AARHUS UNIVERSITY

Cover sheet

This is the accepted manuscript (post-print version) of the article.
The content in the accepted manuscript version is identical to the final published version, although typography and layout may differ.

How to cite this publication

Please cite the final published version:

Marand, A. J., Li, H. J., & Thorstenson, A. (2021). Competing on price, speed, and reliability: How does bounded rationality matter? *Journal of the Operational Research Society*, *72*(9), 2059-2072. https://doi.org/10.1080/01605682.2020.1759384

Publication metadata

Title:	Competing on price, speed, and reliability: How does bounded rationality matter?
Author(s):	Marand, Ata Jalili ; Li, Hongyan Jenny ; Thorstenson, Anders.
Journal:	Journal of the Operational Research Society.
DOI/Link:	10.1080/01605682.2020.1759384
Document version:	Accepted manuscript (post-print)
Document license:	CC-BY-NC-ND

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- *Users may download and print one copy of any publication from the public portal for the purpose of private study or research.*
- *You may not further distribute the material or use it for any profit-making activity or commercial gain*
- *You may freely distribute the URL identifying the publication in the public portal*

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

If the document is published under a Creative Commons license, this applies instead of the general rights.

This cover sheet template is made available by AU Library
Version 2.1, September 2020

Competing on Price, Speed, and Reliability: How Does Bounded Rationality Matter?

Ata Jalili Marand
atajalili@econ.au.dk

Hongyan Li
hojl@econ.au.dk

Anders Thorstenson
anders.thorstenson@econ.au.dk

CORAL, Department of Economics and Business Economics, Aarhus University
Fuglesangs Allé 4, 8210 Aarhus V, Denmark

Abstract

A firm's delivery performance may have significant impact on the satisfaction and purchase behaviour of its customers. Empirical evidence has shown that customers are willing to pay a higher price for a faster and more reliable service. In this study, we address the interactions between the price, promised delivery time, and delivery-reliability level in a competitive setting. We model the problem as competition among an arbitrary number of profit-maximizing firms facing boundedly rational customers who can choose to buy the service from one of the firms or balk. We prove the existence of a unique Nash equilibrium and propose a simple iterative algorithm that converges to the equilibrium. Furthermore, we compare our results with those in the existing literature and report interesting managerial insights. Our results suggest that having a clear understanding of customers' bounded rationality level is crucial for businesses to determine their optimal decisions and position in the market both in monopolistic and competitive settings.

Keywords: Delivery reliability, Competition, Bounded rationality, Queuing, Pricing, Game Theory.

1 Introduction

Empirical evidence accentuates the fact that a superior delivery performance entices customers into buying and paying more. When it comes to what a superior delivery performance is, the focus in the Operations Management literature has mainly been on delivery speed. But does *superior* equal *faster* in this case? In 2017, Bring's[1] research on Scandinavian on-line shoppers reports that 50% of Swedish on-line shoppers cancel their on-line orders on account of delivery times that are not met. Comparison shopping websites, e.g., shopzilla.com and bizrate.com, also report that order fulfillment failures are the main source of customer complaints in on-line retailing (Rao et al., 2011). Analyzing transaction level data from an American personal accessory retailer with both on-line and physical channels, Rao et al. (2014) conclude that there is statistically significant evidence to suggest that delivery reliability

[1] https://www.bring.com/ecommerce/bring-research

does mitigate the risk of product returns. Kelley et al. (1993) explain that customers consider deliveries to be controllable and hence late deliveries are judged harshly by them. In fact, consistent on-time delivery has long been recognized as one of the most critical performance measures for both on-line and off-line businesses in the e-commerce era (Lee & Whang, 2001).

Delivery reliability is a key consideration for business customers too. Speed and reliability are both critical measures in supplier selection, particularly when a long-term relationship is envisioned by the buyer (Benjaafar et al., 2007). While short promised delivery times allow business customers to speed up their operations, on-time (i.e. reliable) deliveries reduce uncertainty and enable them to plan and coordinate their manufacturing activities accurately (Peng & Lu, 2017). In fact, reliability directly impacts customer satisfaction (Rosenzweig et al., 2003). According to a recent report by SeaIntel, a world-leading provider of impartial analyses within the container shipping industry, the shipping companies' overall reliability reached only 66.4% in the first three months of 2018, 9.8% down compared to the end of 2017, which gave rise to criticisms from their customers[2]. The CEO of *DSV*, a Danish logistics service provider, finds it very frustrating for their customers when the planned schedules are not met and emphasizes that the industry is facing a huge challenge in improving the reliability level[3]. The network head of *Maersk*, the world's largest container shipping company, also declares that the industry needs to identify new ways of demonstrating reliability to the customers[4].

As a further example, Peng & Lu (2017) in an empirical study analyze the transaction data collected from a heating, ventilation, and air conditioning (HVAC) control product supply chain to examine the effect of delivery performance on future customer transaction quantities and unit prices. They suggest that delivery performance should not be treated as a single measure. They consider different dimensions of delivery performance, i.e., delivery speed (delivery time), on-time delivery rate (delivery-reliability level), and delivery date inaccuracy (late or early delivery), and find that the price is affected by on-time delivery rate and delivery speed. Although they conclude that early deliveries do not seem to negatively affect the transaction quantity, they find that the delivery-reliability level contributes to profitability by allowing the firms to charge premium prices. The empirical evidence emphasizes the interactions between price, delivery time, and delivery-reliability level. However, only few studies in the OM literature analytically address these interactions.

From an operational point of view, the delivery time and delivery reliability are tightly linked together. As long as the firm's capacity is fixed, a shorter delivery time may attract more customers, but could also lead to a reduced delivery-reliability level as a result of increased congestion which negatively influence customers' purchase behaviour. It has been shown that the trade-off that a firm has to make between a faster or a more reliable service is crucial in gaining competitive advantage in time-based competitions (Shang & Liu, 2011). This trade-off becomes even more complicated when the price can be used as a lever to control the demand rate. Studies of the interplay between price and delivery performance measures are scarce and represented only by Boyaci & Ray (2006), Xiao & Qi (2016), and Marand et al. (2019). In a competitive setting with boundedly rational customers, these interactions have not been addressed in the literature.

In this study, we aim at filling the above-mentioned gap and analyze the competition among an

[2] https://shippingwatch.com/secure/carriers/Container/article10683340.ece

[3] https://shippingwatch.com/secure/Services/article10778476.ece

[4] https://shippingwatch.com/secure/carriers/article10962558.ece

arbitrary number of profit-maximizing service providers that face price, time, and reliability sensitive customers who are boundedly rational. The competition setting captures the commonly seen realistic situations where multiple service providers compete in terms of price and delivery performance. The bounded rationality assumption captures the nature of human beings' limited ability to process information. There are rich empirical evidence supporting human beings' bounded rationality (see, e.g., Huang, Allon & Bassamboo, 2013, and the references therein). It has been shown that people often rely on a limited number of heuristic principles that in general are useful but sometimes lead to severe and systematic errors (Kahneman et al., 1982). Ariely (2009) introduces the bounded rationality as the invisible hand driving human decision making behaviour. As a result of bounded rationality, a customer may not be capable of adequately and/or accurately estimating expected utilities from different options and this might lead to inconsistent comparative judgments. Thus, instead of choosing the optimal/best option, the customer may choose an inferior alternative. In this paper, we are interested to see how the fact that customers might have limited cognitive abilities influences a service provider's pricing strategy and the trade-off with the two dimensions of delivery performance, i.e., speed and reliability, both in monopolistic and in competitive settings.

We make the following main contributions. (i) We study the interactions between price, delivery time, and delivery-reliability level in a competitive setting under the assumption of customers with bounded rationality. (ii) We prove the existence of a unique Nash equilibrium for the price, delivery time, and delivery reliability competition problem and propose a simple iterative algorithm that converges to the equilibrium. (iii) We explore multiple insights on how pricing, delivery time, and delivery reliability decisions should be made when the customers are boundedly rational. For instance, our results suggest that a firm has an incentive to invest in reducing the customers' bounded rationality level only if it can capture a sufficiently high market share at a very low bounded rationality level. A clear understanding of customers' bounded rationality level turns out to be crucial in determining a firm's optimal decisions and position in the market both in monopolistic and competitive settings. Moreover, we show that considering balking as an option that could be chosen by the customers may substantially influence the firms' optimal decisions in the competitive setting.

The remainder of the paper is organized as follows. Section 2 reviews the closely related literature. Section 3 specifies the problem under consideration. Section 4 analyzes the competition problem in two steps: First, Subsection 4.1 formulates and solves the best response problem, and then, Subsection 4.3 analyzes the equilibrium solution. In Section 5, the numerical results are presented. Main managerial insights are summarized in Section 6. Finally, Section 7 concludes the study and suggests future research directions.

2 Literature Review

Our paper is related to a number of streams in the literature, namely, reliability, competition, and bounded rationality in service delivery. In the following, we review the most closely related literature in each branch.

2.1 Reliability

Although both empirical (Baker et al., 2001; Peng & Lu, 2017) and theoretical studies (So, 2000; Shang & Liu, 2011) stress the existence and importance of the interactions between price, delivery time, and delivery reliability, these interactions have been only scarcely analyzed in the literature. To the best of our knowledge, Boyaci & Ray (2006), Xiao & Qi (2016), and Marand et al. (2019) are the only analytical studies with relevant considerations.

Boyaci & Ray (2006) consider a profit-maximizing firm's optimal differentiation strategy in terms of prices, delivery times, and delivery-reliability levels. The firm sells two variants of a product in a capacitated environment. They assume that each product's demand rate linearly depends on both products' prices, delivery times, and delivery-reliability levels. In a two-stage supply chain with one supplier and one manufacturer, Xiao & Qi (2016) study the equilibrium decisions in the supply chain with an all-unit quantity discount contract. Their demand function is linearly dependent on the price, delivery time, and delivery-reliability level. While these two studies employ a market-level aggregate demand function, we directly construct the demand from individual customers' utilities. This approach enables us to account for individual customers' decisions and the bounded rationality of their behaviour. Moreover, our work differs from these studies since they do not consider the horizontal competition among service providers.

Considering the customers' choice behaviour, Marand et al. (2019) seem to be the first to study the interactive impact of price, delivery time, and delivery-reliability level on the equilibrium behaviour of rational customers and the optimal decisions of a revenue-maximizing service provider. They characterize a monopolistic service provider's optimal decisions under customer homogeneity and heterogeneity assumptions. In this study, we extend Marand et al. (2019) by (i) modeling the competition between an arbitrary number of service providers and proving the existence of a unique Nash equilibrium solution to the competition problem, and (ii) investigating the impact of customers' bounded rationality on a service provider's optimal decisions and profitability both in monopolistic and competitive settings. Neither of these analyses are included in Marand et al. (2019).

2.2 Competition

The competition problem in service delivery has been studied extensively, e.g., Chen & Wan (2003), Chen & Wan (2005), Li et al. (2012), Hu & Qiang (2013), Saberi et al. (2014), Behzad & Jacobson (2016), and Chen et al. (2019). In the sequel, we concentrate on the literature that is most closely related to our study and do not provide an exhaustive review of the literature.

So (2000) considers the price and delivery time competition among N profit-maximizing firms. He employs a log-linear function to capture customers' utilities and uses a multinomial logit (MNL) model to capture their choice behaviour. Each firm is characterized by its capacity level (service rate) and unit operating cost. So (2000) proves the existence of a unique Nash equilibrium. He assumes that the delivery-reliability level is exogenously set and does not consider its impact on the customers' choice behaviour. Different from his study, we assume that the customers are also sensitive to the delivery-reliability level. Moreover, while the bounded rationality assumption is implicit in his study, we explicitly assume that the customers are boundedly rational and study the impact of the bounded rationality level on a firm's price, time, and reliability decisions both in monopolistic and competitive settings. We also consider the customers having the balk option and show that it can substantially influence the firms' optimal decisions and positioning in the market.

