Quality and Capacity Decisions in Service Processes

Azin Farzan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Yong-Pin Zhou, Chair

Yong Tan

Hamed Mamani

Program Authorized to Offer Degree:
Michael G. Foster School of Business

University of Washington

Abstract

Quality and Capacity Decisions in Service Processes

Azin Farzan

Chair of the Supervisory Committee:
Professor Yong-Pin Zhou
ISOM

In this research we provide different analytical models that capture the effects of quality of service. We consider quality of service as a measure separate from capacity decision and analyze the effects of these decisions on costumers' behavior.

The first part of research investigates the continuity of the customer experience across various stages. We provide a contract that can coordinate this system. We also analyze the scenarios when the outsourcer pools the demand from multiple clients and show that although pooling in some cases can mitigate the effects of non coordination, when the client's customers are different in attitude, it might be even suboptimal.

The second part analyzes a model that considers the effect of quality of service on repeat purchase, perceived value of quality, and referral. We show that because of the behavior of the customers towards waiting and quality of service, a natural link between the two decisions exist. Depending on the optimal quality level either the two decisions are substitutes or complements. We also analyze a duopoly scenario and investigate the optimal decision of two identical and non-identical firms.

The third part considers optimal brand equity as the measure of gain. We consider the effects of capacity and quality on the customers' value, repeat purchase, and brand choice. We compare this model with the existing models in the literature of marketing and service operations. We also investigate the effects of different marketing strategies on the decision levels and show that the existence of marketing strategy can result in lower quality levels.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express deepest gratitude to my advisor, Professor Yong-Pin Zhou, for the help, continuous support, and inspiration that he has provided throughout the years. His invaluable guidance has been the most essential reference in this dissertation. As a researcher, he advised me on how to do research and as a scholar he taught me how to pursue my questions with avid curiosity. His persistent encouragement, patience, and humor has rescued me from peril more times than I can recall. For all he has done, I am forever indebted.

I am honored to have had the help of a great committee, with the help of whom I have developed this dissertation: Professor Theodore Klastorin, Professor Hamed Mamani, and Professor Yong Tan. Their constructive comments have improved this dissertation tremendously. I appreciate all their help and support.

Many faculty members, colleges, and researchers' help has contributed to the development of this research. I sincerely thank my coauthor Haiyan Wang; Senthil Veeraraghavan, and the reviewers of Management Science for their comments that has improved the research presented in this dissertation.

My appreciation goes to all my friends and family. I immensely appreciate all the friends who are not named here but without whom this dissertation would have never been completed. I would like to offer special thanks to Hossein Ahmadi for his companionship, significant support, and his help in proofreading this document. I owe a great debt to my parents, Shahrzad and Hossein, and my siblings, Azadeh and Arash, whose support has always been with me. Their example has inspired me to seek great goals and their encouragement has given me the confidence and drive to walk towards them.

Chapter 1

INTRODUCTION

The impacts of service quality on the firm's profitability and success has been studied extensively. In 1987, the United States Congress stated: "Poor quality costs companies as much as 20 percent of sales revenues nationally". Malcolm Baldrige National Quality Improvement Act was passed to deal with this problem.([10]). In this dissertation we focus on the impact of service quality.

Although the importance of service quality has been shown by many studies, there is a lack of analytical models on service quality. This is due to the fact that service quality is multidimensional and hard to measure. The service instance could affect customers through many different mechanisms (such as meeting and/or exceeding customers' expectations, value, excellence,...). Although these dimensions are different, they capture different aspects of the quality of service and form a complex model of the customers' experience.

A model that comprehensively measures service quality is the SERVQUAL model ([61]), which measures five significant aspects of service quality: tangible, reliability, responsiveness, assurance, and empathy. The complexity of this model serves as a hinderance to "operationalizing" the concept of service quality and building analytical decision support models around it. It is also a major reason that early research into service quality in the marketing and operations literature has been empirical and qualitative in nature.

Although the SERVQUAL model has inspired a long stream of empirical research on service quality (see [19] for a review), it is not amenable to analytical modeling because there are too many variables, and they are hard to quantify.

A quantitative model of quality of service has been defined and used in location models. Location models consider quality of service as either access or responsiveness. These models consider the quality of service to have a negative relationship with the distance between the customer and the service center. (See[46], [24], [43], and [59]). These models consider the

quality of service to be a direct result of the location decision.

Following the literature, there are several ways through which service quality can affect a customer's behavior: *referral* (e.g. [78], [65]), *loyalty* (e.g., [65], [8], [27]), *perceived value of service* (e.g. [13], [90], [78], [14]), *spending amount* (e.g., [78]), purchase frequency (e.g., [64], [78]), *conversion to sales* (e.g., [68], [3], [45], [52], [84], [25]), and *service frequency* (e.g. [23], [56]).[1]

Recently quality centered models have emerged in the field of service operations. The effect of the quality of the service on the customer behavior and hence the dynamics of the system, is a well known phenomenon. However there is a lack of analytical literature that investigates and analyzes the operational decisions under consideration of the quality of service. The purpose of this dissertation is to, create analytical models that could capture the effects of the quality on the consumer behavior and provides insights on the optimal decisions.

In the following chapters of this dissertation we try to address different problems centering around the quality and capacity decisions of a service provider. In each chapter we create an abstract model to be be used in the analysis of different scenarios. The rest of this dissertation is as follows. In Chapter 2 we discuss a problem of multistage service and analyze the quality and capacity decisions while considering the customer experience. In Chapter 3 we provide a general model of quality and capacity that considers multiple effects of quality of service on the customer behavior and explain the relationship between the quality and capacity decisions under different circumstances. Chapter 4 deals with the effects of quality and capacity decisions on the customer acquisition and retention processes and investigates the effects of the marketing strategy on the optimal quality and capacity decisions. At the end we provide an appendix with the proofs of all the statements in the paper.

---

[1]There is a large literature on each of these topics and we provide some references here to illustrate the point; they are not meant to be comprehensive.

Chapter 2

ANALYSIS OF A TWO-STAGE SERVICE PROCESS: COORDINATION OF STAFFING AND EFFORT

2.1 Introduction

Services are complex, multi-step processes. A visit to an optometrist's office often ends in the adjacent eyewear store, and a service call to the credit card company may be redirected to a sales center once the service is completed. A key feature of such processes is that, while the various tasks are separate and performed by different agents, their outcomes are intricately linked. For such multi-stage service processes, if incentives are not properly designed, firms make suboptimal decisions. In a call center, it is not uncommon for the sales division to complain about the poor service that the customers received in prior stage(s) which makes the sales task harder. This problem has become more commonplace due to the greater emphasis placed on cross- and up-selling by service firms.

Similar incentive misalignment can occur between firms too. A fast-growing outsourcing trend in recent years has been piecemeal outsourcing [79], where a client firm outsources part, but not all, of the service process. IT companies, financial institutions, and healthcare providers are classic examples. When service outsourcing deals fail, it's often due to poor service quality [70]. Customer dissatisfaction, attributed mostly to the incompetence of the customer service representative [55], can potentially lead to huge losses. Therefore, in piecemeal outsourcing , it is crucial to explicitly model and study the effort and service quality provided at all the stages. Since they are inter-linked, the first question we aim to answer is:

1. *How do the staffing and effort decisions made at one stage affect those at the other stages?*

We answer this question in a variety of settings, depending on whether each stage is kept in house or outsourced. Once we characterize how each stage makes its own decisions in

isolation, we recognize that these locally-optimal decisions are system-suboptimal because they do not account for the joint effect of efforts across stages. Hence, we ask whether we can achieve first best for the whole process through the use of contracts. We focus on a two-stage process.

2. *How can we achieve coordination when the second stage of a two-stage process is outsourced?*

We provide an affirmative answer by developing a contract (called QA) that coordinates the two-stage service process. In particular, we are able to identify the systemic cause of suboptimality that goes beyond the classical double marginalization, and propose a general mechanism, a "reverse commission" that the outsourcer pays the client for good first-stage service, as the solution. We view this identification and solution of the root cause to be important contributions of our chapter.

Finally, we note that in practice outsourcers sometimes pool demand from several clients to leverage investments in training, facility, and IT infrastructure. In these cases, not only are all the clients served by the same agent pool, they also receive the same effort level, even if individually, they prefer different effort levels from the outsourcer. Thus, pooling creates complicated interactions among the various clients' actions. We also want to answer the following question:

3. *Can coordination be achieved through contracting when the outsourcer pools the second-stage service across clients? How does outsourcer pooling affect all the players?*

We are able to show that the QA contract we developed for the simpler one-on-one outsourcing setting continues to coordinate even when the outsourcer pools second-stage services across two separate clients. Moreover, when these two clients are identical, pooling always increases profit in a centralized or a coordinated process, and this pooling benefit is naturally shared by all the parties in the process. We can also show that pooling helps to mitigate profit suboptimality in a non-coordinated process. With two non-identical clients, however, we show by example that pooling may lead to profit loss for the whole process.

Therefore, it is not always beneficial for the outsourcer to pool. We provide suggestions on when not to pool.

2.2 Literature Review

To the best of our knowledge, our chapter is the first to model the joint effect of efforts across a multi-stage service process, and propose a proper contract in the outsourcing setting.

The make-or-buy (i.e., whether to outsource) decision has been well studied in both production and service. However, most of the existing service outsourcing literature focuses on the staffing decision and does not consider effort [91]. For example, [29] study call routing and capacity planning in call center outsourcing. [2] study how to divide work among in-house and outsourcer call centers. [4] examine contracts between an outsourcer and its clients who compete on waiting time and price. [37] analyze different contracts while there is information asymmetry about the outsourcer. Only [70] and [69] include service quality as a decision variable independent from the staffing decision, but neither considers the joint effect of efforts across multiple stages of a service process. In the economics and supply chain management literature, there are many chapters that model effort (see [16] for a review), but they focus on the effect of effort on production output or demand, not on service quality.

The extensive literature on service quality has traditionally been empirical or qualitative; only recently have there been analytical models of service quality: [21] and [34] model service quality as service accessibility or availability which is the outcome of the capacity decision; [39] model service quality based on the probability of customer satisfaction; [22] and [70] model service quality as the probability of each service request being resolved. [6] and [50] study the tradeoff between speed and quality. Again, none of these models consider the joint effect of efforts across multiple stages of a service process.

[53] and [47] are the only chapters that model piecemeal outsourcing. [53] consider a two-stage service process where the gatekeeper at the first stage has the option to either handle a request herself or refer it to the second stage. Either or both stages can be outsourced, but the efforts made at the two stages have no joint effect on service quality. [47] address the technical routing issues in a call center where some calls need a second stage service,

but they do not model service quality or investigate contracting issues.

Our chapter models the joint effect of efforts by both service stages, and, as such, is related to the literature on contracting for collaborative services. For research in economics we refer the reader to [41], [12], [48], and [18]. In the supply chain management literature, [88] consider a consulting service where effort can be exerted either by the consultant or by the client. Similar to us, they show that without a proper contract the consultant over-invests in the effort whereas the client under-invests. However, they derive these results for a single-stage process. [72] model a two-stage process where both stages can exert effort and they have a joint effect on the production output. Finally, [11] consider two firms that exert joint effort and incur joint cost to develop a new product. They investigate whether it is more beneficial to collaborate on effort or on cost. All of these chapters differ from our model in that they model the impact of (joint) effort on production output or sales; there is no consideration of quality.

There exists a separate line of literature on contracting for quality. For example, [9] model a production system in which it is costly for the supplier to identify and reduce defective products. The researchers design contracts for various scenarios where the decisions could be contractible or non-contractible. Their process is one-stage, however, and the buyer and supplier do not work on the same product. [75] analyzes a two-stage assembly process where efforts can be exerted at both stages. Moreover, the first stage can exert extra effort to audit the production quality of the second stage. By using a fixed payment contract and a warranty contract, they examine the buy-or-make decision and show that the first stage will perform more audit when the second stage is outsourced than when it is kept in house.

Finally, the outsourcer can often pool work across clients to achieve operating efficiency. It has been widely known in the queueing literature that pooling can help to reduce capacity requirement and customer wait (e.g. [49] and [87]). We add to this literature by showing that outsourcer pooling can also help to reduce system suboptimality in a non-coordinated process.

2.3 Model Setup

We consider a two-stage service process. Each stage is operated by a different agent group, and can represent either a customer service or a sales step. Customers are homogeneous and they go through the two stages sequentially. For example, in a cross-/up-selling setting, customers come to the first stage for customer service, and once that is completed, they are transferred to sales agents in the second stage. For most of our analysis, we assume that all the customers go through both stages (i.e. there is no customer abandonment between stages). In an extension in Section 2.5, we show that, even with customer abandonment, our contract can be enhanced to coordinate the process. Although the basic QA contract no longer coordinates in such a case, we provide numerical evidence in Section 2.7.3 that its performance is very close to the system optimum.

We index the two stages by $i \in \{1, 2\}$ and allow each stage to make its own capacity and effort decisions. The capacity and effort levels at stage i, s_i and e_i, affect waiting time and service quality, respectively, at that stage. Capacity represents staffing level and effort represents investment in quality such as customer service training, IT infrastructure and equipment, and auxiliary support.

We discuss the dynamics of these decisions in more details below.

Capacity Decision Customer arrivals follow a Poisson process with rate λ, their service time at stage i follows *i.i.d.* exponential distribution with rate μ_i, and there is no balking or reneging. Because birth-death Markov processes are time-reversible, the two stages can be modeled as independent $M/M/s$ queueing systems.

Since the close-form solution to $M/M/s$ queue is too complex to be used in analytical models, researchers have developed and tested various approximations [85]. Based on the heavy traffic approximation in the Halfin-Whitt regime [33], [15] show that the optimal staffing level, one that balances waiting and staffing costs, follows a square-root rule: $s = R + y\sqrt{R}$, where $R = \frac{\lambda}{\mu}$ is the offered load and y is the normalized *safety capacity*. The optimal y is a function of both staffing and waiting costs. Due to the additional effort decisions and costs, our model is more complicated, but we still can show that the square-

root staffing rule holds. As in [15], we will approximate the average waiting time in queue by

$$W(y) = \frac{\alpha(y)}{\sqrt{\lambda\mu}}, \tag{2.1}$$

where $\alpha(y) = [y + y^2 \frac{\Phi(y)}{\phi(y)}]^{-1}$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and PDF of standard normal distribution. The same approximation has been used in both [53] and [36].

Effort Decision Effort at each stage can affect service quality at that stage *and* subsequent stages. For example, in a cross-selling process, if the first service stage has not really resolved the customer's problem, it is harder for the second stage agent to make a sale. In our model, we let the first stage effort e_1 affect customer satisfaction not only at the first stage but also at the second stage (jointly with e_2). To the best of our knowledge, our chapter is the first to model this joint effort on customer satisfaction.

Specifically, for a given effort level e_1, each service instance at the first stage will be successfully completed with a random probability $p(e_1)$ whose CDF is $F_1(p|e_1)$. Then the overall expected service quality at the first stage is $\overline{p}(e_1) = \int_0^1 [1 - F_1(p|e_1)]\mathrm{d}p$. Similarly, for given e_1 and e_2, each second stage service will be successfully completed with a random probability $q(e_1, e_2)$ with a CDF $F_2(q|e_1, e_2)$. Then the overall expected second-stage service quality is $\overline{q}(e_1, e_2) = \int_0^1 [1-F_2(q|e_1, e_2)]\mathrm{d}q$. Higher effort should lead to higher service quality, and the marginal return should be diminishing. Thus, we assume

$$\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} \geq 0,\ \frac{\mathrm{d}^2\overline{p}(e_1)}{\mathrm{d}e_1^2} < 0,\ \frac{\partial\overline{q}(e_1, e_2)}{\partial e_1} \geq 0,\ \frac{\partial^2\overline{q}(e_1, e_2)}{\partial e_1^2} \leq 0,\ \frac{\partial\overline{q}(e_1, e_2)}{\partial e_2} \geq 0,\ \frac{\partial^2\overline{q}(e_1, e_2)}{\partial e_2^2} < 0. \tag{2.2}$$

Note that effort at the first stage has a positive[1] effect on the second stage service quality. While it is possible that higher first stage effort sets a higher customer expectation, making her *less* likely to be satisfied at the second stage, this is not the case we study in this chapter. Our assumption of positive effect is more consistent with the cross-selling example that motivated our study.

[1]Unless otherwise noted, we use "positive", "negative", "increasing", and "decreasing" in a non-strict sense.

We also assume that second-stage service quality is submodular[2] in the efforts:

$$\frac{\partial^2 \overline{q}(e_1, e_2)}{\partial e_2 \partial e_1} \leq 0. \tag{2.3}$$

The negative cross-derivative combined with (2.2) ensures that the positive effect of both efforts on the second stage satisfaction has a diminishing return. Finally, we assume the following relations among second order derivatives:

$$\frac{\partial^2 \overline{q}(e_1, e_2)}{\partial e_1 \partial e_2} \geq \frac{\partial^2 \overline{q}(e_1, e_2)}{\partial e_1^2} + \frac{r_1}{r_2} \frac{\partial^2 \overline{p}(e_1)}{\partial e_1^2} \qquad \text{and} \qquad \frac{\partial^2 \overline{q}(e_1, e_2)}{\partial e_1 \partial e_2} \geq \frac{\partial^2 \overline{q}(e_1, e_2)}{\partial e_2^2}. \tag{2.4}$$

The conditions in (2.2) and (2.3) are quite natural and general. Conditions in (2.4), on the other hand, are assumed mainly for analytical tractability – they ensure concavity behavior by the profit function. In particular, (2.4) ensures that the change in the marginal system revenue of e_i is higher due to the change in e_i than $e_j, (j \neq i)$. Although (2.4) is not as intuitive as (2.2) and (2.3), it is not restrictive either. The set of functions for $\overline{p}$ and $\overline{q}$ that we use in the numerical analysis satisfy these conditions.

Costs and Profits We can categorize the various costs as either investment cost or customer cost. Investment in staffing and effort can reduce customer waiting and increase service quality but they are costly. We assume linear staffing cost at stage i $(i = 1, 2)$, $c_{si}s_i$, where c_{si} is the cost rate per agent. Effort cost at stage i, on the other hand, should not be determined by e_i alone; it must also depend on the size of the system s_i. This is because effort costs such as training, equipment, and support often depend on the scale of deployment. Therefore, we let stage-i effort cost be $c_{ei} \cdot e_i \cdot s_i$, where c_{ei} $(i = 1, 2)$ is the unit effort cost per person. The scaling by staffing level is a more realistic way to model effort cost and we view it as another modeling contribution of our chapter. While this more complicated effort cost poses additional technical challenges, it also creates an interesting link between staffing and effort decisions in an intricate way. For example, when making staffing decision, one must consider its impact on both the staffing cost and the effort cost.

Customer experience cost is the sum of waiting cost and (lack of) service quality cost. First, we assume a linear waiting cost at each stage i $(i = 1, 2)$, $\lambda c_{wi} W_i(s_i)$, where $W_i(s_i)$

[2] [71] also considers the case where the joint effect is supermodular.

is the average wait in system and c_{wi} is the waiting cost rate. Next, we assume there is a financial loss to the system whenever a service is not resolved. Specifically, we assume that the system collects a revenue of r_i at stage i only from each *resolved* service. In other words, the system incurs a loss of r_i for each un-resolved customer. This loss may represent lost purchase in the short run, and customer ill-will and bad word-of-mouth in the long run [39]. For given e_1 and e_2, the (lack of) quality costs over all the customers are $r_1[1-\overline{p}(e_1)]\lambda$ and $r_2[1-\overline{q}(e_1,e_2)]\lambda$ at the two stages respectively, where service quality measures $\overline{p}$ and $\overline{q}$ can be estimated by surveys. Many service providers regularly survey the customers for quality of the service by measuring resolution, satisfaction, and the future purchase intent after each service instance. Firms such as Amazon and Sony specifically target resolution in their customer service follow-up surveys. There are also firms that provide quality management tools for call centers, and they can provide objective 3rd-party measurements of service quality at both stages.

2.4 The Joint Effect of Effort

In a decentralized process, each stage performs local optimization so the joint effect of e_1 and e_2 on customer experience is often overlooked. In this section we measure the significance of modeling this joint effect by comparing a centralized process with a decentralized one. We focus on services that have relatively large size (e.g. call centers for customer help, order taking, cross-selling, or marketing). This allows us to apply the square-root staffing rule in heavy traffic limit.

In a decentralized service process, the two stages maximize their own profits independently:

$$\text{Stage 1:} \qquad \max_{s_1,e_1} \Pi_1(s_1,e_1) = r_1\overline{p}(e_1)\lambda - c_{s1}s_1 - c_{e1}e_1s_1 - \lambda c_{w1}W_1 \tag{2.5}$$

$$\text{Stage 2:} \qquad \max_{s_2,e_2} \Pi_2(e_1,s_2,e_2) = r_2\overline{q}(e_1,e_2)\lambda - c_{s2}s_2 - c_{e2}e_2s_2 - \lambda c_{w2}W_2. \tag{2.6}$$

It is important to note that the first stage bears all the costs of investing in e_1, but reaps only part of the benefit, $r_1\overline{p}(e_1)\lambda$; the other part of the benefit, $r_2\overline{q}(e_1,e_2)\lambda$, is instead reaped by the second stage. Compare this with a centralized process where both stages are managed by a single firm and the goal is to maximize profit from the whole process:

$$\max_{s_1,e_1,s_2,e_2} \Pi(s_1,e_1,s_2,e_2) = \Pi_1(s_1,e_1) + \Pi_2(s_2,e_2,e_2). \tag{2.7}$$

In the analysis below we use superscripts ^{DC} and ^{C} for the decentralized and centralized processes when necessary. Substituting the Half-Whitt $M/M/s$ approximation [33, Equation 5.1] in (2.5)-(2.6) and taking first order derivatives, we find

$$s_1^{DC} = R_1 + \sqrt{\frac{c_{w1}\mathrm{P}_{w1}}{(c_{s1}+c_{e1}e_1)}}\sqrt{R_1} \qquad \text{and} \qquad s_2^{DC} = R_2 + \sqrt{\frac{c_w\mathrm{P}_{w2}}{(c_s+c_{e2}e_2)}}\sqrt{R_2}, \tag{2.8}$$

where P_{wi} is the asymptotic delay probability at stage i. These two equations show that the square-root staffing rule in [15] continues to apply, even with the additional effort variables in our model. We simply need to replace c_s in their expression by $c_s + c_e e_1$ to include all the staffing-related costs. We will follow the approach in [15] by using the square-root staffing rule and the wait time approximation (2.1). We refer to this as the BMR regime.

After a change of variable $s_i = R_i + y_i\sqrt{R_i}$, the two stages now maximize the following profits:

$$\begin{aligned}\max_{y_1,e_1} \Pi_1(y_1,e_1) &= r_1\overline{p}(e_1)\lambda - c_{s1}(R_1 + y_1\sqrt{R_1}) - c_{e1}e_1(R_1 + y_1\sqrt{R_1}) \\ &- \sqrt{R_1}c_{w1}\alpha(y_1), \end{aligned}\tag{2.9}$$

$$\begin{aligned}\max_{y_2,e_2} \Pi_2(e_1,y_2,e_2) &= r_2\overline{q}(e_1,e_2)\lambda - c_{s2}(R_2 + y_2\sqrt{R_2}) - c_{e2}e_2(R_2 + y_2\sqrt{R_2}) \\ &- \sqrt{R_2}c_{w2}\alpha(y_2). \end{aligned}\tag{2.10}$$

Since $\alpha(y_i)$ is strictly convex [15], we denote the inverse function of $\frac{\mathrm{d}\alpha(y_i)}{\mathrm{d}y_i}$ by β. Proof of the following lemma, along with all the other proofs, can be found in the appendix.

Lemma 1. *In the BMR regime, for large λ there exist unique (y_1^{DC}, e_1^{DC}) and (y_2^{DC}, e_2^{DC}) that maximize (2.9) and (2.10) respectively. Moreover, they solve the following system of equations:*

$$y_1 = h_1^{DC}(e_1) \qquad \textit{and} \qquad y_1 = h_2^{DC}(e_1), \tag{2.11}$$

$$y_2 = h_3^{DC}(e_2) \qquad \textit{and} \qquad y_2 = h_4^{DC}(e_2), \tag{2.12}$$

where $h_1^{DC}(e_1) = \frac{r_1\sqrt{\lambda\mu_1}}{c_{e1}}\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} - \sqrt{R_1}$, $h_2^{DC}(e_1) = \beta(-\frac{c_{s1}+c_{e1}e_1}{c_{w1}})$, $h_3^{DC}(e_2) = \frac{r_2\sqrt{\lambda\mu_2}}{c_{e2}}\frac{\partial\overline{q}(e_1,e_2)}{\partial e_2} - \sqrt{R_2}$, and $h_4^{DC}(e_2) = \beta(-\frac{c_{s2}+c_{e2}e_2}{c_{w2}})$.

In the proof we show that the Hessian matrices are strictly negative definite in the BMR regime. Thus, (2.9) and (2.10) are concave for sufficiently large λ. In the numerical tests in Section 2.7 we observe that the concavity holds for a wide range of λ.

The next lemma deals with the centralized process where (4.10) is optimized.

Lemma 2. *In the BMR regime, for large λ there exist a unique $(y_1^C, e_1^C, y_2^C, e_2^C)$ that maximizes (4.10). Moreover, they solve the following system of equations:*

$$\begin{aligned} y_1 = h_1^{DC}(e_1) + \frac{r_2\sqrt{\lambda\mu_1}}{c_{e1}}\frac{\partial\overline{q}(e_1,e_2)}{\partial e_1} \qquad & \textit{and} \qquad y_1 = h_2^{DC}(e_1), \\ y_2 = h_3^{DC}(e_2) \qquad & \textit{and} \qquad y_2 = h_4^{DC}(e_2), \end{aligned} \tag{2.13}$$

Equations in (2.13) are identical to those in (2.11)-(2.12) with the exception of the first one: in the decentralized process, e_1 is optimized according to its impact on the first stage only, whereas in the centralized process, e_1's impact on the second stage is also considered. The extra term in (2.13), $\frac{\lambda r_2}{c_{e1}}\frac{\partial\overline{q}(e_1,e_2)}{\partial e_1}$, explicitly and properly accounts for the extra benefit of the first stage effort on second stage quality. Since this term is positive, ignoring it in (2.11) causes the decentralized first stage to under-invest in effort. This idea is formalized in the following theorem:

Theorem 1. *In the BMR regime, for large λ, $e_1^{DC} \le e_1^C$ and $e_2^{DC} \ge e_2^C$. Furthermore, $s_1^{DC} \ge s_1^C$ and $s_2^{DC} \le s_2^C$.*

In the decentralized setting, the first stage under-invests in effort (i.e., $e_1^{DC} \le e_1^C$) because of incentive misalignment. This has a negative impact on the second stage: when customers arrive in a foul mood, the second stage must work extra hard (i.e., $e_2^{DC} \ge e_2^C$) to compensate. As a result, the second stage over-invests in effort. Consequently, because effort cost has the form $c_{ei}e_is_i$, lower (higher) effort level means lower (higher) marginal cost of staffing, so the corresponding staffing level at each stage moves in the opposite direction of the effort level.

As we observed in the introduction, the failure to coordinate across stages is often observed in practice. The academic literature, however, is surprisingly lacking in models that deal with such multi-stage service processes (see [91]). In order to coordinate the decentralized process, a formal mechanism must be established for the first stage to be properly compensated for its effort. We develop a contract to achieve that in the outsourcing setting next.

2.5 Outsourcing the Second Stage

In this section we investigate a piecemeal outsourcing arrangement where the client keeps the first stage in house and outsources the second stage. The closest chapter to ours is [70] which, while not modeling the joint effect of efforts, also studies the staffing and effort decisions in a two-stage service process and develops coordinating contracts. Therefore, we will use the contracts developed in [70] as benchmark for comparison.

In call center outsourcing contracts, pay-per-time, per-per-call, pay-per-agent, pay-per-solve are popular features, but they do not address the double marginalization phenomenon: the profit margin on each served customer is different for the outsourcer than it is for the whole system, which results in the outsourcer taking system sub-optimal actions. To remedy this, [70] design two contract forms.

The first contract is built on the pay-per-solve payment scheme (denoted as PPCR+CS) where the client pays the outsourcer a fixed b for each resolved customer. In addition, the client shares a $(1-\phi)$ portion of the outsourcer's staffing and effort costs. In return, the outsourcer pays a ϕ portion of the customer waiting and service quality costs (the c_g term below). This cost sharing idea is a mirror image of revenue sharing contract used to coordinate inventory supply chains [17]. Under this contract, the client pays the outsourcer the following amount[3]:

$$T = b\overline{q}\lambda + (1-\phi)(c_{s2}s_2 + c_{e2}e_2s_2) - \phi\left\{c_g(1-\overline{q})\lambda + \lambda c_{w2}W_2\right\}. \tag{2.14}$$

The second contract is a partnership (denoted by PART) in which the outsourcer "owns" all the revenue from resolved customers and cost from unresolved customers (the first bracket below), and pays the client a "user fee" (the second bracket) which is a $(1-\phi)$ portion of the first bracket value if the second stage outsourcer provides the centralized optimal service quality. There is an additional cost-sharing term (the third bracket) where the client shares a portion of the outsourcer's staffing, waiting, and effort costs[3]:

$$\begin{aligned} T &= \lambda\left[r_2\overline{q} - c_g(1-\overline{q})\right] - (1-\phi)\lambda\left[r_2\overline{q}^C - (1-\overline{q}^C)c_g\right] \\ &- \left[(1-\phi)c_{s2}s_2 + (1-\phi)c_{e2}s_2e_2^C + \phi\lambda c_{w2}W_2\right]. \end{aligned} \tag{2.15}$$

[3]The original payment expression in [70] is based on the fluid approximation. The equations here have been properly adjusted for the queueing dynamics that we model explicitly (as in [37]).

Both PPCR+CS and PART can coordinate a process without the joint effect of efforts [70], but we can show that they fail to coordinate when the joint effect is present.

Proposition 1. *The PPCR+CS and PART contracts,* (2.14) *and* (2.15)*, cannot coordinate the two-stage outsourcing model represented by* (2.9) *and* (2.10)*.*

[17] make a similar observation that the revenue sharing contract fails to coordinate when demand depends on the costly effort of the client. Therefore, any coordinating contract must account for the effect of the client's effort on the outsourcer. One way is for the outsourcer to pay for the benefit it receives from the client's investment in effort. To be implementable, this payment must be based on objective measurements that can be agreed to by both the client and the outsourcer. To that end, we introduce a concept we call *reverse commission*. It is a simple idea where the outsourcer pays the client a fee for each resolved customer at the first stage (i.e. pay for each "good lead"). In aggregate, this is equivalent to the outsourcer paying the client for the first stage service quality $\overline{p}$. In practice, since $\overline{p}$ is already widely tracked, we can assume it can be objectively monitored and audited.

In revenue sharing type of contracts the client is already paying a commission to the outsourcer. The new fee we propose goes in the reverse direction; hence the name reverse commission. While we have not seen this type of payment in the service outsourcing literature, commission fees based on high-quality referrals have been used in online advertising (pay-per-click can be viewed as paying for all leads, and pay-per-action can be viewed as paying only for high-quality leads), and by some online travel magazines who take a cut if readers buy the travel packages they write about [58]. The following theorem shows that appropriately selected reverse commission can help to coordinate the service process in our model.

Proposition 2. *There exists a $c_p > 0$ such that both PPCR+CS and PART can coordinate the process if the outsourcer pays an additional reversed commission of $c_p\overline{p}$ to the client.*

The adjusted PPRC+CS and PART contracts are quite complicated and they require the outsourcer's effort and staffing levels to be observable, which is not always the case in practice. Below we propose a simpler contract that contains the same reverse commission

term $c_p\overline{p}$ but does not require the outsourcer's decisions to be observable. It's based on the outcomes of these decisions – waiting time and quality – instead. We call it the *Quality Adjusted (QA)* contract.

Definition 1. *Under the Quality Adjusted (QA) contract, the client makes the following payment per customer to the outsourcer:*

$$T(\overline{p}, \overline{q}, W_2) = b\overline{q} - c_q(1-\overline{q}) - c_{w2}W_2 - c_p\overline{p}. \tag{2.16}$$

The first two terms in the QA contract compensate the outsourcer for resolved customers while at the same time penalize it for unresolved customers, and the third term transfers the customer waiting cost to the client. Together, these three terms are designed to incentivize the outsourcer to invest in the right staffing and effort levels. The last term, the reverse commission, is designed to incentivize the *client* to invest in the right effort level. The theorem below gives the parameter values under which the QA contract coordinates the process and splits the profit.

Theorem 2. *Let* $c_p = r_2 \frac{\partial \overline{q}(e_1,e_2)/\partial e_1}{\partial \overline{p}(e_1)/\partial e_1}\Big|_{(e_1^C, e_2^C)}$. *For any* $b + c_q = r_2$, *the QA contract coordinates the service process. Moreover, the profit split between the client and the outsourcer can be adjusted by varying* c_q *(or, equivalently,* b*).*

The implementation of the QA contract requires W_2, $\overline{p}$, and $\overline{q}$ to be common knowledge. This is achievable in a call center setting. The call distributor already records customer waiting time and it can be monitored or audited by a third party. We can thus assume W_2 can be made available to both parties. Almost all call centers and many service firms constantly monitor their own service quality level by using follow-up emails or calls, surveys, and mystery shoppers. The QA contract only requires both parties to agree to have $\overline{p}$ and $\overline{q}$ tracked and monitored by a trusted third party.

The QA contract also requires the cost and revenue parameters to be common knowledge. We assume the outsourcer to operate in a competitive and stable market so the client can estimate the outsourcer's costs. We leave the analysis of asymmetric cost information to future research and refer interested readers to [37] and [69].

Extension: Customer Abandonment Between Stages

Service quality can affect customers' behavior in a number of ways. For example, in a multi-stage process, poor service at one stage not only makes it harder for later stages to satisfy the customer, it can also cause dissatisfied customers to leave between stages (i.e. customer abandonment), reducing the arrival rate to later stages. So far in this chapter, we have not modeled this effect, choosing instead to focus on the impact of service quality on service resolution and customer purchases. In the remainder of this section, we will extend the model by including customer abandonment.

When customer abandonment is a fixed, exogenous proportion (say, $1-t$), we can show that, by replacing the second-stage arrival rate by $t\lambda$, all the results for the no abandonment model continue to hold. In reality, however, customer abandonment rate is often endogenous, and should be modeled as a function of e_1. Once we do that, the analysis quickly becomes complex. We can no longer guarantee the concavity of the objective functions so the use of first order conditions becomes more nuanced. Nevertheless, we will show below that, with some mild, reasonable assumptions to avoid trivial solutions of $e_i = 0$, a revised QA contract continues to coordinate the process.

