\frac{\sqrt{\Delta}}{2} > \frac{c\mu^2}{R}$ we can easily show that both solutions result in the same value of y namely,

$$y = s - t = \sqrt[3]{\frac{c\mu^2}{R} + \frac{\sqrt{\Delta}}{2}} - \sqrt[3]{-\frac{c\mu^2}{R} + \frac{\sqrt{\Delta}}{2}} = Z_0.$$

Thus the optimal ordering rate is $\tilde{\lambda}_I^* = \mu - Z_0$. By substituting, we derive the optimal wholesale price $\tilde{w}^* = R - \frac{c\mu}{Z_0^2}$ and the optimal service fee $\tilde{p}_I^* = R - \frac{c}{Z_0}$.

Given (B.3), the aggregator's profit under her optimal response to λ and $w(\lambda)$ is equal to:

$$\tilde{G}_{A,I}(\lambda) = c\left(\frac{\lambda}{\mu - \lambda}\right)^2 = \tilde{k}_{MC}(\lambda). \tag{B.4}$$

(B.4) is illustrated with the blue line in Fig. 8. It is increasing in λ and $\tilde{k}_{MC}(0) = 0$, $\tilde{k}_{MC}(\lambda_0) = G_0(\lambda_0)$. Therefore, her optimal profit for the rate of incoming orders that maximizes the provider's profit ($\lambda = \mu - Z_0$) is equal to $c\left(\frac{\mu}{Z_0} - 1\right)^2 = k_T$. Taking into account the aggregator's reservation level k, the latter solution is feasible when $R - \frac{c\mu}{Z_0^2} \leq \tilde{w}_{max}(k)$ or equivalently $k \leq k_T$. When $k > k_T$, and due to the monotonicity of the provider's and the aggregator's profit in Fig. 8, the optimal w is equal to $\tilde{w}_{max}(k)$, i.e., the wholesale price that keeps the aggregator's profit to her reservation level k. The corresponding rate of incoming orders now becomes $\tilde{\lambda}_I^*(w) = \mu - Z_1(k)$. By combining the cases we have the form in (11).

The above cases are valid when the market size is not restrictive. When we also take Λ into account, the rate becomes $\min\{\mu - Z_0, \Lambda\}$ or $\min\{\mu - Z_1(k), \Lambda\}$ depending on k, and we easily derive the expressions for the other cases in Theorem 1. □

Proof of Proposition 2. The proof goes along the same lines with the proof of Proposition 1 using the results of Edelson and Hilderbrand (1975) where customers have service value equal to $R - w$ and waiting cost $c - l$. □

Proof of Theorem 2. We first consider the case where the provider is not constrained by the market size, i.e., $\Lambda > \mu$. Using the results from Proposition 2 and the observations from the discussion in Fig. 2, we will solve the problem (13) under the constraint $G_{A,I}^*(w, l) = k$, which is equivalent to $w = w_{max}(k, l)$. Afterwards we will prove that this is the optimal solution.

For $w = w_{max}(k, l)$, the demand function becomes:

$$\lambda = \mu - \frac{\mu\sqrt{c - l}}{\sqrt{c - l} + \sqrt{k}}, \tag{B.5}$$

and the inverse form of (B.5) is:

$$l = c - \frac{k(\mu - \lambda)^2}{\lambda^2} = l_1(k, \lambda). \tag{B.6}$$

Given (B.6), the value of w_{max} with respect to λ is:

$$w_{max} = R - \frac{k\mu}{\lambda^2}, \tag{B.7}$$

and the optimal service fee is:

$$p = R - \frac{k(\mu - \lambda)}{\lambda^2}. \tag{B.8}$$

In order to achieve any input rate $\lambda < \mu$, and at the same time set $G_{A,I} = k$, for any $k \leq G_0(\lambda_0)$, the provider must set l and w according to (B.6) and (B.7) respectively. In this case, the aggregator's optimal response will be p from (B.8).

Under the above, by substituting (B.6) and (B.7) to the provider's profit function, we have:

$$G_{P,I}(p,l,w) = \lambda\left(R - \frac{c}{\mu-\lambda}\right) - k, \tag{B.9}$$

where it is straightforward that the optimal rate of incoming orders is $\lambda^* = \lambda_0$.

Taking into account the market size, it is straightforward that when $\Lambda > \lambda_0$, the optimal rate of incoming orders λ_0 is achievable. Therefore, there exists a unique maximizing strategy $l_I^* = l_1(k,\lambda_0)$.