Allon & Federgruen (2007) consider a model that is generally similar to So (2000), but they include the capacity level decision. They define a service level as the difference between the benchmark upper bound for the waiting time and the expected waiting time (or a fractile of the waiting time distribution). They consider three sequences through which the firms make their choices. They show that in all three sequences an equilibrium pair of prices and service levels (waiting times) exists under some specified conditions. Later, Allon & Federgruen (2009) extend Allon & Federgruen (2007) to a segmented market. However, neither of them consider the impact of the delivery-reliability level on the customer behaviour which is an integral part of our analysis.

Ho & Zheng (2004) and Shang & Liu (2011) appear to be the first to address the time-reliability trade-off under the bounded rationality assumption. Ho & Zheng (2004) explicitly consider the delivery-reliability level along with the delivery time as the two main factors that influence customers' utility and firms' market shares. They employ an MNL model to determine the demand allocation between two competing firms, and assume that the price is exogenous, hence, each firm aims at maximizing its demand rate. The existence of a unique Nash equilibrium is proven for the delivery time and delivery reliability competition. In a similar setting, Shang & Liu (2011) extend Ho & Zheng (2004) to a competition problem among an arbitrary number of firms. Under capacity competition, they prove the existence of Nash equilibria in a duopolistic game and show that this game is similar to a Prisoner's Dilemma when the cost of adding capacity is small.

There are notable differences between these two studies (Ho & Zheng, 2004; Shang & Liu, 2011) and our study: unlike them, the price is assumed to be endogenous in our study. Moreover, whereas the customers are only implicitly assumed to be boundedly rational in their studies, we explicitly consider the customers' bounded rationality. We characterize the impact of the bounded rationality level on the optimal decisions in monopolistic and competitive settings and show that a higher bounded rationality level attenuates competition.

2.3 Bounded rationality

The neo-classic economic literature assumes that customers, as decision makers, are fully rational in the sense that they are perfect utility maximizers. An alternative approach for modelling decision-making behaviour, however, assumes that although customers attempt to optimize, they are either limited in their computational abilities or prone to some unobserved noisy bias (Davis, 2018). Simon coins the term *bounded rationality* to describe such a behaviour (see, e.g., Simon, 1997). Boundedly rational customers' choice behaviour can be modelled using the quantal choice theory (Luce, 1959) that assumes a decision maker who is presented with multiple alternatives, each having a likelihood to be chosen, is more likely to choose better alternatives over worse ones. Experimental studies of rationality in game theory led McKelvey & Palfrey (1995) to conceptualize the quantal response equilibrium to model interactive behaviour of boundedly rational decision makers in a competitive setting. The idea of bounded rationality has also been studied in different contexts in the operations and supply chain management literature, e.g., price contract design (Ho & Zhang, 2008), newsvendor problems (Su, 2008), and capacity allocation (Chen et al., 2012).

The rationality assumption is also a dominant assumption in the service pricing literature (see Hassin & Haviv, 2003; Hassin, 2016, for comprehensive literature reviews). Huang, Allon & Bassamboo (2013) are the first to deviate from this assumption and allow for customers' bounded rationality in service systems. They employ a linear price and time dependent utility function and an MNL model to capture the customers' choice behaviour. In a monopolistic setting, they show that ignoring bounded rationality can result in significant revenue and welfare losses even when the bounded rationality level is low. For an unobservable queue, they conclude that the revenue is strictly increasing in the bounded rationality level when it is sufficiently large. Their study is extended in different directions by, e.g., Huang, Allon & Bassamboo (2013), Huang & Chen (2015), Li et al. (2016), Li et al. (2017), and Ren et al. (2018). Ren & Huang (2018) provide a review of different approaches of modeling customers' bounded rationality in the OM literature. We contribute to this stream of literature by assuming, unlike the above mentioned studies, that the boundedly rational customers' utility is a function of price, delivery time, and also delivery-reliability level, and by considering the competition among an arbitrary number of competing firms.

The interactions between price, delivery time, and delivery reliability and the importance of an integrated framework to analytically study these interactions are emphasized both in empirical and analytical studies. Despite some pioneering research, the subject is still new and deserves more attention. To the best of our knowledge, none of the existing studies have considered the joint price, delivery time, and delivery reliability competition among N horizontally competing firms in a market with boundedly rational customers. In this study, we aim at filling this gap. Table 1 summarizes the key literature, as well as the positioning of our study.

Table 1: Summary of the key literature

Literature	Firm decisions				Market competition	Customer bounded rationality
	Price	Delivery time	Reliability	Capacity		
So (2000)	✓	✓	-	-	✓	(implicitly)
Chen & Wan (2003)	✓	✓	-	-	✓	-
Ho & Zheng (2004)	-	✓	✓	✓	✓	(implicitly)
Chen & Wan (2005)	✓	✓	-	✓	✓	-
Allon & Federgruen (2007)	✓	✓	-	✓	✓	(implicitly)
Shang & Liu (2011)	-	✓	✓	✓	✓	(implicitly)
Huang, Allon & Bassamboo (2013)	✓	✓	-	-	-	✓
Marand et al. (2019)	✓	✓	✓	-	-	-
Our study	✓	✓	✓	-	✓	✓

3 Problem Specification

Consider a market that consists of a homogeneous population of customers and N service providers (henceforth, firms). Upon their arrivals, customers have $N+1$ choices: they can choose between the services provided by these N firms, or they can balk (walk away without being served). Let $U(p_i, t_i, q_i)$ denote a customer's actual (expected) utility from using the service provided by Firm i, $i = 1, \ldots, N$, where p_i is the service price, t_i is the promised delivery time, and q_i is the delivery-reliability level. The price, delivery time, and delivery-reliability level together characterize Firm i's service. It is assumed that there is no moral hazard problem involved.

A major part of the classic queue-pricing literature assumes that a customer's utility is linearly dependent on the price and delivery time (see Hassin & Haviv, 2003, for a comprehensive literature review). Moreover, some more recent studies assume that the customer's utility is linearly dependent on the delivery time and delivery-reliability level (see, e.g., Ho & Zheng, 2004; Shang & Liu, 2011). We extend the literature by assuming that the utility is linearly dependent on price, delivery time, and delivery-reliability level, and define

$$U(p_i, t_i, q_i) = v - c_p p_i - c_t t_i + c_q q_i, \;\; i = 1, \ldots, N, \tag{1}$$

in which v is the reward that customers gain on the completion of service[5], c_p denotes the customer price sensitivity, c_t denotes the customer waiting cost rate (or simply the time sensitivity), and c_q denotes the reliability sensitivity of the customers. If we ignore the effect of the price, i.e., $c_p = 0$, our model is reduced to that of Ho & Zheng (2004) and Shang & Liu (2011), and disregarding the impact of the delivery-reliability level, i.e., $c_q = 0$, results in a model similar to So (2000). To complement their results, we let both c_p and c_q be positive constants, i.e., $c_p, c_q > 0$. Furthermore, we assume that a customer's actual utility of choosing to balk is normalized to zero, i.e., $U_0 = 0$. In the remainder of the paper, subscript 0 denotes the balking option, and for ease of exposition, we define $U(p_0, t_0, q_0) = U_0$.

The queue-pricing literature typically assumes that the customers are fully rational. However, we assume that customers are boundedly rational in the sense that there may be a gap between their perceived utility and their actual utility at the time of making the buying/balking decisions. In order to model the bounded rationality, we add a random noise to the customer's perception of the service utility. Thus, the customers have a perceived utility from the service provided by Firm i as

$$\mathcal{U}(p_i, t_i, q_i; \epsilon_i) = U(p_i, t_i, q_i) + \epsilon_i, \;\; i = 0, 1, \ldots, N.$$

It is assumed that ϵ_i are independent and identically Gumbel-distributed random variables with mean $E(\epsilon_i) = 0$ and variance $Var(\epsilon_i) = \frac{\eta^2\pi^2}{6}$, where $\eta > 0$ is the scale parameter of the corresponding Gumbel distribution (see Talluri & Van Ryzin, 2006, for detailed discussions). The parameter η reflects the bounded rationality level of the customers. As η approaches positive infinity, the customers become completely irrational, and conversely, as η approaches zero, the customers become fully rational. As a result of the customers' bounded rationality, a customer may even choose a service that yields a negative actual utility.

We employ the MNL model to capture the customer's choice behaviour. The MNL model is one of the most commonly used attraction models in the literature (see, e.g., Huang et al., 2013; Strauss et al., 2018). Based on the MNL model, the probability that a customer chooses Firm i's service equals

$$\theta_i = \frac{e^{U(p_i, t_i, q_i)/\eta}}{\sum_{j=0}^{N} e^{U(p_j, t_j, q_j)/\eta}}, \tag{2}$$

[5] All results can easily be adjusted for the case with firm-specific reward, i.e., v_i for $i = 1, \ldots, N$.

and the probability that a customer chooses to balk is

$$\theta_0 = \frac{1}{\sum_{j=0}^{N} e^{U(p_j,t_j,q_j)/\eta}}.$$

We assume that the customers arrive at the market following a Poisson process with the rate Λ (the potential market size). Let λ_i be the demand rate (arrival rate) captured by Firm i. From Eq. (2), the expected market share of Firm i equals $\frac{\lambda_i}{\Lambda} = \theta_i$. Therefore, arrivals to Firm i will also follow a Poisson process with the rate $\lambda_i = \Lambda\theta_i$ or equivalently

$$\lambda_i(p_i, t_i, q_i|\beta_i) = \frac{\Lambda e^{U(p_i,t_i,q_i)/\eta}}{e^{U(p_i,t_i,q_i)/\eta} + \beta_i}, \tag{3}$$

where $\beta_i = \sum_{\substack{j=0 \\ j\neq i}}^{N} e^{U(p_j,t_j,q_j)/\eta} \geq 1$ is the aggregate impact of the other firms' decisions on Firm i's demand rate. Particularly, $\beta_i > 1$ in the competition and $\beta_i = 1$ in the monopolistic setting. It follows from Eq. (3) that

$$U(p_i, t_i, q_i|\beta_i) = \eta \ln \frac{\beta_i\lambda_i}{\Lambda - \lambda_i}. \tag{4}$$

Eq. (4) implies that when $\eta \to 0$, we have $U \to 0$. Expressed in words, when customers are fully rational, the firm extracts all customer surplus at the equilibrium.