Let arrival rate to the second stage be $t(e_1)\lambda$. We assume $t(0) = 0$, $t^{'}(e_1) > 0$, and $t^{"}(e_1) < 0$ (i.e. $t(e_1)$ is an increasing function with diminishing return). The objective functions become:

$$\begin{aligned} \Pi_1(y_1, e_1) &= r_1\overline{p}(e_1)\lambda - c_{s1}(R_1 + y_1\sqrt{R_1}) - c_{e1}e_1(R_1 + y_1\sqrt{R_1}) \\ &- \sqrt{R_1}c_{w1}\alpha(y_1), \qquad (2.17) \\ \Pi_2(e_1, y_2, e_2) &= r_2\overline{q}(e_1, e_2)\lambda t(e_1) - (c_{s2} + c_{e2}e_2)\left[R_2 t(e_1) + y_2\sqrt{R_2 t(e_1)}\right] \\ &- \sqrt{R_2 t(e_1)}c_{w2}\alpha(y_2). \qquad (2.18) \end{aligned}$$

Since all the terms in Π_2 depend on e_1, we cannot prove concavity, so the previous proofs do not directly extend. Moreover, even though we can show numerically (in Sectoin 2.7.3) that the QA contract's performance is very close to the system optimal, it no longer coordinates. Fortunately, we are able to show that when additional payment terms based on e_1 are added to the QA contract, the resulting contract, which we call the *Quality Adjusted with Abandonment (QA-A)* contract, can achieve coordination.

Definition 2. *Under the QA-A contract, the client makes the following payment per cus-*

tomer to the outsourcer:

$$T(\overline{p},\overline{q},W_2,\overline{t}) = b\overline{q} - c_q(1-\overline{q}) - c_w W_2 - c_p\overline{p} - c_1\overline{t} - c_2\ln(\overline{t}). \tag{2.19}$$

The QA-A contract keeps all the QA terms. In addition, it introduces two new terms $c_1\overline{t}$ and $c_2\ln(\overline{t})$ to incentivize the client to exert extra effort to reduce customer abandonment. We specify these two terms in (2.19) as functions of $\overline{t}$ for ease of exposition and understanding. These two terms could just as well be expressed in terms of $\overline{p}$ because both $\overline{p}(e_1)$ and $t(e_1)$ are known functions. Thus they require no extra performance metrics to be monitored and tracked.

Theorem 3. *In the existence of quality dependent abandonment*

a) There exist a finite $\overline{e_1}$ *such that* $r_1\lambda\frac{\mathrm{d}\overline{p}(\overline{e_1})}{\mathrm{d}e_1} + r_2\lambda t(\overline{e_1})\frac{\partial\overline{q}(\overline{e_1},0)}{\partial e_1} + r_2\lambda\frac{\mathrm{d}t(\overline{e_1})}{\mathrm{d}e_1} = c_{e1}R_1$.

b) Assume $\frac{\mathrm{d}\overline{q}(\overline{e_1},0)}{\mathrm{d}e_2} > \frac{c_{e2}}{r_2\mu_2}$. *For* $c_p = r_2 t(e_1^C)\frac{\partial\overline{q}(e_1,e_2)/\partial e_1}{\partial\overline{p}(e_1)/\partial e_1}\Big|_{(e_1^C,e_2^C)}$, $c_1 = \frac{r_2}{2}q(e_1^C,e_2^C) - \frac{c_{s2}+c_{e2}e_2}{2\mu_2}$, $c_2 = \frac{r_2 t(e_1^C)}{2}\overline{q}(e_1^C,e_2^C) - \frac{(c_{s2}+c_{e2}e_2)t(e_1^C)}{2\mu_2} - \frac{(c_{s2}+c_{e2}e_2^C)\sqrt{t(e_1^c)}}{2\sqrt{\lambda\mu_2}}y_2^C - \frac{c_{w2}\sqrt{t(e_1^c)}}{2\sqrt{\lambda\mu_2}}\alpha(y_2^C)$, $c_w = t(e_1^C)c_{w2}$, *and any* $b + c_q = t(e_1^C)r_2$, *the QA-A contract achieves system coordination. Moreover, the profit split between the client and the outsourcer can be adjusted by varying* c_q *(or equivalently* b*).*

The first part of the theorem is used to make the assumption $\frac{\mathrm{d}\overline{q}(\overline{e_1},0)}{\mathrm{d}e_2} > \frac{c_{e2}}{r_2\mu_2}$, which is then used to ensure a reasonable condition that $e_2 = 0$ cannot be optimal. The second part of the theorem gives a particular set of parameters under which the QA-A contract coordinates. The QA-A contract is given only as an example to show that as long as the two parties properly account for $\overline{p}$, the process can be coordinated. There may exist other intuitive, easy-to-use coordinating contracts. We leave the search for better contracts to future research.

Our main insight from Theorems 2 and 3 is that reverse commission is a useful and robust concept: A linear reverse commission can help to coordinate the basic two-stage process, and additional reverse commission terms help to coordinate the system with quality-dependent customer abandonment. We believe that it has the potential to be useful in other situations as well.

2.6 Pooling the Second Stage

So far we have limited our attention to a single two-stage service process where the client-outsourcer relationship is one-on-one. In practice, however, the outsourcer often shares agents, training, IT systems, and facility across different clients (i.e. pooling). This is because the skills, knowledge, and infrastructure that the outsourcer develops for one client is often transferrable to other clients. Thus the outsourcer has a strong incentive to leverage such assets as much as possible, especially when such assets are expensive and the skills are not too firm-specific.

When the outsourcer pools the second stage operations across clients, the operational benefits of pooling (reduced waiting and staffing costs) should improve the overall profit for the whole process, but it is not clear whether all the parties will benefit, and whether the QA contract need to be adjusted. To investigate these questions, we compare the following four systems.

1. *Centralized Dedicated System*: there is no outsourcing or pooling; each client has centralized control over its whole process; this is the centralized system we studied in Section 2.4,

2. *Centralized Pooled System*: there is no outsourcing; a single firm has control over all the service stages of all the processes, and pools all the second stage operations,

3. *Outsourced Dedicated System*: there is outsourcing but no pooling; all clients use the same outsourcer but the outsourcer uses dedicated agents for each client; its relationship with each client is one-one-one (we also refer to the contracts in this setting as the *one-on-one* contracts), and

4. *Outsourced Pooled System*: both outsourcing and pooling are present; all clients use the same outsourcer and the outsourcer pools second stage operations across all the clients.

For simplicity, we will focus on the case of pooling two identical clients in this chapter; our results extend to more than two identical clients. The case of two non-identical clients

will be investigated in the numerical analysis. We use subscript $i = a, b$ to represent the first stage operation in the two processes, and continue to use subscript 2 for the second stage. Moreover, we will indicate the four systems with superscripts ^{CD}, ^{CP}, ^{OP}, and ^{OD} respectively.

Theorem 4. *In a Centralized Pooled system with two identical clients,*

a) Pooling the second stage increases the total system profit.

b) Pooling requires the second stage to exert more effort and reduce staffing. Correspondingly, the two first stages in the pooled system will exert less effort and increase staffing. That is, $e_2^{CP} > e_2^{CD}$, $s_2^{CP} < 2s_2^{CD}$, $e_i^{CP} < e_i^{CD}$, $s_i^{CP} > s_i^{CD}$.

The first part of the theorem is quite straightforward for a centralized process. Since pooling can better utilize the staff in the second stage, there need to be fewer of them. This reduces effort cost for the second stage (there are fewer people to train, supervise, and support) so it increases effort. These results are largely in line with the conventional rationale for consolidation – lower staffing and better training for the consolidated operations.

The second stage's higher effort level benefits the first stage through the joint effort function $\overline{q}$. The first stage responds by reducing its own effort level. Therefore, the effect of joint effort allows the benefit of pooling to be naturally distributed.

Remark 1. *Theorem 4 has an important implication. According to Theorem 1, the first stage in a non-coordinated process will under-invest in effort because it is not fully compensated for the benefits of such investment. Theorem 4 then states that, pooling reduces the optimal effort level at both first-stage clients. Therefore, even when the first stage under-invests in effort, its resultant optimality gap is smaller. In short, pooling helps to mitigate the suboptimality in this case. We further demonstrate this by a numerical example in Section 2.7.4.*

The next theorem shows that the same QA contract that coordinates the OD system also coordinates the OP system. Hence, using this QA contract, the outsourced processes behaves like the centralized processes. Pooling reduces outsourcer staffing cost and that

benefit is shared by the clients. Moreover, when the outsourcer pools, the clients' effort level requirement is lower. So when QA is not used and the process is not coordinated, pooling helps to mitigated profit suboptimality.

Theorem 5. *The QA contract also coordinates the Outsourced Pooled System for two identical clients. Furthermore,* $e_2^{OP} > e_2^{OD}$, $s_2^{OP} < 2s_2^{OD}$, $e_i^{OP} < e_i^{OD}$, $s_i^{OP} > s_i^{OD}$.

In practice some client firms demand dedicated service from the outsourcer. This is sensible when the client is dissimilar to other clients (e.g. different customer types, tasks), requires unique or undivided attention from the outsourcer, or when customer privacy/data confidentiality is a major concern. When none of this is an issue, however, our results suggest that the clients should not object to outsourcer pooling because 1) pooling reduces the operational cost by the outsourcer, and 2) with the use of the QA contract the clients get to share the pooling benefits as well. Moreover, even when coordinating contracts are not used, pooling by the outsourcer can reduce profit loss.

Next we investigate two non-identical clients.

Theorem 6. *For an Outsourced Pooled system with two non-identical clients,*

a) the QA contracts can coordinate the system;

b) the second stage's pooled effort is between the two dedicated effort levels required by the clients independently; that is, $\min\{e_{2a}^{OD}, e_{2b}^{OD}\} < e_2^{OP} < \max\{e_{2a}^{OD}, e_{2b}^{OD}\}$,

c) pooling will not necessarily increase the total system profit.

Theorem 6 demonstrates the QA contract's broad applicability, even to non-identical clients. However, pooling may not always be beneficial. On the one hand, it reduces staffing for the outsourcer. On the other hand, non-identical clients may require different staffing and effort levels by the outsourcer. In the OD system the outsourcer can use separate dedicated agent pools with required staffing and training for each client. In the OP system the outsourcer must use the same pool of agents to serve both clients, so both clients get the same staffing and effort levels. Part b) shows that this common effort level must be a

compromise (i.e. between the two required dedicated effort levels), and it results in over-serving of one client and under-serving of the other. The consequent loss of profit may be stronger than the staffing cost reduction and lead to lower system profits (part c). We will give a numerical example in Section 2.7.5 to illustrate this.

2.7 Numerical Analysis

In this section we use numerical tests to illustrate the various discoveries we have made in previous sections. All of our analytical results are based on general functions of $\overline{p}$ and $\overline{q}$. For the numerical tests in this section, however, we will use these specific functions:

$$\overline{p}(e_1) = 1 - e^{-\gamma e_1} \tag{2.20}$$

$$\overline{q}(e_1, e_2) = \overline{p}(e_1)\overline{p}(e_2 + \delta_r) + [1 - \overline{p}(e_1)]\,\overline{p}(e_2 - \delta_{nr}). \tag{2.21}$$

Function (3.22) has a simple explanation. If a customer's service problem is resolved in stage 1, she is in a better mood to be satisfied by the second-stage service. We model this effect by assuming her second-stage service success probability to be $\overline{p}(e_2 + \delta_r)$, where $\delta_r \geq 0$ represents the the positive effect of good first-stage service on the second stage. In contrast, when a customer's service is unresolved in the first stage, she is already in a bad mood to begin the second stage service. We assume her second-stage service success probability to be $\overline{p}(e_2 - \delta_{nr})$, where $\delta_{nr} \geq 0$ represents the the negative effect of bad first-stage service. The parameters δ_r and δ_{nr} allow us to model the strength of these effects: when $\delta_r = \delta_{nr} = 0$, there is no joint effort effect; and larger (smaller) δ_r and δ_{nr} indicate a stronger (weaker) effect of first-stage effort on the second stage service quality. For simplicity, we set $\delta_r = \delta_{nr} = \delta$ in all the tests except those in Section 2.7.5. Hence,

$$\overline{q}(e_1, e_2) = 1 - e^{-\gamma(e_2+\delta)} + e^{-\gamma(e_2+e_1+\delta)} - e^{-\gamma(e_1+e_2-\delta)}. \tag{2.22}$$

$\overline{p}$ as defined in (2.20) is fairly general and has the concave incresaing property that is required. We can also show that $\overline{q}$ as defined in (2.22) satisfies the conditions in (2.2)-(2.4).

2.7.1 Joint Effect of Efforts

In this section we set up a baseline case to study the joint effect of effort. The parameters are chosen so that either the parameters or the actions they induce conform to observed industry ranges. These parameters will be varied in the senstivity analysis in Section 2.7.2.

Without loss of generality, we let μ =1/min, c_s =\$1/min, and scale all the other parameters accordingly (see Table 2.1 for a summary of the parameters). Offered load of 100

Table 2.1: Baseline Parameters

λ	μ	r	c_s	c_w	c_e	γ	δ
100/min	1/min	\$2.05	\$1/min	\$1/min	\$0.2/server/effort	0.8/effort	1/effort

(λ =100/min) is reasonable for a medium-sized call center; it also allows us room to scale up and down later for sensitivity analysis. We pick γ and δ such that the optimal service quality is about 85-90% for both stages in the centralized system. Finally, all the cost and revenue parameters are chosen so that profit margin for the centralized system is about 10% and $80-90\%$ of the total cost are effort and staffing cost.

The first line in Table 2.2 reports the total profit, and effort and staffing levels at both stages for the centralized process. Corresponding % deviations for the decentralized process are given on the second line.

Table 2.2: Comparison of Centralized and Decentralized Settings

	Π	e_1	e_2	$\overline{p}$	$\overline{q}$	$y_1\sqrt{R_1}$	$y_2\sqrt{R_2}$
Centralized	34.66	2.9449	1.9386	0.9052	0.8690	7.0241	7.4168
Decentralized (%)	-4.04%	-13.63%	6.36%	-3.97%	0.01%	2.12%	-0.70%

As predicted by Theorem 1, the decentralized process under-invests in first stage effort, quite significantly in this case by 13.63%. This results in a first-stage service quality that is 3.97% lower than in the centralized process. The second stage, seeing its customers harder to satisfy, compensates by increasing its own effort. This offsets the lower first stage effort; their joint effect results in a little-changed second-stage quality. Also as predicted by Theorem 1, the staffing level changes in the opposite direction, but the percentage deviation is relatively mild in this case. Overall, by overlooking the joint effect of effort, the decentralized process loses 4.04% in profit, a significant amount considering profit margin is about 10%. The

customers also lose out because they receive 4% lower service quality from the first stage.

We can view Table 2.2 in the outsourcing context as well. When both parties overlook the joint effect of effort, the contract they adopt (e.g. PPCR+CS and PART) can only "coordinate" the process to the decentralized case. Therefore, by adopting QA contract, not only the entire process's profit increases by 4.04%, the customers also benefit by receiving higher service quality.

2.7.2 The Effect of Parameter Values

Results in the previous section are derived for one particular set of parameter values. Although they are reasonably and carefully chosen, it is still important to see if the observations are generally true and how they may change when the parameters vary. In this section, we perform a sensitivity analysis by varying all the parameters systematically. Because of the large number of parameters, we will vary one parameter at a time and fix all the other parameters at their baseline value.

Parameter ranges must be chosen carefully. Take r as an example. If it is too small, the profit will be negative. If it is too large, both stages have high profit margins and their optimal service quality will be very high, so the need for coordination becomes minimal. For this reason, we have chosen five values for each parameter, and they deviate from the baseline values by +10%, +5%, 0%, -5%, and -10% respectively.

Table 2.3 reports the profit difference between the centralized and the decentralized settings as each parameter changes. We observe that the average profit loss for a decentralized process ranges between 2.31% and 9.89%, and is consistently above 4.04% (the baseline value in Table 2.2). The overall picture is quite clear: the impact of joint effect is significant and should not be overlooked.

Table 2.3: Profit Loss for the Decentralized Process as a Function of the Parameters

	r_1	r_2	c_{w1}	c_{w2}	c_{s1}	c_{s2}	c_{e1}	c_{e2}	λ	μ_1	γ	δ
Mean	-5.02%	-4.71%	-4.03%	-4.03%	-4.24%	-4.24%	-4.10%	-4.12%	-4.05%	-4.88%	-4.32%	-4.07%
Max	-9.89%	-8.24%	-4.10%	-4.11%	-5.83%	-5.82%	-5.10%	-5.46%	-4.28%	-8.94%	-6.03%	-5.09%
Min	-2.31%	-2.66%	-3.97%	-3.96%	-3.09%	-3.09%	-3.25%	-2.94%	-3.84%	-2.69%	-3.22%	-3.13%

In addition to profit, Table 2.4 also includes how staffing and effort decisions change with each parameter. It is important to mention that although the magnitude of change varies for each parameter as it changes from value to value within its range, the general directions are consistent for both the centralized and the outsourcing scenarios. For the decentralized setting, when the parameters at the second stage are varied, the decisions at the first stage remain unchanged, but all the other value change with the same direction as shown in Table 2.4.

Although the increase in r_1 and r_2 both increase the total profit, they have opposite impact on the decisions at both stages. A higher r_1 increases the first stage effort level. It then leads to lower second stage effort through the joint effect function $\overline{q}$, and lower (higher) staffing at first (second) stage through the total effort cost function $c_e es$. The effect of r_2 is completely the opposite. We can observe similar dichotomy in other pairs of parameters for the two stages. Therefore it shows the importance of including the effect of these two functions in our model.

Table 2.4: Decision Varibales for the Decentralized Process as a Function of the Parameters

e_1	r_1	r_2	c_{w1}	c_{w2}	c_{s1}	c_{s2}	c_{e1}	c_{e2}	λ	μ_1	γ	δ
Lowest	-3.51%	0.01%	0.13%	-0.04%	-0.08%	0.03%	4.8%	-1.27%	-0.12%	-4.71%	5.17%	-1.79%
Lower	-1.71%	0.00%	0.06%	-0.02%	-0.04%	0.01%	2.34%	-0.63%	-0.06%	-2.29%	2.50%	-0.91%
Baseline	0%											
Higher	1.64%	0.00%	-0.06%	0.02%	0.04%	-0.01%	-2.23%	0.63%	0.05%	2.17%	-2.35%	0.93%
Highest	3.21%	-0.01%	-0.12%	0.03%	0.07%	-0.02%	-4.36%	1.25%	0.1%	4.24%	-4.56%	1.89%
e_2												
Lowest	1.5%	-6.83%	-0.05%	0.20%	0.03%	-0.15%	-1.92%	7.31%	-0.19%	2.04%	7.84%	2.48%
Lower	0.72%	-3.33%	-0.03%	0.10%	0.02%	-0.07%	-0.95%	3.56%	-0.09%	0.97%	3.80%	1.22%
Baseline	0%											
Higher	-0.67%	3.16%	0.02%	-0.09%	-0.02%	0.07%	0.95%	-3.40%	0.09%	-0.89%	-3.57%	-1.19%
Highest	-1.30%	6.18%	0.05%	-0.19%	-0.03%	0.13%	1.88%	-6.64%	0.17%	-1.70%	-6.92%	-2.34%
$y_1\sqrt{R}$												
Lowest	0.53%	0.00%	-4.23%	0.01%	2.67%	0.00%	0.87%	0.19%	-5.12%	6.16%	-0.77%	0.27%
Lower	0.26%	0.00%	-2.08%	0.00%	1.31%	0.00%	0.42%	0.10%	-2.52%	2.95%	-0.37%	0.14%
Baseline	0%											
Higher	-0.25%	0.00%	2.00%	0.00%	-1.26%	0.00%	-0.40%	-0.09%	2.46%	-2.73%	0.36%	-0.14%
Highest	-0.48%	0.00%	3.93%	-0.01%	-2.47%	0.00%	-0.77%	-0.19%	4.86%	-5.26%	0.69%	-0.28%
$y_2\sqrt{R}$												
Lowest	-0.17%	0.77%	0.01%	-4.16%	0.00%	3.02%	0.21%	0.38%	-5.11%	-0.23%	-0.86%	-0.27%
Lower	-0.08%	0.37%	0.00%	-2.04%	0.00%	1.47%	0.11%	0.18%	-2.52%	-0.11%	-0.42%	-0.14%
Baseline	0%											
Higher	0.07%	-0.35%	0.00%	1.96%	0.00%	-1.41%	-0.10%	-0.16%	2.46%	0.10%	0.40%	0.13%
Highest	0.14%	-0.68%	-0.01%	3.85%	0.00%	-2.76%	-0.21%	-0.30%	4.86%	0.19%	0.78%	0.26%

As we argued earlier, the difference between centralized and decentralized processes is identical to the difference in an outsourcing setting between using and not using the QA contract. Therefore, Tables 2.3 and 2.4 also apply to the outsourcing setting. In the QA contract there is an important new term called the reverse commission, denoted by c_p. Table 2.5 shows how the optimal c_p value changes from the baseline case as each parameter moves lower or higher from the baseline case.

Table 2.5: Change in c_p as Parameters Move Lower or Higher from The Baseline Case

	r_1	r_2	c_{w1}	c_{w2}	c_{s1}	c_{s2}	c_{e1}	c_{e2}	λ	μ_1	γ	δ
Lowest	-2.30%	11.17%	0.08%	-0.31%	-0.05%	0.23%	3.02%	-0.71%	0.30%	-3.12%	-9.42%	-15.07%
Lower	-1.12%	5.29%	0.04%	-0.15%	-0.03%	0.11%	1.49%	-5.37%	0.14%	-1.50%	-4.84%	-7.71%
Baseline	0%											
Higher	1.05%	-4.79%	-0.04%	0.15%	0.02%	-0.11%	-1.46%	5.41%	-0.13%	1.34%	5.10%	8.08%
Highest	2.04%	-9.14%	-0.08%	0.29%	0.05%	-0.21%	-2.88%	10.85%	-0.26%	2.67%	10.48%	16.54%

Since the investment in effort has a diminishing return, when first stage already exerts a higher effort, the incentive, in the form of the reverse commission, needs to be higher in order to induce the first stage to exert even higher effort. Conversely when e_1 is low, c_p can be lower. Therefore, the increase in any parameter that will reduce e_1 (e.g. c_{e1}, c_{w1} which increases s_1 and makes training more expensive, r_2 which increases e_2) will also necessitate a lower reverse commission c_p; the parameters that lead to higher e_1 will also need a higher c_p to coordinate the process. Although this general rule holds for all the parameters it is not observable in the case of changing λ and γ. Higher γ increases the effect of the effort exerted on the quality, hence slight increase in the effort exerted at the first stage increases the quality of the second stage much higher and this behavior dominates the diminishing effect of effort. Also when increasing λ,as lambda increases the total amount of reverse commission paid increases which will cause the c_p to decrease. These results are clearly reflected in Table 2.5. It's worth noting that changes in first stage parameters tend to result in a small change in c_p (at most 0.71%). Since c_p is paid by the outsourcer to the client, it means that the choice of c_p is robust with regard to the outsourcer's knowledge of the first stage parameters.

2.7.3 Customer Abandonment Between Stages

In this section, we consider a scenario in which customers abandon if they are dissatisfied with the first-stage service. That is, $t(e_1) = \overline{p}(e_1)$. In the centralized setting, the first stage will take abandonment into consideration and accordingly exert more effort. The QA-A contract in Theorem 3 coordinates the process and will incentivize the first stage to take the centralized optimal action. In the decentralized setting, however, the first stage does not care if customers abandon afterwards so its action does not account for that. Subsequently the second stage will see a lower arrival rate. Table 2.6 presents the numerical results for the baseline case.

As expected, in the centralized (or QA-A coordinated) setting, the first stage exerts even higher effort than in Table 2.2 because it also tries to reduce customer abandonment. In the decentralized case, however, the first stage acts exactly the same as in Table 2.2. Thus

Table 2.6: Comparison of Centralized and Decentralized and Outsourcing Settings in Existence of Abandonment

	Π	e_1	e_2	p	q	$y_1\sqrt{R_1}$	$y_2\sqrt{R_2}$
Centralized/QA-A	31.7637	3.0617	1.9037	0.9136	0.8686	6.9726	7.1035
Decentralized	-7.24%	-16.92%	7.99%	-4.85%	-0.02%	2.72%	-3.30%
QA	-0.28%	-3.50%	1.47%	-0.85%	0.00%	0.54%	-0.58%

the under-investment in effort, 16.92%, is even larger with customer abandonment. This, and the fact that some customers are lost (along with their revenue), leads to a much higher profit gap of 7.24%. A study of all the parameter combinations presented in Section 2.7.2 results in profit gap of 4.06% to 19.82%.

While the QA contract (with parameters calculated assuming no abandonment) no longer coordinates the process, it still achieves near optimal results. The last row in Table 2.6 shows only a 0.28% profit gap (a study of all the cases in Section 2.7.2 results in profit gap of 0.04% to 0.73% with an average of 0.30%). This shows that the QA contract is very robust. It is also much simpler than the coordinating QA-A contract. As such, it can be viewed as a simple, effective heuristic.

2.7.4 The Mitigating Effect of Outsourcer Pooling

Let there be two identical clients who outsource second stage to the same outsourcer, and the outsourcer pools the two demands. We first solve the centralized problem for the baseline case in Section 2.7.1 and report the results on the first line in Table 2.7.

Table 2.7: Comparison of Centralized Pooled and Decentralized Pooled Settings

	Π	e_1	e_2	$\overline{p}$	$\overline{q}$	$y_1\sqrt{R_1}$	$y_2\sqrt{R_2}$
Centralized	78.1025	2.9374	1.9663	0.9046	0.8717	7.0268	10.4723
Decentralized (%)	-3.45%	-13.41%	6.16%	-3.90%	0.01%	2.08%	-0.68%

Comparing with the first line in Table 2.2, we see that pooling allow the outsourcer to

significantly lower its safety staffing level, $y_2\sqrt{R_2}$, (c.f. 2 times the level in Table 2.2). A smaller headcount then allows it to optimally train its staff to a higher level. Accordingly, the first-stage clients take opposite actions but the impact on first-stage staffing and service quality is minimal. By saving on capacity and providing a higher second-stage quality, the entire process is able to obtain a profit that is 12.7% higher than twice the profit in the non-pooled process.

The second line in Table 2.2 reports the profit deviation of the decentralized case. Clearly, non-coordination leads the first (second) stage to under(over)-invest in effort, and the process profit is suboptimal. Comparison with the second line in Table 2.7, however, reveals this suboptimality is mitigated by outsourcer pooling: instead of 4.04%, the process profit is only 3.45% below the centralized one. Simple calculation reveals that the total profit in the decentralized process is 13.4% higher with pooling than without (as opposed to 12.7% difference in the centralized/coordinated setting). This firmly supports our finding that pooling helps to mitigate the negative effect of non-coordination in a two-stage process.

2.7.5 Pooling for Non-Identical Clients

Although we have shown pooling identical clients is beneficial, this does not always hold when clients are nonidentical. Pooling forces the outsourcer to provide the same staffing and effort levels to both clients. When the two clients have completely different requirements, the homogenization of service may result in over-serving one client and under-serving the other. The overall effect may be negative. For example [86] and [89] show that pooling customers with different service time, wait time requirements, or waiting costs may result in a worse system performance. They conclude that when clients require significantly different staffing levels, it may be better not to pool. In addition to staffing, our model also has an important second dimension, service effort. Even when the two clients' staffing requirements are similar, their service quality requirements for the outsourcer could vary significantly. In such a case, we show below that pooling – thus providing the same effort to both clients – may become suboptimal.

Let the two clients' customers be identical except in their sensitivity to good/bad service.

Specifically, let the the two clients, A and B, have identical parameters to those in the baseline case in Table 2.1, except for δ_r and δ_{nr} whose values are given in Table 2.8.

Table 2.8: Customers' Sensitivity to Service

	δ_r	δ_{nr}
Client A	0.1	1.9
Client B	1.9	0.1

Client A's customers' service quality perception is predominantly affected by the low quality of service. They take good service as given so that a good first-stage service has little impact on their second-stage quality (i.e. small δ_r). On the other hand, if they receive bad service at the first stage, they become so peevish that it is very hard for second stage to make them satisfied (i.e. large δ_{nr}). Client B's customers are the complete opposite: they are little influenced by poor service at the first stage (i.e. small δ_{nr}) – maybe they are used to/expecting it – but if they receive good service, they will be so satisfied (i.e. large δ_r) that the lingering effect will make them very easy to satisfy at the second stage as well.

With this setup, we know for the same effort exerted at the first stage, the outsourcer would need to exert a lot more for Client A's customers than for Client B's customers. If the outsourcer pools all the customers, the same effort level will be applied to all the customers – under-serving Client A and over-serving Client B as a result. The numerical results are presented in Table 2.9. The first two rows represent the optimal safety staffing level at both stages, for the two dedicated processes. The third row that reports how much the pooled process deviates from the first two rows. It is clear that pooling allows the outsourcer to optimally reduce staffing level by 30.76%, confirming the classical queueing benefits of pooling. The second block of rows represent the effort level required at each stage. Here the "service homogenization" issue is evident. If the more demanding Client A is served separately, the outsourcer would exerts a high effort level of $e = 2.8431$. For the less demanding Client B, the outsourcer only needs to dedicate a much-lower effort level

Table 2.9: Comparison When Nonidentical Clients Outsource to the Same Outsourcer

		Client a	Client b	Outsourcer	Process Total
$y\sqrt{R}$	Dedicated A	7.0245		7.0609	
	Dedicated B		7.0235	7.8359	
	Pooled (dev%)	-0.99%	1.21%	-30.76%	
e	Dedicated A	2.9437		2.8431	
	Dedicated B		2.9463	1.0333	
	Pooled (dev%)	6.70%	-7.92%	(-21.30%, 116.54%)	
Quality	Dedicated A	0.9052		0.8694	
	Dedicated B		0.9053	0.8685	
	Pooled (dev%)	1.53%	-2.15%	(-8.39%, 9.04%)	
Π	Dedicated A	14.2290		1.1315	69.3948
	Dedicated B		44.4869	9.5475	
	Pooled (dev%)	307.45%	-96.60%	-36.49%	-4.50%

of $e = 1.0333$. When the two clients are pooled, however, the outsourcer can no longer offer differentiated effort level. Instead, it optimally picks an effort level of 2.2375, which is 21.30% lower than what it would dedicate to Client A and 116.54% higher than what it would dedicate to Client B. As a result, Client A's customers receive a second-stage service quality that is 8.38% lower, while Client B's customers get an increase of 9.04% in second-stage service quality. This can be observed in the next three rows. (Note that because the clients have different e_1, their $\overline{q}(e_1, e_2)$ values are different at stage two even though the outsourcer exerts the same effort e_2).

Finally, the last three rows represent total process profit. While pooling allows staffing reduction, it also results in a compromised outsourcer effort level that's optimal for neither client. This causes profit loss. The overall profit change is the sum of the positive and negative impacts of pooling. In this given case, however, the total profit is 4.50% lower with

pooling than without. Thus, it is better off for the outsourcer to use dedicated agents to serve the two clients. We know one U.S. outsourcer that does exactly this: it pools agents for similar clients but uses dedicated services for clients that require unique or very different amount of training.

Chapter 3

SETTING QUALITY AND SPEED IN SERVICE INDUSTRY WITH REPEATED IMPATIENT CUSTOMERS

3.1 Introduction and Literature Review

While there are many quantitative models for quality control and management in the manufacturing setting, most research on service quality is qualitative. It generates important insights but does not help to quantitatively justify the necessary investment in service quality. In this chapter, we aim to build an analytical model to quantify the impact of service quality on customer behavior and hence on the service provider's profitability, and to help the service provider decide the most appropriate level of investment in service quality and speed.

Recently, the analytical modeling of service quality has become an emergent field of research interest and study. In [20] and [35], service quality is modeled as service accessibility and product availability, and is a direct outcome of the capacity decision (staffing, inventory, etc.). [28] models an oligopoly game where customers follow a multi-arm bandit decision model to switch among suppliers in response to their quality variation.

[40] models the effect of service quality on customer life-time value. In particular, they assume that after each service encounter, a customer is either satisfied or dissatisfied, and service quality is modeled as the long-run proportion of satisfied customers. On the other hand, [23] and [68] define service quality as the service resolution probability. In a customer-service environment, [23] use service resolution to measure the probability that a customer's issue is successfully resolved during the service encounter and s/he will not return for the same problem. In a sales environment, [68] uses service resolution to measure a successful sales transaction during each service encounter. While these definitions of service quality are not comprehensive, they are representative of many practical situations, and have the advantage of being easy to define, measure, and model.

More recently, [7] explicitly model the tradeoff between speed and quality in "customer-intensive services." They measure service quality by customers' valuation of the service, and since service quality relies heavily on real-time human interactions between the service provider and the customer, it has a negative relationship with the speed at which service is provided. Based on this they analyze the service provider's optimal speed and price decisions. [51] extend the tradeoff between service speed and quality to a dynamic setting. Faster speed will reduce congestion in that particular period but will incur an immediate cost and degrade service quality so that demand in the next period becomes lower. They analyze how the service provider's optimal speed and price decisions change over time. Both [7] and [51] impose an exogenously fixed (linear) tradeoff between speed and quality that reduces the degree of freedom in decision making by one.

We also examine how a firm should optimally invest in service quality and speed. As in [40], we measure service quality by the likelihood of a customer being satisfied after each service encounter. This way, we allow the firm to make an investment decision in service quality directly, separate from its investment decision in service quality. This is different from [7] and [51] where service quality is a direct outcome of the service speed decision. By allowing the two decisions to be separate, we are thus able to model practical situations where a firm can invest in both aspects of its service separately – for example, hiring can increase service capacity while training or providing supporting infrastructure can increase service quality. This also makes the model general because now fixing investment level in one doesn't preclude the ability to make an additional investment decision in the other. In Section 3.7 we extend the model by imposing a general constraint between the two variables.

Our chapter differs from existing literature in that we explicitly model several ways in which service quality impacts a customer's interaction with the firm.

Apparently, not all of them apply to all the situations (for example, *spending amount* mostly concerns revenue-generating operations while *service frequency* mostly concerns customer-service operations). In this chapter, we have explicitly modeled three important effects of service quality. To our knowledge, we are the first in the literature to do so.

Customer loyalty Service quality can affect customers' loyalty in their constancy

of preference over a certain length of time, their tenure length, and their reaction to service failure, etc. In our model, we focus on the customers' repurchase behavior (which determines the customer's tenure length). Dissatisfied customers defect while satisfied ones stay and make repurchases.