When $\Lambda \leq \lambda_0$, from (B.9), the optimal for the provider is to capture the market. Then, the provider's function is $G_{P,I}(\Lambda) = \Lambda\left(w_{max}(k,l) - \frac{l}{\mu-\Lambda}\right) = \Lambda\left(R - \frac{c}{\mu-\Lambda}\right)$ under any l. Therefore, the role of l is to ensure that the market is captured. Along the same lines with the discussion in Fig. 2, this happens when $l \geq l_1(k,\Lambda)$ and we have the following cases:

i. When $\Lambda \in (\mu - Z_1(k), \lambda_0]$, any $l \in [l_1(k,\Lambda), c]$ is an optimal delay compensation strategy.

ii. When $\Lambda \in [G_0^{-1}(k), \mu - Z_1(k)]$, this is satisfied for any $l \in [0,c]$.

□

Proof of Theorem 3. We will explore the decentralized and the centralized profit for the four regions in Fig. 4.

i. When $G_0(\Lambda) < \min\{S,k\}$: Both provider and channel will choose no service. Therefore, the profit of both the decentralized and the centralized settings are equal to k, which implies coordination.

ii. When $S < \min\{k, G_0(\Lambda)\}$: Both provider and channel will choose direct mode. Therefore, the profit for both is $G_0(\Lambda) + k - S$, which implies coordination.

iii. When $S > \tilde{S}_0$ and $k \leq G_0(\Lambda)$: Both provider and channel will choose indirect mode. The profit for the centralized setting is $G_0(\Lambda) + k$. When the market is not captured, it is easy to show that the decentralized profit is strictly less than the centralized. On the other hand, when it is captured, then it is equal. Therefore, coordination occurs when the market is captured.

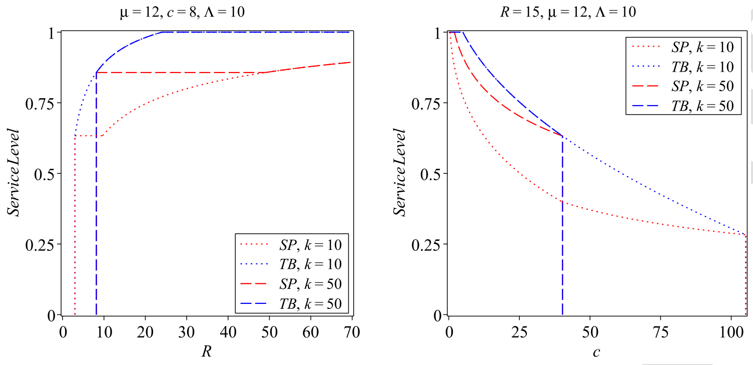
iv. When $S \in [S_0, \tilde{S}_0]$ and $k \leq G_0(\Lambda)$: The centrally optimal mode is indirect, and the provider's direct. Therefore, $G_C^* = G_0(\Lambda) + k > G_0(\Lambda) + k - S$, which implies that coordination is not attainable.

□

Proof of Theorem 4. When $S > k$ both the provider and the channel chooses indirect mode with profit $G_0(\Lambda)+k$. Otherwise they both choose direct with profit $G_0(\Lambda)+k-S$. Therefore, in both cases coordination is attained. □

C. Additional Figures

Supplementary to the numerical analysis of subsection 7.1, Fig. 9 illustrates the sensitivity of the service level with respect to the customer service value R (Fig. 9a) and the waiting cost c (Fig. 9b). Analogous insights to those in Fig. 6 can also be derived here.



(a)

(b)

Fig. 9: Service level with respect to R and c

References

Abdel-Malek, L., Kullpattaranirun, T., Nanthavanij, S., 2005. A framework for comparing outsourcing strategies in multi-layered supply chains. International Journal of Production Economics 97, 318–328.

Afèche, P., Baron, O., Kerner, Y., 2013. Pricing time-sensitive services based on realized performance. Manufacturing & Service Operations Management 15, 492–506.

Ahmadi, R., Iravani, F., Mamani, H., 2017. Supply chain coordination in the presence of gray markets and strategic consumers. Production and Operations Management 26, 252–272.

Arshinder, K., Kanda, A., Deshmukh, S., 2008. Supply chain coordination: Perspectives, empirical studies and research directions. International Journal of Production Economics 115, 316–335.

Bai, J., So, K.C., Tang, C.S., Chen, X., Wang, H., 2019. Coordinating supply and demand on an on-demand service platform with impatient customers. Manufacturing & Service Operations Management 21, 556–570.