Furthermore, it is assumed that Firm i's service times are exponentially distributed with the parameter μ_i. Thus, we model the operations of each firm as an $M/M/1$ queueing system similar to most of the related literature. The delivery time is measured by the sojourn time. The sojourn time in an $M/M/1$ queue is exponentially distributed (Asmussen, 2008). In fact, even for a $G/G/1$ queue, the tail distribution of the sojourn time can be accurately approximated by an exponential distribution for high delivery-reliability levels (Abate et al., 1996). Thus, setting a reasonably high lower bound, $\underline{q}$, for the delivery-reliability level, our results approximately hold for more general queueing systems, as well. The lower bound $\underline{q}$ can be interpreted as a benchmark or market entrance requirement. This benchmark is, for instance, 91%, 88.7%, and 90.2% for electronics, personal care, and pharmaceutical products, respectively (Shang & Liu, 2011). We assume $\underline{q} \geq 0.64$ to be able to derive our analytical results. Let s be the random variable denoting the delivery time. According to our definition, Firm i's delivery-reliability level equals

$$q_i = Pr(s \leq t_i) = 1 - e^{-(\mu_i-\lambda_i)t_i}, \tag{5}$$

for $0 \leq \lambda_i < \mu_i$. From Eq. (5), the arrival rate to Firm i can be expressed as

$$\lambda_i = \mu_i - \frac{1}{t_i}\ln(\frac{1}{1-q_i}), \tag{6}$$

where $t_i \geq 0$ and $\underline{q} \leq q_i < 1$. According to Eq. (6), we have $\lambda_i < \mu_i$ and the stability condition is

Table 2: Notation

Parameter	Description	Parameter	Description
c_p	price sensitivity of customers	Π_i	profit of Firm i
c_t	time sensitivity of customers	U	actual expected utility
c_q	reliability sensitivity of customers	$\mathcal{U}$	perceived utility
Λ	potential market size	p_i	price charged by Firm i
$\underline{q}$	market entrance reliability requirement	t_i	delivery time quoted by Firm i
γ_i	unit operating cost of Firm i	q_i	delivery reliability quoted by Firm i
v	customer reward upon service completion	λ_i	demand rate at Firm i
μ_i	capacity of Firm i	x_i	strategy profile of Firm i
ϵ_i	Gumbel-distributed random noise variable	θ_0	probability of choosing to balk
η	bounded rationality measure	θ_i	probability of choosing Firm i's service
β_i	competition impact on Firm i's problem	N	Number of competing firms

always met. The profit maximization problem of Firm i is therefore defined as follows:

$$\max_{p_i,t_i,q_i} \quad \Pi_i(p_i,t_i,q_i|\beta_i) = \lambda_i(p_i,t_i,q_i|\beta_i)(p_i-\gamma_i) \tag{7}$$

$$\text{s.t.} \quad \gamma_i \le p_i, \tag{8}$$

$$0 \le t_i, \tag{9}$$

$$\underline{q} \le q_i < 1, \tag{10}$$

in which λ_i and q_i are defined by Eq. (3) and Eq. (5), respectively, and γ_i is the unit operating cost rate.

In the described setting, we consider the competition problem between N profit-maximizing firms in a Nash game, where all firms simultaneously decide on their prices, delivery times, and delivery-reliability levels. Let $x_i = (p_i, t_i, q_i)$ denote the strategy profile of Firm i, $i = 1, ..., N$. A strategy profile $\phi^* = (x_1^*, ..., x_N^*)$ is a Nash equilibrium if $\Pi_i(p_i^*, t_i^*, q_i^*|\beta_i^*) \ge \Pi_i(p_i, t_i, q_i|\beta_i^*)$ for any feasible (p_i, t_i, q_i) and for all $i = 1, ..., N$. In other words, a Nash equilibrium solution to the competition problem is a vector consisting of the individual firms' decisions such that no firm has an incentive to unilaterally deviate from the equilibrium solution. In the next section, we analyze the Nash game. The notation is summarized in Table 2.

4 Competition Game Analysis

4.1 Best Response Problem

To solve Firm i's best response problem, we assume that the combined impact of other firms' decisions on Firm i, i.e., β_i, is given and fixed. Substituting Eq. (6) and Eq. (1) in Eq. (4) results in

$$p_i(t_i,q_i|\beta_i) = \frac{1}{c_p}\big(v_i - c_t t_i + c_q q_i - \eta \ln \frac{\beta_i(\mu_i - \frac{1}{t_i}\ln(\frac{1}{1-q_i}))}{\Lambda - \mu_i + \frac{1}{t_i}\ln(\frac{1}{1-q_i})}\big). \tag{11}$$

The price $p_i(t_i, q_i|\beta_i)$ is defined for $\frac{1}{\mu_i}\ln(\frac{1}{1-q_i}) < t_i < \frac{1}{\mu_i - \Lambda}\ln(\frac{1}{1-q_i})$ when $\mu_i > \Lambda$, and for $\frac{1}{\mu_i}\ln(\frac{1}{1-q_i}) < t_i$ when $\mu_i \le \Lambda$. Eq. (11) expresses the price in terms of the delivery time and delivery-reliability level. Substituting Eq. (6) and Eq. (11) into the objective function (7) results in the following optimization problem:

$$\max_{t_i,q_i} \ \Pi_i(t_i,q_i|\beta_i) = (\mu_i - \frac{1}{t_i}\ln(\frac{1}{1-q_i}))(p_i(t_i,q_i|\beta_i)-\gamma_i) \tag{12}$$
$$\text{s.t. } 8-10.$$

Proposition 1 characterizes the optimal solution to Firm i's response problem. For the proofs, refer to the on-line supplement.

Proposition 1. *Given β_i,*

(*a*) *there exist a unique optimal price, p_i^*, delivery time, t_i^*, and delivery-reliability level, q_i^* that maximize Firm i's profit function, $\Pi_i(p_i,t_i,q_i|\beta_i)$.*

(*b*) *If*

(i) $\Lambda > \mu_i$, $\underline{q} < 1 - \frac{c_t}{c_q\mu_i}$ *and* $H(\underline{q}|\beta_i) > 0$, *or*

(ii) $\Lambda < \mu_i$, $1 - \frac{c_t}{c_q(\mu_i-\Lambda)} < \underline{q} < 1 - \frac{c_t}{c_q\mu_i}$ *and* $H(\underline{q}|\beta_i) > 0$, *or*

(iii) $\Lambda < \mu_i$, $\underline{q} < 1 - \frac{c_t}{c_q(\mu_i-\Lambda)}$,

then $p_i^* = p_i(t_i^*, q_i^*|\beta_i)$, $t_i^* = \frac{c_q(1-q_i^*)}{c_t}\ln\frac{1}{1-q_i^*}$, *and* q_i^* *is the unique root of* $H(q|\beta_i) = 0$, *where*

$$H(q|\beta_i) = \frac{c_q^2\mu_i(1-q)^2}{c_t}\ln\frac{1}{1-q} + \frac{\eta\Lambda}{\Lambda-\mu_i+\frac{c_t}{c_q(1-q)}} + \eta\ln\frac{\mu_i - \frac{c_t}{c_q(1-q)}}{\Lambda-\mu_i+\frac{c_t}{c_q(1-q)}} - v_i - c_q q + \eta\ln\beta_i + c_p\gamma_i.$$

(*c*) *Otherwise,* $p_i^* = p_i(t_i^*, q_i^*|\beta_i)$, $q_i^* = \underline{q}$, *and* t_i^* *is the unique root of* $G(t|\beta_i) = 0$, *where*

$$G(t|\beta_i) = \frac{c_t\mu_i t^2}{\underline{k}} + \frac{\eta\Lambda}{\Lambda-\mu_i+\frac{\underline{k}}{t}} + \eta\ln\frac{\mu_i-\frac{\underline{k}}{t}}{\Lambda-\mu_i+\frac{\underline{k}}{t}} - v_i - c_q\underline{q} + \eta\ln\beta_i + c_p\gamma_i,$$

in which $\underline{k} = \ln(\frac{1}{1-\underline{q}})$.

Proposition 1 uniquely determines Firm i's optimal decisions. Moreover, although no closed-form solution exists, finding the optimal solution is not challenging, because $H(q|\beta_i)$ and $G(t|\beta_i)$ are monotone functions of q and t, respectively, and the roots of $H(q|\beta_i) = 0$ and $G(t|\beta_i) = 0$ can easily be found using efficient numerical methods such as the bisection method.

4.2 Impact of Bounded Rationality in Monopoly

Solving the best response problem also amounts to solving a monopolistic firm's decision problem. Clearly, $\beta_i = 1$ for a monopolistic firm. In this subsection, we investigate the impact of the bounded rationality level on a monopolistic firm's optimal decisions and profitability. In particular, we would like to know if firms can benefit from reducing the customers' bounded rationality level by, for instance, providing them with some information that reduces the variability of the utility evaluation process. We are also interested in analyzing the impact of customers' bounded rationality level on each of the price, time, and reliability decisions made by the firm. Consider a tagged Firm i. Let $\hat{\theta}$ denote the unique root of $\frac{-1}{1-\theta} - \ln \frac{\theta}{1-\theta} = 0$.[6] Proposition 2 summarizes our results for this part.

Proposition 2. *The sensitivity of the monopolistic Firm i's optimal decisions to the bounded rationality level is schematically depicted in Figure 1.*

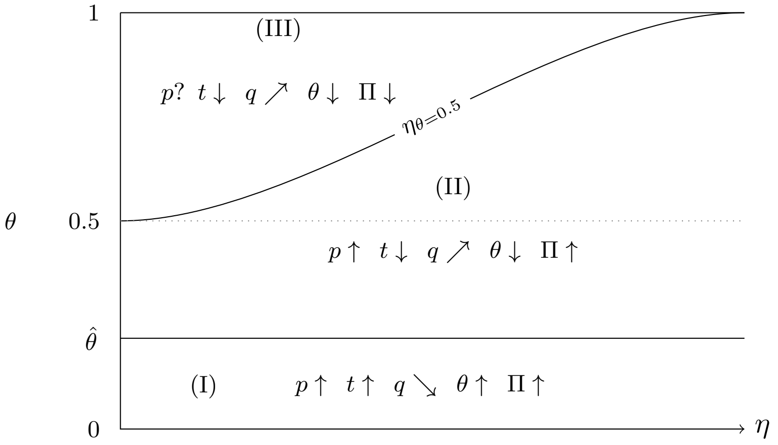

1
(III)
$p?\ t\downarrow\ q\nearrow\ \theta\downarrow\ \Pi\downarrow$
$\eta_{\theta=0.5}$
(II)
θ
0.5
$p\uparrow\ t\downarrow\ q\nearrow\ \theta\downarrow\ \Pi\uparrow$
$\hat{\theta}$
(I)
$p\uparrow\ t\uparrow\ q\searrow\ \theta\uparrow\ \Pi\uparrow$
0
η

Figure 1: A monopolistic firm's sensitivity to η ($\uparrow$: increasing, $\nearrow$: non-decreasing, $\downarrow$: decreasing, $\searrow$: non-increasing, ?: undetermined)

To elaborate the results presented in Proposition 2, we start by defining the following categorization of firms based on their market shares at a very small bounded rationality level: (i) a *large firm* for which $0.5 < \theta$, (ii) a *medium firm* for which $\hat{\theta} < \theta \leq 0.5$, and (iii) a *small firm* for which $\theta \leq \hat{\theta}$. Note that given the market-specific parameters, a large firm has a sufficiently high capacity level and/or low unit operating cost, and a small firm suffers from a low capacity level and/or a high unit operating cost. In Figure 1, $\eta_{\theta=0.5}$ denotes the bounded rationality level at which a large firm's market share decreases to 0.5.

As shown in Figure 1, the market share of a large firm decreases as the bounded rationality level increases (within area III). At a sufficiently high bounded rationality level, i.e., $\eta_{\theta=0.5} < \eta$, the market share drops below 0.5 and the firm becomes a medium firm (moving from area III to area II). However, from that point on, as η increases, it never crosses the threshold $\hat{\theta}$ for η although the firm's market share continually decreases. In other words, as a result of the increased bounded rationality level a large firm turns into a medium firm, but never becomes a small firm. A large firm's profit, however, is first decreasing (for $\eta < \eta_{\theta=0.5}$) and then increasing (for $\eta_{\theta=0.5} < \eta$) in the bounded rationality level. Huang, Allon & Bassamboo (2013) also show that a revenue-maximizing service provider is better off in markets where the bounded rationality level is sufficiently high. However, they assume that the delivery-reliability level is fixed and do not show the impact of the bounded rationality level on the reliability decision. As demonstrated in Figure 1, a large firm keeps improving its delivery performance as the bounded rationality level increases. When the bounded rationality level is sufficiently high, the firm charges a premium price for its superior delivery performance.