Referral High service quality increases customer demand not only by inducing existing customers to repurchase but also by attracting new customers. More favorable media coverage, better brand image, and better word-of-mouth all play an important role.

Perceived value of service A customer values the same service higher when she is satisfied than when she is dissatisfied. In a stochastic service environment, customers experience wait and they will join a service only if the expected value exceeds the expected admission and waiting costs. Thus, service quality influences a customer's perceived value which in turn affects whether she chooses to join the service. (There is a rich literature on how congestion impacts the customers' joining behavior and the service provider's capacity and/or pricing decisions. [38] provide an excellent review on this topic.)

Finally, we also examine competition based on service quality. This differs from the service competitions base on congestion and price in [5] and [7]. It is also different from the retailer competition based on price and service in an inventory setting in [83]. Although our model considers exogenous prices, we explicitly model customer waiting and the various impact of service quality, especially that on customer repeat purchase.

3.2 The Base Model

We use a stylized $M/M/1$-type queueing system to represent the service provider. New customers arrive to the system following a stationary Poisson process with a rate λ, there is only one server, and service time follows *i.i.d.* exponential distributions with rate μ. Here μ represents the service provider's speed and is a decision variable.

We assume that each service instance has an *i.i.d.* random outcome with only two possibilities: customer is either completely satisfied or dissatisfied. Service quality is measured by the probability q that any given customer is satisfied after each service, $q \in [0, 1]$. This definition is similar to the service resolution concept used for service quality in [23] and [70]. This measure of service quality can also be interpreted as the long-run proportion of

customers who are satisfied after the service. The latter interpretation of q has been used in [40].

Throughout the chapter we use q as the measure of quality, which is the firm's decision variable and affects the customer behavior in the following manner: If a customer is satisfied, she perceives the value of service to be v. Moreover, she will remain a loyal customer and returns for another possible purchase after a random delay. On the other hand, if a customer is dissatisfied, she will perceive the value of service to be zero and will not return in the future. To model the effect of quality on referral, we further let the arrival rate of new customers, $\lambda(q)$, to be increasing in service quality q.

The service provider does not differentiate between new and repeat customers and utilizes a FCFS queue. Customers are homogeneous (See Section 3.4 for discussion about heterogeneity): All the customers have the same value of service , v, in the event of satisfaction, incur the same cost of c per unit time waiting. We assume the service provider charges a fixed p for each service ($p < v$).

When each customer arrives, she will join the queue only if her utility, defined as her expected value of the service minus price and the expected waiting cost, is non-negative. Otherwise, customer will not join the queue and balks. We assume that all the customers are patient so that there is no abandonment – once the customer decides to join the queue, she will wait until she has received the service. The system dynamics are illustrated in Figure 3.1.

The service provider is a risk-neutral profit maximizer. He can influence revenue by investing in service quality q and service speed μ. Following [40], we assume that the service quality cost function $\alpha(q)$, and the service speed cost function $\beta(\mu)$, are convex increasing, and twice differentiable over the feasible values, i.e. $0 < q < 1, 0 \geq \mu$. The assumption that the service provider can invest in service quality and speed separately and independently is beneficial when customers' perception of service quality is dominated by factors other than waiting time (e.g. problem resolution, courteousness). In such cases, the service provide can invest in infrastructure, training, IT support, and culture, etc. to improve service quality. In this section we assume the price, p, to be fixed and exogenous. This assumption is appropriate when price is set at a higher strategic level than the operating capacity (see,

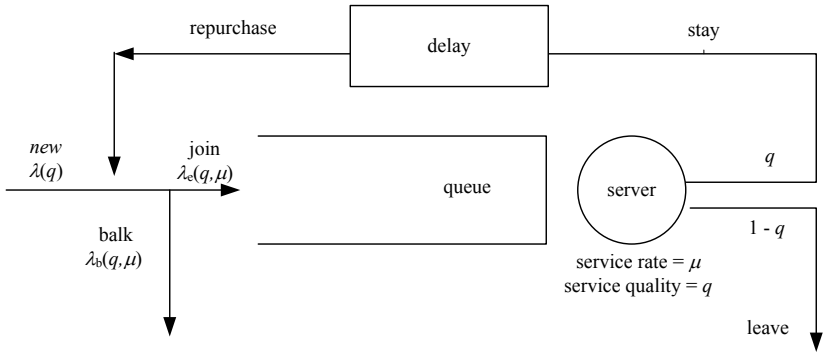


Figure 3.1: The Model

e.g., [80], p. 176) or when the dynamics of the market forces the service provider to be a price taker. Later, in Section 3.3 we generalize the model and consider the cases where price is also a decision variable.

Although the service provider is free to invest in quality and speed independently, we will show in this section that their *optimal* investment levels are inherently linked due to their joint impact on customer joining and repurchasing behaviors. In Section 3.3 we consider cases where price can be determined by the firm as well.

3.2.1 Customer Joining and Repurchasing Behaviors

Because the outcome of each service encounter is random, a customer does not know whether she will receive a satisfactory service *a priori*. With probability q she will get a value of v and with probability $(1-q)$ she will get a value of zero. Therefore, her expected value of the service upon arrival is qv. A customer will join the queue only if she derives a non-negative utility, i.e. $qv - p - c\mathbf{EW} \geq 0$, where $\mathbf{EW} = 1/(\mu - \lambda_e(q))$ is the expected waiting time in the equilibrium. This customer choice model in a queueing setting follows the seminal work of [57] and is commonly used in the literature ([38]). To avoid trivial cases, the service provider must provide a high enough quality level so that $q > p/v$.

We assume that the firm's q and μ are common knowledge to the customers. Moreover, real-time wait information is not available to the customers so their joining strategy depends

only on the expected wait time. This happens when the queue is invisible to the customers or when the customers make their joining decision before visiting the service provider. When service speed (quality) is high enough such that $\frac{c}{\mu-\lambda_e(q)} \leq qv - p$, all the customers will join. Otherwise, a low service speed (quality) will result in negative expected utility if all the customers join the system. However, if a large enough portion of the customers balk, then the utility gained by the rest of the customers will be non-negative. In such a case, each arrival adopts a mixed strategy and joins the system with a random probability. Since the customer population is homogeneous, we consider symmetric Nash equilibrium of the customer joining game. The equilibrium joining probability induces a waiting time that satisfies $\frac{c}{\mu-\lambda_e(q,\mu)} = qv - p$.

Satisfied customers make a repurchase attempt after a random amount of delay. We assume that the delay follows an *i.i.d.* exponential random variable. Thus, the delay module in Figure 3.1 can be modeled by an $M/M/\infty$ queue. Since the arrival of new customers is a Poisson process, each arrival joins the queue with a probability independent of the state of the system, service time is exponential, and each served customer returns for a repurchase attempt with probability q and after an exponential delay, we conclude that the model presented in Figure 3.1 is an open Jackson network ([30]) and the simple $M/M/1$ formulas can be used to calculate the performance measures.

First, we derive the effective arrival rate to the system, $\lambda_e(q, \mu)$, which is essential to the analysis. There are two possible cases:

- When service capacity is high, all the customers join the queue, $\lambda_e(q, \mu) = \lambda(q) + q\lambda_e(q)$. Thus, $\lambda_e(q, \mu) = \frac{\lambda(q)}{1-q}$.

- When service capacity is not high enough, some customers balk. From above we know that the equilibrium joining probability q satisfies $\frac{c}{\mu-\lambda_e(q,\mu)} = qv - p$, hence $\lambda_e(q, \mu) = [\mu - \frac{c}{qv-p}]^+$.

Considering these two cases jointly, we obtain the following Lemma. The proof of this Lemma and all the other results can be found in the appendix.

Lemma 3. $\lambda_e(q,\mu) = \min\left\{\frac{\lambda(q)}{1-q},\ [\mu - \frac{c}{qv-p}]^+\right\}$. *Moreover,* λ_e *is non-decreasing in* q *and* μ.

While our model allows for many reasonable $\lambda(q)$ function, for simplicity we will assume the new customers arrival to be linear function in quality level, i.e. $\lambda(q) = \lambda_0 + wq$, where λ_0 is the base rate and w is the sensitivity of the potential customers' to the quality level. Hence, Lemma 3 can be represented as:

$$\lambda_e(q,\mu) = \min\left\{\frac{\lambda_0 + w}{1-q} - w,\ \mu - \frac{c}{qv-p}\right\}. \tag{3.1}$$

In the first case, when $\lambda_e(q,\mu) = \frac{\lambda_0+w}{1-q} - w$, capacity is sufficiently high and there is no balking; the overall system throughput is restricted only by how many new and repeat customers can be generated. Thus service quality affects revenue only through customer loyalty and referral and perceived value of service plays no role. However, in the second case, capacity is insufficient and some customers will balk. As the system cannot serve all the arrivals, the impact of service quality on attracting new customers and inducing repurchases loses its importance. Conversely, the perceived value effect of service quality becomes prominent, as higher perceived value leads to customers' higher willingness to wait, thus increasing the system throughput.

3.2.2 Service Provider's Decisions

The service provider maximizes its expected profit, denoted by $V(q,\mu)$, as follows:

$$\max_{q,\mu} V(q,\mu) = p\lambda_e(q,\mu) - \alpha(q) - \beta(\mu) = p\min\left\{\frac{\lambda_0 + w}{1-q} - w,\ [\mu - \frac{c}{qv-p}]^+\right\} - \alpha(q) - \beta(\mu). \tag{3.2}$$

Our approach to solving this problem is to divide the (q,μ) space into two regions according to the minimization function in (4.4). The boundary of the two regions is therefore defined by the function $f(q) \triangleq \mu = \frac{\lambda_0+w}{1-q} - w + \frac{c}{qv-p}$.

Now depending on the region we can break down the objective function in (3.2) into two different functions:

$$V(q,\mu) = \begin{cases} V^I(q,\mu) \triangleq p[\mu - \frac{c}{qv-p}]^+ - \alpha(q) - \beta(\mu) & \text{if } 0 < \mu \le f(q) \text{ (i.e. area I)}, \\ V^{II}(q,\mu) \triangleq \frac{p(\lambda_0+w)}{1-q} - pw - \alpha(q) - \beta(\mu) & \text{if } \mu \ge f(q) \text{ (i.e. area II)}. \end{cases} \tag{3.3}$$

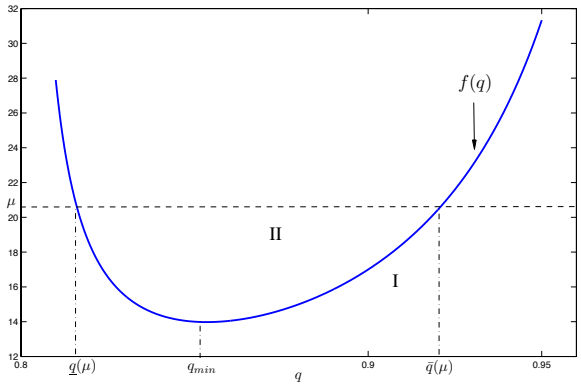


Figure 3.2: q_{min}

Note that when $\mu = f(q)$, $V(q,\mu) = V^I(q,\mu) = V^{II}(q,\mu)$; hence we let the boundary $\mu = f(q)$ belong to both areas. For any point below the curve the capacity is insufficient to serve all the customers for the given q value, as capacity μ causes long enough wait that some customers balk. On the other hand, for any point above the curve there is extra capacity available so that for the given quality level, customers can expect to receive a strictly positive net value. Only on the curve $\mu = f(q)$ is the investment in capacity just high enough so that no customer balks but they all expect to receive zero net benefit. Thus we refer to $\mu = f(q)$ as the *balanced curve*.

The balanced curve, $f(q)$, is not monotone. It's easy to see that $f(q)$ is convex in q, and we denote its minimum on $(p/v, 1]$ by q_{min} (see Figure 3.2). Using first order condition, we find

$$q_{min} = \frac{p + \sqrt{\frac{cv}{\lambda_0 + w}}}{v + \sqrt{\frac{cv}{\lambda_0 + w}}}. \tag{3.4}$$

In particular, to the left of q_{min}, quality level is low and the capacity needed to balance a given quality level q, $f(q)$, is decreasing. This corresponds to the case where the effect of perceived value of service dominates the other two. When quality increases, customers are willing to wait longer and the system can balance with a lower capacity level. On the contrary, to the right of q_{min}, quality level is high so the effects of loyalty and referral dominate. Consequently, higher service quality means more customers to the system that

need to be served. To balance, the system need higher capacity investment, and $f(q)$ is increasing. We say that capacity and quality are *substitutes* to the left of q_{min} and *complements* to the right of q_{min}.

Lemma 4. *In area II, the optimal $V^{II}(q,\mu)$ value is achieved on the boundary $\mu = f(q)$.*

Because $V^I(q,\mu) = V^{II}(q,\mu)$ on the boundary $\mu = f(q)$, Lemma 4 implies that (q^*, μ^*) must lie in area I (which includes the boundary $\mu = f(q)$).

Proposition 3. *The optimization problem* (3.2) *is equivalent to:*

$$\max V^I(q,\mu) \quad s.t. \quad 0 < \mu \leq f(q). \tag{3.5}$$

However, the optimal investment (q^*, μ^*) does not necessarily reside on the balanced curve. Depending on the economic parameters, we will show later that it may not be optimal for the firm to serve all the potential customers. That is, it may indeed be better to deliberately under-invest in service capacity and be in area I. The following lemma shows that it is never optimal to over-invest in capacity.

That (q^*, μ^*) must be in area I is actually quite intuitive: when capacity is already high enough to serve all the potential customers (for a fixed quality level), any further investment in μ will only decrease customers' waiting time without increasing the firm's revenue. In other words, customers reap all the benefit by getting positive expected value while the firm pays. This is sub-optimal for the firm.

Proposition 3 allows us to focus on a much simpler objective function at the expense of a more restricted feasible region. Please note that if $\mu - \frac{c}{qv-p} < 0$, no customer enter the system and problem is trivial. Hence we can use $\mu - \frac{c}{qv-p}$ in our analysis. The advantage of analyzing (3.5) is that $V^I(q,\mu)$ is both separable and concave in q and μ. Let $(\hat{q}, \hat{\mu})$ be its unique unconstrained maximizer:

$$\hat{q} = \arg\max_q \left\{ -\frac{pc}{qv-p} - \alpha(q) \right\}, \tag{3.6}$$

$$\hat{\mu} = \arg\max_\mu \left\{ p\mu - \beta(\mu) \right\}. \tag{3.7}$$

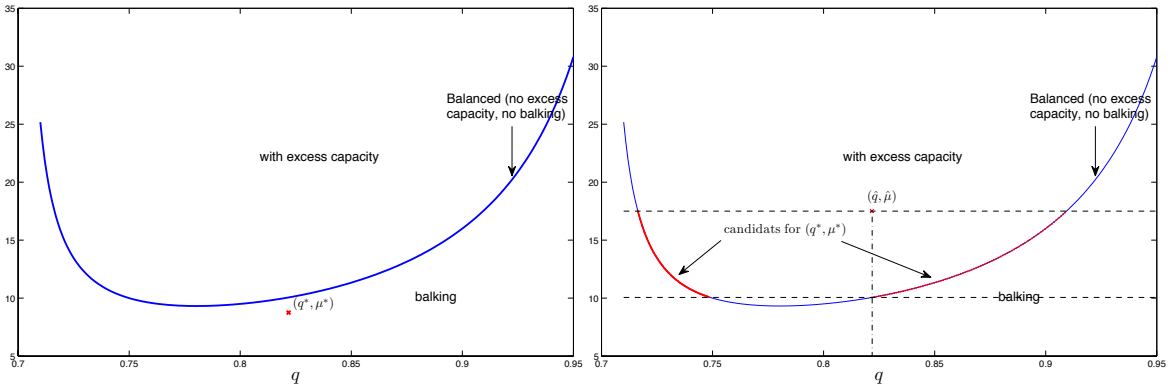
Because $\hat{q}$ and $\hat{\mu}$ are based on the V^I objective function, they are "biased". Specifically, $\hat{q}$ balances the tradeoff between loss of revenue due to customer balking and the quality

investment cost; it ignores the repeat customer dynamics. $\hat{\mu}$ balances the potential revenue tradeoff between loss of revenue due to customer balking and the quality investment cost; it ignores the limit of customer arrival rate. Therefore, $\hat{\mu}$ is clearly an upper bound on the constrained optimal μ^*. The relationship between $\hat{q}$ and q^* is less clear. The two following propositions investigate these relationships.

Proposition 4.

(A) If $\hat{\mu} \leq f(\hat{q})$, then (q^*, μ^*) is unique and $q^* = \hat{q}$ and $\mu^* = \hat{\mu}$;

(B) If $\hat{\mu} > f(\hat{q})$, any optimal solution (q^*, μ^*) satisfies $\mu^* = f(q^*)$ and $f(\hat{q}) \leq \mu^* \leq \hat{\mu}$.



(A) When $(\hat{q}, \hat{\mu})$ is in area I (B) When $(\hat{q}, \hat{\mu})$ is in area II

Figure 3.3: Mapping $(\hat{q}, \hat{\mu})$ into (q^*, μ^*)

Figure 3.3(A) and (B) depict Cases (A) and (B) of Proposition 4 respectively. In case (A), the unconstrained maximizer $(\hat{q},\ \hat{\mu})$ is already in the feasible region (area I), hence it is also the constrained maximizer. When the unconstrained maximizer $(\hat{q},\ \hat{\mu})$ lies in the infeasible region (interior of area II), however, it must be mapped to area I. Case (B) of the proposition shows the constrained (q^*, μ^*) that is mapped from $(\hat{q},\ \hat{\mu})$ must be on the balanced curve $\mu = f(q)$. To further characterize the location of (q^*, μ^*), we need the following lemma.

Lemma 5. *Let* $V(q, f(q)) = \frac{p(\lambda_0+w)}{1-q} - pw - \alpha(q) - \beta(f(q))$. *Then we have,*

1. $\lim_{q \to \frac{p}{v}^+} \frac{dV(q, f(q))}{dq} > 0$;

2. $\lim_{q \to 1^-} \frac{dV(q, f(q))}{dq} \leq 0$.

Lemma 5 ensures that profit function's maximizer on the balanced curve $\mu = f(q)$ lies on the interior. This allows us to search for the optimal quality level using the first order conditions.

Theorem 7. *For the case of* $\hat{\mu} > f(\hat{q})$, $\mu^* = f(q^*)$ *where* q^* *is the solution of*

$$\frac{p(\lambda_0 + w)}{(1-q)^2} - \alpha'(q) - \beta'(f(q))f'(q) = 0, \tag{3.8}$$

that achieves the highest $V(q, f(q))$. *Moreover,* $f(\hat{q}) \leq \mu^* \leq \hat{\mu}$. *When* $\hat{q} \geq q_{min}$, $q^* \geq \hat{q}$; *when* $\hat{q} < q_{min}$, $q^* < \hat{q}$.

Figure 3.3(B) shows the two segments of $\mu = f(q)$ on which (q^*, μ^*) can reside. It is worth noting that when the operational balance between capacity and quality is considered, investment level in capacity is *always* lowered (i.e. $\mu^* \leq \hat{\mu}$) but quality may be either increased or decreased. This holds generally, regardless of the cost functions and other parameters. We believe this result stems from the fact that in our model service capacity plays a straightforward role, but quality affects multiple behaviors in the system dynamics and its various effects have different implications.

Assumption 1. $\frac{dV(q,f(q))}{dq} = \frac{p(\lambda_0+w)}{(1-q)^2} - \alpha'(q) - \beta'(f(q))f'(q) = 0$ *has a unique solution.*

Although from Theorem 7 it is clear that the optimal solution will be on either the complement side or the substitute side of the balanced curve, Theorem 7 does not guarantee uniqueness. Incidently, with the general assumptions for the quality and capacity cost functions, uniqueness cannot be proven. Assumption 1 assumes a unimodal first order conditions that will achieve the unique solution of the problem. Later in the section we provide some examples of the functions that satisfy Assumptions 1.

Definition 3. *We define S as a set of parameter values of $\alpha(\cdot)$ and $\beta(\cdot)$, such that for $\forall x \in S$; $\frac{\mathrm{d}V^{II}}{\mathrm{d}q}$ is twice differentiable in x and there exists an upper bound M such that $\frac{\partial}{\partial x}\frac{\mathrm{d}V^{II}}{\mathrm{d}q} \leq M$ for $x \in [x^o - \delta/2, x^o + \delta/2]$ and $q \in [q^*(x^o) - \gamma/2, q^*(x^o) + \gamma/2]$.*

The conditions in Definition 3 are not very restrictive. In Section 2.7 we offer special cost functions and show set S for these cost functions.

Proposition 5. *If Assumption 1 holds. When $\hat{\mu} > f(\hat{q})$, as a parameter $x \in S$ changes continuously, q^* moves continuously on $f(q)$.*

Proposition 5 shows that changes in the parameters will result in smooth changes in the optimal quality value. This result implies that as the parameters change, optimal quality level will change smoothly and as the capacity level is determined by $f(q)$, so does capacity level. This implies that small changes in the parameter values will result in the optimal solution to move on the balanced curve.

On the balanced curve, μ and q are substitutes of each other to the left of q_{min} and complements to the right of q_{min}. Theorem 7 further divides area II into two parts: when $\hat{q}$ is to the left of q_{min}, $(\hat{q},\ \hat{\mu})$ will be mapped to the "substitutes" portion of the balanced curve; when $\hat{q}$ is to the right of q_{min}, $(\hat{q},\ \hat{\mu})$ will be mapped to the "complements" portion of the balanced curve. Thus, q_{min} serves as an important threshold, and we only need to check whether the unconstrained optimal quality level $\hat{q}$ is above or below q_{min}. Since it is the quality cost function $\alpha(q)$, not the capacity cost function $\beta(\mu)$ – see (3.6) – that determines $\hat{q}$, we conclude that it is the quality cost function that determines the nature of relationship between capacity and quality investment. The capacity cost does not play a role in this determination.

Theorem 8. *When $\alpha^{'}(q_{min}) \leq \frac{pcv}{(q_{min}v-p)^2}$, $q^* \geq \hat{q} \geq q_{min}$ and the investments in capacity and quality are complements with respect to changes in any $x \in S$. Conversely, when $\alpha^{'}(q_{min}) > \frac{pcv}{(q_{min}v-p)^2}$, $q^* < \hat{q} < q_{min}$ and the investments in capacity and quality are substitutes.*

An important observation is that whether investments are substitutes or complements depends solely on the customer's cost and value parameters, and the firm's service quality

cost function. As long as the optimal solution lies on the balanced curve, operational parameters such as the arrival function and the service capacity have no effect on the complementarity or substitutability of the quality and capacity decisions. Furthermore, comparatively when the customer valuation for the product is high with respect to the price (v/p) and/or costs for waiting (c) is low, investments are more likely to be complements and need to be increased or decreased together. This result matches our original intuition as for lower v/p and higher c, the customers get lower utility and thus the perceived value of service effect of q will be more pronounced. As it was stated earlier, in such situations the investments are substitutes. On the contrary, higher v/p and lower c mean loyalty and repeat purchase effects are more dominant and investments are complements.

We also observe that as long as the optimal solution is on the balanced curve, the capacity cost function plays no role in the relationship between the two decisions. It is the quality cost function $\alpha(q)$ that dictates this relationship. In particular, investments in service and quality are more likely to be complements when $\alpha(q)$ is *less* convex. This is intuitive because a more convex $\alpha(q)$ means a higher marginal cost for investing in quality so the optimal investment in quality tends to be lower. From the shape of the balanced curve, it is clear that capacity and quality are substitutes for lower quality levels.

3.3 Price as a decision

Previous sections provide an insight on quality and capacity decisions in a case where price is predetermined. Although the analysis above is valuable for situations where the service providers would have to forgo the pricing aspect due to specific market structures or limited decision making scope, there are service processes where the service provider can change the price (as well as capacity and quality levels) to gain higher profits.

The analysis in the previous section provides a detailed description of the behavior of the optimal solution for any given price decision. The natural next step in the analysis is to design a model which can provide some insight on the optimal price decision in the markets where firms are in control of the prices. In this section we demonstrate the effects of pricing decisions on the optimal solution and complete the original model proposed in Section 3.2

by including price as a decision variable. We redefine the profit function as,

$$V(q,\mu,p) = \begin{cases} V^I(q,\mu,p) \triangleq p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu) & \text{if } 0 < \mu \leq f(q,p) \text{ (i.e. area I)}, \\ V^{II}(q,\mu,p) \triangleq \frac{p(\lambda_0+w)}{1-q} - pw - \alpha(q) - \beta(\mu) & \text{if } \mu \geq f(q,p) \text{ (i.e. area II)}. \end{cases} \tag{3.9}$$

where, $f(q,p) \triangleq \frac{\lambda_0+w}{1-q} - w + \frac{c}{qv-p}$ is the balanced surface that determines the boundary between the area I and II. It is important to note that in (3.9), the balanced surface, $f(q,p)$, is dependent on both the quality and price decisions. Figure 3.4 shows the shape of the $f(q,p)$ function. $f(q,p)$ is jointly convex in q and p. As it is expected for fixed price, p, the balanced surface degenerates to the balanced curve $f(q)$ introduced in the previous section. However as price increases the $f(q,p)$ curve created by $(q, f(q))$ will move up and right. (i.e. if $p_2 > p_1$, then $f(q,p_2) > f(q,p_1)$, and $q_{min}(p_2) > q_{min}(p_1)$.)

Because of the complex nature of the model when price decision is considered, analysis of the problem is more difficult. It is important to note that although in the new analysis,

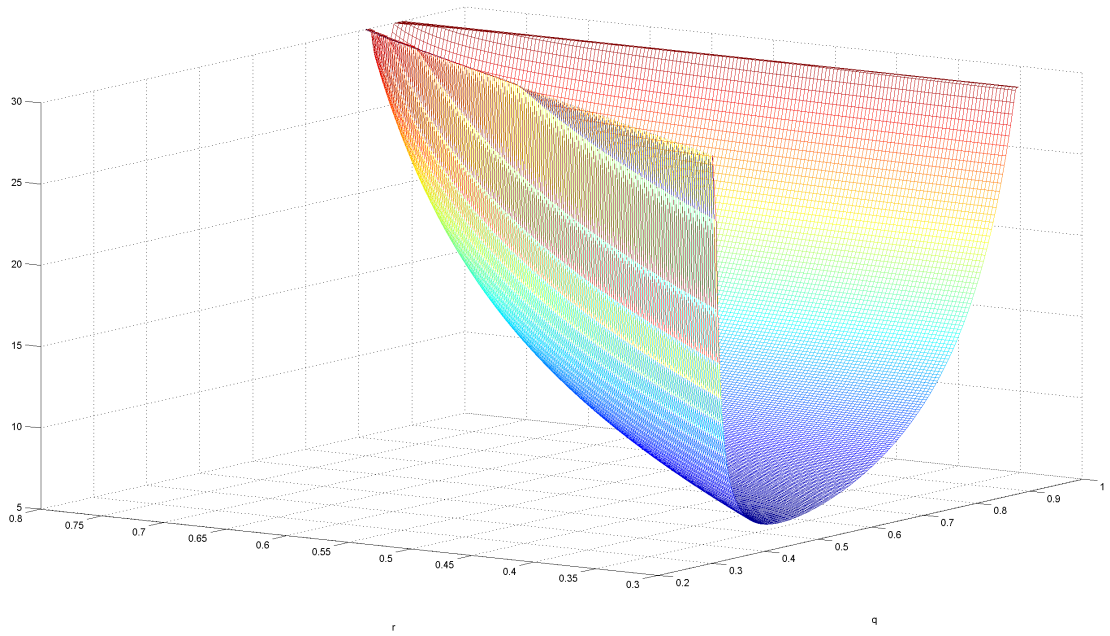


Figure 3.4: $f(q,p)$

pricing is also another dimension that needs to be considered, when fixing the price at the optimal level, analysis performed in previous sections can be applied to this section as well. Specifically, the result of Lemma 4 holds at the optimal price level, and thus can be generalized: In area II, the optimal $V^{II}(q,\mu,p)$ value is achieved on the boundary $\mu = f(q,p)$. Within the area II the capacity is high enough that it is optimal for all the customers to join the system. From (3.9) it is clear that within area II profit increases with increasing p and decreases with increasing μ. This implies that at any point within area II (except for the boundary) by either increasing the price (hence increasing the revenue) or decreasing the capacity (hence decreasing the cost), the service provider can increase the total profit without changing the total number of customers whom are served. As a result the optimal solution set can lie in area II if and only if $\mu = f(q,p)$. Hence the optimization problem is equivalent to

$$\max V^I(q,\mu,p) \quad s.t. \quad 0 < \mu \leq f(q,p). \tag{3.10}$$

Although the joint concavity of the problem above is dependent on the specific $\alpha(\cdot)$ and $\beta(\cdot)$ functions. We can show that optimal solution of the unconstraint problem (which may be unique or not) is achieved at the local optima.

Lemma 6. *There exists a point $(\check{q},\check{\mu},\check{p})$ that is the global maximizer of $V^I(q,\mu,p)$, if and only if*

$$\begin{cases} \frac{\check{p}cv}{(\check{q}v-p)^2} - \alpha'(\check{q}) = 0, \\ \check{p} - \beta'(\check{\mu}) = 0, \\ \check{\mu} - \frac{c\check{q}v}{(\check{q}v-\check{p})^2} = 0, \end{cases} \tag{3.11}$$

and has the maximum value of $V^I(q,\mu,p)$ among all the other points that satisfy (3.11).

The conditions of Lemma 6 limit the search for the optimal point to only the extremum points. Lemma 6 finds the extremum point $(\check{q},\check{\mu},\check{p})$ as the solution that generates the highest $V^I(q,\mu,p)$ for all the solutions to Lemma 6. This solution is the solution to the unconstrained problem and plays an analogous role to that of $(\hat{q},\hat{\mu})$.

Similar to the scenario when we had no pricing decision, in cases where $\check{\mu} > f(\check{q},\check{p})$, the solution to the constrained problem will lie on the balanced surface.

Proposition 6. *Let $(\check{q}, \check{\mu}, \check{p})$ be the solution achieved from Lemma 6; we have $(q^*, \mu^*, p^*) = (\check{q}, \check{\mu}, \check{p})$ iff,*

$$\alpha'(\check{q}) \leq v\left(\frac{\lambda_0 + w}{1-\check{q}} - w\right). \tag{3.12}$$

Otherwise, (q^*, μ^*, p^*) can be found by solving the following,

$$\begin{cases} \frac{p(\lambda_0+w)}{(1-q)^2} - \alpha'(q) - \beta'(f(q,p))\left(\frac{\lambda_0+w}{(1-q)^2} - \frac{cv}{(qv-p)^2}\right) = 0, \\ \frac{(\lambda_0+w)}{(1-q)} - w - \beta'(f(q,p))\frac{c}{(qv-p)^2} = 0. \end{cases} \tag{3.13}$$

The term in (3.12) is the condition that ensures that the solution lies below the balanced surface, as long as marginal cost of quality is less than the maximum value that the service that is provided can generate if all the customers where satisfied. It is obvious from (3.12) that if the arrival rate of the new customers is high, or $\alpha(q)$ is less convex then it is optimal to only serve a portion of the customers and allow the rest of the customers to balk. It is important that both of these conditions ensure higher traffic to the system and hence they increase the chance of having an overloaded system.

The optimal solution above the balanced surface, similar to Section 3.2, has to be mapped on the balanced surface and the second part of the Proposition 6 provides the optimal solution on the surface. It is clear that for any point above the balanced surface, $f(q,p)$, the total profit increases by either increasing p or decreasing μ. Naturally, we expect that in the optimal solution, the price to be higher than the price solution to the unconstrained problem and the capacity to be lower.

Proposition 7. *When the optimal solution lies on the boundary, $p^* \geq \check{p}$, $f(\check{q}, \check{p}) \leq \mu^* \leq \check{\mu}$, and when $\check{q} < q_{min}(p^*)$, $q^* < \check{q}$, when $\check{q} > q_{min}(p^*)$, $q^* > \check{q}$. Furthermore, when $\check{q} < q_{min}(\check{p})$, then $q^* < \check{q}$.*

Proposition 7 shows that depending on the optimal value of the price, for a fixed price, the quality and capacity decisions could be either complements or substitutes. As the optimal price increases, the threshold for which the quality and capacity decisions become complements, $q_{min}(p)$, also increases, and hence it is more likely for quality and capacity to be substitutes for a fixed value of price. It is important to note that this result is consistent with the analysis that was performed in Section 3.2. As price increases, v/p decreases, and

the capacity and quality decisions tend to be substitutes. However, this analysis is only valid when price is considered fixed. If the price can change as well as quality and capacity, then from Proposition 7 it is unclear how these decisions will change with respect to each other. Theorem 9 investigates the relationship between the three decision variables in the optimal decision. In the previous section, for the cases where the solution was located in area I, because of the separable property of quality and capacity decisions we were unable to provide any insight on the relationship between these two variables. In this section, the consideration of the price creates a natural link between all the variables.

Theorem 9. *In consideration of the price, for any $x \in S$, as x changes, , in the optimal solution set, quality and capacity decisions are always complements. Furthermore,*

1. *if $(q^*, \mu^*, p^*) \neq (\check{q}, \check{\mu}, \check{p})$ and , $\beta''(f(q,p)) > \frac{v + \frac{2v\beta'(f(q,p))}{(qv-p)^2}}{\frac{\lambda_0+w}{(1-q)^2} - \frac{cv}{(qv-p)^2}}$, and $\frac{\lambda_0+w}{(1-q)^2} - \frac{cv}{(qv-p)^2} > 0$ the price decision is substitutes to both quality and capacity decisions.*

2. *Otherwise, price decision is also complements.*

Theorem 9 shows that regardless of where the solution is located, quality and capacity decisions are complements. In this case, as quality increases, price will either increase or decrease so that, capacity and quality decisions are substitutes. For the first part, when optimal decision set is below the balanced curve, for any given set of variables, changes in parameters that would result in an increase in one of them, subsequently increases the other two decision variables as well. This shows that in a system where there is no excess capacity, an increase in price should be followed by increases in capacity and quality as well, similarly for quality and capacity. This result emerges from the fact that in this system if price is increased the value gained by the customers will decrease and the portion of the customers who will not join the system will increase, subsequently an increase in quality and capacity is needed to regain a portion of the lost customers due to increase in price. Same type of argument also holds for increasing quality and capacity levels.