Banerjee, S., Riquelme, C., Johari, R., 2015. Pricing in ride-share platforms: A queueing-theoretic approach. Available at SSRN 2568258 .

Bardi, E.J., Tracey, M., 1991. Transportation outsourcing: A survey of US practices. International Journal of Physical Distribution & Logistics Management 21, 15–21.

Benioudakis, M., Burnetas, A., Ioannou, G., 2021. Lead-time quotations in unobservable make-to-order systems with strategic customers: Risk aversion, load control and profit maximization. European Journal of Operational Research 289, 165–176.

Bernstein, F., Federgruen, A., 2007. Coordination mechanisms for supply chains under price and service competition. Manufacturing & Service Operations Management 9, 242–262.

Cachon, G., Terwiesch, C., 2020. Matching supply with demand: An introduction to operations management, 4th edition. McGraw-Hill Publishing.

Cachon, G.P., 2003. Supply chain coordination with contracts. Handbooks in operations research and management science 11, 227–339.

Cachon, G.P., Daniels, K.M., Lobel, R., 2017. The role of surge pricing on a service platform with self-scheduling capacity. Manufacturing & Service Operations Management 19, 368–384.

Cachon, G.P., Lariviere, M.A., 2005. Supply chain coordination with revenue-sharing contracts: strengths and limitations. Management Science 51, 30–44.

Carreiro, A.M., Jorge, H.M., Antunes, C.H., 2017. Energy management systems aggregators: A literature survey. Renewable and Sustainable Energy Reviews 73, 1160–1172.

Chen, M., Hu, M., Wang, J., 2022. Food delivery service and restaurant: Friend or foe? Management Science 68, 6539–6551.

Choi, T.M., Cheng, T.E., 2011. Supply chain coordination under uncertainty. Springer Science & Business Media.

Choi, T.M., Guo, S., Liu, N., Shi, X., 2020. Optimal pricing in on-demand-service-platform-operations with hired agents and risk-sensitive customers in the blockchain era. European Journal of Operational Research 284, 1031–1042.

Chopra, S., Meindl, P., 2019. Supply Chain Managemen: Strategy, planning, and operation. Pearson New York, NY, USA.

Edelson, N.M., Hilderbrand, D.K., 1975. Congestion tolls for poisson queuing processes. Econometrica 43, 81–92.

Espino-Rodríguez, T.F., Padrón-Robaina, V., 2006. A review of outsourcing from the resource-based view of the firm. International Journal of Management Reviews 8, 49–70.

Feng, J., Zhang, M., 2017. Dynamic quotation of leadtime and price for a make-to-order system with multiple customer classes and perfect information on customer preferences. European Journal of Operational Research 258, 334–342.

Feng, T., Ren, Z.J., Zhang, F., 2019. Service outsourcing: Capacity, quality and correlated costs. Production and Operations Management 28, 682–699.

Gans, N., Zhou, Y.P., 2007. Call-routing schemes for call-center outsourcing. Manufacturing & Service Operations Management 9, 33–50.

Ghosh, S., Hassin, R., 2021. Inefficiency in stochastic queueing systems with strategic customers. European Journal of Operational Research 295, 1–11.

Gurvich, I., Lariviere, M., Moreno, A., 2019. Operations in the on-demand economy: Staffing services with self-scheduling capacity, in: Sharing economy: Making supply meet demand. Springer, pp. 249–278.

Hassin, R., 2016. Rational queueing. Chapman & Hall book. CRC press.

Hassin, R., Haviv, M., 2003. To queue or not to queue: Equilibrium behavior in queueing systems. volume 59. Springer Science & Business Media.

Iria, J., Soares, F., Matos, M., 2018. Optimal supply and demand bidding strategy for an aggregator of small prosumers. Applied Energy 213, 658–669.

Legros, B., Jouini, O., Koole, G., 2020. Should we wait before outsourcing? Analysis of a revenue-generating blended contact center. Manufacturing & Service Operations Management 23, 1118–1138.

Li, T., Yu, M., 2017. Coordinating a supply chain when facing strategic consumers. Decision Sciences 48, 336–355.

Li, X., Li, Y., Cai, X., 2013. Double marginalization and coordination in the supply chain with uncertain supply. European Journal of Operational Research 226, 228–236.

Lin, Y.T., Parlaktürk, A.K., Swaminathan, J.M., 2018. Are strategic customers bad for a supply chain? Manufacturing & Service Operations Management 20, 481–497.