A medium firm's market share is always decreasing in the bounded rationality level, but it never drops below the threshold level $\hat{\theta}$, i.e., it remains a medium firm (remains within area II). As the

[6] $\hat{\theta} \approxeq 0.218$.

bounded rationality level increases, a medium firm continually improves its delivery performance and charges a higher price for its superior delivery performance. Figure 1 shows that a medium firm always benefits from a more boundedly rational market.

Only a small part of the customers buy from a firm offering a service with a strictly negative utility. As the customers become more boundedly rational, they choose a small firm more often. Therefore, in contrast to a large or medium firm, a small firm's market share increases as the bounded rationality level increases, but it never exceeds the threshold value $\hat{\theta}$, i.e, it never becomes a medium firm (it remains within area I). A small firm always benefits from customers' bounded rationality by charging a higher price even for an inferior delivery performance, i.e., a longer delivery time and a less reliable service. Figure 1 shows that the optimal delivery time, delivery reliability, and market share are monotone in the bounded rationality level for any firm type. Moreover, as the bounded rationality level increases, the market share approaches the threshold value $\hat{\theta}$.

4.3 Equilibrium Analysis

In this section, we consider the Nash game between N competing firms, where all the firms simultaneously decide on their prices, delivery times, and delivery-reliability levels. As mentioned in Section 3, a Nash equilibrium solution to the competition problem is a vector consisting of individual firms' decisions such that no firm has an incentive to unilaterally deviate from the equilibrium solution. Below, we propose an iterative procedure to solve the competition problem and show the existence of a unique Nash equilibrium.

Based on the optimal solution of the best response problem presented in Proposition 1, Lemma 1 further addresses the sensitivity of Firm i's decisions to β_i.

Lemma 1. *For Firm i, the optimal price and delivery time are strictly decreasing, and the optimal delivery-reliability level is non-decreasing in β_i.*

Lemma 1 implies that under increased competition, i.e., at a higher β_i, a firm not only offers a lower price and shorter delivery time, but it also never decreases its delivery-reliability level. For a higher level of competition, So (2000) also shows that the price and delivery time decrease. Correspondingly, Ho & Zheng (2004) show, through numerical studies, that the delivery time would be tighter for a higher attraction level of competitors. However, these two analyses are not inclusive in terms of considering customer sensitivity to all three factors of price, delivery time, and delivery-reliability level. Lemma 1 also implies that the competition decreases the firm's market share, i.e., demand rate, since from Eq. (6) we have $\frac{\partial \lambda_i}{\partial \beta_i} = \frac{1}{t_i^2}\ln(\frac{1}{1-q_i})\frac{\partial t_i}{\partial \beta_i} - \frac{1}{t_i(1-q_i)}\frac{\partial q_i}{\partial \beta_i} < 0$. For the same reason, the firm's profit also decreases with the competition.

Based on Lemma 1, we propose the following iterative procedure, adopted from So (2000) and modified to fit our problem, to solve the competition problem.

(1) Initialization: For Firm i, choose $p_i = \gamma_i$, $t_i = \frac{1}{\mu_i}\ln(\frac{1}{1-\underline{q}})$, and $q_i = 1$.

(2) Iterative step: Start with $i = 1$. Apply the results in Proposition 1 to find Firm i's optimal decisions given the current decisions of other firms. Repeat this for all $i = 2, ..., N$.

(3) Termination condition: Repeat step (2) until the difference between the decisions of two successive steps fall below a pre-specified tolerance level ε, i.e., $\Delta\beta_1 < \varepsilon$, where $\Delta\beta_1$ is the difference of β_1 in two successive iterations.

Furthermore, Proposition 3 can be proven.

Proposition 3. *There exists a unique Nash equilibrium to the N-firm competition problem, and the above-mentioned iterative procedure converges to this equilibrium solution.*

Proposition 3 ensures the existence as well as the uniqueness of the Nash equilibrium. This result extends the finding in Chen & Wan (2003) who show that the existence and uniqueness of equilibrium is even not necessarily guaranteed in the price and time competition under the full rationality assumption In addition, the following corollary is found to hold.

Corollary 1. *Suppose $\mu_N \leq ... \leq \mu_1$ and $i, j \in \{1, ..., N\}$ such that $i \leq j$. With a linear utility function, if $\frac{c_t}{c_q} > (1 - \underline{q})\mu_i$, then Firm j cannot differentiate in the delivery-reliability level and $q_j = \underline{q}$.*

Corollary 1 implies that if $\frac{c_t}{c_q} > (1 - \underline{q})\mu_1$, then our model would be reduced to the price and time competition with a fixed delivery-reliability level. It also shows how the availability of capacity and market parameters together influence a firm's offered delivery-reliability level; when the ratio of the time sensitivity to the reliability sensitivity, $\frac{c_t}{c_q}$, is high, no firm has an incentive to offer a reliability greater than the market entrance requirement. As this ratio decreases, however, more firms may benefit from differentiating in reliability.

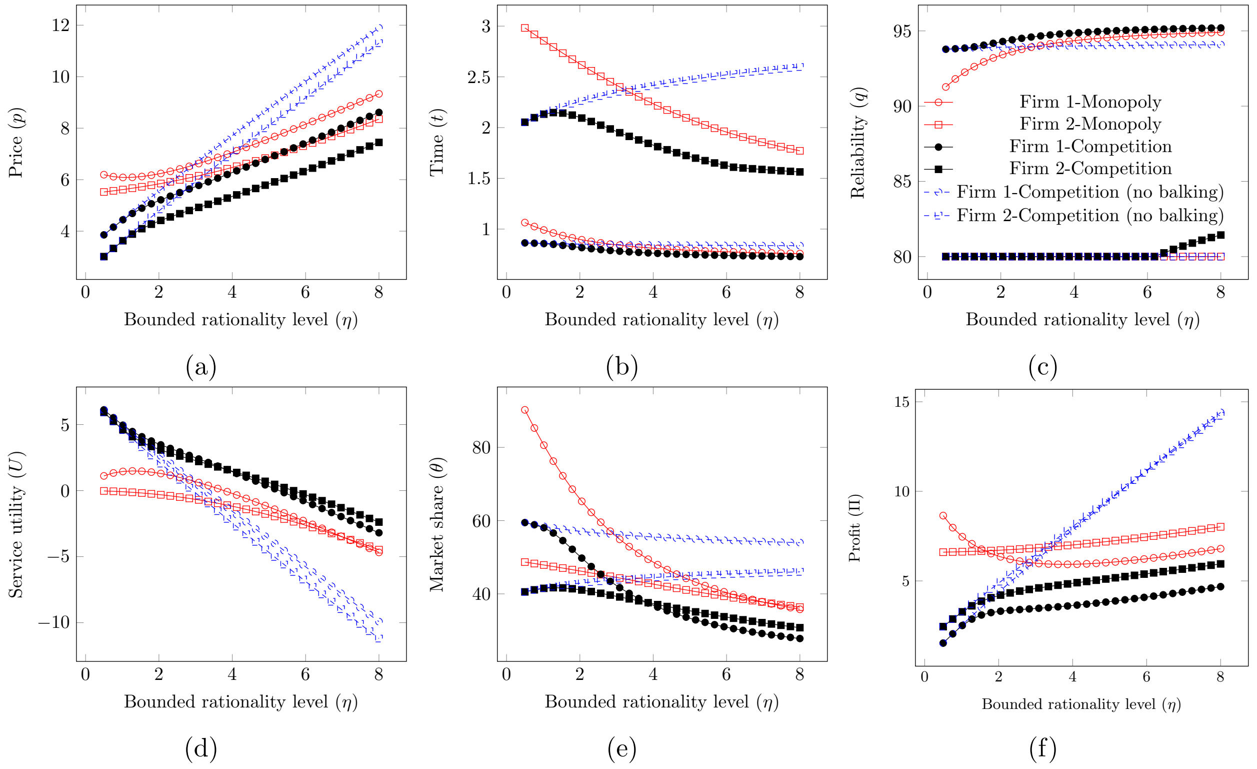
5 Numerical Study

In a competitive setting, the dynamics of the problem are too complicated to be studied analytically. In this section, we therefore study the competition problem through a number of numerical examples. The analysis mainly focuses on the impacts of the bounded rationality level, the balking option, the price sensitivity and the delivery reliability sensitivity on firms' optimal decisions and profitabilities under competition. We also compare our results with those presented in the literature. Particularly, we compare our price, time, and reliability (PTR) competition with the price and time (PT) competition in So (2000) and time and reliability (TR) competition in Ho & Zheng (2004) and Shang & Liu (2011). Our results emphasize the impact of the bounded rationality level and the balking option on the competition outcome and highlight the importance of integrating the pricing and delivery performance decisions in service delivery.

Consider a two-player game. A firm with a higher capacity level (lower unit operating cost) compared to its competitor will have a capacity (cost) advantage in the competition. It is obvious that a firm with the capacity (cost) advantage and without the cost (capacity) disadvantage has the dominating power in the competition. As a result, it gains a higher market share and also makes more profit compared to its competitor. More interesting cases, however, occur when one of the firms has the capacity advantage and the other one has the cost advantage. We focus on such cases and consider two-player games in order to keep the results interpretable. Numerical instances are chosen to illustrate the interesting findings.

5.1 Impact of Bounded Rationality and Balking Option

From Eq. (2), it can be noted that the customers' choice behaviour becomes less sensitive to the utilities of the services provided by the firms as the bounded rationality level increases. The bounded rationality level parameter, i.e., η, scales down the impact of the firms' decisions on the customers' choice. At a sufficiently high bounded rationality level, therefore, the differences between the firms start to diminish, and at the extreme case, i.e., $\eta \to \infty$, the customers randomly choose between different options. As the bounded rationality level increases, the service utilities sharply decrease even to strictly negative values at optimality. When the customers are given the balking option, the firms have to tune their decisions and consequently service utilities with the constant utility of the balking option. In fact, in this case, the firms cannot decrease the customers' utilities so much as they could if the customers did not have the balking option. The balking option assumption is consistent with the competition and demand elasticity assumptions. They imply that the product/service is not essential, and it has at least a number of substitutes one of which is not buying it (Parkin, 2019).