On the other hand for Part 2, when the solution lies on the balanced surface, if the changes in the parameter values are so that the quality increases, then the capacity will

increase. This is regardless of on what side of the balanced surface the optimal solution is located. Specifically, when the optimal solution is located on the decreasing side of the $f(q, p^*)$ curve, as quality increases, the price will also increase, which will result in a total increase in the capacity as well. When the solution is located on the increasing side of the $f(q, p^*)$, then depending the location of p^* and q^*, with an increase in q^*, p^* could either increase or decrease, but the capacity will increase regardless as the direction of change in p^*.

One interesting aspect of this result is that although the changes in the parameters could result in an increase in quality and capacity, still the price could be lowered. Specifically when $\beta''(f(q,p)) > \frac{v+\frac{2v\beta'(f(q,p))}{(qv-p)^2}}{\frac{\lambda_0+w}{(1-q)^2}-\frac{cv}{(qv-p)^2}}$ and $\frac{\lambda_0+w}{(1-q)^2} - \frac{cv}{(qv-p)^2} > 0$, an increase in quality and capacity will be followed by a decrease in price. This conditions could be satisfied due to several reasons, some of which are, very convex capacity cost function, and low $\frac{\lambda_0+w}{(1-q)^2} - \frac{cv}{(qv-p)^2}$. In the case where the capacity cost function is very convex since, $\frac{\lambda_0+w}{(1-q)^2}-\frac{cv}{(qv-p)^2} > 0$, the effect of attracting new customers and repeat customers is dominant so the total arrival to the system will increase, yet the increase in marginal cost of capacity is very steep, and the capacity cannot be increased as much to satisfy all the customers, hence in order to keep the customers in the system, the price should be decreased. On the other hand if $\frac{\lambda_0+w}{(1-q)^2} - \frac{cv}{(qv-p)^2}$ is low, the changes in capacity with respect to just the change in quality are low, so in order to capture the extra demand, price needs to be decreased.

It is important to emphasize the differences between this section and the section where the price was considered fixed by the market. In this section we explain that if pricing is considered as a decision, then the levels of quality and capacity decisions are always complements. In the cases where the firm has more demand that the firm can (or should) meet, then the price is also a complement but in the cases where all the customers are served there are conditions under which when quality is increased, the price could be lowered. This section provides a great comparison between the scenarios where the decision maker is allowed to change all the three decisions versus the cases where the decision maker only has the ability to control quality and capacity decisions.

3.4 Heterogenous Customers

In the base model, we assume that all customers are homogenous. In this section we extend our base model to consider the situation where customers' valuation on the service is heterogenous. We investigate the dependence of the intuitions achieved in the previous sections on the homogeneity assumption. In particular, we assume that customers' valuation on the service when they are fully satisfied, v, is uniformly distributed over $[\underline{v}, \bar{v}]$, where $V = \bar{v} - \underline{v}$. Each customer has private information about his/her valuation.

For a pair of quality level and capacity level (q, μ), the expected throughput λ_e is given in the following lemma.

Lemma 7. $\lambda_e = \min\left\{\frac{\Lambda}{1-q},\ \frac{1}{2}\left[\mu + \frac{(\bar{v}q-p)\Lambda}{Vq(1-q)}\right] - \sqrt{\frac{1}{4}\left[\mu - \frac{(\bar{v}q-p)\Lambda}{Vq(1-q)}\right]^2 + \frac{c\Lambda}{q(1-q)V}}\right\}$. *where,* $\Lambda = \lambda_0 + wq$. *Furthermore* λ_e *is nondecreasing in* q, *and* μ.

Although it is difficult to theoretically prove that λ_e is concave in q and μ, we can show that solution is on a critical point. This implies that the insights found in the previous sections also can be applied to this model. Specifically, we can denote the balanced curve as $f_a(q) = \frac{\Lambda}{1-q} - \frac{c}{q\underline{v}-p}$, where for any fixed value of price, p, the $f(q)$ is convex in q. Similar to the result of the base model whenever it is optimal to serve all the customers (with any valuation of service) then the solution will lie on this balanced curve. For any point above below the balanced curve a number of customers whom will not receive a positive value from the service will balk. As it is clear from this balanced curve the same properties as the ones in Theorem 8 also hold for this blanched curve. Please note that this result shows that again, whenever all the customers are served, if providing lower quality is optimal, then the quality and capacity decisions tend to be substitutes, versus when high quality is provided, the two decisions are complements, where the extra capacity is required to capture the new and repeat customers arriving due to the high quality.

Although the analysis above is similar to the results that we achieved for homogenous customers, the relationship between quality and capacity decisions is unclear when the firm only serves customers with higher valuations, and lets go of some of the customers. If the optimal decision is not to serve all the customers, then there exists a customer valuation,

v_e (, $\underline{v} < v_e \leq \bar{v}$), that customers with valuations of v_e and higher are served and the customers with lower valuation will balk. This value is a by product of the quality and capacity decisions made by the firm and can be determined using these values; $v_e^* = \bar{v} - \lambda_e^*(1-q^*)V/\Lambda^*$. For any given v_e, the exists a curve, $f_e(q) = \frac{\Lambda}{1-q} - \frac{c}{qv_e-p}$, below which all the customers with valuation of v_e will receive will balk. Although this function, f_e, has a similar structure to the balanced curve, it cannot be used to determine the relationship between the quality and capacity decisions. In this case, changes in quality or capacity decisions could result in changes in the value of v_e, and as a result changing f_e. Thus the result of Theorem 8 cannot be directly applied when all the customers are not served. Although a clear balanced curve does not exist in this scenario, we can analyze the relationship between the two decision variables. Here assumption of heterogeneity creates a natural link between the quality and capacity decisions. The following proposition provides the conditions under which the two decisions are substitutes versus complements.

Proposition 8. *The quality and capacity decisions are substitutes with respect to changes in any $x \in S$ if and only if, (i) When $\mu^* = \frac{\lambda_0+wq^*}{1-q^*} + \frac{c}{q^*\underline{v}-p}$ and $q^* < \frac{p+\sqrt{\frac{c\underline{v}}{\lambda_0+w}}}{\underline{v}+\sqrt{\frac{c\underline{v}}{\lambda_0+w}}}$, or (ii) When $q^* < \frac{\sqrt{\lambda_0^2+w\lambda_0}-\lambda_0}{w}$, and $\frac{\bar{v}(\lambda_0+wq^*)(\lambda_0+2wq^*-wq^{*2})}{V(\lambda_0-2\lambda_0q^*-wq^{*2})(1-q^*)} + \frac{p(\lambda_0+wq^*)}{q^*(1-q^*)V} < \mu^* < \frac{\lambda_0+wq^*}{1-q^*} + \frac{c}{q^*\underline{v}-p}$. In all the other cases, they are complements.*

Proposition 8, shows conditions under which the optimal quality and capacity decisions are substitutes or complements. Part (i) of the conditions refers for the cases where all the customers are served and is the direct result of Theorem 8. However, Part (ii) refers to the scenarios when only a portion of the customers are served. Although these conditions are different from those is Theorem 8, the intuitions are similar. As it is clear from the conditions, whenever, quality level is low, (please note that $\frac{\sqrt{\lambda_0^2+w\lambda_0}-\lambda_0}{w} < 0.5$), the quality and capacity decisions are substitutes and as quality level increases the two decisions become complements. Please note that in this case, the capacity level has to be higher than a certain level for the decisions to be substitutes. (see Figure 3.5).

A major difference between the relationship between quality and capacity decisions on and below the balanced curve is feasibility. Although the substitution area on the balanced curve is always feasible, below the balanced curve, there could exist parameter value sets

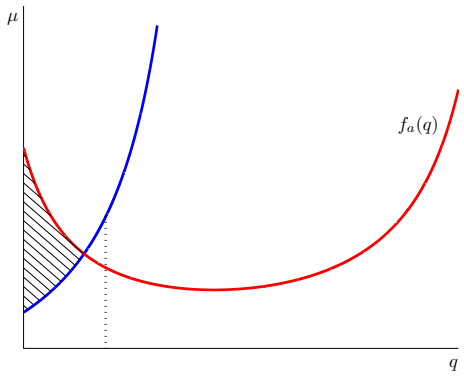


Figure 3.5: Shaded area shows the substitution area

such that the two decisions can be only complements and no substitution area exists. One clear example where quality and capacity decisions can never be substitutes below the balanced curve is when $\frac{\sqrt{\lambda_0^2+w\lambda_0}-\lambda_0}{w} \leq \frac{r}{\underline{v}}$. Again the conditions in Part ii of Proposition 8 are not always feasible. Specifically we can show that, $\frac{\bar{v}(\lambda_0+wq^*)(\lambda_0+2wq^*-wq^{*2})}{V(\lambda_0-2\lambda_0 q^*-wq^{*2})(1-q^*)} > \frac{\lambda_0+wq^*}{1-q^*}$, hence as long as $\frac{c}{q^*\underline{v}-p}$, is small, there is no substitution area under the curve. It is important to note that when the solution lies below the balanced curve, as the waiting cost decreases, similar to the cases where the optimal decision lies on the balanced curve, the area in which the decisions are substitutes will shrink.

As it is clear from the discussion, it is less likely for the decisions to be substitutes under the balanced curve. This is due to the fact that when the waiting cost is low, by increasing the quality, in this case the service provider attracts customers with lower valuations, as well as new and repeat customers of higher valuations. Hence by increasing quality, the firm has to increase capacity as well, to be able to hold a portion of the new customers that have lower valuation of the service, and hence are not willing to wait long to receive the service. On the other hand, as the customer waiting cost increases, by increasing the quality, the firm cannot convince the customers with lower evaluations to join, hence by reducing the capacity, will extract the rent from the existing customers who are receiving higher values due to higher quality. This is consistent with the results that we achieve for the balanced curve; below the balanced curve when the effects of attracting new and repeat customers is

more pronounced the decisions are complements, and when (due waiting cost is high to the customers), the effects of quality on the value received by the customer is more pronounced, the quality and capacity decisions are substitutes. Naturally, below the balanced curve, the opportunity for attracting new customers is also attainable compared to the cases where the solution lies on the balanced curve, and hence the area of substitution is relatively smaller when the solution is below the balanced curve compare to on the balanced curve.

3.5 Duopoly Model

In this section we study a differentiated Bertrand duopoly competition on service quality to attract new customers. Let λ_i, q_i, μ_i $(i = 1, 2)$ be firm i's potential customer arrival rate, quality level, and capacity level, respectively. Similar to [83], we assume a linear demand model based on service quality competition as follows:

$$\lambda_1(q_1, q_2) = \lambda_{01} + aq_1 - bq_2, \tag{3.14}$$

$$\lambda_2(q_1, q_2) = \lambda_{02} + aq_2 - bq_1, \tag{3.15}$$

where λ_{0i} represents the exogenous base arrival rate of new customers for firm i $(i = 1, 2)$. Moreover, a reflects the responsiveness of firm i's its own service quality and b reflects the responsiveness of firm i's customers to the competitor (firm j, $j \neq i$)'s service quality. Where the clientele of each firm are more sensitive towards the quality level offered by the firm compared to the competitors, $a > b$. As it is clear, the arrival rate is increasing the quality investment of the both firms and the portion of the demand that is dedicated to each firm is increasing in the quality of the firm and decreasing in the quality level of the competitor. It is important to note that although we assume fixed values for a and b, these values just determine the average rate of customer arrival. Hence arrival in this case is still stochastic and only the rate is determined by the parameters.

The total arrival rate to the system is determined by $\lambda_{01} + \lambda_{02} + (a - b)(q_1 + q_2)$. It is clear that increasing the quality level by each firm increases the total arrival rate of the system. However the rate of increase in arrival rate of the firm, a, is higher than the rate of increase in the total number of customers, $a - b$. Hence by increasing the quality level, a firm both captures new customers that are attracted to the market because of higher

quality, and also "steals" some of the customers of the competitor. This feature that allows the firms to capture the demand of the competitor is what differentiates the competition scenario from the monopoly case.

In this duopoly model, the two firms simultaneously determine their own quality level q and capacity level μ, based on which customers then make joining or balking decisions. Given firm j's quality decision, firm i's ($i \neq j$) objective is maximizing the following function:

$$\max_{q_i,\mu_i} V(q_i,\mu_i|q_j) = \min\left\{\frac{p(\lambda_{0i}+aq_i-bq_j)}{1-q_i}, p\mu_i - \frac{pc}{q_i v - p}\right\} - \alpha(q_i) - \beta(\mu_i), i = 1, 2. \quad (3.16)$$

We start our analysis for the symmetric Nash equilibrium when the two firms are identical (e.g. the same λ_0, a, b, p, c, $\alpha(\cdot)$, and $\beta(\cdot)$). Later in the section we discuss some insights in the cases where the two firms are non-identical. When a symmetric equilibrium exists, we denote it by (q_D^*, μ_D^*). To understand the effect of competition, we will compare (q_D^*, μ_D^*) with the monopoly investment levels (q_M^*, μ_M^*) that maximizes the profit function in (3.2) with $w = a - b$. This value is chosen so that when the two firms' are identical, the arrival to each firm, $\lambda_0 + (a-b)q$, has the same structure in both the monopoly and the duopoly models. Any subsequent difference in the market division is then purely due to competition in the duopoly model.

We continue to let $\hat{q}$ and $\hat{\mu}$ be the solution to (3.6) and (3.7).

Proposition 9. *Let* $\hat{\lambda}_0 = (1-\hat{q})\left(\hat{\mu} - \frac{c}{\hat{q}v-r}\right) - \hat{q}\,(a-b)$. *There exists a symmetric Nash equilibrium* (q_D^*, μ_D^*).

1) If $\lambda_0 \geq \hat{\lambda}_0$, *then* $(q_D^*, \mu_D^*) = (\hat{q}, \hat{\mu})$.

2) If $\lambda_0 < \hat{\lambda}_0$, *then* q_D^* *is the solution of*

$$\begin{aligned} &\frac{r(\lambda_0 + a - bq)}{(1-q)^2} - \alpha'(q) \\ - \quad &\beta'\left(\frac{\lambda_0 + (a-b)q}{1-q} + \frac{c}{qv-r}\right)\left[\frac{\lambda_0 + a - bq}{(1-q)^2} - \frac{cv}{(qv-r)^2}\right] = 0 \end{aligned} \quad (3.17)$$

that maximizes $V(q_D^*, \mu_D^*|q_D^*)$ *where* $\mu_D^* = \frac{\lambda_0 + (a-b)q_D^*}{1-q_D^*} + \frac{c}{q_D^* v - r}$.

The first case of Proposition 9 is not surprising. Even without facing competition both firms choose to serve only a portion of the available customers. In this case it is suboptimal for the firms to serve all the customers hence all new customers that can be attracted from the other firm through quality competition will bring no benefit. Therefore the two firms operate like two monopolies and $(q_D^*, \mu_D^*) = (\hat{q}, \hat{\mu})$.

In the second case, total market size is small enough that the two firms have to compete for customers. Even though in the symmetric equilibrium neither firm gains more customers by stealing the other firm's customers(i.e. $b(q_i - q_j) = 0$), the *threat* of competition in the duopoly setting forces them to invest in higher quality levels than in the monopoly setting. More specifically, marginal investment in quality will increase customer demand by $w = a - b$ in the monopoly setting, but by $a = w + b$ in the duopoly setting. Therefore, both firms have more incentive to invest in higher service quality due to competition. Furthermore, the more responsive customers are to quality competition (i.e. the higher b), the higher the pressure of competition and both firms will invest in higher quality. The following proposition shows the effects of the competition on various decisions and measures of the system.

Proposition 10. *For identical firms in the symmetric equilibrium in the duopoly case compared to the monopoly case,*

1. *The firms will provide higher quality levels:* $q_D^* \geq q_M^*$.

2. *The total clientele for each firm is larger:* $\lambda_0 + (a - b)q_D^* \geq \lambda_0 + wq_M^*$.

3. *Both firms receive lower profits:* $V_D^* \leq V_M^*$.

It is important to note that the equal signs in the Proposition 10 refer to the cases where $q_D^* = q_M^* = \hat{q}$, in the other cases the signs are strict. As it is expected when the other firm is capable of stealing some of the demand both firms will increase their quality levels to gain back the customers that were taken by the other firm. Changes in the capacity decision is dependent on the relationship between the quality and capacity decisions. If the two decisions of the duopoly case are located on the side of complements of the monopoly balanced curve, then not only the quality levels will increase because of the competition,

but also the capacity levels will increase. However, when the decisions are located on the side of substitutes of the monopoly balanced curve, then competition will result in lower capacity levels. Although the capacity levels increase when the decisions are complements, also does the total number of the customers that are served by the firms. As a result the customers' average wait time increases in both cases. Regardless of the increase in wait time and quality all the customers will receive zero value from the service.

Please note that the threshold for complement for the duopoly case, $q_{i,min}$, is lower than the threshold for the monopoly case, q_{min}. Hence there are conditions under which the quality and capacity decisions are complements but the solution lies on the substitute area of the monopoly balanced curve. This also implies that we could have conditions in which the quality and capacity decisions are substitutes in the monopoly case but complements in the existence of competition, (however, the reverse is not possible). This result can be explained by the intuition that was used in the monopoly scenario. In the competition the firms could lose part of the customers due to competition so the effects of attracting customers becomes more pronounced than the increased value of service and hence the area of complement is relatively larger in competition scenario.

Although for the cases where the firms will allow some of the customers to balk, the symmetric solution is the unique equilibrium of the problem. For the cases where the solution lies on the balanced curve, the equilibrium is not necessarily symmetric, although we are unable to achieve the specific asymmetric equilibria. We can analyze the reactions of the firms with respect to the changes in the decisions of the other firm. Although Proposition 10 shows that the *threat* of competition will increase the quality levels compared to the monopoly case, the reaction of the firms to an increase in the quality of the competitor is unclear. To analyze these conditions we consider two non-identical firms, with optimal quality levels q_1 and q_2.

Lemma 8 provides some grounds for analyzing the relationship between the quality decisions for two nonidentical firms. We define $g(q_i|q_j)$ to be the balanced curve for firm i given the quality level of firm j, q_j. For this Lemma we define $\bar{q_i}(q_j)$, and $\underline{q_i}(q_j)$ as the two solutions to $\hat{\mu} = g(q_i|q_j)$, where $\underline{q_i}(q_j) \leq \bar{q_i}(q_j)$.

Lemma 8. *Let q_{Li} be the solution to $\beta''\left(g(q_i|q_j)\right)\frac{\mathrm{d}g(q_i|q_j)}{\mathrm{d}q_i} = \frac{p-\beta'(g(q_i|q_2))}{1-q}$ for $q \in \{q : (g(q|q_j) \leq \hat{\mu}_i\}$. Then $q_{Li}(q_j)$ exists and is unique. Furthermore, $q_{i,min}(q_j) < q_{Li}(q_j) \leq \bar{q}_i(q_j)$.*

For the cases where $q_1 = \hat{q_1}$ or $q_2 = \hat{q_2}$, comparative statics is uninteresting. In these cases the change in the parameter values of one of the firms will result in no change in the unconstrained solution of the other firm. Without loss of generality let's assume $q_1 = \hat{q_1}$. As long as the unconstrained solution is the optimal solution of firm 1, the changes in the parameters of firm 2 will have no effect on the optimal decisions of firm 1. The optimal decisions of firm 1 remain the same up to the point where the changes in the decisions of firm 2 are so drastic that the unconstrained solution does not lie below the *new* balanced curve of firm 1 anymore. Hence we limit out analysis to the cases where the optimal decision of the firm lies on the balanced curve.

Proposition 11. *Let q_{L1} be determined by Lemma 8 when q_1 and q_2 are on the balanced curve. If $q_1 > q_{L1}$ then q_1 and q_2 are strategic complements. Otherwise they are strategic substitutes.*

Proposition 11 provides some insights on the relationship between the values of the decision variables of the two firms. It specifically shows the limits of the values of quality under which the two decisions are strategic complements vs substitutes (see [82]). It is important to note that Proposition 11 determines that if the level of quality provided by the firm is low, as the competitor increases the quality, the firm will surrender some of the market to the competitor and reduces the costs by decreasing quality. On the other hand, when the quality level of the firm is high, the firm will compete for the market share by increasing the quality level. Please note that q_{L1} is always greater than $q_{i,min}$, so whenever the quality and capacity decisions are substitute for a firm, the quality decision is substitute to the quality decisions of the other firms as well.

3.6 Discussions and Numerical Studies

3.6.1 Special cost functions

So far we have used very general cost functions $\alpha(q)$ and $\beta(\mu)$. Next, we consider a family of power cost functions to further investigate the properties of (q^*, μ^*). Specifically, let $\alpha(q) = \frac{k}{(1-q)^l}, k > 0, l \geq 1$ and $\beta(\mu) = m\mu^2, m > 0$ for the rest of this section. For technical reasons we have limited the quality investment function to $l \geq 1$ and the capacity investment function to quadratic. In the Section 3.6.2 we show that the qualitative results continue to hold for $0 < l < 1$ and higher orders of the β function. To avoid the trivial case where the firm makes no profit, we further assume $k < p(\lambda_0 + w)$.

Lemma 9. *$\alpha(q) = \frac{k}{(1-q)^l}, k > 0, l \geq 1$ and $\beta(\mu) = m\mu^2, m > 0$ satisfy Lemma 5.*

Proposition 12. *Let $\alpha(q) = \frac{k}{(1-q)^l}, k > 0, l \geq 1$ and $\beta(\mu) = m\mu^2, m > 0$.*

(A) $\hat{\mu} = \frac{p}{2m}$ and $\hat{q}$ is the unique solution to $\frac{pcv}{(qv-p)^2} = \frac{lk}{(1-q)^{l+1}}$.

(B) If $\hat{\mu} \leq f(\hat{q})$, then $q^* = \hat{q}$ and $\mu^* = \hat{\mu}$. If $\hat{\mu} > f(\hat{q})$, q^* is the unique solution to

$$\begin{aligned} p \quad & (\lambda_0 + w) + 2mc\left[-\frac{\lambda_0 + w}{qv - p} + \frac{v(\lambda_0 + w)(1-q)}{(qv-p)^2} + \frac{cv(1-q)^2}{(qv-p)^3}\right] \\ = \quad & \frac{lk}{(1-q)^{l-1}} + \frac{2m(\lambda_0 + wq)(\lambda_0 + w)}{1-q}, \end{aligned} \tag{3.18}$$

and $\mu^* = f(q^*)$.

(C) There exists a threshold $\tilde{l} > 1$ such that $q^* \geq \hat{q} \geq q_{min}$ for $l \leq \tilde{l}$, and $a^* \leq \hat{q} < q_{min}$ for $l > \tilde{l}$.

Part (A) of the Proposition gives an explicit expression for $\hat{q}$ and $\hat{\mu}$; Part (B) then converts the unconstrained solution $(\hat{q}, \hat{\mu})$ to the constrained solution (q^*, μ^*). Part (C) confirms our earlier intuition that when the marginal investment cost in quality is bigger (i.e., when l is big) the investments in capacity and quality tend to be substitutes, and when it is smaller (i.e., when l is small) the investments tend to be complements. There exists a threshold on l that delineates this boundary of these two regions.

Please note that for the cost functions assumed, only parameters l and m satisfy the conditions in Definition 3. In this case we can define $S = \{l, m\}$.

3.6.2 Sensitivity Analysis

In this section, we report how the firm's quality and capacity strategy, profit vary as the initial customer arrival rate λ_0, customer valuation on the fully satisfied service v, and the price p change.

We first look at how the firm's quality and capacity change as the potential customer arrival rate increases. As λ_0 increases, $f(q)$ moves up in the $\mu - q$ space. Note that $(\hat{q}, \hat{\mu})$ remains the same as λ_0 changes.

From Figure 3.6 and Figure 3.8, we can see that the optimal capacity increases with λ and finally remains at $\hat{\mu}$.

Define $\bar{\lambda}_0 = (1-\hat{q})\left(\hat{\mu} - \frac{c}{\hat{q}v-p}\right) - wq$. That is, at $\bar{\lambda}_0$, $\hat{\mu} = f(\hat{q})$. For $\lambda_0 > \bar{\lambda}_0$, the optimal quality level and capacity level are $\hat{q}$, $\hat{\mu}$, respectively, and balking happens. As λ_0 increases above $\bar{\lambda}_0$ the balanced curve moves so that the unconstrained solution will lie below the balanced curve. Therefore, the firm's capacity level will remain at $\hat{\mu}$ for any $\lambda_0 > \hat{\lambda_0}$. Please note that $\hat{\mu}$ is the optimal capacity level without any restrictions and for any $\mu > \hat{\mu}$ the marginal benefit of capacity is negative. As we discussed earlier, the quality level and the capacity level must be balanced. Hence, the quality level will remain at $\hat{q}$.

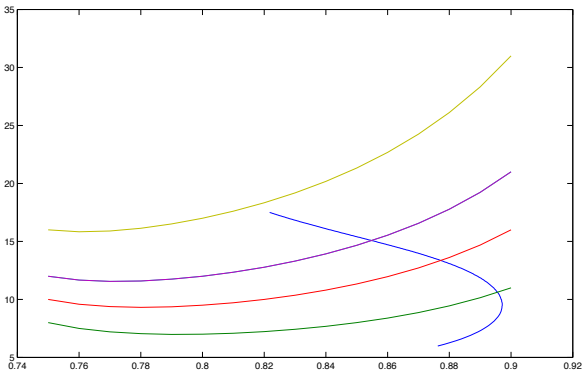


Figure 3.6: Optimal Quality and Capacity as λ Changes for Small l

As the potential arrival rate increases, our first thought would be that the firm may either

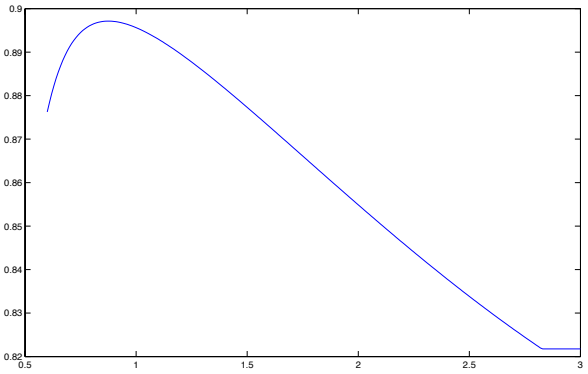


Figure 3.7: Optimal Quality as λ Changes for Small l

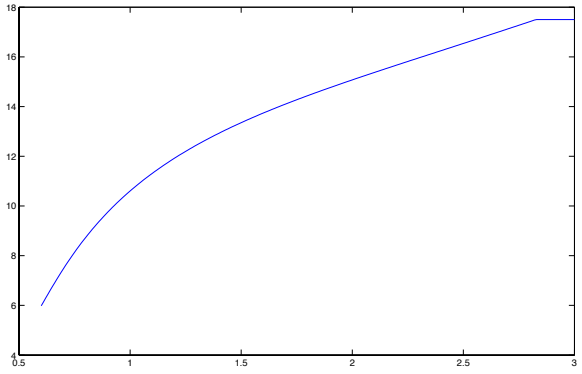


Figure 3.8: Optimal Capacity as λ Changes for Small l

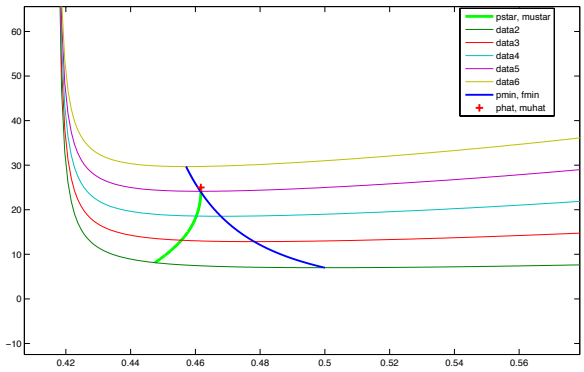


Figure 3.9: Optimal Quality and Capacity as λ Changes for Large l

increase the capacity level or increase the quality level to accommodate more customers. By increasing the capacity level, the firm may reduce the congestion caused by more customer arrivals. By increasing the quality level, the firm may increase customers' perceived value of the service and increase their patience (customers are willing to wait longer). Our numerical examples show that the firm will increase capacity as the potential arrival rate increases for $\lambda_0 < \bar{\lambda}_0$.

As customer's valuation v increases, it is clear that the service provider's profit increases as customers are willing to wait longer if they need to. Technically, as v increases, the curve $f(p)$ moves downward. Note that $\hat{\mu}$ does not change with v. It can be shown that $\hat{p}$ decreases with v.

Our numerical analysis shows that the optimal quality level does not monotonically change with v. The optimal quality level deceases with v and then increases with v. See an example shown in Figures 3.10 through 3.12. Figure 3.11 particularly shows how the quality decision changes as v increases. Figure 3.10 shows how the optimal set (q^*, μ^*) changes in the $q - \mu$ space as v changes.

The optimal capacity always decreases with v. See Figure 3.12.

It is also interesting to find that the optimal throughput also decreases with v and then increases with v. As customer's valuation v increases, customers are willing to wait for a

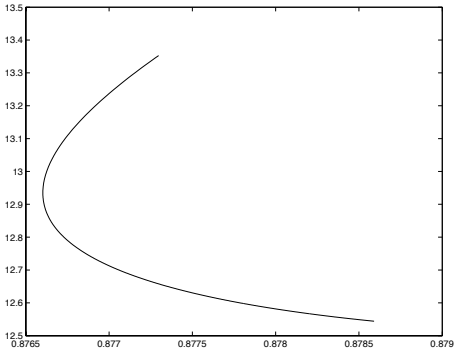


Figure 3.10: (q^*, μ^*) as v Changes

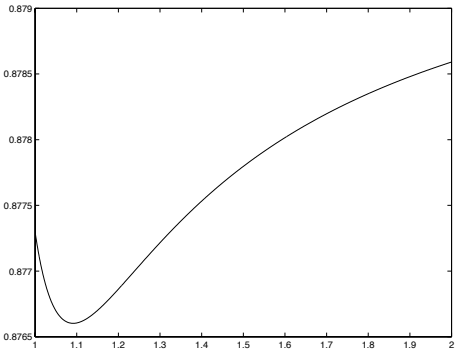


Figure 3.11: Optimal Quality as v Changes

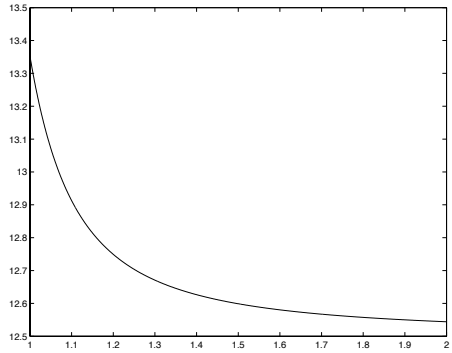


Figure 3.12: Optimal Capacity as v Changes

longer time. The service provider cuts down both capacity and quality investments first, then increases quality investment but decrease capacity investment.

As the revenue rate p increases, the service provider may increase the capacity investment. Our numerical analysis confirms this intuition. The quality level is also increasing in the revenue rate.

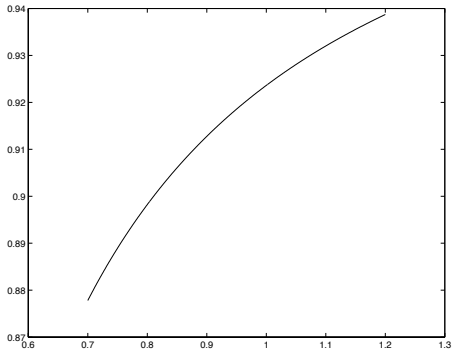


Figure 3.13: Optimal Quality as p Changes

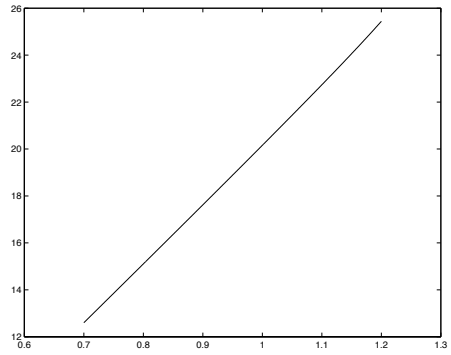


Figure 3.14: Optimal Capacity as r Changes

3.6.3 Complement Versus Substitute Decisions

We study both the continuity of the optimal decisions with respect to changes in $x \in S$ where in this case $S = \{l, m\}$. We can show that as l and m move smoothly the solution

moves smoothly and we can show the complementarity and substitution effects as well.

Base Model

For this case by increasing the values of either l and m in Figures 3.15 and 3.16 respectively. Please note that these figures show the optimal solution set (q^*, μ^*) on the $q - \mu$ space.

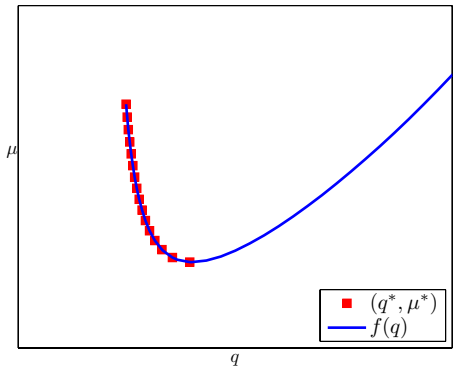


Figure 3.15: Optimal Decision as l Changes

As l increases, quality becomes more costly so the optimal quality level decreases. Figure 3.15 shows that as l increases in the substitutes scenario, the capacity level increases and the optimal solution moves on the balanced curve. Similarly, Figure 3.16 shows that as the capacity cost function increases the capacity level decreases and in the substitution region the quality level increases.

Although Figures 3.15 and 3.16 specifically show the continuity of the optimal solution only for a given set of parameters, our numerical analysis shows that as long as the parameters are set so that $\hat{\mu} > f(\hat{q})$, the continuity holds and the solution moves on the balanced curve.

Price as a Decision

In this case we also investigate the changes in the optimal decision set as l and m change. In this case price is also a decision variable. It is clear that although in the previous case

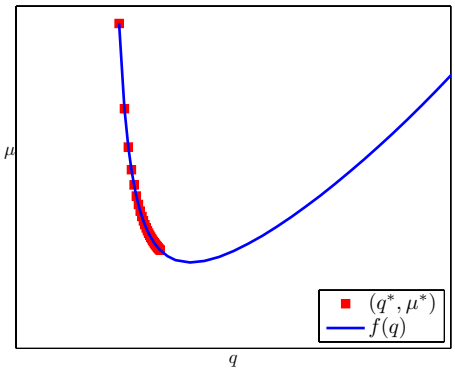


Figure 3.16: Optimal Decision as m Changes

the quality and capacity decision were substitutes here they are always complements. As l

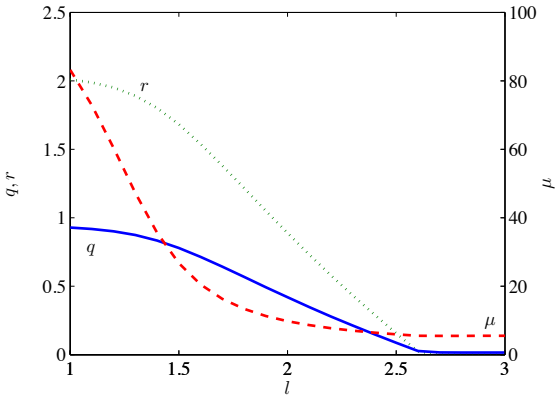


Figure 3.17: Optimal Decision as l Changes

increases, quality becomes more costly so the optimal quality level decreases. Figure 3.17 shows that as l increases, the capacity level decreases with quality. In this scenario parameters are such that, price is also complements and as it is clear from Figure 3.17 price decreases with the capacity and quality as well.