Liu, L., Parlar, M., Zhu, S.X., 2007. Pricing and lead time decisions in decentralized supply chains. Management Science 53, 713–725.

Lu, T., Chen, Y.J., Tomlin, B., Wang, Y., 2019. Selling co-products through a distributor: The impact on product line design. Production and Operations Management 28, 1010–1032.

Min, H., 2013. Examining logistics outsourcing practices in the united states: From the perspectives of third-party logistics service users. Logistics Research 6, 133–144.

Myerson, R.B., 1989. Mechanism design, in: Allocation, information and markets. Palgrave Macmillan, London, pp. 191–206.

Pu, X., Sun, S., Shao, J., 2020. Direct selling, reselling, or agency selling? Manufacturer's online distribution strategies and their impact. International Journal of Electronic Commerce 24, 232–254.

Taylor, T.A., 2018. On-demand service platforms. Manufacturing & Service Operations Management 20, 704–720.

Viswanathan, S., Wang, Q., 2003. Discount pricing decisions in distribution channels with price-sensitive demand. European Journal of Operational Research 149, 571–587.

Vohra, R.V., 2012. Dynamic mechanism design. Surveys in Operations Research and Management Science 17, 60–68.

Vosooghidizaji, M., Taghipour, A., Canel-Depitre, B., 2020. Supply chain coordination under information asymmetry: A review. International Journal of Production Research 58, 1805–1834.

Wang, Y., Niu, B., Guo, P., Song, J.S., 2020. Direct sourcing or agent sourcing? Contract negotiation in procurement outsourcing. Manufacturing & Service Operations Management 23, 294–310.

Yu, Z., Razzaq, A., Rehman, A., Shah, A., Jameel, K., Mor, R.S., 2022. Disruption in global supply chain and socio-economic shocks: a lesson from covid-19 for sustainable production and consumption. Operations Management Research 15, 233–248.

Zissis, D., Ioannou, G., Burnetas, A., 2020. Coordinating lot sizing decisions under bilateral information asymmetry. Production and Operations Management 29, 371–387.

Service provision on an aggregator platform with time-sensitive customers: Pricing strategies and coordination

Myron Benioudakis[a,b], Dimitris Zissis[a,c], Apostolos Burnetas[d,*] and George Ioannou[a]

[a]*Department of Management Science and Technology, Athens University of Economics and Business, Athens, 10434, Greece*
[b]*Department of Mechanical Engineering, University of Thessaly, Volos, 38334, Greece*
[c]*Norwich Business School, University of East Anglia, Norwich, NR4 7TJ, UK*
[d]*Department of Mathematics, National and Kapodistrian University of Athens, Athens, 15774, Greece*

ARTICLE INFO

Keywords:
Demand aggregator platform
Strategic queueing
Time-based pricing
Double marginalization
Outsourcing
Reselling/Agency selling

ABSTRACT

The increasing tendency to fulfill customer needs via virtual platforms has led to a rapid growth of sharing economy. This practice allows non-entrepreneurs to set up a business and firms to focus on their core operations by outsourcing tasks related to attracting, finding, contracting, and invoicing customers. Hence, potential entrepreneurs and firms face the question whether to penetrate a market directly or through a platform, and to what extent. In this work, we focus on providers who offer unique services and make a choice between enlisting in a demand aggregator's platform and reaching the market directly; due to the unique services, we assume that the providers may have sufficient power to set the wholesale price that is paid to the platform. A game-theory model in a queueing framework is developed to address the questions of mode selection and pricing strategies. Such settings allow to include customers who are sensitive in delays in product/services delivery and exhibit strategic behavior. We show that under single-price contracts channel profits are adversely affected due to double marginalization. The latter effect can be mitigated by time-dependent pricing involving delay compensation or a revenue sharing contract, resulting in system coordination. We identify the equilibrium strategies and the provider's optimal policy. We also derive insights on the combined effects of key parameters such as the market size, the direct cost of customer service, and the aggregator's reservation level on the optimal pricing strategies, and quantify their impact.

[*]Corresponding author

benioudakis@aueb.gr (M. Benioudakis); dzisis@aueb.gr (D. Zissis); aburnetas@math.uoa.gr (A. Burnetas); ioannou@aueb.gr (G. Ioannou)

http://scholar.uoa.gr/aburnetas (A. Burnetas)

ORCID(s): 0000 - 0003 - 3295 - 3343 (M. Benioudakis); 0000 - 0002 - 6957 - 3494 (D. Zissis); 0000 - 0002 - 9365 - 9255 (A. Burnetas)