(a)

(b)

(c)

(d)

(e)

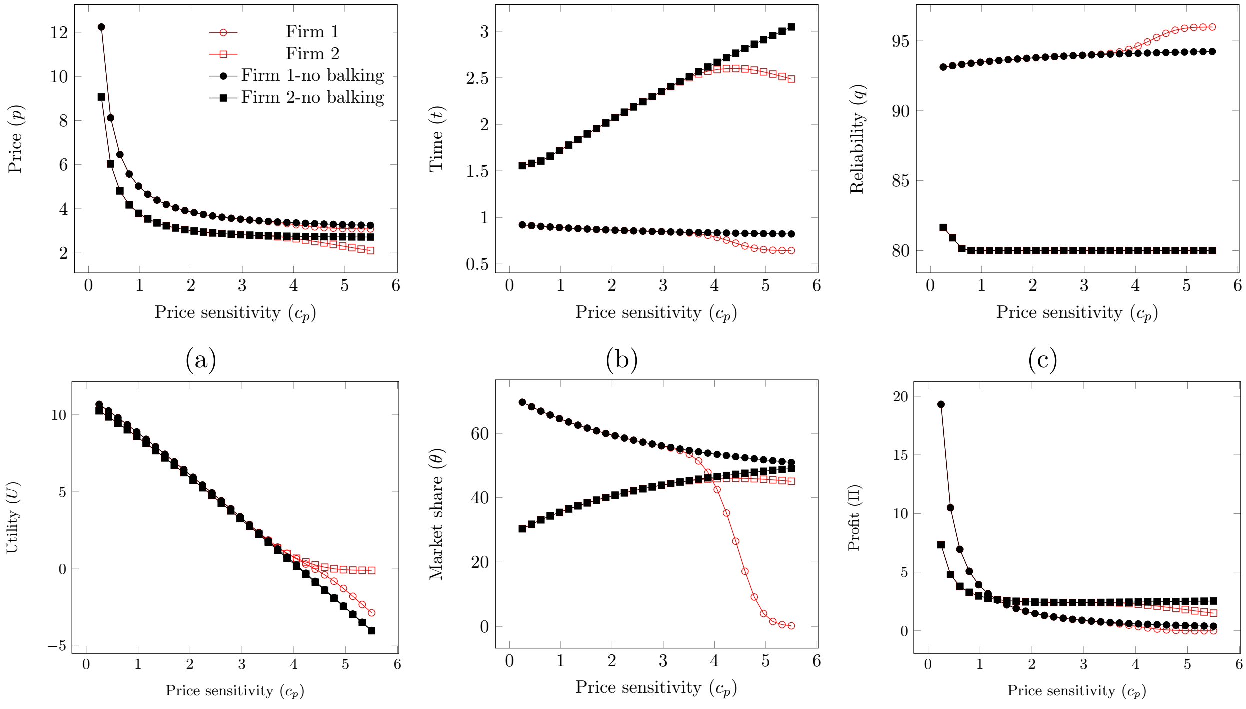
(f)

Figure 2: Example 1; $(\mu_1, \gamma_1) = (5, 3)$, $(\mu_2, \gamma_2) = (2, 1)$, $v = 10$, $c_p = 2$, $c_t = 1$, $c_q = 5$, $\Lambda = 3$, and $\underline{q} = 0.8$

Example 1. Consider two firms with $(\mu_1, \gamma_1) = (5, 3)$ and $(\mu_2, \gamma_2) = (2, 1)$ such that Firm 1 (Firm 2) has the capacity (cost) advantage over Firm 2 (Firm 1). Figure 2 shows the two firms' optimal decisions, service utilities, market shares, and profits in a monopoly and in a competition with and without the balking option. The following observations are made: (i) The competition decreases the profits and market shares of the firms even though they provide services with higher utilities, i.e., lower prices and superior delivery performances. It aligns with the intuition that customers are on average better off in a competitive market than in a monopolistic market as shown by Lemma 1. (ii) In contrast to a time-based competition, a firm with a higher capacity level does not necessarily have the dominating power in the market in a price- and time-based competition. In particular, a firm with a capacity advantage may gain a lower market share and be less profitable compared to a competitor with a lower capacity level but a cost advantage. This can be observed in Figures 2e and 2f at, for example, $\eta = 8$. (iii) Ceteris paribus, the bounded rationality level can be a determining factor in firms' positioning in a competitive market. For instance, while $\theta_1 > \theta_2$ at $\eta = 2$, $\theta_1 < \theta_2$ at $\eta = 8$ (see Figure 2e). (iv) The impact of the customers' bounded rationality on the firms' optimal decisions and profitabilities is better understood when customers have the balking option. The balking option acts as a hidden competitor that moderates the impact of the other firms. The balking option may significantly influence the firms' optimal decisions and profitabilities as shown in Figure 2. Moreover, without the balking option, the firms share the whole market such that an increase in one firm's market share coincides with a decrease in the competitor's market share. However, with the balking option, it can be the case that both firms' market shares increase or decrease, because some balking customers may choose to buy or some buying customers may choose to balk.

5.2 PTR Competition vs TR Competition

In a TR competition, a firm with a capacity advantage always wins a higher market share (and gains a higher profit for a fixed price) in the market (Shang & Liu, 2011). However, with an endogenous price, the two firm-specific parameters, i.e., μ and γ, together determine a firm's position in the market and its profitability. The TR competition is in fact a special case of the PTR competition. The following example shows the impact of the price sensitivity on the competition.



(a)

(b)

(c)

(d)

(e)

(f)

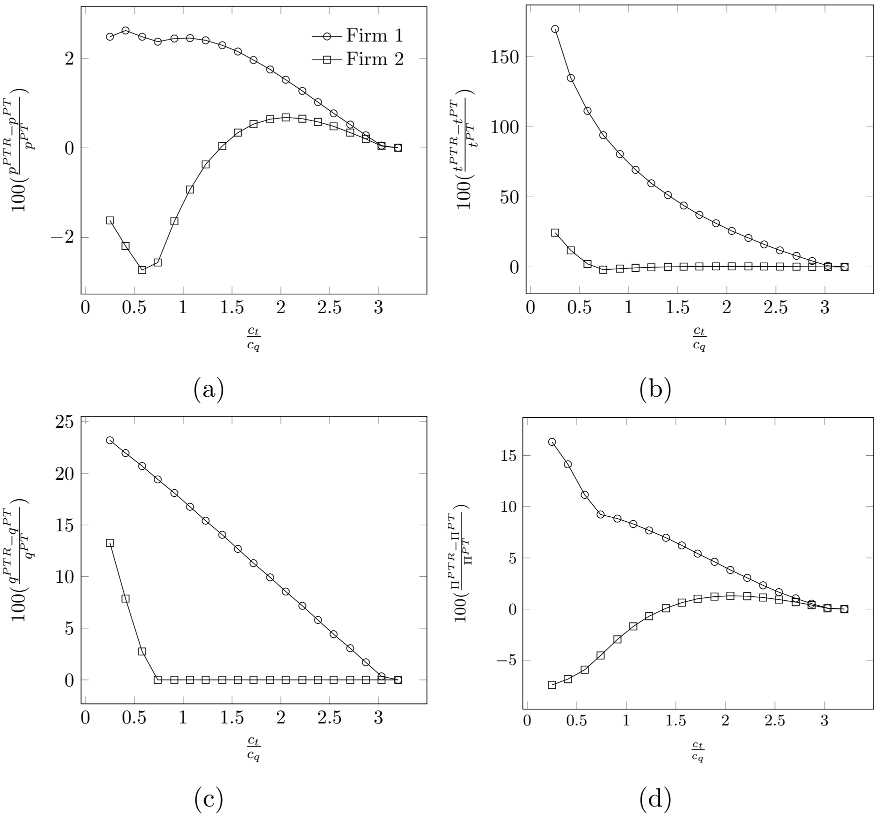
Figure 3: Example 2; $(\mu_1, \gamma_1) = (5, 3)$, $(\mu_2, \gamma_2) = (2, 1)$, $v = 10$, $\eta = 1$, $c_t = 1$, $c_q = 5$, $\Lambda = 3$, and $\underline{q} = 0.8$

Example 2. Consider the same two firms as in Example 1. We fix $\eta = 1$ as in Ho & Zheng (2004) and Shang & Liu (2011). Figure 3 reports the results for different values of c_p. An extremely low c_p represents the TR competition. At a low c_p, Firm 1 exploits its capacity advantage over Firm 2 to capture a higher market share. In other words, Firm 1's superior delivery performance outweighs the higher price it charges as a result of its cost disadvantage, because the customers are relatively more sensitive to the delivery performance at a low c_p. As the price sensitivity increases, the capacity advantage of Firm 1 diminishes, and Firm 2 with a better performance in the price dimension wins a higher market share and profit. At $c_p = 5.5$, Firm 1 practically leaves the competition. Note that, at a low level c_p, both firms provide their services with strictly positive utilities (see Figure 3d). As a result, only a *very* small fraction of customers chooses to balk and, consequently, the results of the competitions with and without the balking option coincide as shown in the figure. At a high c_p, however, the firms cannot provide their services with positive utilities. Therefore, a larger fraction of customers chooses to balk and correspondingly the results for the two competition cases (with and without the balking option) differ. The following observations are made: (i) While in a TR competition a firm with a capacity advantage always captures a higher market share, a firm with a capacity disadvantage but a cost advantage may capture a higher market share and be even more profitable in a PTR competition. (ii) The customers' sensitivity to price may significantly influence the firms' optimal decisions both in price, time, and reliability dimensions. In other words, a firm's optimal delivery time and delivery-reliability decisions in a TR competition may significantly differ from those in a PTR competition. This fact highlights the interplay between the pricing and operational decisions.

5.3 PTR Competition vs PT Competition

The PT competition is also a special case of the PTR competition. In Section 4, we mention that if the customers' reliability sensitivity is relatively small compared to their time sensitivity, the firms have no incentive to quote a delivery reliability level higher than the market entrance requirement. In this case, the PTR competition is reduced to the PT competition. We are interested in the impact of the relative sensitivity of the delivery-reliability on the competition outcome.

Example 3. Consider the same two firms as in Example 1. Figure 4 compares the two firms' optimal decisions in the PTR and PT competitions for different values of $\frac{c_t}{c_q}$. The following observations are made: (i) For a sufficiently high $\frac{c_t}{c_q}$, the firms compete only in terms of price and time and the PTR competition is reduced to the PT competition. (ii) A firm with a capacity advantage is better off when $\frac{c_t}{c_q}$ is lower. But a firm with a capacity disadvantage is not necessarily worse off in such a case. For instance, at $\frac{c_t}{c_q} = 2$, both firms are more profitable in PTR competition. Note that in a monopolistic setting, a firm is never worse off with an endogenous reliability decision. (iii) Figure 4 shows the fact that also considering delivery-reliability as a strategic tool in the competition can have a substantial impact on the firms' optimal decisions and profitabilities.



(a)

(b)

(c)

(d)

Figure 4: Example 3; $(\mu_1, \gamma_1) = (5, 3)$, $(\mu_2, \gamma_2) = (2, 1)$, $v = 10$, $\eta = 1$, $c_p = 3$, $c_q = 5$, $\Lambda = 3$, and $\underline{q} = 0.8$

6 Managerial Insights

There are multiple interesting findings which have been presented in different sections throughout the paper. In this section, we summarizes them as main managerial insights of the study.

(i) Only if the customers' reliability sensitivity is sufficiently large compared to their time sensitivity, the firm may have an incentive to quote a delivery reliability level higher than the market entrance requirement. In particular, if $\frac{c_t}{c_q} > (1 - \underline{q})\mu$, then a firm will not quote a delivery reliability level higher than the minimum requirement.

(ii) Decisions on price, delivery time, and delivery-reliability level are interdependent so that deciding on one or two of them without considering the other(s) can have a detrimental impact on profitability. Particularly, when the customers are highly price and reliability sensitive, outcomes of an integrated competition model may significantly differ from those in a pure price and time competition or a pure time and reliability competition.

(iii) In contrast to the time and reliability competition where a firm with a higher capacity has the dominating power in the market, in a price, time, and reliability competition both capacity level and unit operating cost along with market-specific parameters determine a firm's position in the market. Thus, even a firm that suffers from a low capacity level can gain a competitive advantage from reducing its operating cost, for example, by implementing efficiency improvements and waste reduction techniques.

(iv) In the monopolistic setting, the bounded rationality level significantly influences a firm's optimal decisions and profitability. A firm has an incentive to invest in reducing the bounded rationality level of the customers only if it can reach a sufficiently high market share in a sufficiently small bounded rationality level. Otherwise, the firm is always better off in markets with more boundedly rational customers.

(v) In the competition setting, the bounded rationality level is a determining factor in firms' positioning and profitability. In particular, our results show that the effect of firm-specific characteristics diminishes as the bounded rationality level increases, i.e., a high bounded rationality level attenuates the competition.