Similarly, in as the cost of capacity increases, both capacity and quality levels will

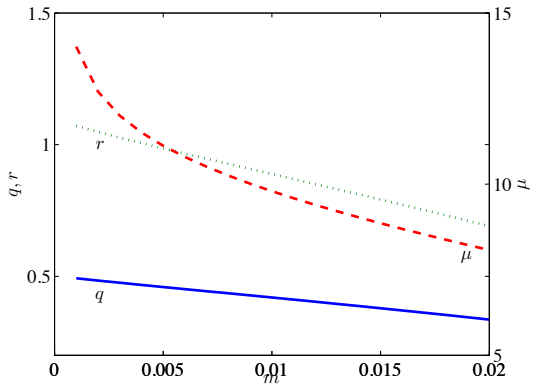


Figure 3.18: Optimal Decision as m Changes

decrease, and again as the conditions for complementarity of price hold, price also decreases (see Figure 3.18).

In the extensive numerical analysis that we performed we investigated the results of Theorem 9 and show that quality and capacity decisions are always complements when price is also considered a decision.

3.6.4 Benchmark Case: without Consideration of Repeat Purchasing Behavior

This subsection briefly studies the situation where the firm ignores customers' repeat purchasing behavior and show how the firm sets up the optimal quality level and the corresponding capacity level, denoted by q^N and μ^N, respectively (superscript N means "no repeat purchasing behavior"). Denote the throughput as λ_e^N.

Lemma 10. *If* $\mu \geq \lambda_0 + wq + \frac{c}{qv-p}$, $\lambda_e^N = \lambda_0 + wq$. *Otherwise,* $\lambda_e^N = \mu - \frac{c}{qv-p}$. *In short,* $\lambda_e^N = \min\{\lambda_0 + wq, \mu - \frac{c}{qv-p}\}$.

Let $f^N(q) = \lambda_0 + wq + \frac{c}{qv-p}$. $f^N(q)$ is convex and is decreasing for small q. If $w \leq \frac{cv}{(v-p)^2}$, $f^N(q)$ is decreasing. Otherwise, $f^N(q)$ is decreasing first and then increasing. Note that

$f^N(q) < f(q)$ for any $q \in (0,1)$. The firm's profit function is as follows:

$$V^N(q,\mu) = \begin{cases} p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu), & 0 < \mu \le f^N(q), \\ p(\lambda_0 + wq) - \alpha(q) - \beta(\mu), & \mu > f^N(q). \end{cases} \tag{3.19}$$

We can find the optimal quality level and capacity in this case using the similar approach as we do in our base model. Comparing equation (3.19) with equation (3.2), we can see that $V^N(q,\mu)$ has the same expression as $V(q,\mu)$ when $\mu \le f^N(q)$. $\hat{q}$ and $\hat{\mu}$ defined in Section 3.2.2 are still technically critical here. We summarize the "optimal" decision when the firm ignores customer repeat purchasing behavior in the following lemma.

Lemma 11.

(i) If $\hat{\mu} \le f^N(\hat{q})$, $q^N = \hat{q}$, $\mu^N = \hat{\mu}$.

(ii) If $\hat{\mu} > f^N(\hat{q})$, q^N *is the solution of the following equation:*

$$pw - \alpha'(q) - \beta'(f^N(q))\left[w - \frac{cv}{(qv-p)^2}\right] = 0. \tag{3.20}$$

$\mu^N = \lambda_0 + wq^N + \frac{c}{q^N v - p}, \mu^N \le \hat{\mu}$. *If* $w \le \frac{cv}{(\hat{q}v-p)^2}$, $q^N \le \hat{q}$. *Otherwise,* $q^N > \hat{q}$.

Lemma 11 indicates that if $(\hat{q},\hat{\mu})$ is below the curve $f^N(q)$, i.e., in the area I in Figure 3.19, then the firm would set the optimal quality and capacity level at $(\hat{q},\hat{\mu})$. In this case, customer balking happens and the throughput $\lambda_e^N = \mu - \frac{c}{qv-p}$. If $(\hat{q},\hat{\mu})$ is above the curve $f^N(q)$, i.e., in the area II and area III, the quality and capacity level that the firm would set up is on the curve $\mu = f^N(q)$. In this case, no customer balks and the throughput $\lambda_e^N = \lambda_0 + wq$. q^N can actually be obtained through the first order condition since $p(\lambda_0 + wq) - \alpha(q) - \beta(f^N(q))$ is concave in q. Although the capacity level must be lower than $\hat{\mu}$, the quality level may be higher or lower than $\hat{q}$, depending on the reputation sensitivity w. In particular, if the reputation sensitivity is high ($w > \frac{cv}{(\hat{q}v-p)^2}$), the quality level $q^N > \hat{q}$.

Recall that when the firm does consider customers repeat purchasing behavior, if $(\hat{q},\hat{\mu})$ is in area I and area II in Figure 3.19, the firm will set the quality level and capacity level

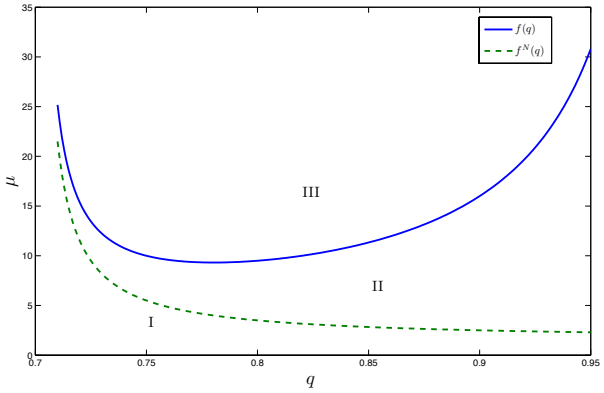


Figure 3.19: No Repeat

at $\hat{q}$, $\hat{\mu}$, respectively. The corresponding throughput $\lambda_e = \mu - \frac{c}{qv-p}$. So, if $(\hat{q}, \hat{\mu})$ is in area I, ignoring the repeat purchasing behavior does not affect the quality and capacity level and the throughput. The firm makes the same profit as it would when it does consider the customer repeat purchasing behavior.

If $(\hat{q}, \hat{\mu})$ is in area II in Figure 3.19, $\mu^N < \mu^* = \hat{\mu}$, that is, the firm underinvests in capacity if the firm ignores customer repeat purchasing behavior. Depending on the value of the reputation sensitivity w, the firm may overinvest or underinvest in quality (see Lemma 11). That is, if $w < \frac{cv}{(\hat{q}v-p)^2}$ ($w > \frac{cv}{(\hat{q}v-p)^2}$), the firm underinvests (overinvests) in quality regardless of the value of the base arrival rate λ_0. (Note that $\hat{q}$ is independent of the base arrival rate λ_0.)

If $(\hat{q}, \hat{\mu})$ is in area III in Figure 3.19, when $w < \frac{cv}{(\hat{q}v-p)^2}$, we can show that $q^N < q^*$. If $w > \frac{cv}{(\hat{q}v-p)^2}$, our numerical examples show that depending on the base arrival rate λ_0, the firm may underinvests or overinvests in quality. Figures 3.20, 3.21, 3.22 illustrate three different scenarios in which $q^N < q^*$, $q^N = q^*$ and $q^N > q^*$, respectively. The three scenarios have the same parameters except the base arrival rate λ_0: $p = 0.7, c = 0.2, p = 1.2, m = 0.01, k = 0.1, l = 1, w = 5$. In these examples, $\hat{q} = 0.8$ and $w > \frac{cv}{(\hat{q}v-p)^2}$.

Without considering the customer repeat purchase behavior, the firm always under invest in capacity. So even when the firm overinvests in quality, because of shortage of capacity,

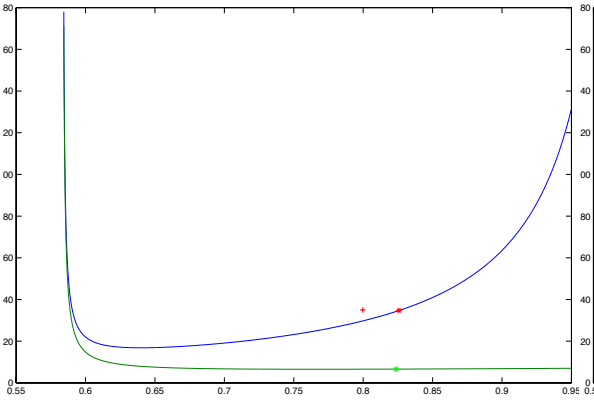


Figure 3.20: $q^N < q^*$ with $\lambda_0 = 1.8$

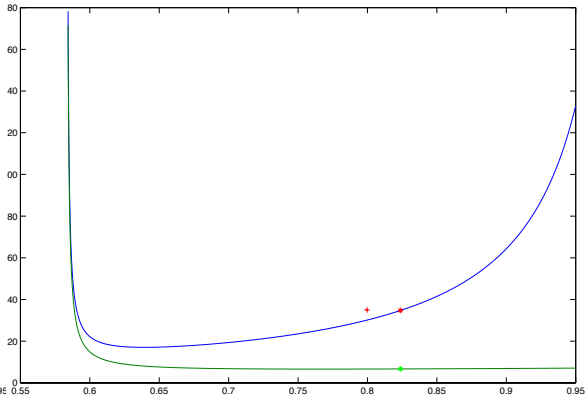


Figure 3.21: $q^N = q^*$ with $\lambda_0 = 1.88$

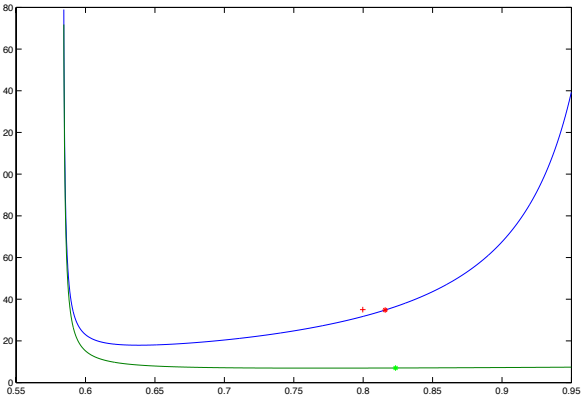


Figure 3.22: $q^N > q^*$ with $\lambda_0 = 2.2$

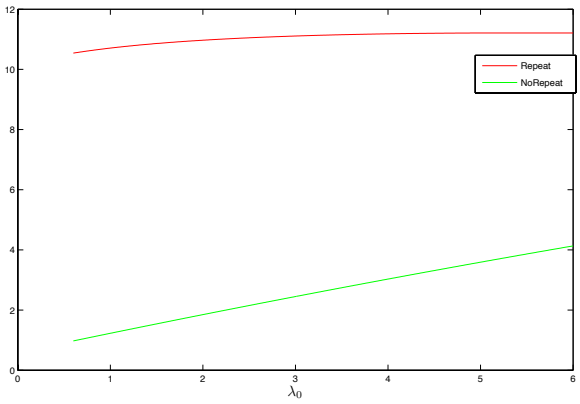


Figure 3.23: The firm's profits as the base arrival rate λ_0 varies

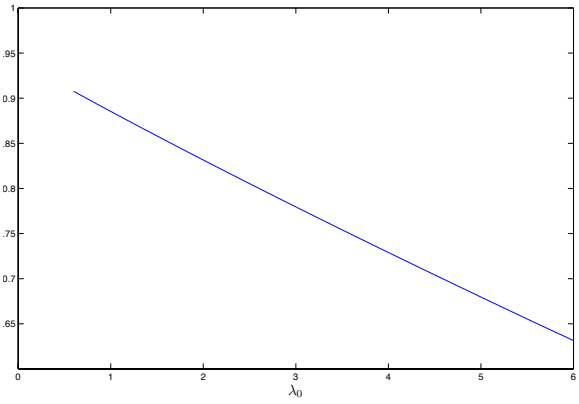


Figure 3.24: The percentage of lost profit due to ignoring customer repeat purchase

more customers balk and the firm ends up with smaller profit.

To investigate the effect of ignorance of customer repeat purchasing behavior on firm's profit as well as on customers' experiences, we run extensive numerical experiments and get some interesting findings.

Observation 1: When the base arrival rate λ_0 is *smaller*, ignoring customer repeat purchase behavior results in *larger* profit loss. Figure 3.23 illustrates firm's profits in the model with considering customer repeat purchase behavior and the model ignoring customer repeat purchase behavior as the base arrival rate changes. Figure 3.24 shows the percentage of lost profit due to ignoring customer repeat purchase behavior. (The parameters of these two figures are as follows: $p = 0.7, c = 0.2, v = 1.2, m = 0.01, k = 0.1, l = 1, w = 2$ and λ_0 varies from 0.6 to 6.)

This observation is intuitive. As the base arrival rate is small, the firm's profit is constrained by the customer arrivals and it is economically favorable to invest in a relatively high capacity level (i.e., $\hat{\mu}$). Due to the ignorance of customer repeat purchase behavior, the quality level is set lower than it should be and then the customer arrival rate is lower.

Observation 2: When customers are less sensitive to the quality, i.e., w is **smaller**, ignoring customer repeat purchase behavior results in **larger** profit loss.

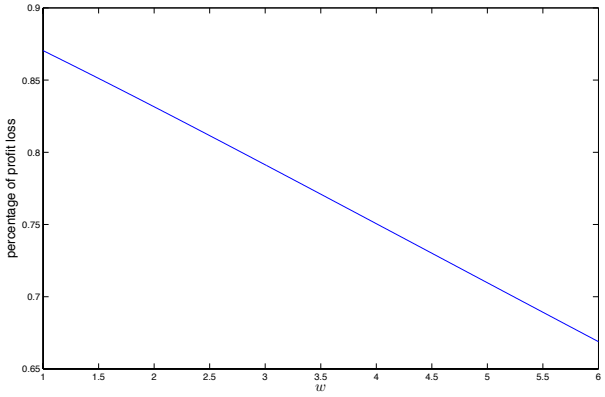


Figure 3.25: The percentage of lost profit due to ignoring customer repeat purchase as w varies

Remark 2. *The service provider makes decision based on the benchmark model. However, the market mechanism is still working like the real model.*

3.6.5 The effect of competition

In this subsection, we report the effect of competition on two firms' profits, quality and capacity decisions as well as customers' experiences.

We first vary the base arrival rate λ_0, and find that due to competition, more customers may get served. In Figure 3.26, the green curve represents the firm's throughput in the competition model while the red curve represents the firm's throughput in the comparative monopoly model. From this figure, we can see that the as λ_0 is below certain value, the throughput in the competition model is higher. That is, more customers get served when there is competition.

Figure 3.27 reports the ratios of firm's profit, quality and capacity in the competition model over firm's profit, quality and capacity, respectively. From this figure, we can see that as λ_0 increases, the profit loss due to competition decreases (i.e., the ratio increases) and the firm's quality and capacity level decrease. Moreover, the change of quality level is relative slower than the change of capacity.

As we discussed earlier in the competition model, b can be interpreted as the competi-

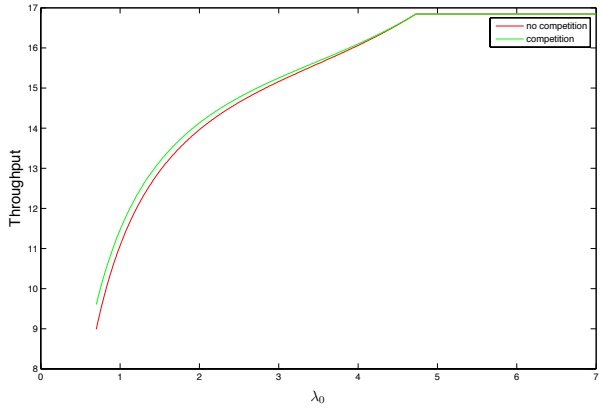


Figure 3.26: Firm's throughput changes in monopoly and competition models as λ_0 varies

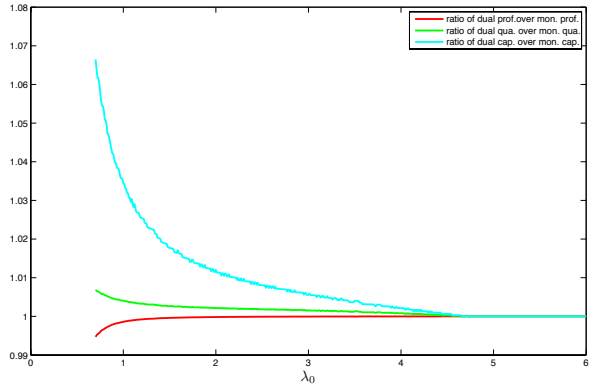


Figure 3.27: Ratios of firm's profit, quality and capacity in competition model over firm's profit, quality and capacity in monopoly model as λ_0 varies

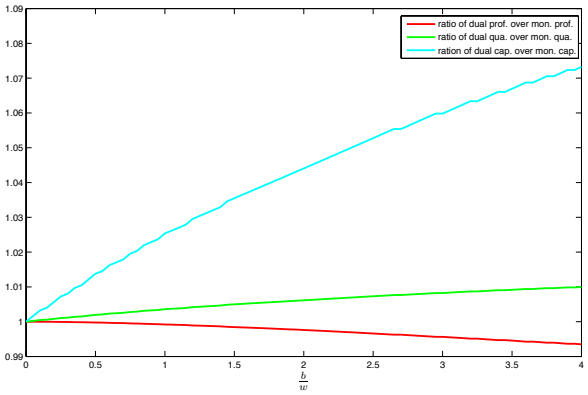


Figure 3.28: Ratios of firm's profit, quality and capacity in competition model over firm's profit, quality and capacity in monopoly model as λ_0 varies

tion intensity. We run numerical experiments to see how b affects firms' quality, capacity decisions and profit. Figure 3.28 shows the ratio of firms' profit(quality, capacity) in the competition model over firms' profit(quality, capacity) in the monopoly model as $\frac{b}{w}$ varies. Unsurprisingly, as $\frac{b}{w}$ increases, firms' profit decreases in the competition model. Again, we notice that as competition intensity varies, firms' quality level change is relative smaller than the change of capacity.

Strategic Complements and Substitutes

In this case we also investigate the changes in the optimal decision set as l changes. Please note that two scenarios are possible. Either the conditions of Theorem 7 will result in the decisions of the two firms to be substitutes or complements. The result for changes in m are also similar.

We first investigate the scenarios in which the optimal decisions of firm 1 q_1 is substitutes with the optimal decision of firm 2 q_2. As l_2 increases, quality becomes more costly for firm 2 so the optimal quality level of firm 2 decreases. On the other hand firm 1 will increase its quality level to capture the market that firm 2 loses. Figures 3.29 and 3.30 shows that as l_2 increases, how the quality and capacity decisions of the two firms will change. Please note

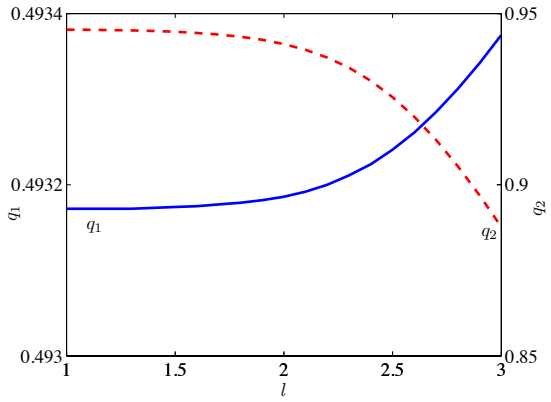


Figure 3.29: Strategic Substitutes: Optimal Quality as l_2 Changes

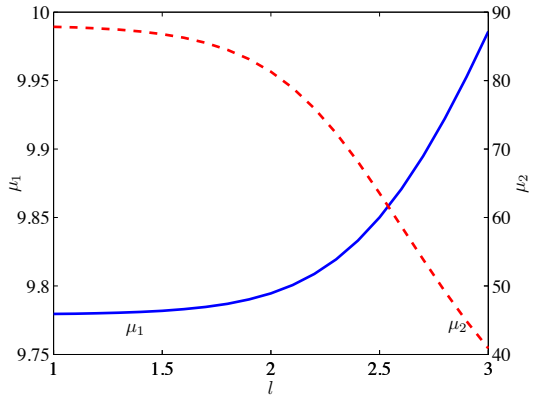


Figure 3.30: Strategic Substitutes: Optimal Capacity as l_2 Changes

that although the optimal solution of firm 1 lies on the substitution area of $g(q_1|q_2)$ as q_2 changes so does the balanced curve and hence in this case, capacity decision increases with the quality decision.

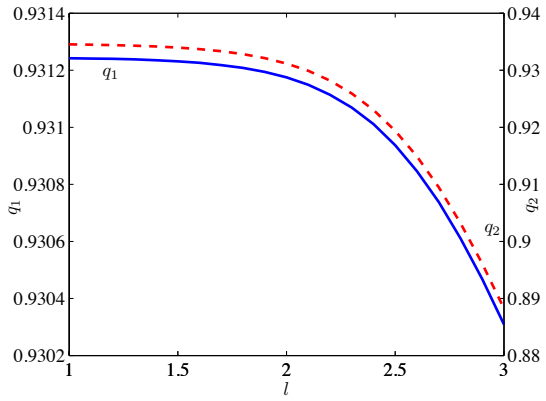


Figure 3.31: Strategic complements: Optimal Quality as l_2 Changes

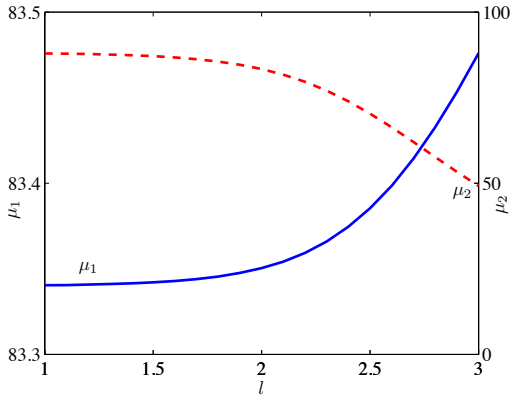


Figure 3.32: Strategic complements: Optimal Capacity as l_2 Changes

We now investigate the scenarios in which the optimal decisions of firm 1 q_1 is complements with the optimal decision of firm 2 q_2. As l_2 increases, quality becomes more costly for firm 2 so the optimal quality level of firm 2 decreases. As a result firm 1, will increase its

quality level as the competition intensity decreases. Figures 3.31 and 3.32 shows that as l_2 increases, how the quality and capacity decisions of the two firms will change. Please note that in this case the optimal solution of firm 1 lies on the complements area of $g(q_1|q_2)$ but still as q_2 changes so does the balanced curve but in this case, the result is so that capacity decision of firm 1 increases with the quality decision of firm 1.

We investigate different parameter sets and can show that the results follow the general intuitions of the one of the cases above.

3.7 The Capacity-Quality Tradeoff

In our model investments in capacity and quality are separate decisions. Nevertheless, the balanced curve $\mu = f(q)$ and Proposition 12 clearly demonstrate an implicit tradeoff between the two decision variables. In this section, we will investigate the impact of an explicit relationship (constraint) between these two decision variables. For example, there may exist a budget constraint so that the investments in capacity and quality negatively limit each other. As another example, [7] and [51] both study the quality-speed tradeoff where a higher service speed leads to higher service capacity but also lower service quality.

To model these possible relationships between capacity and quality, we impose an additional constraint $\mu \leq R(q)$ or $\mu = R(q)$. In the budget example, it is not necessary to always invest to the limit so the inequality relationship is more appropriate. When the direct relationship between q and μ is clear, then the equality constraint should be used. In either case, we assume the following.

Assumption 2. *$R(q)$ is concave decreasing in q.*

That $R(q)$ is decreasing is quite natural and easily satisfied by either the budget or the quality-speed constraint. The concavity assumption is common in literature. The linear quality-speed tradeoff studied in [7] and [51] are special cases. In the budget scenario, let b be the total budget limit; then $\alpha(q) + \beta(\mu) \leq b$ leads to $\mu \leq \beta^{-1}(b - \alpha(q))$. As long as both $\alpha(\cdot)$ and $\beta(\cdot)$ are convex increasing, the concavity will also be satisfied. For the rest of this section we assume that Assumption 2 holds. We will analyze $\mu \leq R(q)$ first, and then $\mu = R(q)$.

3.7.1 $\mu \leq R(q)$

With the additional constraint the problem becomes:

$$\max_{q,\mu} \quad V^I(q,\mu) \;=\; p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu) \tag{3.21}$$

$$s.t. \quad p/v < q \leq 1 \tag{3.22}$$

$$0 \leq \mu \leq f(q) \tag{3.23}$$

$$0 \leq \mu \leq R(q). \tag{3.24}$$

We continue to denote the optimal solution to (3.21)-(3.23) by (q^*, μ^*). In addition, we denote the optimal solution to (3.21)-(3.24) by (q^{**}, μ^{**}). Clearly, when (q^*, μ^*) satisfies (3.24) we have $(q^{**}, \mu^{**}) = (q^*, \mu^*)$. In the analysis below we focus on the more interesting case of $\mu^* > R(q^*)$.

Because $f(q)$ is convex and $R(q)$ is concave, they intersect at most twice on the (q, μ) space. When they have zero or one intersection point, the concave curve $R(q)$ lies below the convex curve $f(q)$: $R(q) \leq f(q)$, $\forall q \in (p/v, 1]$. When they have two intersection points, we denote the intersections by $q_L < q_H$. We call these two cases Case 1 and Case 2 respectively, and show them in Figure 3.7.1 and Figure 3.33 separately.

In Case 1, we solve the problem (3.21, 3.22, 3.24). That is, we replace $f(q)$ in the original problem by $R(q)$. Since $\mu^* > R(q^*)$, from Proposition 4 we know $(\hat{q}, \hat{\mu})$ must be above the $R(q)$ curve also (i.e., $\hat{\mu} > R(\hat{q})$). Moreover, with a proof similar to that of Proposition 4, we can show that (q^{**}, μ^{**}) must be on the $R(q)$ curve (i.e., $\mu^{**} = R(q^{**})$). Hence, it suffices to optimize along the $R(q)$ curve, where the objective function $V^I(q, \mu)$ simplifies to

$$V_R(q) = pR(q) - \frac{pc}{qv-p} - \alpha(q) - \beta(R(q)). \tag{3.25}$$

Lemma 12. *$V_R(q)$ is concave for $q \in [R^{-1}(\hat{\mu}), 1)$ and has a unique maximizer for $q \in (p/v, 1)$.*

We denote this unique maximizer by $q_R^* = \arg\max_{v/p<q<1} V_R(q)$ and $\mu_R^* = R(q_R^*)$.

Proposition 13. *Suppose $\mu^* > R(q^*)$ and $R(q) \leq f(q)$, $\forall q \in (p/v, 1]$.*

- *$q^{**} = q_R^*$ and $\mu^{**} = \mu_R^*$.*

- $q^{**} \leq q^*$ *and* $\mu^{**} \leq \mu^*$.

Proposition 13 not only shows that when the additional $R(q)$ constraint is non-trivial, the optimal point (q^{**}, μ^{**}) resides on the $R(q)$ curve, it is also dominated by the original optimal point (q^{**}, μ^{**}) (i.e. in Figure 3.7.1, it must lie between points X and Y). This makes intuitive sense because if the additional constraint makes the original optimal solution infeasible, one would expect that investment in *both* capacity and quality to be scaled back. A similar observation will be made in Case 2 (i.e., $R(q)$ intersects with $f(q)$) only if q^* is large. When q^* is small, however, it may be optimal to decrease capacity but *increase* quality in order to satisfy the additional constraint. We will make these statements precise in Proposition 14.

Now we turn our attention to Case 2 where $f(q)$ and $R(q)$ intersect twice. From Proposition 4 we know that (q^*, μ^*) can only lie in Region I of Figure 3.2; we further divide it into three areas in Figure 3.33. In area 2, the additional constraint is not binding so the original solution remains optimal; we trivially have $q^{**} = q^*$, $\mu^{**} = \mu^*$. Analysis of the other two areas are more intricate and interesting. We summarize results for all three areas in the following proposition:

Proposition 14. *Consider* $\mu^* > R(q^*)$ *and Case 2 where* $R(q)$ *and* $f(q)$ *intersect twice at* $q_L < q_H$.

(i) $\mu^{**} = R(q^{**})$.

(ii) When $q^* \geq q_H$, $q^{**} = \max\{q_H, q_R^*\}$. *Moreover,* $q^{**} \leq q^*$ *and* $\mu^{**} \leq \mu^*$.

(iii) When $q^* \leq q_L$, $q^{**} = \min\{q_L, q_R^*\}$. *Moreover,* $q^{**} \leq \hat{q}$ *and* $\mu^{**} \leq \mu^*$.

The results in Proposition 14 are quite straightforward. First, the additional constraint $\mu \leq R(q)$ must be binding. Next, if the original optimal solution lies in area 1 or 3 of Figure 3.2 Case 2 (items (i) and (ii) in Proposition 14), then the constrained optimal solution must remain in the same area (hence the min and max operators).

The more important part of Proposition 14 is how the investment levels change from (q^*, μ^*) to (q^{**}, μ^{**}) due to the additional constraint. We are able to show that when the q^*

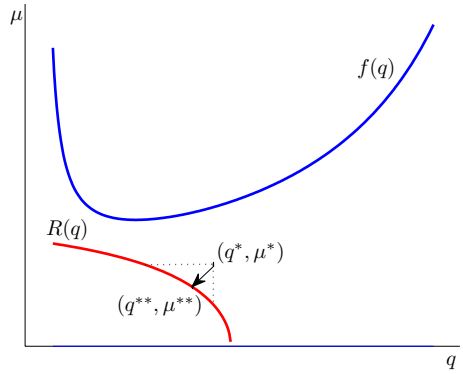


Figure 3.33: $R(q)$ and $f(q)$ have no intersections

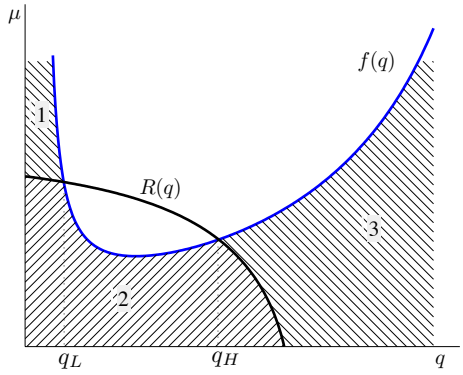


Figure 3.34: $R(q)$ and $f(q)$ intersect twice

is large (i.e., point (ii) above), the additional constraint causes the firm to reduce investment in both quality and capacity. When q^* is low (i.e., point (iii) above), we know capacity will be reduced, but quality may not be. We only know $q^{**} \leq \hat{q}$, and since q^* is also no more than $\hat{q}$ (Theorem 7), q^{**} can be either higher or lower than q^*. This has to do with the fact that at low quality level, quality and capacity are substitutes. Therefore, when investments need to be scaled back, depending on the shape of the various functions, it may make sense to increase the investment in one variable just so that the other variable can be scaled back even more. It is interesting, however, to observe that it is quality that may be increased. Capacity, on the other hand, is always reduced. This echos the observation we made in Theorem 7, albeit in a different context. Once again, we beleive these phenomena have to do with all the various effects of service quality that we model.

Special Case: Budget Constraint

As an illustration, we consider the a special case where the additional constraint is a budget constraint $\alpha(q)+\beta(\mu) \leq b$, where b is the maximum amount that can be invested in capacity and quality. To obtain sharper analytical results, we will also use the special cost functions discussed in §3.2.2. That is:

$$\alpha(q) = \frac{k}{(1-q)^l} \ (l \geq 1), \qquad \beta(\mu) = m\mu^2, \qquad \mu \leq R(q) = \sqrt{\frac{b}{m} - \frac{k}{m(1-q)^l}}.$$

Due to budget b, we must have $q \leq q_b \stackrel{\Delta}{=} 1 - \sqrt[l]{\frac{k}{b}}$. It can be verified that $R(q)$ is concave and decreasing in q. Moreover, $V_R(q)$ from (3.25) can be simplified to:

$$V_R(q) = p\sqrt{\frac{b}{m} - \frac{k}{m(1-q)^l}} - \frac{pc}{qv-p} - b. \tag{3.26}$$

It can be shown that $\lim_{q\to\frac{p}{v}} \frac{dV_b(q)}{dq} = +\infty$ and $\lim_{q\to q_b} \frac{dV_b(q)}{dq} = -\infty$. Therefore the maximum must be achieved in the interior and satisfy the first order condition:

$$\frac{dV_R(q)}{dq} = -\frac{plk}{2m(1-q)^{l+1}\sqrt{\frac{b}{m} - \frac{k}{m(1-q)^l}}} + \frac{pcv}{(qv-p)^2} = 0.$$

from which we obtain

$$\mu_R^* = \frac{lk(q_R^* v - p)^2}{2mcv(1-q_R^*)^2} \stackrel{\Delta}{=} b(q_R^*). \tag{3.27}$$

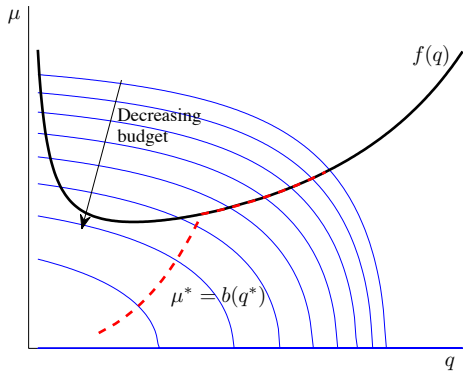


Figure 3.35: Optimal Quality and Capacity as Budget Varies

Figure 3.35 illustrates how the optimal investments change with the available budget. We see that $(\hat{q}, \hat{\mu})$ is above the balanced curve and is mapped to the right for the optimal (q^*, μ^*) on the balanced curve. Now impose the budget constraint. As long as b is above $\alpha(q^*) + \beta(\mu^*)$, it's not a binding constraint and $(q^{**}, \mu^{**}) = (q^*, \mu^*)$. The optimal investments are not influenced by budget. When b drops below that level, however, the investments must be scaled down. In this example, q^* is high, so both q^* and μ^* decrease with b, as Proposition 14 (ii) indicates. Note that during this adjustment, (q^{**}, μ^{**}) equals $(q_H, f(q_H))$, thus remaining on the balanced curve. When b decreases further, (q^{**}, μ^{**}) starts to leaves the balanced curve to to on the $b(q)$ curve. This is to be expected from Proposition 14, as (q^{**}, μ^{**}) should equal either (q_R^*, μ_R^*) or $(q_H, f(q_H))$. Finally, as b becomes very small, $R(q)$ and $f(q)$ no longer intersect, and Proposition 13 indicates that $(q^{**}, \mu^{**}) = (q_R^*, \mu_R^*)$ and it follows the $b(q)$ curve. In other words, when budget is not a concern, the firm's investments in capacity and quality must be operationally balanced, but when budget is tight the investment levels must be dictated by available budget.

The point from which (q^{**}, μ^{**}) switches from $(q_H, f(q_H))$ to (q_R^*, μ_R^*) is interesting. From Figure 3.35 we see that when b is slightly below that point, $R(q)$ and $f(q)$ still have two intersection points and it is possible to stay on the balanced curve, but the optimal investments should deviate from the balanced curve and choose (q_R^*, μ_R^*) over $(q_H, f(q_H))$.

Since $q_R^* > q_H$ and $\mu_R^* < f(q_H)$), essentially the firm opts to invest more in quality and less in capacity, and the system is not balanced – customer balking starts to occur. In this case the firm choose to serve fewer customers but serve them with higher quality.

3.7.2 $\mu = R(q)$

We have studied the situation where the $q - \mu$ relationship is defined by $\mu \leq R(q)$. Now we consider the situation where there is a one-to-one relationship between q and μ, i.e., $\mu = R(q)$.

In the case where $\mu \leq R(q)$, we have seen that the optimal quality and capacity level will be on the constraint boundary if $\mu^* \geq R(q^*)$. Therefore, for the problem with constraint $\mu = R(q)$, when $\mu^* \geq R(q^*)$, the optimal quality level and capacity level are the same as in the case with constraint $\mu \leq R(q)$.

When $\mu^* < R(q^*)$, the two types of constraints have different solutions. Apparently, (q^*, μ^*) is the optimal solution when the constraint is the type of $\mu \leq R(q)$. But for the constraint $\mu = R(q)$, it is not. We can still study two situations separately: 1) $\mu = R(q)$ and $\mu = f(q)$ do not intersect with each other on the $\mu - q$ space; 2) $\mu = R(q)$ and $\mu = f(q)$ intersect with each other at two points on the $\mu - q$. In case 1), the optimal quality level can be found by solving the following problem:

$$\max_{\frac{p}{v} < q \leq R^{-1}(0)} pR(q) - \frac{pc}{qv - p} - \alpha(q) - \beta(R(q)). \tag{3.28}$$

Due to Assumption 2, we can find the unique optimal quality level using the first order condition, denoted as q_R^*. The optimal capacity level is $R(q_R^*)$.

In case 2), the profit function on the constraint in the interval $[q_L, q_H]$ is

$$\frac{p(\lambda_0 + w)}{1 - q} - \alpha(q) - \beta(R(q)) - pw, \tag{3.29}$$

which is continuous. Thus, there must exist a maximizer over the interval $[q_L, q_H]$. Denote this maximizer as q^E (the superscript E indicates that no customer would balk and there is even extra capacity).

The profit function on the constraint outside of the interval $[q_L, q_H]$ is $pR(q) - \frac{pc}{qv-p} - \alpha(q) - \beta(R(q))$. We need to find the maximizer for this function outside of the interval

$[q_L, q_H]$, denoted as q^B (the superscript B indicates that there is customer balking).

The higher of the two profit values achieved at $(q^E, R(q^E))$ and $(q^B, R(q^B)$ will thus be the overall optimal profit for the firm. Numerical test results suggest that either can be the optimal value.

When the quality level and capacity level has a one-to-one relationship, investing in one parameter (either the quality level or the capacity level), the other will be fixed. In the following, we study a model where $\mu = R(q)$ and there is only quality investment cost. The firm's profit function $V(q)$ can be expressed as follows:

$$V(q) = \min\left\{pR(q) - \frac{pc}{qv-p} - \alpha(q), \frac{p(\lambda_0 + wq)}{1-q} - \alpha(q)\right\}. \tag{3.30}$$

If $R(q)$ and $f(q) = \frac{\lambda_0+wq}{1-q} + \frac{c}{qv-p}$ do not intersect with each other, $V(q) = pR(q) - \frac{pc}{qv-p} - \alpha(q)$, which is concave in q (we assume $R(q)$ is concave).

If $R(q)$ and $f(q)$ intersect with each other at two points, say, at q_L, q_H, then

$$V(q) = \begin{cases} pR(q) - \frac{pc}{qv-p} - \alpha(q), & q \notin (q_L, q_H) \\ \frac{p(\lambda_0+w)}{1-q} - \alpha(q) - pw, & q \in [q_L, q_H] \end{cases} \tag{3.31}$$

3.7.3 Budget constraint in duopoly model

How would the two firms set up the quality and capacity levels when they are budget constraints? It is interesting that when the two firms are budget constrained, the equilibrium results are different. In particular, when the two firms have the same budget level and the budget is short so that the firms cannot implement q_D, μ_D derived above, in the equilibrium, the two firms would set up the the same quality and capacity as in the comparable monopoly model.

Proposition 15. *If $\alpha(q_D) + \beta(\mu_D) > B$, let (q_M^B, μ_M^B) be the quality and capacity decision in the comparable monopoly model (in which $\lambda(q) = \lambda_0 + (a-b)q$) with the same budget level B. Let (q_D^B, μ_D^B) be the quality and capacity decision in the duopoly model. Then $q_D^B = q_M^B,\ \mu_D^B = \mu_M^B$.*

Proposition 15 shows that the two firms set up the same quality and capacity levels as in the benchmark case with budget constraint when the two firms have the same short budget

level. However, if the two firms have different budgets while all the other parameters remain the same, the equilibrium results are different. Denote B_1, B_2 as the firm 1 and firm 2's budget levels, respectively. Without loss of generality, we assume that $B_1 < B_2$ Denote firms' decisions in the equilibrium as $(q_D^{B_1}, \mu_D^{B_1})$, $(q_D^{B_2}, \mu_D^{B_2})$, respectively. In this situation, we consider two comparable monopoly models. The first comparable case, corresponding to firm 1 in the competition model, the arrival rate function is $\lambda(q) = \lambda_0 + (a-b)q$ and the budget level is B_1. Similarly, the second comparable case corresponds to firm 2 in the competition model, in which the arrival rate is $\lambda(q) = \lambda_0 + (a-b)q$ and the budget level is B_2. Denote the optimal decisions in these two monopoly models as $(q_M^{B_1}, \mu_M^{B_1})$, $(q_M^{B_2}, \mu_M^{B_2})$, respectively.

As far as the budget levels are concerned, if $\alpha(q_D) + \beta(\mu_D) \leq B_1$, then both firms will implement (q_D, μ_D) derived above in the case without budget constraint. Here, we mainly study the situation where at least one firm is budget constrained in the competition. Therefore, we study the situation where $\alpha(q_D) + \beta(\mu_D) > B_1$.

According to our numerical analysis, in most scenarios, the optimal quality level is non-decreasing in the budget level. For example, if $q^* > q_{min}$, then the optimal quality level is non-decreasing in the budget level. Therefore, we consider this case below for the competition model with different budget levels.

Lemma 13. *Under the assumptions that $B_1 < B_2$ and $\alpha(q_D) + \beta(\mu_D) > B_1$, $q_D^{B_1} < q_D^{B_2}$ in the equilibrium.*

3.7.4 Numerical analysis for Budget Constraint

In this subsection, we explore the effect of budget constraint on firm's profit as well as customers' experiences through extensive numerical experiments.

Figure 3.36 illustrates how the optimal quality and optimal capacity vary with the budget level through an example. There are two main segments of the quality-capacity changing track. First, the curve is below curve $\mu = f(q)$ and balking happens. As the budget level increases the quality level and capacity level vary along the curve $\mu = f(q)$ and no balking happens then.

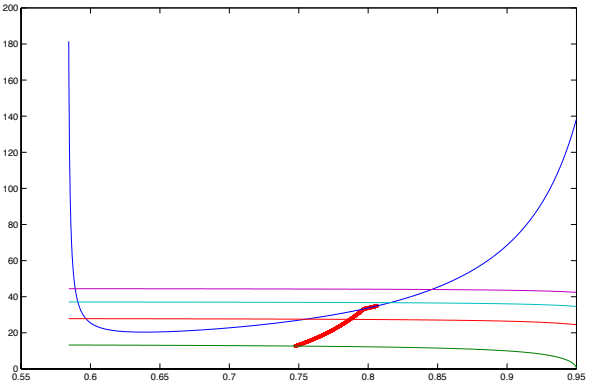


Figure 3.36: Firm's quality and capacity decisions as budget level B varies

We did extensive numerical tests and report in Table (to be inserted) and Figure 3.37 the average quality, capacity, throughput and profit reduction as the budget level deducts from the minimum budget level that can support q^* and μ^*.

From Figure 3.37, we can see that the quality level is less sensitive to the budget reduction. That is, as the budget is tight, the firm would like to still keep a relative high quality.

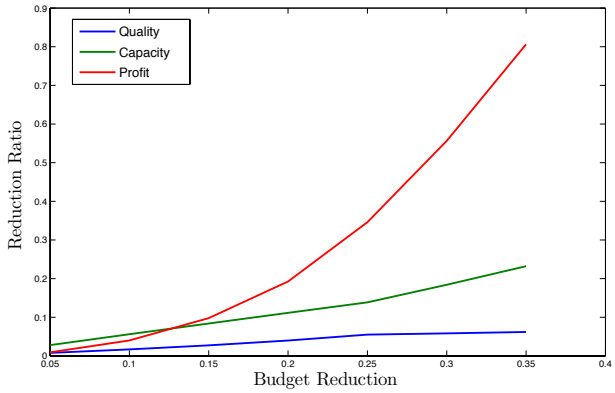


Figure 3.37: Firm's quality, capacity, throughput, and profit reductions as budget level B reduces

Chapter 4

INCORPORATING CUSTOMER ACQUISITION IN CUSTOMER LIFETIME VALUE ANALYSIS

4.1 Introduction and Literature Review

Consideration of customer's lifetime value in the analysis of a service process is essential. Each firm generates revenue from two sources: by repurchasing of the existing customers base, or by attracting sales from the new customers. Customers of service processes, similar to the products, have a limited lifetime. The limited span of decision making is not just due to the limited lifetime of the customers, but also due to the fact that the service providers will revisit the decisions made and will readjust them periodically.

Customer Lifetime Value (CLV) models in the marketing literature are designed to deal with the limited lifetime of the customers. Most of these models fail to consider the acquisition of the new customers (see [32]). As a result, if used in decision making, these measures can result in suboptimal results. Biased evaluation of profitability can occur for different reasons. The discrepancy can be due to the wrong calculation of the value of the new customers, existing customers, or failing to consider the interdependency between the two. Hence to analyze or estimate the profitability of the firms we need to consider both the customer acquisition and the customer retention capabilities. In this chapter, we assess the customer acquisition and retention problems.

Extant research has addressed the effect of different factors on the customer acquisition; brand choice and attraction models are among these models. In brand selection models, a set of characteristics, determined by the firm, are defined which can affect the consumer's choice of brand. Although these models do not inherently model the multiple purchase behavior, there are several research chapters which model brand loyalty and(/or) brand switching or examine the underlying reasons for these behaviors (see [42]). While most of the consumer choice or brand loyalty models (designed as a function of the firm characteristics) are too

complex to be used within other models, the Inertia/Variety Seeking models define brand loyalty/switching as the tendency to stick with/leave the brand. [44] assume the brand loyalty to directly affect the customer's realization of value of the product or service and define inertia as a term added to the customer's realization of value.

There are numerous academic studies in the area of customer retention. [32] provide a review of these models. The main concern with customer retention is to measure the lifetime value of a customer. Customer lifetime value models measure the potential value of the customer for a fixed (or infinite) amount of time. Pareto/NBD, introduced by [76], is a strong model for computing the customer lifetime value. It also provides the infrastructure for estimation of the future sales. However, this model is hard to implement empirically and is too complex to be used as an input into other models. [26] present BG/NBD model to simplify the Pareto/NBD model. In BG/NBD, unlike the the Pareto/NBD, the probability of leaving the firm is modeled as a bernoulli function at the end of each service instance; with probability of r a customer will repeat the purchase. Hence the dropout process, is modeled using beta-geometric function instead of Pareto of the second kind.

Although customer retention and acquisition have been analyzed extensively in the Marketing literature, the two topics rarely have been analyzed as interdependent. [81] show that the assumption of independence between customer retention and acquisition could result in biases in the customer retention analysis. [67] present an econometrics model for customer acquisition, retention, and profitability, but they ignore the effect of the time of acquisition on the profitability. [77] consider a bivariate timing model to show that the time of acquisition affects both the profitability and the retention. Another related research is [54]; they define the customer equity as the total lifetime value of all the customers. [74] then use the definition of the customer equity and use a Markov switching matrix to model customer acquisition and retention processes. Using this model they investigate that investing in which marketing aspect can generate higher values. Except [77] and [74] none of the chapters above provide general models for the specific relationship between the acquisition and retention behavior. (Also see [73], [31], and [66] for surveys.) These are the only models that consider the different effects of the decisions made, but fail to consider the operational aspects. [62] and [60] are the only chapters that consider a capacity constraint but they fail

to consider the effects of the queueing operations and service quality on the acquisition and retention processes.

Our model contributes to the literature of marketing by considering the effect of the customer interactions on the acquisition and retention processes by modeling the quality and capacity decisions. We analyze the customer acquisition within the framework of Pareto/NBD model for customer lifetime value analysis. We provide a model that includes the customer acquisition as a separate random process. Although we define the arrival process and the repeat purchase process to be independent, we capture the inter dependencies of the customer lifetime value and the customer acquisition by modeling these processes based on the consumer choice models. Both the retention and acquisition processes are dependent on the operation aspects of the model such as capacity and quality. There are several ways that quality of service can affect the customer. We model the effect of the quality and capacity decisions on loyalty, perceived value of service, referral, and by including the retention process we also analyze some aspects of service frequency.

Our model also contributes to the literature in service operations by both utilizing marketing aspects, such as customer lifetime value and customer equity in the determination of the operational decisions; and developing an elaborate model that incorporates the effects of the quality and capacity decisions on the customer behavior and consequently the optimal decision of the firm. The chapter closest to ours is [1], which considers a customer lifetime value model to analyze the effect of the pricing and the capacity decisions on the customer retentions and acquisition. They consider abandonment as measure of lack of quality and do not consider an independent quality measure in their analysis. They study different scenarios of investing in marketing vs operations aspects and the routing scenarios. Most of the chapters within the service operations literature either consider the operations aspects to not affect the demand, or affect the demand through affecting the new and repeat customer processes (see [22] and [63]). Still except for [1] no other chapter considers the effect of the operation aspects on the customer retention processes.

The rest of this chapter goes as follows: first we generalize the customer lifetime value introduced in [26] to include acquisition processes as well as the retention process. Then by considering the operational aspects we model the effects of quality and capacity decisions on

both the acquisition and retention process and use this approach to analyze the CLV model. We provide explanations on how considering the two processes as endogenous will affect the decisions. We investigate different marketing strategies and investigate the benefits of one versus the other. At the end, we provide a model of competition between two service providers and explain the dynamics of competition within this model.

4.2 Model and Results

In this section we set up different models to analyze the customer acquisition and retention models. We start with the original model of customer lifetime value introduced in [26]. First we generalize the model introduced in [26] to consider also the lifetime value of the customers who will enter within the considered time frame. Then we generalize the model to cases where the firm can make decisions about the capacity and the quality of service provided. BG/NBD model is based on the model introduced in [76]. Both Pareto/NBD and BG/NBD models evaluate a customer base of fixed size by measuring the life time value of these customers during a fixed period of time of length T. It is assumed that customers repeat their purchase following a Poisson distribution with rate θ until they leave the firm. While [76] model the customer's lifetime with an exponential function independent of the purchase instance, [26] assume the dropout to happen after the purchase instance with a fixed probability, $1 - r$. Although the model presented in [76] is more generalized than the model in [26], the latter is a better fit for analysis of the retentions that are the result of the firm's decisions. Hence throughout this chapter we use the general assumptions of BG/NBD model. Although Pareto/NBD and BG/NBD models have been used in the marketing literature extensively in the analysis of consumer behavior data, these models are designed for very specific scenarios. Both of these models track a single customer within a time frame, and only consider the customers that are existing at the beginning of the period. If this measure of customer lifetime value is used to calculate the total brand equity, then all the opportunities of attracting new customers within the time frame are ignored. To our belief the biggest restriction of the BG/NBD model is the assumption of the fixed customer base. New customers enter the market, repeat their purchase and leave the firm. This assumption not only restricts the calculation of the brand equity, it also ignores

another vast area of interest which relates to the new customer acquisition. Although in the literature customer acquisition and retention have been associated to the same features of the product or service, and brand switching models are used to capture these dependencies, they have rarely been analyzed together to capture brand equity.

In this chapter we try to tackle this problem by introducing two separate models. The scenario denoted by "^{E}" (for Existing) is similar to the BG/NBD model introduced by [26] and assumes a fixed number of customers, N, to exist at the beginning of the time period. The scenario denoted by "^{N}" (for New) considers the new customers' arrivals within the time period of T. By considering both of these arrival sources we provide an elaborate model that considers the acquisition of new customers, and the retention of the existing and new customers. In this scenario the demand for the service has two sources, the fist source is the new demand and the second source is the repeated purchase made by the customers who have experienced the service before the beginning of the time frame and are still "alive".

BG/NBD Model without New Arrivals This model is similar to the model analyzed in [26]. The number of purchases made by an existing customer in the pool can be calculated as follows ([26]).

$$\mathrm{E}(x^E) = \frac{1}{1-r}(1 - e^{-\theta(1-r)T}) \tag{4.1}$$

Hence the total number of purchases by the existing pool of customers can be calculated as

$$\mathrm{E}(X^E) = \frac{N}{1-r}(1 - e^{-\theta(1-r)T}). \tag{4.2}$$

In the following sections we analyze the models that consider customer acquisition and analyze how the firms decisions and strategies can affect the customer acquisition and retention processes.

4.2.1 BG/NBD Model with New Arrivals

This model is a generalized version of the model introduced in [26]. We consider a scenario in which new customers arrive to the system following a Poisson distribution with rate λ. After receiving the service, similar to the existing customers, if satisfied with the service,

with probability r, the new customers will join the customer pool and require the service again, Otherwise they will leave the system. The individual customer repurchase process follows a Poisson process with rate θ. Definition of the parameters and random variables can be found in Table 4.1. Because of the assumption of infinite capacity server and instant

Table 4.1: Definition of Parameters and Variables

θ	Repurchase rate for a given customer who is alive
λ	Arrival rate of new customers to the firm
μ	The service rate of the firm
r	Probability of repeating the purchase after a service instance
N	Number of existing customers in the system
$\mathrm{E}(x^i)$	Number of purchases by a given customer of type i in $[0,T]$
$\mathrm{E}(X^i)$	Number of purchases by all the customers of type i in $[0,T]$

service we can analyze this system for individual customers.

Lemma 14. *The expected number of purchases from new customers during the period* $[0,T]$ *can be obtained as:*

$$\mathrm{E}(X^N) = \frac{\lambda T}{1-r} + \frac{r\lambda}{\theta(1-r)^2}(1 - e^{-\theta(1-r)T}). \tag{4.3}$$

As it is clear the total number of the purchases made in this scenario is the total number of purchases made by the new customers and the loyal customers who were present at the system at time 0. The number of loyal customers existing in the system at time 0, can be derived from the number of people existing in an M/M/∞ system with arrival rate of λ_r and service rate of θ. For the system to be balanced we should have $\lambda_e = \lambda_r + \lambda$ and $\lambda_r = r\lambda_e$, hence $\lambda_r = \frac{r}{1-r}\lambda$. By using this arrival rate expected size of the existing customer pool is, $N = \frac{r}{1-r}\frac{\lambda}{\theta}$.

Theorem 10. *For a firm with a stream of new customers, the total purchase instances during the period* $[0, T]$ *can be derived as*

$$\mathrm{E}(X) = \frac{\lambda T}{1-r} + \frac{2r\lambda}{\theta(1-r)^2}(1 - e^{-\theta(1-r)T}). \tag{4.4}$$

By comparing (4.4) with (4.2), it is easy to show that consideration of only the existing customers and ignoring the new entries to the firm will result in an underestimation of the customer value. This gap not only affects the analysis of the lifetime value, but also distorts the decisions that are made by the firm using this performance measure. In the next section we model the firm's and customers' decision making process using the result of this section.

4.2.2 Analysis of Acquisition and Retention

There are several models of acquisition or retention which model the customers' behavior with respect to the features of the service provided by the firm. Although across different research studies customer retention and acquisition have been associated with the same features, most of the literature analyzes the two behaviors separately. In this section we set up a model to analyze the effects of these behaviors on the two basic features of service processes, i.e. capacity and quality. The firm can invest in quality and capacity levels which will affect both the new customer arrival rate and the retention probability. Capacity determines the customers' waiting time and quality reflects the quality of the service received by the customer. We use the general attraction model to model the effect of the features on the arrival rate, and we use the utility theory to model the retention probability. To include all the aspects of the customer retention behavior, we consider the customer to become loyal to the service after the first service instance. We model loyalty as a stickiness that the customer experiences to the firm. This stickiness can result from the convenience of familiarity with the service or inconvenience of searching for a new service provider.

Firm's Decisions We model the service process as an infinite server queue with exponential service with rate μ. In order to make the model tractable we assume the service time to be very small compared to the period of analysis, $\mu \ll T$. This assumption will ensure that the repurchase process can be analyzed as a process independent from the service which

will allow us to assume the repurchase rate to be not only a renewal process but also to follow a Poisson distribution. As a result the firm makes a decision on the service rate for the customers which will determine the customer wait within the system. The firm can also change the features of the service relating to the quality of the service observed by the customers. We define the quality decision as the average of a Bernoulli random variable; that is with probability q, $\tilde{q} = 1$, when the customer is satisfied with the service; and with probability $1 - q$, $\tilde{q} = 0$ when the customer is dissatisfied with the service. The firms decides on the average quality of the service provided and customers experience the random variable showing their individual resolution. We consider these decisions to be costly for the firm. Firm charges the customer a price p for every service instance, and incurs the costs for capacity and quality. We define a profit function for the firm as below.

$$\Pi = p\mathrm{E}(X) - \beta(\mu) - \alpha(q) \tag{4.5}$$

Where both $\alpha(\cdot)$, and $\beta(\cdot)$ are convex increasing functions, with $\lim_{q \to 1} \alpha(q) \to \infty$. These two assumptions ensure that the investment in quality and capacity has positive and diminishing returns.

Customer's Behavior The firms decisions will affect the customer behavior through two separate channels. First the customer analyzes the firm from outside and decides whether to acquire the service, and after observing the service provided by the firm, the customer decides to either stay or leave the firm. Customers' decision at both levels depends on the decision of the firm and how it will affect the utility of the customer from the service. Using the basics of utility theory we consider the customer to gain a value V from the service in each instance but the actual value that the customer receives from the service is reduced by the costs that the customer incurs to receive the service. Price of the service, quality and waiting are some of these costs. We model quality as a reward to the costumer, a high quality service increases the customer gained value by c_q. The waiting on the other hand is undesirable and for each unit of time that the customer waits within the system s/he incurs a cost of c_w. Hence the utility gained by customer at a service instance can be calculated

as:

$$U = \underbrace{qV}_{\text{Value of the Service Based on Quality Level}} - \underbrace{p}_{\text{Price}} - \underbrace{c_w w}_{\text{Waiting Cost}} + \underbrace{I}_{\text{Stickiness to the Brand}} \tag{4.6}$$

where I is the stickiness index of the customers to the firm, after experiencing the service the customer gains a convenience value $0 \leq I \leq p$ from staying with the firm.

This utility can be used to model both acquisition and retention rates. To model the customer decision to join the system we use the basics of the attraction model. Because our analysis does not include the competition, we use the general approach of the attraction model and define the arrival to the firm to be a portion of general arrivals of the customer to the market. We use the expected utility as a decision making function for the customer.

$$\overline{U} = qV - p - \frac{c_w}{\mu} \tag{4.7}$$

Before the customer acquires the firm as the service provider the customer has no stickiness to the firm. Based on this average utility the arrival rate can be calculated as

$$\lambda = \frac{\Lambda}{1 + e^{-\delta \bar{U}}} \tag{4.8}$$

where δ is a normalizing factor calculated for a specific market. We assume the firm to record a measure of intent of next purchase from the customers. Every customer will either continue the service with the firm after a service instance or will leave the firm. In our earlier analysis r, probability of repeating the purchase, determines this value. In order to account for the effect of the decisions made by the firm on the customers' decision, we used the realized utility of the customers at each instance as the measure of rejoining the system. Specifically, if the realized utility of the customer from the service is greater than zero, the customer will stay with the firm; otherwise the customer leaves the firm. Hence:

$$r = \mathrm{P}(\tilde{q}V - p - c_w \tilde{w} + I > 0) = q(1 - e^{-\frac{V-p+I}{c_w}\mu}). \tag{4.9}$$

We consider the decision of staying with the firm to be dependent on the observation of the utility from the service received. This decision is made after the customer has received a service from the firm so the loyalty index in this case is a fixed value $I > 0$. Based on the model provided in this section we can model the profit function of the firm based on the values of the decision variables μ, q.

4.2.3 Capacity and Quality Decisions

Capacity and quality decisions are made by the firm to maximize the total profit. These decisions are costly but directly affect customers' behavior and hence the total demand of the firm. Thus the detailed analysis of the effects of these decisions on the customers' behavior plays a major role in making the right decisions. In this section we generate a decision making framework for when the effect of these decisions on customer acquisition and retention is considered. We also compare our results to the result in operations literature when either acquisition or retention process is considered exogenous.

In order to perform the analysis we first need to formulate the profit function of the service provider. Given (4.4) and (4.5) we can calculate the firm's profit function as

$$\Pi = p\frac{\lambda T}{1-r} + p\frac{2r\lambda}{\theta(1-r)^2}(1-e^{-\theta(1-r)T}) - \beta(\mu) - \alpha(q) \tag{4.10}$$

Using these functions we can calculate the optimal levels of investment in capacity and quality decision.

Theorem 11. *The optimal quality and capacity levels of a firm with a profit function in* (4.10) *satisfies the first order conditions:*

$$\begin{cases} \frac{\partial \Pi}{\partial q} = 0, \\ \frac{\partial \Pi}{\partial \mu} = 0. \end{cases} \tag{4.11}$$

Please note that not every solution to the above system of equations is optimal. The solution to the above system of equations can be a saddle point or a local minimum. Yet we can prove that the global maximizer is at a point where the first order condition are satisfied.

Assumption 3. *The solution to* (4.11) *is unique.*

The assumption above ensures that the profit function is unimodal, and there is a unique solution that is also the global maximizer.

Please note that in the above both λ and r are functions of q and μ (see (4.8) and (4.9)) so in the consideration of the optimal solution the effect of these decisions on both the acquisition and retention processes is considered. On the other hand the operations literature

mostly considers the effects of these decision variables either on the acquisition process or retention process (see Section 4.1). Theorem 12 explains the biased decision making due to these considerations. We refer to the optimal solution when only retention is considered as (q^r, μ^r), and when only acquisition is considered as (q^a, μ^a).

Theorem 12. *Considering the model of lifetime value above, if Assumption 3 holds,*

a) If the customer acquisition process is considered exogenous of service provider's decisions, the firm underinvests in both the quality level and the capacity levels. Furthermore, $\alpha'(q^*) - \alpha'(q^r) > \frac{\partial \lambda}{\partial q}(q^r, \mu^r)\frac{\partial \Pi}{\partial \lambda}(q^r, \mu^r)$ *and* $\beta'(q^*) - \beta'(q^r) > \frac{\partial \lambda}{\partial \mu}(q^r, \mu^r)\frac{\partial \Pi}{\partial \lambda}(q^r, \mu^r)$.

b) If the customer retention process is considered exogenous of service provider's decisions, the firm underinvests in both the quality level and the capacity levels. Furthermore, $\alpha'(q^*) - \alpha'(q^a) > \frac{\partial r}{\partial q}(q^a, \mu^a)\frac{\partial \Pi}{\partial r}(q^a, \mu^a)$ *and* $\beta'(q^*) - \beta'(q^a) > \frac{\partial r}{\partial \mu}(q^a, \mu^a)\frac{\partial \Pi}{\partial r}(q^a, \mu^a)$.

In Theorem 11, both repeat purchase probability and arrival rate are increasing in quality and capacity decisions. This implies that marginal benefit of these decisions is higher compared to the cases presented in Theorem 12. The result of Theorem 12 stems from the fact that part of the benefits of investment is ignored, as q and μ increase, the arrival of new customers to the system increases and the probability of receiving a satisfactory service also increases. Both these aspects will result in higher demand, and hence higher revenue.

The other contribution of our model is the inclusion of the new customer potential to the calculation of brand equity. Hence the next natural step is the comparison of the scenarios when the traditional customer lifetime value model is used to find the optimal capacity and quality levels. To analyze this scenario we assume the steady state average number of people to be within the system at time zero and use this customer pool for the analysis only. Given (4.2) and (4.5) we can calculate the firm's profit function as

$$\Pi^E = p\frac{r\lambda}{\theta(1-r)^2}(1 - e^{-\theta(1-r)T}) - \beta(\mu) - \alpha(q) \tag{4.12}$$

In this scenario we consider both the effects of the quality and capacity decisions on the existing customer pool size (using λ) and the effect of these decisions on the retention probability. We refer to the optimal solution of this model as (q^E, μ^E).

Proposition 16. *Suppose Assumption 3 holds. Modeling the customer lifetime value without considering the arrival of the new customers within the period* $[0, T]$ *will result in lower quality and capacity investment than the optimal. Furthermore,* $\alpha^{'}(q^*) > 2\alpha^{'}(q^E)$

This theorem not only emphasizes the result in Section 4.2.1, but also quantifies the error in the calculation of capacity and quality decisions in the existing customer lifetime value models in the literature. This important because it proves that ignoring new arrivals could lead to suboptimal results. Even if we do not include the first-time purchases in the lifetime value calculation, the revenue from the repeated purchase is more than twice as much when we include the repeated purchases by the new customers. Please note that since $\alpha(\cdot)$ is an increasing convex function, the condition in Proposition 16 only guarantees an increase of less than twice in q^E.

4.3 The Effect of Marketing Strategy

In the previous section we modeled the effects of the investment on the operational aspects on the consumer behavior. Another well known area of investment that affects the consumer behavior is the marketing strategy. In this section we analyze different marketing strategies and investigate the effects of these strategies compared to each other and on the operational strategy.

4.3.1 Price Promotions for New Customers

In this section we analyze the effect of discount promotion for the new customers who are entering the system for the first time. Price promotions for the first time customers will ensure an increase in the arrival of the new customers. Please note that after experiencing the service for the first time, customers will experience convenience with the firm and could be eventually locked in due to search and switch costs. As a result not only price promotions will increase the one time purchase of the new customers, but it could also increase the long run repeat purchases as well. In this section we analyze price promotions and determine the optimal discounted price. We assume the promotion offered to the new customers is a price $p_d < p$. Hence the revenue of the customers will divide into two parts: the revenue

from the first purchase of the new customers, and the revenue from the repeated purchase of the customers. The total profit with this price promotion can be calculated as follows:

$$\Pi_d = p_d\lambda_d T + p\frac{\lambda_d T}{1-r} + p\frac{2r\lambda_d}{\theta(1-r)^2}(1-e^{-\theta(1-r)T}) - \beta(\mu) - \alpha(q) \tag{4.13}$$

where $\lambda_d = \frac{\Lambda}{1+e^{-\delta U_d}} > \lambda$.

Proposition 17. *The optimal value of the discounted price, given fixed quality and capacity levels, is*

$$\begin{aligned} p_d &= \left\{\frac{rp}{1-r} + \frac{2rp}{\theta T(1-r)^2}(1-e^{-\theta(1-r)T})\right\} \\ &+ \frac{1}{\delta}\left\{1 - W(-e^{1-\delta(qV-\frac{c_W}{\mu}+\frac{rp}{1-r}+\frac{2rp}{\theta T(1-r)^2}(1-e^{-\theta(1-r)T}))})\right\}, \end{aligned} \tag{4.14}$$

where $W(\cdot)$ is the LambertW function.

This proposition determines the optimal discounted price to achieve the optimal total profit while the quality and capacity levels are fixed at the current level. Although in Proposition 17 we consider the quality and capacity decisions to be fixed, it is clear that the optimal values of q, μ and p_d are interdependent.

Theorem 13. *If Assumption 3 holds, in the optimal price promotion and operational strategy, with offering price discount, the optimal capacity and quality levels will decrease.*

Theorem 13 shows that as the discounted price is offered, a higher arrival level is achieved. Hence effect of investment in quality and capacity on the arrival of the new customers will diminish. This implies a decrease in the marginal benefit of investment in quality and capacity, and as a result a decrease in the optimal values of quality and capacity.

This result is particularly interesting for several reasons. First, because of the marketing strategy the original operational decisions should be revisited. This implies that the marketing decision and the operations decision should be considered jointly and optimized together. Second, although the price promotions are offered to attract new customers, in the optimal setting the service provider should decrease the quality and capacity levels and as a result let a higher portion of the customers to either not enter the system, or leave the system. This result is interesting because sometimes operational decisions are blamed

for low retention rates while the price promotions are in place to attract new customers. Theorem 13 shows optimally the retention rates should decrease with the offer of price promotions.

4.3.2 Loyalty Programs for Repeat Customers

In this section we analyze the effect of loyalty programs for the repeat customers. Loyalty programs offer perks and sometimes financial incentives to the repeat customers to stay with the firm to avoid brand switching. These strategies are useful when the cost of switching for the customers is low and the competition in the market it high. In contrast with the previous section, the focus of investment in loyalty programs is the increase of repeat customers and increase in the retention rate.