(vi) Customers' balking option acts as a hidden competitor that moderates the firms decisions. The assumption of having a balking option is consistent with the common competition and demand elasticity assumptions. The balking behaviour of the customers has a significant impact on the firms' optimal decisions, profitabilities, and positions in the market.

7 Conclusion

There is abundant empirical evidence supporting the fact that a better delivery performance influences customers' satisfaction and willingness to pay, and thereby influences firms' profitability. But what does better delivery performance mean? Two dimensions of the delivery performance are generally acknowledged in the literature: promised delivery time and delivery-reliability level. Although the existing interactions between price, time, and reliability have been widely recognized in empirical studies, they have been very scarcely addressed in the theoretical OM literature. No previous research has analytically considered their interactions in a competitive setting with boundedly rational customers.

In this study, we define, model, and analyze competition among an arbitrary number of profit-maximizing firms facing boundedly rational customers who are sensitive to price, promised delivery time, and delivery-reliability level. We use an MNL model to capture the customers' bounded rationality and stochastic choice behaviour. We prove the existence of a unique Nash equilibrium to the price, delivery time, and delivery reliability competition problem and propose a simple iterative algorithm that converges to this equilibrium. Furthermore, we explore managerial insights on how pricing, delivery time, and delivery reliability decisions should be made when the customers are boundedly rational.

We have considered a linear functional form for the utility function. It would be interesting to consider other utility functions in future research. Moreover, we have assumed that the customers are homogeneous. Taking the customers' heterogeneity into account is another possible direction for future studies. In this case, the market can also be segmented, and price and delivery performance differentiation can be included in the model. Additionally, we model the operations of each firm as an $M/M/1$ queueing system. More general queueing systems, i.e., $G/G/k$, could also be interesting to study. Another direction is to consider the capacity competition. The delivery time, delivery-reliability level, and capacity competition is studied by Shang & Liu (2011). However, they assume that the price is exogenous. An integrated model would be challenging to solve but may provide interesting results.

References

Abate, J., Choudhury, G. L., & Whitt, W. (1996). Exponential approximations for tail probabilities in queues ii: Sojourn time and workload. *Operations Research*, *44*(5), 758–763.

Allon, G. & Federgruen, A. (2007). Competition in service industries. *Operations Research*, *55*(1), 37–55.

Allon, G. & Federgruen, A. (2009). Competition in service industries with segmented markets. *Management Science*, *55*(4), 619–634.

Ariely, D. (2009). The end of rational economics. *Harvard Business Review*, *87*(7-8), 78–84.

Asmussen, S. (2008). *Applied probability and queues*, volume 51. Springer Science & Business Media.

Baker, W., Marn, M., & Zawada, C. (2001). Price smarter on the net. *Harvard Business Review*, *79*(2), 122–7.

Behzad, B. & Jacobson, S. H. (2016). Asymmetric bertrand-edgeworth-chamberlin competition with linear demand: A pediatric vaccine pricing model. *Service Science*, *8*(1), 71–84.

Benjaafar, S., Elahi, E., & Donohue, K. L. (2007). Outsourcing via service competition. *Management Science*, *53*(2), 241–259.

Boyaci, T. & Ray, S. (2006). The impact of capacity costs on product differentiation in delivery time, delivery reliability, and price. *Production and Operations Management*, *15*(2), 179–197.

Chen, H. & Wan, Y.-W. (2003). Price competition of make-to-order firms. *IIE Transactions*, *35*(9), 817–832.

Chen, H. & Wan, Y.-w. (2005). Capacity competition of make-to-order firms. *Operations Research Letters*, *33*(2), 187–194.

Chen, W., Zhang, Z. G., & Hua, Z. (2019). Analysis of price competition in two-tier service systems. *Journal of the Operational Research Society*, 1–13.

Chen, Y., Su, X., & Zhao, X. (2012). Modeling bounded rationality in capacity allocation games with the quantal response equilibrium. *Management Science*, *58*(10), 1952–1962.

Davis, A. M. (2018). *Biases in Individual Decision-Making*, chapter 5, (pp. 149–198). John Wiley & Sons, Ltd.

Hassin, R. (2016). *Rational queueing*. CRC press.

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

Ho, T.-H. & Zhang, J. (2008). Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Management Science*, *54*(4), 686–700.

Ho, T. H. & Zheng, Y.-S. (2004). Setting customer expectation in service delivery: An integrated marketing-operations perspective. *Management Science*, *50*(4), 479–488.

Hu, Y. & Qiang, Q. (2013). An equilibrium model of online shopping supply chain networks with service capacity investment. *Service Science*, *5*(3), 238–248.

Huang, J., Leng, M., & Parlar, M. (2013). Demand functions in decision modeling: A comprehensive survey and research directions. *Decision Sciences*, *44*(3), 557–609.

Huang, T., Allon, G., & Bassamboo, A. (2013). Bounded rationality in service systems. *Manufacturing & Service Operations Management*, *15*(2), 263–279.

Huang, T. & Chen, Y.-J. (2015). Service systems with experience-based anecdotal reasoning customers. *Production and Operations Management*, *24*(5), 778–790.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: heuristics and b iases.* Cambridge University Press.

Kelley, S. W., Hoffman, K., & Davis, M. A. (1993). A typology of retail failures and recoveries. *Journal of Retailing*, *69*(4), 429 – 452.

Lee, H. L. & Whang, S. (2001). Winning the last mile of e-commerce. *MIT Sloan Management Review*, *42*(4), 54–54.

Li, L., Jiang, L., & Liu, L. (2012). Service and price competition when customers are naive. *Production and Operations Management*, *21*(4), 747–760.

Li, X., Guo, P., & Lian, Z. (2016). Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management*, *25*(11), 1885–1901.

Li, X., Guo, P., & Lian, Z. (2017). Price and capacity decisions of service systems with boundedly rational customers. *Naval Research Logistics*, *64*(6), 437–452.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* Wiley.

Marand, A. J., Tang, O., & Li, H. (2019). Quandary of service logistics: Fast or reliable? *European Journal of Operational Research*, *275*(3), 983 – 996.

McKelvey, R. D. & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and economic behavior*, *10*(1), 6–38.

Parkin, M. (2019). *Economics, Global Edition. 13th edition.* Pearson Education Limited.

Peng, D. X. & Lu, G. (2017). Exploring the impact of delivery performance on customer transaction volume and unit price: Evidence from an assembly manufacturing supply chain. *Production and Operations Management*, *26*(5), 880–902.

Rao, S., Griffis, S. E., & Goldsby, T. J. (2011). Failure to deliver? linking online order fulfillment glitches with future purchase behavior. *Journal of Operations Management*, *29*(7), 692–703.

Rao, S., Rabinovich, E., & Raju, D. (2014). The role of physical distribution services as determinants of product returns in internet retailing. *Journal of Operations Management*, *32*(6), 295–312.

Ren, H. & Huang, T. (2018). Modeling customer bounded rationality in operations management: A review and research opportunities. *Computers & Operations Research*, *91*, 48–58.

Ren, H., Huang, T., & Arifoglu, K. (2018). Managing service systems with unknown quality and customer anecdotal reasoning. *Production and Operations Management*, *27*(6), 1038–1051.

Rosenzweig, E. D., Roth, A. V., & Dean, J. W. (2003). The influence of an integration strategy on competitive capabilities and business performance: an exploratory study of consumer products manufacturers. *Journal of Operations Management*, *21*(4), 437–456.

Saberi, S., Nagurney, A., & Wolf, T. (2014). A network economic game theory model of a service-oriented internet with price and quality competition in both content and network provision. *Service Science*, *6*(4), 229–250.

Shang, W. & Liu, L. (2011). Promised delivery time and capacity games in time-based competition. *Management Science*, *57*(3), 599–610.

Simon, H. (1997). *Models of Man*. John Wiley & Sons, New York.

So, K. C. (2000). Price and time competition for service delivery. *Manufacturing & Service Operations Management*, *2*(4), 392–409.

Strauss, A. K., Klein, R., & Steinhardt, C. (2018). A review of choice-based revenue management: Theory and methods. *European Journal of Operational Research*, *271*(2), 375–387.

Su, X. (2008). Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management*, *10*(4), 566–589.

Talluri, K. T. & Van Ryzin, G. J. (2006). *The theory and practice of revenue management*, volume 68. Springer Science & Business Media.

Xiao, T. & Qi, X. (2016). A two-stage supply chain with demand sensitive to price, delivery time, and reliability of delivery. *Annals of Operations Research*, *241*(1-2), 475–496.

On-line Supplement for “Competing on Price, Speed, and Reliability: How Does Bounded Rationality Matter?”

Proof of Proposition 1. For the sake of simplicity, we drop the subscript i for the remainder of this proof as long as it does not cause any confusion. We start by proving part (b). From the first-order conditions, we have

$$\begin{aligned}\frac{\partial \Pi(t,q|\beta_i)}{\partial t} &= \frac{k}{t^2}\left(\frac{t^2}{k}(\mu-\frac{k}{t})\frac{\partial p(t,q|\beta_i)}{\partial t}+p(t,q|\beta_i)-\gamma\right),\\ \frac{\partial \Pi(t,q|\beta_i)}{\partial q} &= \frac{-1}{t(1-q)}\left(-t(1-q)(\mu-\frac{k}{t})\frac{\partial p(t,q|\beta_i)}{\partial q}+p(t,q|\beta_i)-\gamma\right)\end{aligned}$$

where

$$\begin{aligned}\frac{\partial p(t,q|\beta_i)}{\partial t} &= -\frac{1}{c_p}\left(c_t+\frac{k\Lambda}{t^2(\Lambda-\mu+k/t)(\mu-k/t)}\right)<0,\\ \frac{\partial p(t,q|\beta_i)}{\partial q} &= \frac{1}{c_p}\left(c_q+\frac{\Lambda}{t(1-q)(\Lambda-\mu+k/t)(\mu-k/t)}\right)>0.\end{aligned}$$

In the system of equations ($\frac{\partial \Pi(t,q|\beta_i)}{\partial t}=0, \frac{\partial \Pi(t,q|\beta_i)}{\partial q}=0$) equating $\left(\frac{t^2}{k}(\mu-\frac{k}{t})\frac{\partial p(t,q|\beta_i)}{\partial t}+p(t,q|\beta_i)-\gamma\right)$ and $\left(-t(1-q)(\mu-\frac{k}{t})\frac{\partial p(t,q|\beta_i)}{\partial q}+p(t,q|\beta_i)-\gamma\right)$ results in

$$t(q)=\frac{c_q(1-q)}{c_t}\ln\frac{1}{1-q}. \tag{13}$$

To have an interior delivery-reliability level solution, there should exist a solution to the first-order conditions. Substituting Eq. (13) in $\frac{\partial \Pi(t,q|\beta_i)}{\partial t}=0$ results in

$$\frac{\partial \Pi(t(q),q|\beta_i)}{\partial t}=\frac{k}{(t(q))^2}H(q|\beta_i)=0.$$