In this section we analyze the optimal investment level in the loyalty programs. To capture the effects of Loyalty programs we assume the loyalty program investment to increase the stickiness level such that, $0 \leq I \leq I_l \leq p$. We consider the investment in the loyalty programs to increase the stickiness from its original level, I, to I_l to have a cost of $g(I_l)$, where $g(\cdot)$ is an increasing convex function. The rate of customer arrival remains unchanged but the repeat purchase rate, r_l, is affected.

$$\Pi_l = p\frac{\lambda T}{1-r_l} + p\frac{2r_l\lambda}{\theta(1-r_l)^2}(1-e^{-\theta(1-r_l)T}) - \beta(\mu) - \alpha(q) - g(I_l) \tag{4.15}$$

where $r_l = q(1-e^{-\frac{V-p+I_l}{c_w}\mu})$.

Proposition 18. *The optimal value of the stickiness given fixed quality and capacity levels is the solution to the following,*

$$\frac{q\mu}{c_w}e^{-\frac{V-p+I_l}{c_w}\mu}\frac{\partial \Pi_l}{\partial r_l} - g^{'}(I_l) = 0. \tag{4.16}$$

This proposition determines the optimal loyalty investment to achieve the optimal total profit while the quality and capacity levels are fixed at the current level. Although in Proposition 18 we have considered the quality and capacity decisions to be fixed, it is clear that the optimal values of q, μ and I_l are interdependent.

Theorem 14. *If Assumption 3 holds. In the optimal loyalty investment and operational decisions, with offering programs, the optimal quality level decreases, but the optimal capacity level could either decrease or increase.*

Theorem 14 shows that the loyalty investment and the quality investment behave like substitutes. As the loyalty investment increases to ensure more loyal customers the marginal benefit of the quality decreases. Hence effect of investment in quality on the retention rate diminishes. On the other hand, investment in capacity is dependent on the problem specifics. If the loyalty programs ensure high traffic to the system, and it is optimal to capture the extra demand, then the capacity will be increased. On the other hand if it is not optimal to capture all the extra demand created by the loyalty program, the capacity level might be decreased as the loyalty program ensures that the retention rate does not fall drastically with this decrease.

Please note that in this scenario similar to the price promotion scenario, we could have cases in which the firm invests in loyalty programs but decreases the quality and capacity levels. One interesting aspect of this theorem is that although capacity could be decreased or increased, the quality decision behaves as a substitute to the marketing decision. As we invest in the customers to be more loyal, providing high quality is not required any more. This result is specifically interesting because it provides insight for cases when the stickiness is imposed on the customer because of the subscription or contracts. In these scenarios customers are locked in after entering the system, so the firm can provide lower quality service to the customers who are under contracts compared to the customers who are experiencing the service on a trial basis every time.

4.4 Duopoly Model

The models of acquisition and brand switching are especially interesting when multiple firms are competing with each other. The existence of other firms will allow the customers to choose between the value that they receive at each service provider and if necessary, pick one over the other, and switch between the two if their expectations were not satisfied. In this section we assume that two firms are competing for the same service. In order to develop a

brand switching model we first need to redefine the probability of joining and retention to consider the existence of other firms.

We consider a Logit choice model with variables that vary over alternatives to analyze the new customer arrival process. Again before the customer acquires the firm as the service provider the customer has no loyalty to the firm. Based on this average utility, the arrival rate is,

$$\lambda_D = \frac{\Lambda}{1+\exp^{-\delta(U_i-U_j)}} = \frac{\Lambda}{1+\exp^{-\delta(V(q_i-q_j)-c_w(\frac{1}{\mu_i}-\frac{1}{\mu_j}))}}, \tag{4.17}$$

where $i \neq j$. Please note that the difference between this case and the previous case is that the customers now would choose between joining the two firms, where before they would join the system based on the expectation of the value received. Similarly the retention process is also affected by the competition. Previously the reservation value of the customers was zero, that is the customers would stay in the system if the last service that they received provided them with positive value, now the expectation of the value received from the competitor is the reservation value, and if the customers receive a valuation that is below what they expect to receive from the competitor, then they will switch and leave the firm.

$$r_D = \mathrm{P}(\tilde{q_i}V - r - c_w\tilde{w_i} + I_i > q_jV - r - c_ww_j) = q_i(1 - e^{-\frac{\mu_i}{\mu_j} - \frac{V(1-q_j)+I}{c_w}\mu_i}) \tag{4.18}$$

Lemma 15. *The new customer's arrival rate and repeat purchase probability are decreasing in the other firm's quality and capacity decisions.*

Lemma 15 shows that as the other firm invests more in the quality and capacity, both the new customer arrival rate and the repeat purchase rate will decrease. This result is expected from other models of competition and also the brand switching models.

For the purpose of this chapter we analyze only identical firms and the symmetric equilibria and ignore the cases for non-identical firms or asymmetric equilibria. In the previous section we considered the customers to not enter or to not repeat the purchase if the service is not at the satisfactory level. However, the existence of competition will allow the customer to switch between the firms if the customer expects to receive higher values from the competitor. As a result we expect the equilibrium in the duopoly scenario to be different

from the monopoly scenario. We refer to the optimal solution in monopoly and duopoly scenarios respectively as (q^D, μ^D) and (q^M, μ^M).

Theorem 15. *Competition will result in higher quality and less waiting for the customers; i.e.* $q^D \geq q^M$ *and* $\mu^D \geq \mu^M$

This theorem shows that although all the parameters are kept at the monopoly level, and both firms are identical, the *threat* of competition will result in both firms to increase their capacity and quality levels. Please note that in this case, at the equilibrium each firm receives half of new customer who enter the system. Also the retention rate of both firms is equal, hence the firms will gain customers from the other firm with the same rate that they lose the customers to the other firm. Hence the total demand of the firms is equal and is equal to the half of the total demand. Yet, the mere existence of the other firm in the market will provide incentive to each firm to increase capacity and quality levels.

4.5 Conclusion

In this chapter we develop a brand equity model that considers both the customer acquisition process and the customer retention process based on the BG/NBD customer lifetime value model. We then model the total firm profit using the customer equity measures. We consider the firms profit over a fixed period of time and incorporate both acquisition and retention processes within our decision making scope. We analyze the optimal quality and capacity decisions and show that existing models in the literature of both operations and marketing achieve biased results by ignoring either the effects of operational decisions on one of the acquisition or retention processes, or by ignoring one of these processes.

We also investigate the effects of different marketing strategies on the quality and capacity levels and show that investment in marketing can relieve the firm from investment in quality and sometimes even capacity. We also consider a symmetric duopoly scenario and show that the existence of competition will force the service providers to increase their quality and capacity levels.

Chapter 5

APPENDICES

5.1 Proofs of Chapter 2

Proof of Lemma 1 The Hessian matrix for (2.9) is:

$$H_1 = \begin{vmatrix} r_1\lambda\frac{\mathrm{d}^2\overline{p}(e_1)}{\mathrm{d}e_1^b} & -c_{e1}\sqrt{R_1} \\ -c_{e1}\sqrt{R_1} & -c_{w1}\sqrt{R_1}\frac{\mathrm{d}^2\alpha(y_1)}{\mathrm{d}y_1^2} \end{vmatrix}.$$

When λ goes to infinity, the determinant is positive. Therefore, when λ is large enough the Hessian is negative definite and there must exist a unique pair (y_1^{DC}, e_1^{DC}) that maximizes (2.9). After some algebraic manipulation, first order conditions of (2.9) become (2.11). Similarly, for any given e_1^{DC}, the Hessian for (2.10) is negative definite; therefore, $(y_2^{DC}(e_1), e_2^{DC}(e_1))$ is the solution to(2.12), the first order conditions in (2.10).

Proof of Lemma 2

We first prove that H_2, the Hessian for (4.10), is negative definite.

$$H_2 = \begin{vmatrix} r_1\lambda\frac{\mathrm{d}^2\overline{p}(e_1)}{\mathrm{d}e_1^2} + r_2\lambda\frac{\partial^2\overline{q}(e_1,e_2)}{\partial e_1^2} & -c_{e1}\sqrt{R_1} & r_2\lambda\frac{\partial^2\overline{q}(e_1,e_2)}{\partial e_1\partial e_2} & 0 \\ -c_{e1}\sqrt{R_1} & -c_{w1}\sqrt{R_1}\frac{\mathrm{d}^2\alpha(y_1)}{\mathrm{d}y_1^2} & 0 & 0 \\ r_2\lambda\frac{\partial^2\overline{q}(e_1,e_2)}{\partial e_1\partial e_2} & 0 & r_2\lambda\frac{\partial^2\overline{q}(e_1,e_2)}{\partial e_2^2} & -c_{e2}\sqrt{R_2} \\ 0 & 0 & -c_{e2}\sqrt{R_2} & -c_{w2}\sqrt{R_2}\frac{\mathrm{d}^2\alpha(y_2)}{\mathrm{d}y_2^2} \end{vmatrix}. \quad (5.1)$$

To show H_2 is negative definite, we use Sylvester's criterion and show that all of the leading principal minors of H_2 have determinants with alternating signs: From (2.2), it is easy to see that $D_1 = r_1\lambda\frac{\mathrm{d}^2\overline{p}(e_1)}{\mathrm{d}e_1^2} + r_2\lambda\frac{\partial^2\overline{q}(e_1,e_2)}{\partial e_1^2} < 0$. Moreover, we note that in the heavy traffic limit, $\lambda \to \infty$. Therefore, the determinant of the 2x2 leading principle minor $D_2 = -\lambda\sqrt{R_1}\left(r_1\frac{\mathrm{d}^2\overline{p}(e_1)}{\mathrm{d}e_1^2} + r_2\frac{\partial^2\overline{q}(e_1,e_2)}{\partial e_1^2}\right)\left(c_{w1}\frac{\mathrm{d}^2\alpha(y_1)}{\mathrm{d}y_1^2}\right) - R_1c_{e1}^2 \to \infty$. Similarly, for the determinant of the 3x3 leading principle minor, using (2.4) we have $D_3 = r_2\lambda\frac{\partial^2\overline{q}(e_1,e_2)}{\partial e_2^2}D_2 +$

$c_{w1}\sqrt{R_1}\left(\lambda r_2 \frac{\partial^2 \overline{q}(e_1,e_2)}{\partial e_1 \partial e_2}\right)^2 \frac{\mathrm{d}^2\alpha(y_1)}{\mathrm{d}y_1^2} \to -\infty$. Finally, for the determinant of the 4x4 leading principle minor, we have

$$\begin{aligned} D_4 &= \left\{-c_{w2}\sqrt{R_2}\left(\lambda r_2 \frac{\partial^2 \overline{q}(e_1,e_2)}{\partial e_2^2}\right)\frac{\mathrm{d}^2\alpha(y_2)}{\mathrm{d}y_2^2} - R_2 c_{e2}^2\right\} D_2 \\ &- c_{w2}\sqrt{R_2}\left(\lambda r_2 \frac{\partial^2 \overline{q}(e_1,e_2)}{\partial e_1 \partial e_2}\right)^2 \frac{\mathrm{d}^2\alpha(y_2)}{\mathrm{d}y_2^2}\frac{\mathrm{d}^2\alpha(y_1)}{\mathrm{d}y_2^1} \to \infty. \end{aligned} \tag{5.2}$$

Therefore, in the heavy traffic regime, when λ is large enough, we have $D_1 < 0, D_2 > 0, D_3 < 0, D_4 > 0$. H_2 is thus negative definite and we can find the solution to (4.10) by solving (5.3)-(5.6), which imply (2.13).

$$\frac{\partial \Pi(y_1,e_1,y_2,e_2)}{\partial e_1} = r_1\lambda\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} + r_2\frac{\mathrm{d}\overline{q}(e_1,e_2)}{\mathrm{d}e_1}\lambda - c_{e1}(R_1 + y_1\sqrt{R_1}) = 0 \tag{5.3}$$

$$\frac{\partial \Pi(y_1,e_1,y_2,e_2)}{\partial y_1} = -c_{s1}\sqrt{R_1} - c_{e1}e_1\sqrt{R_1} - c_{w1}\frac{\mathrm{d}\alpha(y_1)}{\mathrm{d}y_1}\sqrt{R_1} = 0 \tag{5.4}$$

$$\frac{\partial \Pi(y_1,e_1,y_2,e_2)}{\partial e_2} = r_2\lambda\frac{\mathrm{d}\overline{q}(e_1,e_2)}{\mathrm{d}e_2} - c_{e2}(R_2 + y_2\sqrt{R_2}) = 0 \tag{5.5}$$

$$\frac{\partial \Pi(y_1,e_1,y_2,e_2)}{\partial y_2} = -c_{s2}\sqrt{R_2} - c_{e2}e_2\sqrt{R_2} - c_{w2}\frac{\mathrm{d}\alpha(y_2)}{\mathrm{d}y_2}\sqrt{R_2} = 0, \tag{5.6}$$

Proof of Theorem 1

It is straightforward to verify that $\lim_{\lambda\to\infty}\left(\frac{\partial h_2^{DC}(e_1)}{\partial e_1} - \frac{\partial h_1^{DC}(e_1)}{\partial e_1}\right) = -\infty$. So for large enough λ, $h_2^{DC}(e_1) - h_1^{DC}(e_1)$ is decreasing in e_1. Moreover,

$$h_2^{DC}(e_1^C) - h_1^{DC}(e_1^C) \overset{\text{by def}}{=} h_2^C(e_1^C) - h_1^{DC}(e_1^C) \tag{5.7}$$

$$\overset{\text{from (2.13)}}{=} h_1^C(e_1^C) - h_1^{DC}(e_1^C) \overset{\text{by def}}{\geq} 0, \tag{5.8}$$

$$h_2^{DC}(e_1^{DC}) - h_1^{DC}(e_1^{DC}) \overset{\text{from (2.11)}}{=} 0. \tag{5.9}$$

Therefore, $e_1^C \geq e_1^{DC}$. The negative cross-derivative in (2.3) immediately implies that $e_2^C \leq e_2^{DC}$. Finally, because $h_2(\cdot)$ and $h_4(\cdot)$ are both decreasing in e, the inequalities on s follow from the inequalities on e.

Proof of Proposition 1

For the PPCR+CS contract, the client and outsourcer profit functions are:

$$\begin{aligned}
\Pi_1(s_1,e_1) &= r_1\overline{p}\lambda - c_{s1}s_1 - c_{e1}s_1e_1 - c_{w1}\lambda w_1 + (r_2 - b - \phi c_g)\lambda\overline{q} - (1-\phi)c_{w2}\lambda W_2 \\
&- (1-\phi)(c_{s2}s_2 + c_{e2}e_2) + \phi\lambda c_g, \\
\Pi_2(s_2,e_1,e_2) &= (b+\phi c_g)\overline{q}\lambda - \phi(c_{s2}s_2 + c_{e2}e_2 + c_{w2}w_2\lambda) - \phi\lambda c_g.
\end{aligned}$$

First order condition for $\Pi_1(s_1,e_1)$ with respect to e_1 is

$$\frac{\partial\Pi_{PPRC+CS}(s_1,e_1,s_2,e_2)}{\partial e_1} = r_1\lambda\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} + (r_2 - b - \phi c_g)\frac{\partial\overline{q}(e_1,e_2)}{\partial e_1}\lambda - c_{e1}s_1 = 0. \quad (5.10)$$

To coordinate the entire process in PPRC+CS, we must give the first stage client additional incentive to increase its effort level. From (5.10), this can be accomplished only by setting $c_g = b = 0$, which leads to the outsourcer to not to exert any effort. Therefore, the PPCR+CS contract cannot coordinate our two-stage process under piecemeal outsourcing.

Similarly, for the PART contract, the profit functions are

$$\begin{aligned}
\Pi_1(s_1,e_1) &= r_1\overline{p}\lambda - c_{s1}s_1 - c_{e1}s_1e_1 - c_{w1}\lambda W_1 + (1-\psi)[(r_2\lambda\overline{q}^C - c_g\lambda(1-\overline{q}^C) \\
&- c_{w2}\lambda W_2 - c_{s2}s_2 - c_{e2}s_2e_2)], \\
\Pi_2(s_2,e_1,e_2) &= (r_2+c_g)\overline{q}\lambda - (1-\psi)(r_2+c_g)\overline{q}^C\lambda - \psi c_{s2}s_2 \\
&+ (1-\psi)c_{e2}s_2e_2^C - c_{e2}s_2e_2 - \psi c_{w2}W_2 - \psi c_q\lambda.
\end{aligned}$$

From the first order condition,

$$\frac{\partial\Pi_{PART}(s_1,e_1,s_2,e_2)}{\partial e_1} = r_1\lambda\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} - c_{e1}s_1 = 0, \quad (5.11)$$

it is evident that the client is not incentivized to exert the right amount of effort, hence PART is noncoordinating.

Proof of Proposition 2

With the additional reversed commission the profit functions can be written as:

$$\begin{aligned}
\Pi_1^{PPCR+CS}(s_1,e_1) &= (r_1+c_p)\overline{p}\lambda - c_{s1}s_1 - c_{e1}s_1e_1 - c_{w1}L_1 + (r_2-b-\phi c_g)\lambda\overline{q} \\
&- (1-\phi)c_{w2}L_2 - (1-\phi)(c_{s2}s_2 + c_{e2}e_2) + \phi\lambda c_g, \\
\Pi_2^{PPCR+CS}(s_2,e_1,e_2) &= (b+\phi c_g)\overline{q}\lambda - \phi(c_{s2}s_2 + c_{e2}e_2 + c_{w2}w_2\lambda) - \phi\lambda c_g - c_p\overline{p}\lambda. \\
\Pi_1^{PART}(s_1,e_1) &= (r_1+c_p)\overline{p}\lambda - c_{s1}s_1 - c_{e1}s_1e_1 - c_{w1}L_1 \\
&+ (1-\psi)[(r_2\lambda\overline{q}^C - c_g\lambda(1-\overline{q}^C) - c_{w2}L_2 - c_{s2}s_2 - c_{e2}s_2e_2)], \\
\Pi_2^{PART}(s_2,e_1,e_2) &= r_2\overline{q}\lambda - c_g(1-\overline{q})\lambda - (1-\psi)[r_2\overline{q}^C\lambda - (1-\overline{q}^C)c_g\lambda] - \psi c_{s2}s_2 \\
&+ (1-\psi)c_{e2}s_2e_2^C - c_{e2}s_2e_2 - \psi c_{w2}L_2\} - c_p\overline{p}\lambda.
\end{aligned}$$

The two first order condition for $\Pi_1(s_1,e_1)$ with respect to e_1 are

$$\begin{aligned}
\frac{\partial\Pi_{PPRC+CS}(s_1,e_1,s_2,e_2)}{\partial e_1} &= (r_1+c_p)\lambda\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} + (r_2-b-\phi c_g)\frac{\partial\overline{q}(e_1,e_2)}{\partial e_1}\lambda \\
&- c_{e1}s_1 = 0, \qquad (5.12) \\
\frac{\partial\Pi_{PART}(s_1,e_1,s_2,e_2)}{\partial e_1} &= (r_1+c_p)\lambda\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} - c_{e1}s_1 = 0. \qquad (5.13)
\end{aligned}$$

It is easy to show that by setting $r_2 = \frac{b}{\phi} + c_g$, and $c_p = \phi r_2 \frac{\frac{\partial\overline{q}(e_1,e_2)}{\partial e_1}}{\frac{d\overline{p}(e_1)}{de_1}}\Big|_{(e_1^C,e_2^C)}$ for PPRC+CS contract the centralized solution can be achieved. Also by setting $c_g = 0$, and $c_p = r_2\frac{\frac{\partial\overline{q}(e_1,e_2)}{\partial e_1}}{\frac{d\overline{p}(e_1)}{de_1}}\Big|_{(e_1^C,e_2^C)}$ for PART the coordination results can be achieved.

Proof of Theorem 2

Adjusting for (2.16), the profit functions of the first and second stages become

$$\begin{aligned}
\Pi_1(y_1,e_1) &= (r_1+c_p)\overline{p}(e_1)\lambda - c_{s1}(R_1 + y_1\sqrt{R_1}) - c_{e1}e_1(R_1 + y_1\sqrt{R_1}) \\
&- c_{w1}\sqrt{R_1}\alpha(y_1) + \lambda c_q \\
\Pi_2(e_1,y_2,e_2) &= \lambda r_2\overline{q}(e_1,e_2) - c_{w2}\sqrt{R_2}\alpha(y_2) - c_p\overline{p}(e_1)\lambda - c_{s2}(R_2 + y_2\sqrt{R_2}) \\
&- c_{e2}e_2(R_2 + y_2\sqrt{R_2}) - \lambda c_q.
\end{aligned}$$

Similar to Lemma 1, there exists a unique optimal solution that satisfies the first order conditions we can verify that $(e_1^C, e_2^C, y_1^C, y_2^C)$ satisfies the conditions above, so the QA contract coordinates.

Proof of Theorem 3

The first order conditions for the centralized scenario are as follows:

$$\begin{aligned}
\frac{\partial \Pi}{\partial e_1} &= r_1\lambda\frac{\mathrm{d}\overline{p}}{\mathrm{d}e_1} + r_2\lambda t\frac{\partial \overline{q}}{\partial e_1} + \left[r_2\overline{q}\lambda - (c_{s2}+c_{e2}e_2)R_2 - \frac{\sqrt{R_2}}{2\sqrt{t}}[(c_{s2}-c_{e2}e_2)y_2 - c_{w2}\alpha]\right]\frac{\mathrm{d}t}{\mathrm{d}e_1} \\
&- c_{e1}(R_1 + y_1\sqrt{R_1}) = 0 \qquad (5.14)\\
\frac{\partial \Pi}{\partial y_1} &= -c_{s1}\sqrt{R_1} - c_{e1}e_1\sqrt{R_1} - c_{w1}\frac{\mathrm{d}\alpha}{\mathrm{d}y_1}\sqrt{R_1} = 0 \qquad (5.15)\\
\frac{\partial \Pi}{\partial e_2} &= r_2\lambda t\frac{\mathrm{d}\overline{q}}{\mathrm{d}e_2} - c_{e2}(R_2 t + y_2\sqrt{R_2 t}) = 0 \qquad (5.16)\\
\frac{\partial \Pi}{\partial y_2} &= -c_{s2}\sqrt{R_2 t} - c_{e2}e_2\sqrt{R_2 t} - c_{w2}\frac{\mathrm{d}\alpha}{\mathrm{d}y_2}\sqrt{R_2 t} = 0. \qquad (5.17)
\end{aligned}$$

Next, we prove that the maximum of Π can be achieved at a finite point. As is clear from the definition of the quality functions, as $e_1 \to \infty$, $\frac{\mathrm{d}\overline{p}}{\mathrm{d}e_1}, \frac{\mathrm{d}t}{\mathrm{d}e_1}, \frac{\partial \overline{q}}{\partial e_1} \to 0$. Because for any e_2, $\frac{\partial \overline{q}(e_1,e_2)}{\partial e_1} \le \frac{\partial \overline{q}(e_1,0)}{\partial e_1}$ and $\overline{q}(e_1,e_2) \le 1$, we have $\frac{\partial \Pi}{\partial e_1} \le r_1\lambda\frac{\mathrm{d}\overline{p}(e_1)}{\mathrm{d}e_1} + r_2\lambda t(e_1)\frac{\mathrm{d}\overline{q}(e_1,0)}{\mathrm{d}e_1} + r_2\lambda\frac{\mathrm{d}t(e_1)}{\mathrm{d}e_1} - c_{e1}R_1$ for all e_2. The right hand side term goes to $-c_{e1}R_1 < 0$ as $e_1 \to \infty$. Therefore, there exists a finite $\overline{e_1}$ such that

$$r_1\lambda\frac{\mathrm{d}\overline{p}(\overline{e_1})}{\mathrm{d}e_1} + r_2\lambda t(\overline{e_1})\frac{\partial \overline{q}(\overline{e_1},0)}{\partial e_1} + r_2\lambda\frac{\mathrm{d}t(\overline{e_1})}{\mathrm{d}e_1} - c_{e1}R_1 = 0, \qquad (5.18)$$

and the maximum profit is achieved at $e_1 \le \overline{e_1}$ for all e_2. The same can be applied to show that there exists an $\overline{e_2}$ such that the search for optimal e_2 can be limited to $e_2 < \overline{e_2}$.

It can also be shown that as $y_i \to \infty$, $\frac{\mathrm{d}\alpha(y_i)}{\mathrm{d}y_i} \to 0$. Hence there exists some $\overline{y_i}$ for which $-c_{si} - c_{wi}\frac{\mathrm{d}\alpha(\overline{y_i})}{\mathrm{d}y_i} < 0$. Thus the search for optimal y_i can be limited to $y_i \le \overline{y_i}$.

Since the objective function is continuous and the variable space is now limited to $[0,\overline{e_1}]\times[0,\overline{e_2}]\times[0,\overline{y_1}]\times[0,\overline{y_2}]$, we know the maximum must exist. It is clear $y_i = 0$ can not be optimal. Moreover, $e_1 = 0$ cannot be optimal since it ensures no arrivals to the second stage. With the extra assumption $\frac{\partial \overline{q}(\overline{e_1},0)}{\partial e_2} > \frac{c_{e2}}{r_2\mu_2}$, we can also guarantee that $e_2 = 0$ is not optimal. Thus, the maximum must be achieved in the interior and it must satisfy the first order conditions (5.14)-(5.17).

Using the parameter values in the Theorem 3 , we can show that the first order conditions for QA-A are identical to (5.14)-(5.17). Hence, the solution(s) of the contract must also include the system optimal solution.

Under the QA-A contract, the first-stage objective function has the following Hessian matrix:

$$H_1 = \begin{vmatrix} (r_1+c_p)\lambda\frac{\mathrm{d}^2\overline{p}(e_1)}{\mathrm{d}e_1^2} + \lambda c_1\frac{\mathrm{d}^2t(e_1)}{\mathrm{d}e_1^2} + \frac{\lambda c_2}{t(e_1)}\frac{\mathrm{d}^2t(e_1)}{\mathrm{d}e_1^2} - \frac{\lambda c_2}{t(e_1)^2}(\frac{\mathrm{d}t(e_1)}{\mathrm{d}e_1})^2 & -c_{e1}\sqrt{R_1} \\ -c_{e1}\sqrt{R_1} & -c_{w1}\sqrt{R_1}\frac{\mathrm{d}^2\alpha(y_1)}{\mathrm{d}y_1^2} \end{vmatrix}.$$

Its $D_1 < 0$ and in the BRM regime $D_2 > 0$, so first stage has a unique optimal solution. Once first stage takes action (i.e. e_1 and y_1 are fixed), we can show the optimal solution for the second stage is also unique. (Note that in our decentralized two-stage process, the first-stage optimal solution is not affected by the second-stage actions.) Therefore, the QA-A contract leads to a unique solution which is the global optimum identified by (5.14)-(5.17). Again, c_q (or equivalently b) can be adjusted to allow sharing of the total profit

Proof of Theorem 4

Part a) If we assume that the first stage decision variables remain unchanged the proof is as provided in the literature, hence by readjusting the decision variables we can achieve equal or higher total profits.

Part b) The Centralized Dedicated System has been analyzed in Section 2.4. The first order conditions for the Centralized Pooled System are below:

$$\begin{array}{ll} y_i^{CP} = h_1^C(e_i), & y_i^{CP} = h_2^C(e_i), \\ y_2^{CP} = h_3^{CP}(e_2) = \frac{r_2\sqrt{2\lambda\mu_2}}{c_{e2}}\frac{\partial\overline{q}(e_i,e_2)}{\partial e_2} - \sqrt{\frac{2\lambda}{\mu_2}}, & y_2^{CP} = h_4^C(e_2). \end{array} \tag{5.19}$$

From (5.19), when λ doubles for the outsourcer, h_3 will only increase by a factor of $\sqrt{2}$. Also, from (5.19), we see that h_4 remains the same. Similar to the proof of Theorem 1, from $h_3^{CP} < h_3^{CD}$, we can show $e_2^{CP} > e_2^{CD}$ and $y_2^{CP} < 2y_2^{CD}$. When e_2 increases, the negative cross-derivative in $h_1^C(e_i)$ means that both $e_i (i = a, b)$ will decrease. As before, the downward slope of h_2 results in an increase in y_i. That is, $e_i^{CP} < e_i^{CD}$, $y_i^{CP} > y_i^{CD}$.

Proof of Theorem 5

Because the two clients are identical, we will suppress the i subscript for the contract parameters. Let $R_2 = 2\lambda/\mu_2$. With the QA contract specified in (2.16), the profit functions

for client i and outsourcer in the Outsourced Pooled System are:

$$\begin{aligned}
\Pi_i(y_i, e_i) &= \{r_i + c_{pi}\}\,\overline{p}(e_i)\lambda - c_{si}\left(R_i + y_i\sqrt{R_i}\right) - c_{ei}e_i\left(R_i + y_i\sqrt{R_i}\right) \\
&- c_{wi}\sqrt{R_i}\alpha(y_i) + \lambda c_q, \\
\Pi_2(e_i, y_2, e_2) &= \lambda r_2[\overline{q}(e_a, e_2) + \overline{q}(e_b, e_2)] - c_p[\overline{p}(e_a) + \overline{p}(e_b)]\lambda \\
&- (c_{s2} + c_{e2}e_2)\left(R_2 + y_2\sqrt{R_2}\right) - c_{w2}\sqrt{R_2}\alpha(y_2) - 2\lambda c_q,
\end{aligned}$$

From Lemma 1 we know that there exists a unique optimal solution and it satisfies the first order conditions and we can verify that $(e_i^{CP}, e_2^{CP}, y_i^{CP}, y_2^{CP})$ satisfy these conditions, so the QA contract, designed for the Outsourced Dedicated System, also coordinates the Outsourced Pooled System.

After some algebraic manipulation, the first order conditions for the OP system, become

$$y_i^{OP} = h_1^{OP}(e_i) = \frac{r_i\sqrt{R_i}}{c_{ei}}\frac{\mathrm{d}\overline{p}(e_i)}{\mathrm{d}e_i} + \frac{r_2\sqrt{R_i}}{c_{ei}}\left.\frac{\frac{\partial\overline{q}(e_i,e_2)}{\partial e_i}}{\frac{\mathrm{d}\overline{p}(e_i)}{\mathrm{d}e_i}}\right|_{(e_i^{CP}, e_2^{CP})}\frac{\mathrm{d}\overline{p}(e_i)}{\mathrm{d}e_i} - \sqrt{R_i}, \quad (5.20)$$

$$y_i^{CP} = h_2^O(e_i), \quad y_2^{OP} = h_3^O(e_2), \quad y_2^{CP} = h_4^C(e_2). \quad (5.21)$$

When λ doubles for the outsourcer, h_3 increases by a factor of $\sqrt{2}$, and h_4 remains unchanged. Similar to the proof of Theorem 1, we can show that $e_2^{OP} > e_2^{OD}$ and $y_2^{OP} < 2y_2^{OD}$. On the other hand when e_2 increases ($e_2^{CP} > e_2^C$), the negative cross-derivative in (5.21) means that e_i will decrease. As before, the downward slope of h_2 in (5.21) means that y_i increases. That is, $e_i^{OP} < e_i^{OD}$, $y_i^{OP} > y_i^{OD}$.

Proof of Theorem 6

Part a) Under the QA contract, the profit functions for clients $i = a, b$ and the outsourcer are:

$$\begin{aligned}
\Pi_i(y_i, e_i) &= \{r_i + c_{pi}\}\,\overline{p}(e_i)\lambda_i - c_{si}(R_i + y_i\sqrt{R_i}) \\
&- c_{ei}(R_i + y_i\sqrt{R_i})e_i - c_{wi}\sqrt{R_i}\alpha(y_i) + \lambda_i c_{qi}, \\
\Pi_2(e_a, e_b, , y_2, e_2) &= r_2[\lambda_a\overline{q}(e_a, e_2) + \lambda_b\overline{q}(e_b, e_2)] - c_{pa}\overline{p}(e_a)\lambda_a - c_{pb}\overline{p}(e_b)\lambda_b \\
&- c_{s2}(R_2 + y_2\sqrt{R_2}) - c_{e2}(R_2 + y_2\sqrt{R_2})e_2 - c_{w2}\sqrt{R_2}\alpha(y_2) \\
&- \lambda_a c_{qa} - \lambda_b c_{qb},
\end{aligned}$$

where $R_2 = \frac{\lambda_a+\lambda_b}{\mu_2}$. From Lemma 1 we can verify that the solution to the Centralized Pooling system, $(e_{1a}^{CP}, e_{1b}^{CP}, e_2^{CP}, y_a^{CP}, y_b^{CP}, y_2^{CP})$, satisfies the first order conditions. Hence, the two QA contracts designed for the individual Outsourced Dedicated systems can work together to coordinate the Outsourced Pooled system.

Part b) Consider two Centralized Dedicated Systems, the first system includes client a, and the second system includes client b. Without the loss of generality we can assume $e_{2a}^{CD} > e_{2b}^{CD}$. Now if we consider a Centralized Pooled Systems where the first system includes two clients identical to client b, then it is easy to show $e_2^{CP} > e_{2b}^{CP}$. From this equation and Theorem 4 it is clear that $e_2^{CP} > e_{2b}^{CD}$. From $e_{2a}^{CD} > e_{2b}^{CD}$, it can be shown that $h_{4a}^{CD} < h_{4b}^{CD}$, hence, $h_{3a}^{CD} < h_{3b}^{CD}$. So,

$$\sqrt{\lambda_a}\frac{\partial \overline{q}(e_{1a}, e_{2a})}{\partial e_2} + \frac{c_{e2}(\sqrt{\lambda_b} - \sqrt{\lambda_a})}{r_2\mu_2} \leq \sqrt{\lambda_b}\frac{\partial \overline{q}(e_{1b}, e_{2b})}{\partial e_2}. \tag{5.22}$$

Since $\sqrt{\lambda_a + \lambda_b} - \sqrt{\lambda_a} - \sqrt{\lambda_b} \leq 0$ we can show,

$$\begin{aligned} \frac{\sqrt{\lambda_a}}{\sqrt{\lambda_a + \lambda_b}}\frac{\partial \overline{q}(e_{1a}, e_{2a})}{\partial e_2} \qquad & [\sqrt{\lambda_a + \lambda_b} - \sqrt{\lambda_a} - \sqrt{\lambda_b}] \\ \leq \quad & c_{e2}\frac{\sqrt{\lambda_a} - \sqrt{\lambda_a + \lambda_b}}{r_2\mu_2} + c_{e2}\frac{\lambda_b - \sqrt{\lambda_a\lambda_b}}{r_2\mu_2\sqrt{\lambda_a + \lambda_b}}. \end{aligned} \tag{5.23}$$

Now using (5.22) and (5.23), we can get,

$$\begin{aligned} \sqrt{\lambda_a}\frac{\partial \overline{q}(e_{1a}, e_{2a})}{\partial e_2} \quad - \quad & \frac{\lambda_a}{\sqrt{\lambda_a + \lambda_b}}\frac{\partial \overline{q}(e_{1a}, e_{2a})}{\partial e_2} - \frac{\lambda_b}{\sqrt{\lambda_a + \lambda_b}}\frac{\partial \overline{q}(e_{1b}, e_{2b})}{\partial e_2} \\ \leq \quad & c_{e2}\frac{\sqrt{\lambda_a} - \sqrt{\lambda_a + \lambda_b}}{r_2\mu_2}. \end{aligned} \tag{5.24}$$

From (5.24) it is clear that $h_{3a}^{CD} < h_3^{CP}$, and hence, $e_{2a}^{CD} > e_2^{CP} > e_{2b}^{CD}$.