$H(q|\beta_i)$ is well-defined for $1-\frac{c_t}{c_q(\mu-\Lambda)}<q<1-\frac{c_t}{c_q\mu}$ when $\mu>\Lambda$, and for $q<1-\frac{c_t}{c_q\mu}$ when $\mu<\Lambda$. $H(q|\beta_i)$ is decreasing in q, since

$$\begin{aligned}\frac{\partial H(q|\beta_i)}{\partial q} &= \frac{c_q^2\mu}{c_t}(-2q\ln\frac{1}{1-q}+1-q)-\frac{c_t\eta\Lambda}{c_q(1-q)^2(\Lambda-\mu+\frac{c_t}{c_q(1-q)})^2}\\ &\quad -\frac{c_t\eta\Lambda}{c_q(1-q)^2(\Lambda-\mu+\frac{c_t}{c_q(1-q)})(\mu-\frac{c_t}{c_q(1-q)})}-c_q<0,\end{aligned}$$

for $q\geq\underline{q}\geq 0.64$

For (i) $\Lambda>\mu$, the acceptable region for q is $\underline{q}<q<1-\frac{c_t}{c_q\mu}$. We have $\lim_{q\to(1-\frac{c_t}{c_q\mu})^-}H(q|\beta_i)=-\infty$. If $\lim_{q\to\underline{q}}H(q|\beta_i)>0$, then there is a unique solution to $H(q|\beta_i)=0$. For $\Lambda<\mu$, the acceptable region for q is $\max\{1-\frac{c_t}{c_q(\mu-\Lambda)},\underline{q}\}<q<1-\frac{c_t}{c_q\mu}$. We have $\lim_{q\to(1-\frac{c_t}{c_q\mu})^-}H(q|\beta_i)=-\infty$. For (ii) $1-\frac{c_t}{c_q(\mu-\Lambda)}<\underline{q}$, if $\lim_{q\to\underline{q}}H(q|\beta_i)>0$, then there is a unique solution to $H(q|\beta_i)=0$. For (iii) $1-\frac{c_t}{c_q(\mu-\Lambda)}>\underline{q}$, there exists a unique solution to $H(q|\beta_i)=0$, because $\lim_{q\to(1-\frac{c_t}{c_q(\mu-\Lambda)})^+}H(q|\beta_i)=+\infty$. Firm i's profit function is unimodal in q_i (first increasing and then decreasing). Due to this fact, when an interior solution for delivery-reliability level exists, the boundary delivery-reliability level, i.e. $\underline{q}$, cannot be optimal. This completes the proof to part (b).

When there is no interior solution for the delivery-reliability level, the firm has to choose the boundary value, $\underline{q}$. Substitute $q = \underline{q}$ in $p(t, q|\beta_i)$ and λ in Eq. (11) and Eq. (6), respectively. Taking the first derivative of the objective function with respect to t results in

$$\frac{\partial \Pi(t, \underline{q})}{\partial t} = \frac{\underline{k}}{t^2} G(t|\beta_i) = 0.$$

$G(t|\beta_i)$ is increasing in t, because

$$\frac{\partial G(t|\beta_i)}{\partial t} = \frac{2c_t \mu t}{\underline{k}} + \frac{\eta \Lambda \underline{k}}{t^2(\Lambda - \mu + \frac{\underline{k}}{t})^2} + \frac{\eta \Lambda \underline{k}}{t^2(\mu - \frac{\underline{k}}{t})(\Lambda - \mu + \frac{\underline{k}}{t})} > 0.$$

Moreover, $\lim_{t\to(\frac{\underline{k}}{\mu})^+} G(t|\beta_i) = -\infty$ and $\lim_{t\to+\infty} G(t|\beta_i) = +\infty$ when $\Lambda > \mu$, and $\lim_{t\to(\frac{\underline{k}}{\mu-\Lambda})^-} G(t|\beta_i) = +\infty$ when when $\Lambda < \mu$. Therefore, there is a unique solution to $G(t|\beta_i) = 0$. It also shows that for any given delivery-reliability level the optimal delivery time can be uniquely determined from the first-order condition with respect to t. This completes the proof to part (c).

Note that from $\frac{\partial \Pi(t,q)}{\partial q} = 0$ and Eq. (13), the price can be found as

$$p(q) = \frac{1}{c_p}\Big(\frac{c_q^2(1-q)^2}{c_t}(\mu - \frac{c_t}{c_q(1-q)}) \ln \frac{1}{1-q} + \frac{\eta \Lambda}{(\Lambda - \mu + \frac{c_t}{c_q(1-q)})} + c_p \gamma\Big), \tag{14}$$

and from $\frac{\partial \Pi_{(t,q)}}{\partial t} = 0$ the price is equal to

$$p(t) = \frac{1}{c_p}\Big(\frac{c_t t^2}{\underline{k}}(\mu - \frac{\underline{k}}{t}) + \frac{\eta \Lambda}{(\Lambda - \mu + \frac{\underline{k}}{t})} + c_p \gamma\Big). \tag{15}$$

In either case of an interior or a boundary delivery-reliability level, according to Eq. (14) and Eq. (15), the optimal price is greater than γ, and therefore constraint (8) is met.

Parts (b) and (c), together, uniquely determine the optimal solution. This completes the proof. □

Proof of Proposition 2. The optimal delivery-reliability level is either an interior point, i.e., $q \in (\underline{q}, 1)$, or on the boundary, i.e., $q = \underline{q}$.

Interior delivery-reliability level: If the optimal delivery-reliability level is an interior point, according to Proposition 1, it is the unique root of $H(q|\beta_i = 1) = 0$. We have

$$\frac{\partial q}{\partial \eta} = -\frac{\frac{\partial H(q|\beta_i=1)}{\partial \eta}}{\frac{\partial H(q|\beta_i=1)}{\partial q}}.$$

The sign of $\frac{\partial q}{\partial \eta}$ is determined by the sign of $\frac{\partial H(q|\beta_i=1)}{\partial \eta}$ since $\partial H(q|\beta_i = 1)\partial q < 0$ as shown in the proof of Proposition 1. Furthermore, we have

$$\frac{\partial H(q|\beta_i = 1)}{\partial \eta} = 1 + \frac{\theta}{1-\theta} + \ln(\frac{\theta}{1-\theta}),$$

in which θ is implicitly dependent on q. $\frac{\partial H(q|\beta_i=1)}{\partial \eta}$ is increasing in θ, since $\frac{\partial^2 H(q|\beta_i=1)}{\partial \eta \partial \theta} = \frac{1}{(1-\theta)^2} + \frac{1}{\theta(1-\theta)} > 0$. In addition, $\lim_{\theta \to 0} \frac{\partial H(q|\beta_i=1)}{\partial \eta} = -\infty$ and $\lim_{\theta \to 1} \frac{\partial H(q|\beta_i=1)}{\partial \eta} = +\infty$. Hence, there is a unique market share, i.e., $\hat{\theta} \approxeq 0.218$ at which $\frac{\partial H(q|\beta_i=1)}{\partial \eta} = 0$, and $\frac{\partial H(q|\beta_i=1)}{\partial \eta}$ is negative (positive) for $\theta < \hat{\theta}$ ($\theta > \hat{\theta}$). It can thereby be concluded that $\frac{\partial q}{\partial \eta} < 0$ ($\frac{\partial q}{\partial \eta} > 0$) for $\theta < \hat{\theta}$ ($\theta > \hat{\theta}$).

For an interior delivery-reliability solution, we have $t(q) = \frac{c_q(1-q)}{c_t} \ln \frac{1}{1-q}$ and $\frac{\partial t(q)}{\partial q} = \frac{c_q}{c_t}(1 - \ln(\frac{1}{1-q})) < 0$ for $q > \underline{q} \geq 0.69$. Thus, $\frac{\partial t(q)}{\partial \eta} = \frac{\partial t(q)}{\partial q} \frac{\partial q}{\partial \eta} > 0$ ($\frac{\partial t(q)}{\partial \eta} < 0$) for $\theta < \hat{\theta}$ ($\theta > \hat{\theta}$). For an interior delivery-reliability solution, we have $\lambda = \mu - \frac{c_t}{c_q(1-q)}$ and $\frac{\partial \lambda}{\partial q} = -\frac{c_t}{c_q(1-q)^2} < 0$ or equivalently $\frac{\partial \theta}{\partial q} < 0$. Thus, $\frac{\partial \theta}{\partial \eta} = \frac{\partial \theta}{\partial q} \frac{\partial q}{\partial \eta} > 0$ ($\frac{\partial \theta}{\partial \eta} < 0$) for $\theta < \hat{\theta}$ ($\theta > \hat{\theta}$). Moreover, $\frac{\partial \Pi}{\partial \eta} = -(\mu - \frac{c_t(1-q)}{c_q}) \ln(\frac{\theta}{1-\theta})$. For $\theta < 0.5$ ($\theta > 0.5$), it is easy to show that $\ln(\frac{\theta}{1-\theta}) < 0$ ($\ln(\frac{\theta}{1-\theta}) > 0$), and consequently $\frac{\partial \Pi}{\partial \eta} > 0$ ($\frac{\partial \Pi}{\partial \eta} < 0$).

It can be concluded that for $\hat{\theta} < \theta < 0.5$, the optimal delivery time is decreasing, and the optimal delivery reliability is increasing. Moreover, the optimal profit is increasing, even though the market share (or equivalently, demand rate) is decreasing. This indicates that the optimal price is increasing for $\hat{\theta} < \theta < 0.5$.

From Eq. (14), the following equation is obtained: $\frac{\partial p(q)}{\partial \eta} = \frac{\partial p(q)}{\partial q} \frac{\partial q}{\partial \eta} + \frac{\Lambda}{c_p(\Lambda - \mu t + \frac{c_t}{c_q(1-q)})}$. It can be shown that $\frac{\partial p(q)}{\partial q} < 0$ (see the proof of Lemma 1). We have also shown that $\frac{\partial q}{\partial \eta} < 0$ for $\theta < \hat{\theta}$. Furthermore, $\frac{\Lambda}{c_p(\Lambda - \mu t + \frac{c_t}{c_q(1-q)})} > 0$. Hence, it can be concluded that $\frac{\partial p(q)}{\partial \eta} = \frac{\partial p(q)}{\partial q} \frac{\partial q}{\partial \eta} + \frac{\Lambda}{c_p(\Lambda - \mu t + \frac{c_t}{c_q(1-q)})} > 0$ for $\theta < \hat{\theta}$.

Boundary delivery-reliability level: If the optimal delivery-reliability level is a boundary solution, it is the unique root of $G(t|\beta_i = 1) = 0$ according to Proposition 1. We have

$$\frac{\partial t}{\partial \eta} = -\frac{\frac{\partial G(t|\beta_i=1)}{\partial \eta}}{\frac{\partial G(t|\beta_i=1)}{\partial t}}.$$

The sign of $\frac{\partial t}{\partial \eta}$ is the opposite of the sign of $\frac{\partial G(t|\beta_i=1)}{\partial \eta}$ since $\frac{\partial G(t|\beta_i=1)}{\partial t} > 0$ as shown in proof of Proposition 1. Furthermore, we have

$$\frac{\partial G(t|\beta_i = 1)}{\partial \eta} = 1 + \frac{\theta}{1-\theta} + \ln(\frac{\theta}{1-\theta}),$$

which is the same as $\frac{\partial H(q|\beta_i=1)}{\partial \eta}$. One can obtain similar results for the boundary delivery reliability case following the same steps that we have taken for the interior delivery reliability case.

□

Proof of Lemma 1. The optimal delivery-reliability level is either an interior point, i.e., $q \in (\underline{q}, 1)$, or on the boundary, i.e., $q = \underline{q}$.

Interior delivery-reliability level: If the optimal delivery-reliability level is an interior point, according to Proposition 1, it is the unique root of $H(q|\beta_i) = 0$. We have

$$\frac{\partial q}{\partial \beta} = -\frac{\frac{\partial H(q|\beta_i)}{\partial \beta}}{\frac{\partial H(q|\beta_i)}{\partial q}} > 0,$$

because $\frac{\partial H(q|\beta_i)}{\partial \beta} = \frac{1}{\beta} > 0$, and $\frac{\partial H}{\partial q} < 0$ as shown in Proof of Proposition 1. In addition, the optimal delivery time is found from Eq. (13). We have

$$\frac{\partial t(q)}{\partial q} = \frac{c_q}{c_t}(1 - \ln \frac{1}{1-q}) < 0,$$

because $\ln \frac{1}{1-q} > 1$ for $q \geq \underline{q} \geq 0.64$. Thus,

$$\frac{\partial t(q)}{\partial \beta} = \frac{\partial t(q)}{\partial q} \frac{\partial q}{\partial \beta} < 0.$$

According to Eq. (14), for all $q \geq \underline{q} \geq 0.64$, we have

$$\frac{\partial p(q)}{\partial q} = \frac{1}{c_p}(\frac{c_q^2}{c_t}(1 - \ln \frac{1}{1-q})(1-q)(\mu - \frac{c_t}{c_q(1-q)}) - \frac{c_q^2 \mu}{c_t}(1-q) \ln \frac{1}{1-q} - \frac{\eta \Lambda c_t}{c_q(1-q)^2(\Lambda - \mu + \frac{c_t}{c_q(1-q)})^2}) < 0.$$

Therefore,

$$\frac{\partial p(q)}{\partial \beta} = \frac{\partial p(q)}{\partial q} \frac{\partial q}{\partial \beta} < 0.$$

Boundary delivery-reliability level: If the optimal delivery-reliability level is determined by $\underline{q}$, according to Proposition 1, the optimal delivery time is the unique root of $G(t|\beta_i) = 0$. We have,

$$\frac{\partial t}{\partial \beta} = -\frac{\frac{\partial G(t|\beta_i)}{\partial \beta}}{\frac{\partial G(t|\beta_i)}{\partial t}} < 0,$$

because $\frac{\partial G(t|\beta_i)}{\partial \beta} = \frac{1}{\beta} > 0$, and $\frac{\partial G(t|\beta_i)}{\partial t} > 0$. According to Eq. (15), we have

$$\frac{\partial p(t)}{\partial t} = \frac{1}{c_p}(\frac{2c_t}{\underline{k}} t(\mu - \frac{\underline{k}}{t}) + c_t + \frac{\Lambda \underline{k}}{t^2(\Lambda - \mu + \frac{\underline{k}}{t})}) > 0.$$

Therefore,

$$\frac{\partial p(t)}{\partial \beta} = \frac{\partial p(t)}{\partial t} \frac{\partial t}{\partial \beta} < 0.$$

Thus, it can be concluded that the optimal price and delivery time are strictly decreasing in β. If conditions stated in Proposition 1 part (a) are not satisfied, then the optimal delivery-reliability level would be the boundary point and invariant with respect to β. The optimal delivery-reliability level is therefore increasing in β.

□

Proof of Proposition 3. We prove the results in two parts: proof of the convergence and proof of the uniqueness.

Proof of the convergence: We use an inductive reasoning to prove the convergence of the procedure. Define $A_i = e^{U_i/\eta}$ as the attraction of Firm i. Let $p_i^{(j)}$, $t_i^{(j)}$ and $q_i^{(j)}$ denote the price, delivery time and delivery-reliability level of Firm i in the j-th iteration. Set $p_i^{(0)} = \gamma_i$, $t_i^{(0)} = \frac{k}{\mu_i}$ and $q_i^{(0)} = 1$. Obviously, $p_i^{(j)} > \gamma_i$, $t_i^{(j)} > \frac{k}{\mu_i}$ and $q_i^{(j)} < 1$ for all j. Particularly, $p_i^{(1)} > p_i^{(0)}$, $t_i^{(1)} > t_i^{(0)}$ and $q_i^{(1)} < q_i^{(0)}$. Assume $p_i^{(j)} \geq p_i^{(j-1)}$, $t_i^{(j)} \geq t_i^{(j-1)}$ and $q_i^{(j)} \leq q_i^{(j-1)}$. Consider $j = n$. $\beta_i^{(j)}$ denotes the combined impact of the other firms' actions on Firm i's problem at the j-th iteration. For Firm 1 we have,

$$\begin{aligned}
\beta_1^{(n)} &= \sum_{m=2}^{N} e^{U_m/\eta} + 1 \\
&= \sum_{m=2}^{N} e^{(v - c_p p_m^{(n-1)} - c_t t_m^{(n-1)} + c_q q_m^{(n-1)})/\eta} + 1 \\
&\leq \sum_{m=2}^{N} e^{(v - c_p p_m^{(n-2)} - c_t t_m^{(n-2)} + c_q q_m^{(n-2)})/\eta} + 1 = \beta_1^{(n-1)}.
\end{aligned}$$

Therefore, according to Lemma 1, we have $p_1^{(n)} \geq p_1^{(n-1)}$, $t_1^{(n)} \geq t_1^{(n-1)}$ and $q_1^{(n)} \leq q_1^{(n-1)}$, because $\beta_1^{(n)} \leq \beta_1^{(n-1)}$.

Now, suppose that $p_i^{(n)} \geq p_i^{(n-1)}$, $t_i^{(n)} \geq t_i^{(n-1)}$ and $q_i^{(n)} \leq q_i^{(n-1)}$ for all $i = 1, 2, .., s-1$. For Firm s we have

$$\begin{aligned}
\beta_s^{(n)} &= \sum_{m=1}^{s-1} e^{U_m/\eta} + \sum_{m=s+1}^{N} e^{U_m/\eta} + 1 \\
&= \sum_{m=1}^{s-1} e^{(v - c_p p_m^{(n)} - c_t t_m^{(n)} + c_q q_m^{(n)})/\eta} + \sum_{m=1}^{s-1} e^{(v - c_p p_m^{(n-1)} - c_t t_m^{(n-1)} + c_q q_m^{(n-1)})/\eta} + 1 \\
&\leq \sum_{m=1}^{s-1} e^{(v - c_p p_m^{(n-1)} - c_t t_m^{(n-1)} + c_q q_m^{(n-1)})/\eta} + \sum_{m=1}^{s-1} e^{(v - c_p p_m^{(n-2)} - c_t t_m^{(n-2)} + c_q q_m^{(n-2)})/\eta} + 1 = \beta_s^{(n-1)}.
\end{aligned}$$

Similarly, according to Lemma 1, we have $p_s^{(n)} \geq p_s^{(n-1)}$, $t_s^{(n)} \geq t_s^{(n-1)}$ and $q_s^{(n)} \leq q_s^{(n-1)}$, because $\beta_s^{(n)} \leq \beta_s^{(n-1)}$.

From Proposition 1, for a given β_i the decisions of Firm i are uniquely determined. We also know that through the above-mentioned iterative procedure, β_i decreases for each firm in each step. Moreover, the balking option implies that there is a lower bound for $\beta_i > 1$. Thus, the iterative procedure converges.

Proof of the uniqueness: We prove the uniqueness by contradiction. Suppose there exist two equilibrium solutions denoted by $\phi = (x_1, \ldots, x_N)$ and $\phi' = (x'_1, \ldots, x'_N)$ with corresponding optimal solutions (p_i, t_i, q_i) and (p'_i, t'_i, q'_i). Let the attractions of each firm in the two equilibrium solutions be expressed as $A'_i = (1 + r_i)A_i$. Without loss of generality, the firms can be numbered such that $r_1 \geq r_2 \geq \ldots \geq r_N$ and $r_1 > 0$. Next, we show that r_2 must be positive, i.e., $r_2 > 0$.

Assume $r_2 \leq 0$. It implies $A'_i \leq A_i$ for $i = 2, \ldots, N$, and subsequently $\beta'_1 \leq \beta_1$. From Lemma 1, we have $p_1' \geq p_1$, $t_1' \geq t_1$ and $q_1' \leq q_1$. It implies

$$A_1 = e^{(v - c_p p_1 - c_t t_1 + c_q q_1)/\eta} \geq e^{(v - c_p p_1' - c_t t_1' + c_q q_1')/\eta} = A_1' = (1 + r_1)A_1,$$

that is in contradiction with $r_1 > 0$. Therefore, we must have $r_2 > 0$.

Consider Firm 1's problem. Assume $r_i = r_2$ for $i = 2, \ldots, N$. Denote the corresponding equilibrium solution by $\phi'' = (x_1'', \ldots, x_N'')$ where $A_i'' = (1 + r_2)A_i$ for $i = 2, \ldots, N$. Let (p_i'', t_i'', q_i'') represent the corresponding optimal solution. From Eq. (3) and Eq. (6), for ϕ and ϕ'' we have

$$\mu_1 - \frac{\Lambda A_1}{A_1 + \sum_{i=2}^{N} A_i} = \frac{k_1}{t_1}$$

and

$$\begin{aligned} \mu_1 - \frac{\Lambda A_1''}{A_1'' + \sum_{i=2}^{N} A_i''} &= \mu_1 - \frac{\Lambda A_1''}{A_1'' + (1 + r_2)\sum_{i=2}^{N} A_i} \\ &= \mu_1 - \frac{\Lambda A_1''/(1 + r_2)}{A_1''/(1 + r_2) + \sum_{i=2}^{N} A_i} \\ &= \frac{k_1''}{t_1''}. \end{aligned}$$

Since $r_2 > 0$, $A_i'' > A_i$ for $i = 2, \ldots, N$, and subsequently $\beta_1'' > \beta_1$. From Lemma 1, we have $p_1'' < p_1$, $t_1'' < t_1$ and $q_1'' \geq q_1$ (equivalently $k_1'' \geq k_1$) which gives $\frac{k_i''}{t_i''} > \frac{k_i}{t_i}$. It implies

$$\mu_1 - \frac{\Lambda A_1''/(1 + r_2)}{A_1''/(1 + r_2) + \sum_{i=2}^{N} A_i} > \mu_1 - \frac{\Lambda A_1}{A_1 + \sum_{i=2}^{N} A_i}$$

or equivalently

$$A_1'' < (1 + r_2)A_1. \tag{16}$$

Again, consider Firm 1's problem. Under equilibrium solution ϕ', we have $A_i' \leq A_i''$ for $i = 2, \ldots, N$, because $r_i \leq r_2$ for $i = 3, \ldots, N$. It implies $\beta_1' \leq \beta_1''$. From Lemma 1, we know $p_1'' \leq p_1'$, $t_1'' \leq t_1'$ and $q_1'' \geq q_1'$. Thus, we have

$$A_1'' = e^{(v - c_p p_1'' - c_t t_1'' + c_q q_1'')/\eta} \geq e^{(v - c_p p_1' - c_t t_1' + c_q q_1')/\eta} = A_1' = (1 + r_1)A_1. \tag{17}$$

From Eq. (16) and Eq. (17), we have

$$(1 + r_1)A_1 = A_1' \leq A_1'' < (1 + r_2)A_1, \tag{18}$$

that implies $r_1 < r_2$ and contradicts the assumption $r_1 \geq r_2$. Therefore, the equilibrium solution must be unique.

□

Proof of Corollary 1. In the equilibrium the firms would offer either the lowest delivery-reliability level, i.e., $q = \underline{q}$, or an interior delivery-reliability level, i.e., $q > \underline{q}$. From Proposition 1, if an interior optimal delivery-reliability level exits for firm i, then we have $\mu_i - \lambda_i = \frac{k_i}{t_i} = \frac{c_t}{c_q(1-q_i)}$ and consequently $\mu_i > \frac{c_t}{c_q(1-q_i)}$. Since $\frac{c_t}{c_q(1-q_i)}$ is increasing in q_i, if $\mu_i < \frac{c_t}{c_q(1-\underline{q})}$, then for no $q_i > \underline{q}$ the condition $\frac{c_t}{c_q(1-q_i)} < \mu_i$ holds, and there exists no interior delivery-reliability level. □