Part c) Proof by counter example in Section 2.7.5.

5.2 Proofs of Chapter 3

Proof of Lemma 3

When no customer balks, $\lambda_e = \lambda(q) + q\lambda_e$, so $\lambda_e = \frac{\lambda(q)}{1-q}$. Moreover, every customer gets a non-negative utility, so $qv \geq \frac{c}{\mu-\lambda_e} + p$, or equivalently, $\lambda_e \leq \mu - \frac{c}{qv-p}$.

When customer balking happens, the throughput is lower than the arrival of all the potential customers: $\lambda_e < \frac{\lambda(q)}{1-q}$. Moreover, all those who join get zero utility so $qv = \frac{c}{\mu-\lambda_e}+p$, or equivalently, $\lambda_e = \mu - \frac{c}{qv-p}$.

Combining the two cases, we see that $\lambda_e = \min\left\{\frac{\lambda(q)}{1-q}, \mu - \frac{c}{qv-p}\right\}$. Since both $\frac{\lambda(q)}{1-q}$ and $\mu - \frac{c}{qv-p}$ are non-decreasing in μ and q, so is λ_e.

Proof of Lemma 4

This follows from the fact that $V^{II}(q,\mu)$ is decreasing in μ for all fixed q and the fact that $\mu = f(q)$ defines the lower boundary of area II.

Proof of Proposition 3

This follows from Lemma 4 and the fact that $V^I(q,\mu) = V^{II}(q,\mu)$ on the boundary $\mu = f(q)$.

Proof of Proposition 4

$(\hat{q},\hat{\mu})$ maximizes $V^I(q,\mu)$ subject to $p/v < q \leq 1$ and $0 < \mu < \infty$ while (q^*,μ^*) solves the same problem with the additional constraint $\mu \leq f(q)$. Therefore, if $(\hat{q},\hat{\mu})$ satisfies this additional constraint (i.e. $\hat{\mu} \leq f(\hat{q})$), we must have $q^* = \hat{q},\ \mu^* = \hat{\mu}$.

For $\hat{\mu} > f(\hat{q})$, we provide the proof in two steps. First, we use concavity of $V^I(q,\mu)$ in Lemma 16 (below) to show that (q^*,μ^*) must lie on the balanced curve $\mu = f(q)$. Then, we show that any $\mu > \hat{\mu}$ or $\mu < f(\hat{q})$ won't be optimal.

Lemma 16. *If $\hat{\mu} > f(\hat{q})$ then there is a (q^*,μ^*) on the balanced curve (that is, $\mu^* = f(q^*)$).*

Proof. Since $\hat{\mu} > f(\hat{q})$ it's in area II of the (q,μ) space. Take any (q,μ) is in the interior of area I (i.e. $\mu < f(q)$) and connect it with $\hat{\mu} > f(\hat{q})$. The line must intersect the boundary $\mu = f(q)$. Let the cross point be $(q',\mu') = \gamma(\hat{q},\hat{\mu}) + (1-\gamma)(q,\mu)$. By concavity we have $V^I(q',\mu') \geq \gamma V^I(\hat{q},\hat{\mu}) + (1-\gamma)V^I(q,\mu)$. Since $(\hat{q},\hat{\mu})$ maximizes V^I, we must have $V^I(q',\mu') \geq V^I(q,\mu)$. Therefore, (q^*,μ^*) can be found on the balanced curve. □

Now we can focus our attention on $\mu = f(q)$. We note that $V^I(q,\mu)$ is the sum of two strict uni-variate concave functions in q and μ separately. Therefore, the closer a point

(q, μ) is to $(\hat{q}, \hat{\mu})$ along either dimension, the higher the profit function value. It follows that any (q, μ) on the balanced curve such that $\mu > \hat{\mu}$ is dominated by $(q, \hat{\mu})$ which is in the interior of area I, so it can't be optimal. Similarly, any (q, μ) on the balanced curve such that $\mu < f(\hat{q})$ is dominated by $(\hat{q}, f(\hat{q}))$. Therefore, the optimal (q^*, μ^*) on the part of the balanced curve where $f(\hat{q}) \leq \mu \leq \hat{\mu}$.

Proof of Lemma 5

The proof of first statement, $\lim_{q \to \frac{p}{v}^+} \frac{dV(q, f(q))}{dq} > 0$, stems from $\lim_{q \to \frac{p}{v}^+} \frac{p(\lambda_0 + w)}{(1-q)^2} - \alpha^{'}(q) > -\infty$ and $\lim_{q \to \frac{p}{v}^+} \beta^{'}(f(q))[\frac{(\lambda_0 + w)}{(1-q)^2} - \frac{cv}{(qv-p)^2}] \to -\infty$ and this is due to the fact that, $\lim_{q \to \frac{p}{v}^+} \frac{cv}{(qv-p)^2} \to \infty$ and $\lim_{q \to \frac{p}{v}^+} \beta^{'}(f(q)) \to \infty$.

Similarly the proof of the $\lim_{q \to 1^-} \frac{dV(q, f(q))}{dq} \leq 0$ stems from $\lim_{q \to 1^-} \frac{dV(q, f(q))}{dq} \to -\infty$ as $\lim_{q \to 1^-} \beta^{'}(f(q)) \to \infty$, and $\lim_{q \to 1^-} \frac{(\lambda_0 + w)}{(1-q)^2} \to \infty$.

Proof of Theorem 7

When $\hat{\mu} > f(\hat{q})$, according to Proposition 4, q^* is the maximizer of the following problem:

$$\max_{\frac{p}{v} < q < 1} V(q, f(q)) = \frac{p(\lambda_0 + wq)}{1-q} - \alpha(q) - \beta(f(q)).$$

q^* is a point on $(\frac{p}{v}, 1)$ such that $\frac{dV(q,f(q))}{dq}|_{q=q^*} = 0$. This yields (3.8). Now to achieve the maximizer all the critical points should be compared.

At q_{min}, $f'(q) = 0$ and $\frac{\lambda_0 + w}{(1-q_{min})^2} = \frac{cv}{(q_{min}v-p)^2}$, so

$$\begin{aligned} \frac{dV(q, f(q))}{dq}|_{q=q_{min}} &= \frac{p(\lambda_0 + w)}{(1-q_{min})^2} - \alpha'(q_{min}) - \beta'(f(q_{min}))f'(q_{min}) \\ &= \frac{pcv}{(q_{min}v-p)^2} - \alpha'(q_{min}). \end{aligned}$$

Note that (3.6) gives us

$$\frac{pcv}{(\hat{q}v-p)^2} - \alpha'(\hat{q}) = 0. \tag{5.25}$$

When $\hat{q} \geq q_{min}$, $\frac{pcv}{(q_{min}v-p)^2} \geq \alpha'(q_{min})$, so $\frac{dV(q,f(q))}{dq}|_{q=q_{min}} \geq 0$ and $q^* \geq q_{min}$. Similarly, when $\hat{q} < q_{min}$, $q^* < q_{min}$.

Proof of Proposition 5

We define $H_x(q) \equiv \frac{\mathrm{d}V^{II}}{\mathrm{d}q}$. Now for a fixed x^o, we define q^o as the unique solution of the optimization, $H_{x^o}(q^o) = 0$.

We pick $q_r = q^0 + \epsilon/2$ and $q_l = q^0 - \epsilon/2$. It is clear that $H_{x^o}(q_l) > 0$ and $H_{x^o}(q_r) < 0$.

We define $\delta = Min\left\{\frac{H_{x^o}(q_l)}{2M}, \frac{-H_{x^o}(q_r)}{2M}, \gamma\right\}$ where γ and M are from the definition of S. Then $\forall k; |x - x^o| < \delta$ we have $H_{x^o}(q_l) - H_x(q_l) \leq \delta M$ and $H_x(q_r) - H_{x^o}(q_r) \leq \delta M$. On the other hand it is clear that $\delta M \leq Min\left\{\frac{H_{x^o}(q_l)}{2}, \frac{-H_{x^o}(q_r)}{2}\right\}$, so we have $H_x(q_l) \geq H_{x^o}(q_l) - \delta M \geq \frac{H_{x^o}(q_l)}{2} > 0$ and $H_x(q_r) \leq H_{x^o}(q_r) + \delta M \leq \frac{H_{x^o}(q_r)}{2} < 0$. Now it is clear that $q^*(x) \in [q_l, q_r]$, so $|q^*(x) - q^o| \leq |q_r - q_l| = \epsilon$.

Proof of Theorem 8

In light of Theorem 7, we need only determine when $\hat{q} \geq q_{min}$. Note that $\hat{q}$ is given in (5.25) whose left hand side is decreasing in q. Therefore, $\hat{q} \geq q_{min}$ is equivalent to

$$\frac{pcv}{(q_{min}v - p)^2} - \alpha^{'}(q_{min}) \geq 0. \tag{5.26}$$

This, together with Theorem 7, leads to the conclusion that when (5.26) holds, $q^* \geq \hat{q} \geq q_{min}$. This is the part of the balanced curve that is decreasing and investments in capacity and quality are substitutes. Conversely, when $\frac{pcv}{(q_{min}v-p)^2} - \alpha^{'}(q_{min}) < 0$, $q^* \leq \hat{q} < q_{min}$ and investments in capacity and quality are complements.

Proof of Lemma 6

$V^I(q, \mu, p)$ is a jointly continuous function. The solution cannot be achieved at $\mu = 0, q = p/v$, or $p = qv$ since these points ensure no arrival to the system and zero profits. Also, $p = 0$, generates no revenue for the firm and hence is not optimal. On the other hand, because of the convex nature of $\beta(\mu)$, there exists a $\bar{\mu}$, for which $\beta'(\bar{\mu}) > p$, hence the optimal capacity level is not achieved at infinity. Also $\lim_{q \to -1} \alpha(q) \to \infty$. Hence the optimal solution does not lie on the boundary and can be achieved on the local optima. The conditions mentioned in the problem will ensure that this local optimal point is a maximum and not a minimum or a saddle point.

Proof of Proposition 6

The first part of the problem follows from boundary condition on the constraint of (3.10) and the first order conditions in (3.11). The second part follows from the fact that once the solution to the unconstraint problem lies above the boundary, the optimal solution lies on the boundary.

Proof of Proposition 7

Similar to the proof of Theorem 7.

Proof of Theorem 9

Part 1

$p = \beta'(\mu)$ follows from the first order conditions in (3.11). Also from (3.11) we have $\mu = \frac{cqv}{(qv-p)^2}$ and $\alpha'(q) = \frac{crv}{(qv-p)^2}$. This implies $\frac{\mu}{q} = \frac{\alpha'(q)}{p}$. The rest follows from these equations.

This implies that if change in x result in an increase in one.

Part 2

As the quality level increases, price will change so that the result of first order condition, $\frac{\partial V^I I(p,q)}{\partial p}$ in (3.13) is still unchanged and is equal to zero. So we can denote price as a function of quality level, $p(q)$, that ensures that with changes in q, first order conditions hold. So we should have,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}q}\frac{\partial V^{II}(p,q)}{\partial p} &= \frac{\lambda_0 + w}{(1-q)^2}\{1 - \beta^{''}\frac{c}{(qv-p)^2}\} \\ &+ \{v - \frac{\mathrm{d}p}{\mathrm{d}q}\}\{2\beta^{'}(f(q,p))\frac{c}{(qv-p)^3} + \beta^{''}(f(q,p))\frac{c^2}{(qv-p)^4}\} = 0 (5.27)\end{aligned}$$

This will provide us with

$$\frac{\mathrm{d}p}{\mathrm{d}q} = v + \frac{\lambda_0 + w}{(1-q)^2}\frac{\{1 - \beta^{''}(f(q,p))\frac{c}{(qv-p)^2}\}}{2\beta'(f(q,p))\frac{c}{(qv-p)^3} + \beta^{''}(f(q,p))\frac{c^2}{(qv-p)^4}} \tag{5.28}$$

Now using (5.29) we can calculate the changes in the capacity level as,

$$\frac{\mathrm{d}\mu}{\mathrm{d}q} = \frac{\lambda_0 + w}{(1-q)^2} - \frac{cv}{(qv-p)^2} + \frac{\mathrm{d}p}{\mathrm{d}q}\frac{c}{(qv-p)^2} \tag{5.29}$$

$$= \frac{\lambda_0 + w}{(1-q)^2}\{\frac{2\beta^{'}(f(q,p))\frac{1}{(qv-p)} + 1}{2\beta^{'}(f(q,p))\frac{1}{(qv-p)} + \beta^{''}(f(q,p))\frac{c}{(qv-p)^2}}\} > 0 \tag{5.30}$$

This implies that no matter the change in p, μ increases with q. The condition for p, can be achieved from investigating $\frac{\partial^2 V^I I(p,q)}{\partial p \partial q}$.

Proof of Lemma 7

First, notice that $q > \frac{p}{\bar{v}}$. Otherwise, no customer would join the queue. Because of customers' repeat purchasing behavior, when customers with valuation $v \geq v_e$ join the queue, the effective arrival rate to the firm is $\lambda_e = \frac{\bar{v}-v_e}{V}\frac{\Lambda}{1-q}$. The net benefit of those customers with valuation v_e must be nonnegative, i.e., $qv_e - p - \frac{c}{\mu-\lambda_e} \geq 0$.

i) For $q > \frac{p}{\underline{v}}$, if $\mu \geq \frac{\Lambda}{1-q} + \frac{c}{q\underline{v}-p}$, the net benefit of those customers with valuation equal to $\underline{v}$ if no customers balk is nonnegative ($q\underline{v} - p - \frac{c}{\mu-\frac{\Lambda}{1-q}} \geq 0$). That is, all customers join and hence $\lambda_e = \frac{\Lambda}{1-q}$.

If $\mu < \frac{\Lambda}{1-q} + \frac{c}{q\underline{v}-p}$, balking happens and $\underline{v} < v_e \leq \bar{v}$. The customers with valuation v_e has 0 expected benefit $qv_e - p - \frac{c}{\mu-\lambda_e} = 0$. Plugging in from $v_e = \bar{v} - \lambda_e(1-q)V/\Lambda$, we get $\lambda_e^2 - \left[\mu + \frac{(q\bar{v}-p)\Lambda}{q(1-q)V}\right]\lambda_e + \frac{(q\bar{v}-p)\mu\Lambda}{q(1-q)V} - \frac{c\Lambda}{q(1-q)V} = 0$. Solving this equation, we get $\lambda_e = \frac{(\bar{v}-v_e)\Lambda}{V(1-q)} = \frac{1}{2}\left[\mu + \frac{(q\bar{v}-p)\Lambda}{Vq(1-q)}\right] - \sqrt{\frac{1}{4}\left[\mu - \frac{(q\bar{v}-p)\Lambda}{Vq(1-q)}\right]^2 + \frac{c\Lambda}{q(1-q)V}}$. The other root, $M = \frac{1}{2}\left[\mu + \frac{(q\bar{v}-p)\Lambda}{Vq(1-q)}\right] + \sqrt{\frac{1}{4}\left[\mu - \frac{(q\bar{v}-p)\Lambda}{Vq(1-q)}\right]^2 + \frac{c\Lambda}{q(1-q)V}}$, can not be the solution. For any value of the parameters, $M \geq \mu$, hence it predicts higher arrival rate than the service rate. In this scenario, the waiting time would be so high that all the customers would balk the system. So M cannot determine the arrival rate.

ii) For $q \leq \frac{p}{\underline{v}}$, then there is always balking no matter how large μ is. The net benefit of those customers with valuation v_e is 0. Using the same technique in the case where $q > \frac{p}{\underline{v}}$ and $\mu < \frac{\Lambda}{1-q} + \frac{c}{q\underline{v}-p}$, we get $\lambda_e = \frac{1}{2}\left[\mu + \frac{(q\bar{v}-p)\Lambda}{Vq(1-q)}\right] - \sqrt{\frac{1}{4}\left[\mu - \frac{(q\bar{v}-p)\Lambda}{Vq(1-q)}\right]^2 + \frac{c\Lambda}{q(1-q)V}}$. [In this case, $v_e > \frac{p}{q}$ and $\lambda_e < \frac{(q\bar{v}-p)\Lambda}{Vq(1-q)}$.]

It is trivial to show that $\frac{\Lambda}{1-q}$ is increasing in q and nondecreasing is μ. For the case

where customers are balking, we set $A = \frac{(q\bar{v}-p)\Lambda}{q(1-q)V}$, $B = \frac{c\Lambda}{q(1-q)V}$, and $C = \frac{1}{4}\left[\mu - A\right]^2 + B$

$$\frac{\partial \lambda_e}{\partial \mu} = \frac{1}{2} - \frac{1}{4}\frac{\mu - A}{\sqrt{C}}, \tag{5.31}$$

$$\frac{\partial \lambda_e}{\partial q} = \frac{1}{2}A^{'} + \frac{1}{4}\frac{A^{'}(\mu - A) - 2B^{'}}{\sqrt{C}}. \tag{5.32}$$

It is clear that since for all the value of the parameters, $B \geq 0$, hence $\sqrt{C} \geq \frac{\mu - A}{2}$, and subsequently, $\frac{\partial \lambda_e}{\partial \mu} \geq 0$. It is not easy to prove monotonicity in q, from (5.32). However (5.32), is the result of the solution of the following equation: $\bar{v} - \lambda_e \frac{(1-q)V}{\lambda_0 + wq} - \frac{c}{q(\mu - \lambda_e)} - \frac{p}{q} = 0$. Now the partial derivative of the both sides of the equation can be calculated as as, $-\frac{\partial \lambda_e}{\partial q}\left\{\frac{c}{q(\mu-\lambda_e)^2} + \frac{(1-q)V}{\lambda_0 + wq}\right\} + \lambda_e \frac{(\lambda_0 + w)V}{(\lambda_0 + wq)^2} + \frac{c}{q^2(\mu - \lambda_e)} + \frac{p}{q^2} = 0$. Thus, $\frac{\partial \lambda_e}{\partial q} \geq 0$.

Proof of Proposition 8

Case (i) refers to the cases where all the customers with all the valuations are served. In this scenario the solution lies on the balanced curve with the valuation as the lowest customer evaluation, $\frac{\Lambda}{1-q} + \frac{c}{q\underline{v} - p}$. The result of the proposition in this case follows from Theorem 8.

For the Case (ii), a portion of the customers balk and the arrival rate to the system is determined by the customers who will balk the system. For this case similar to the proof of Lemma 7, we set $A = \frac{(q\bar{v}-p)\Lambda}{q(1-q)V}$, $B = \frac{c\Lambda}{q(1-q)V}$, and $C = \frac{1}{4}\left[\mu - A\right]^2 + B$, hence the profit function can be written as,

$$V(q, \mu) = p\left\{\frac{1}{2}(\mu + A) - \sqrt{\frac{1}{4}(\mu - A)^2 + B}\right\} - \alpha(q) - \beta(\mu). \tag{5.33}$$

From (5.33), we can show,

$$\begin{aligned}
\frac{\partial^2 V}{\partial q \partial \mu} &= -\frac{p}{4}\frac{-A^{'}\sqrt{C} - (\mu - A)\frac{1}{4}\frac{A^{'}(\mu - A) - 2B^{'}}{\sqrt{C}}}{C} \\
&= \frac{p}{4}\frac{A^{'}\left\{\frac{1}{4}\left[\mu - A\right]^2 + B\right\} - \frac{1}{4}A^{'}(\mu - A)^2 + \frac{1}{2}B^{'}(\mu - A)}{C\sqrt{C}} \\
&= \frac{p}{8C\sqrt{C}}(2A^{'}B + B^{'}(\mu - A)) = \frac{p}{8C\sqrt{C}}(B^{'}(\mu + A) + 2B^2\frac{\bar{v}}{c}) \\
&= \frac{p}{8C\sqrt{C}}\frac{c}{Vq^2(1-q)^2} \\
&\quad \left\{(-\lambda_0 + 2\lambda_0 q + wq^2)(\mu + \frac{(q\bar{v} - p)\Lambda}{q(1-q)V}) + \frac{2\bar{v}(\lambda_0 + wq)^2}{V}\right\}
\end{aligned}$$

It is clear from (5.34), that the necessary condition for the quality and capacity to be substitutes is that $B^{'} = c\frac{-\lambda_0+2\lambda_0 q+wq^2}{Vq^2(1-q)^2} < 0$, hence, $-\lambda_0 + 2\lambda_0 q + wq^2 < 0$. Since $w > 0$, the left hand side is negative only between the roots, $\frac{-\sqrt{\lambda_0^2+w\lambda_0}-\lambda_0}{w} < 0 < q^* < \frac{\sqrt{\lambda_0^2+w\lambda_0}-\lambda_0}{w}$. Now if $B^{'} < 0$, then the necessary condition for the decisions to be substitutes is, $\mu > \frac{2\bar{v}\Lambda^2}{V(\lambda_0-2\lambda_0 q-wq^2)} - \frac{(q\bar{v}-p)\Lambda}{q(1-q)V} = \frac{\bar{v}\Lambda(\lambda_0+2wq-wq^2)}{V(\lambda_0-2\lambda_0 q-wq^2)(1-q)} + \frac{p\Lambda}{q(1-q)V} > 0$

Proof of Proposition 9

Before proving the proposition, we need to extend some functions defined for the monopoly setting to the duopoly setting first. Let $g(q_i|q_j) = \frac{\lambda_0+aq_i-bq_j}{1-q_i} + \frac{c}{q_i v-p}$ be the balanced curve for firm i when the other firm $j \neq i$ chooses quality level q_j, and denote its minimum by $q_{i,min}(q_j) = \arg\min_{q_i} g(q_i|q_j)$. Recall that for fair comparison, we have set the $w = a - b$, where w is the new customer demand sensitivity to the monopoly service quality, and a and b are firm i's new customer demand sensitivity to service quality provided by firms i and j ($j \neq i$) respectively. Therefore, whenever $b \neq 0$ (hence $a \neq w$), $g(q_i|q_j) \neq f(q_i)$ except for $q_i = q_j$. Furthermore, we define firm i's profit function, given q_j, to be

$$U(q_i, \mu_i|q_j) = p \min\left\{\frac{\lambda_0 + aq_i - bq_j}{1-q_i}, \mu_1 - \frac{c}{q_i v - p}\right\} - \alpha(q_i) - \beta(\mu_i).$$

Again, whenever $b \neq 0$ (hence $q \neq w$), $U(q_i, \mu_i|q_j) \neq V(q_i, \mu_i)$ except for $q_i = q_j$.

Without loss of generality we assume that firm 2 sets the quality level at q_2. Then depending on the value of q_2 we can have two outcomes for the balanced curve. We can either have $\hat{\mu} \geq g(\hat{q}|q_2)$, or $\hat{\mu} \leq g(\hat{q}|q_2)$. Now following from the monopoly case, we can show that in the first case the optimal quality and capacity levels for firm one can be determined. For the case where $\hat{\mu} \leq g(\hat{q}|q_2)$, the optimal solution for the firm 1 are $(q_1^*, \mu_1^*) = (\hat{q}, \hat{\mu})$. In the other case, when $\hat{\mu} \geq g(\hat{q}|q_2)$,then the optimal solution for the firm 1, lies on the balanced curve for the value of q_2, $g(q_1|q_2)$, so the optimal q_1^* is the solution to

$$U^{'}(q, g(q)|q_2) = \frac{p(\lambda_0 + a - bq_2)}{(1-q)^2} - \alpha^{'}(q) - \beta^{'}(g(q|q_2))g^{'}(q|q_2) = 0. \tag{5.34}$$

Part 1: $\lambda_0 \geq \hat{\lambda}_0$

We first show that $(\hat{q}, \hat{\mu})$ is a symmetric Nash equilibrium. Now we know that in response to quality level of q_2 firm 1 sets the quality level at $\hat{q}$ and capacity level at $\hat{\mu}$. Because

$\lambda_0 \geq \hat{\lambda}_0$, $(\hat{q}, \hat{\mu})$ lies on or below firm 1's balanced curve (i.e. $g(\hat{q}|\hat{q}) \geq \mu$). Therefore, from above we know it is optimal for firm 1 to choose $(\hat{q},\ \hat{\mu})$. Conversely when firm 1's quality level is set at $\hat{q}$, $(\hat{q},\ \hat{\mu})$ is firm 2's optimal decision. Hence, $(\hat{q}, \hat{\mu})$ for both firms is a Nash equilibrium.

Now consider the case where $q_2 \neq \hat{q}$, as long as $\hat{\mu} \leq g(\hat{q}|q_2)$, firm 1 responds by setting the quality and capacity at $(\hat{q},\ \hat{\mu})$. (See above). If firm 1, chooses $(\hat{q},\ \hat{\mu})$, from above, we know that firm 2 will also choose $(\hat{q},\ \hat{\mu})$. And since $\lambda_0 \geq \hat{\lambda}_0$, $(\hat{q},\ \hat{\mu})$ is the symmetric equilibrium.

On the other hand, if $\hat{\mu} \geq g(\hat{q}|q_2)$, (which implies $q_2 > \hat{q}$), then from above, we know that firm 1's best response decision (q_1^*, μ_1^*) in such a case q_1^* is the solution to (5.34) and $\mu_1^* = g(q_1^*|q_2)$. Then two cases can happen:

- If $\hat{q} < q_{1,min}(q_2)$, then the monopoly analysis indicates that $q_1^* < \hat{q}$. In this case, $\hat{\mu} \leq g(\hat{q}|q_1)$ so firm 2's optimal response is $q_2 = \hat{q}$; this contradicts $q_2 > \hat{q}$.

- If $\hat{q} \geq q_{1,min}(q_2)$, then both q_2 and $\hat{q}$ are to the right of $q_{1,min}(q_2)$ on firm 1's balanced curve $g(q|q_2)$, where it is increasing. Therefore, $g(q_1|q_2) > g(\hat{q}|q_2)$. Moreover, $\lambda_0 \geq \hat{\lambda}_0$ implies $\beta^{'}(g(q_1|q_2)) \geq \beta^{'}(\hat{\mu}) = p$. Now, from the left side of (5.34) we have:

$$\begin{aligned} U^{'}(q_1, g(q_1)|q_2) &= \frac{p(\lambda_0 + a - bq_2)}{(1-q_1)^2} - \alpha^{'}(q_1) - \beta^{'}(g(q_1|q_2))g^{'}(q_1|q_2) \\ &\leq \frac{p(\lambda_0 + a - bq_2)}{(1-q_1)^2} - \alpha^{'}(q_1) - p\left(\frac{\lambda_0 + a - bq_2}{(1-q_1)^2} - \frac{cv}{(q_1 v - p)^2}\right) \\ &= -\alpha^{'}(q_1) + \frac{pcv}{(q_1 v - p)^2}. \end{aligned}$$

  $-\alpha^{'}(q) + \frac{pcv}{(qv-p)^2}$ is negative for any $q > \hat{q}$. So we conclude $q_1^* \leq \hat{q}$.

In summary, we have shown that , when $\lambda_0 \geq \hat{\lambda}_0$, $(\hat{q}, \hat{\mu})$ is the unique Nash equilibrium.

Part 2: $\lambda_0 < \hat{\lambda}_0$

Consider any symmetric Nash equilibrium in which both firms set (q_D^*, μ_D^*). We already know from the monopoly analysis that a firm's optimal (q, μ) must lie on or below its balanced curve, so we have $\mu_D^* \leq g(q_D^*|q_D^*)$. However, if $\mu_D^* < g(q_D^*|q_D^*)$ then (q_D^*, μ_D^*) lies below the balanced curve then and it must equal $(\hat{q}, \hat{\mu})$ from our monopoly analysis. This

contradicts $\lambda_0 < \hat{\lambda}_0$; so for any symmetric Nash equilibrium we must have $\mu_D^* = g(q_D^*|q_D^*)$. Applying the first order condition to $U(q, g(q)|q_2)$ we obtain (3.17).

Proof of Proposition 10

For the cases where $\lambda_0 \geq \hat{\lambda}_0$, the solution for both cases is exactly equal to the monopoly case. For the case where $\lambda_0 < \hat{\lambda}_0$, the signs are strict and proof is as follows.

Part 1) Substituting $w = a - b$ into (3.8), we obtain that q_M is the unique solution of the following equation:

$$\frac{p(\lambda_0 + a - b)}{(1-q)^2} - \alpha'(q) - \beta' \left(\frac{\lambda_0 + (a-b)q}{1-q} + \frac{c}{qv - p} \right) \left[\frac{\lambda_0 + a - b}{(1-q)^2} - \frac{cv}{(qv-p)^2} \right] = 0. \quad (5.35)$$

Let $LD(q)$ and $LM(q)$ denote the left side of equations (3.17) and (5.35) respectively. Then

$$LD(q_D) - LM(q_D) = \frac{b}{1-q_D} \left[p - \beta' \left(\frac{\lambda_0 + (a-b)q_D}{1-q_D} + \frac{c}{q_D v - p} \right) \right].$$

From $\lambda_0 < \hat{\lambda}_0$, we find that $\mu_D < \hat{\mu}$ which is $\beta' \left(\frac{\lambda_0 + (a-b)q_D}{1-q_D} + \frac{c}{q_D v - p} \right) < \hat{\mu}$. So, $\beta'(f(q_D)) < \beta'(\hat{\mu}) = p$. Hence, $0 = LD(q_D) > LM(q_D)$, and, due to the unimodality, $q_D > q_M$.

Part 2) This part is the direct result of Part 1.

Part 3) The total profit expression for both of the cases, duopoly and monopoly, are exactly the same.

$$\begin{aligned} U_D &= p\frac{\lambda_0 + (a-b)q_D^*}{1-q_D^*} - \alpha(q_D^*) - \beta(f(q_D^*)), \\ U_M &= p\frac{\lambda_0 + (a-b)q_M^*}{1-q_M^*} - \alpha(q_M^*) - \beta(f(q_M^*)). \end{aligned}$$

It is clear from the result of monopoly case that for any $q > q_M^*$, the total profit is lower than the total monopoly case. Hence for $q_D^* > q_M^*$, $V_D^* \leq V_M^*$.

Proof of Lemma 8

As it is clear that for any $q_i \leq q_{i,min}(q_j)$, $\frac{\mathrm{d}g(q_i|q_j)}{\mathrm{d}q_i} \leq 0$. On the other hand as long as $\hat{\mu}_i > g(\hat{q}_i|q_j)$, $p > \beta'(g(q_i|q_j))$. Hence for any $q_i \leq q_{i,min}(q_j)$, the right hand side of the equality is always negative while the left hand side is always positive so the solution can only exist when $q_{Li}(q_j) > q_{i,min}(q_j)$.

For any $q_{Li}(q_j) > q_{i,min}(q_j)$, the LHS is positive and increasing while the RHS is positive and decreasing, so for $q_{mini}(q_j) < q_{Li}(q_j) < \bar{q}_i(q_j)$, the solution is unique.

To prove existence please note that both sides of the equation are continuous, and both LHS and RHS range between zero and a positive number, while, the LHS is increasing and the RHS decreasing. Hence the difference, LHS-RHS, is negative at $q_{i,min}(q_j)$, and is positive at $\bar{q}_i(q_j)$, and continuous. This implies that at some point, $q_{Li}(q_j)$ these two values would be equal and a solution exists.

Proof of Proposition 11

Before the proof starts please note that from the result of the monopoly solution we have that $q_i > \bar{q}_i(q_j)$ and $q_i < \underline{q_i}(q_j)$ are uninteresting cases that will never occur. Hence we limit our analysis to $\underline{q_i}(q_j) \geq q_i \leq \bar{q}_i(q_j)$.

The objective function for firm 1 is determined in (3.17). Now the first order conditions that determine the value of q_1, at the equilibrium is the solution to (5.34). The changes in q_1, is dependent on q_2 as follows.

$$\frac{\mathrm{d}}{\mathrm{d}q_2} \quad \left\{ \frac{p(\lambda_0 + a - bq_2)}{(1-q_2)^2} - \alpha^{'}(q_1) - \beta^{'}(g(q_1|q_2))g^{'}(q_1|q_2) \right\} \tag{5.36}$$

$$= \quad \frac{bp}{(1-q_1)^2} + \beta^{''}(g(q_1|q_2))g^{'}(q_1|q_2)\frac{b}{1-q_1} + \beta^{'}(g(q_1|q_2))\frac{b}{(1-q_1)^2} \tag{5.37}$$

$$= \quad \frac{b}{1-q_1}\left\{ \beta^{''}(g(q_1|q_2))g^{'}(q_1|q_2) - \frac{p - \beta^{'}(g(q_1|q_2))}{1-q_1} \right\} \tag{5.38}$$

Now as long as $b \neq 0$, if $\beta^{''}(g(q_1|q_2))g^{'}(q_1|q_2) > \frac{p-\beta^{'}(g(q_1|q_2))}{1-q_1}$, (or namely $q_1 > q_{L1}(q_2)$,) then the first order condition for q_1 increases with an increase in q_2. Since q_1 was a local maximizer at the point, and changes in q_2 does not affect the increasing or decreasing nature of the first order condition for firm 1, then q_1 would have to increase to satisfy the first order condition. Hence q_1 will increase with an increase in q_2. Otherwise, $\beta^{''}(g(q_1|q_2))g^{'}(q_1|q_2) < \frac{p-\beta^{'}(g(q_1|q_2))}{1-q_1}$, (or namely $q_1 < q_{L1}(q_2)$,) q_1 decreases with an increase in q_2.

Proof of Lemma 10

Without customer repeat purchasing behavior, the expected waiting time a customer will incur is $\frac{1}{\mu-\Lambda}$ when the customer joining rate is Λ and the expected benefit of customer if she choose to join is $qv-p-\frac{c}{\mu-\Lambda}$. When $\mu \geq \Lambda+\frac{c}{qv-p} = \lambda_0+wq+\frac{c}{qv-p}$, the expected profit of joining is nonnegative even all customers choose to get the service. So the throughput $\lambda_e^N = \Lambda = \lambda_0 + wq$.

When $\mu < \Lambda+\frac{c}{qv-p}$, balking happens. At equilibrium, all customers who join the queue has 0 expected net benefit. So the throughput in this case is $\lambda_e^N = \mu - \frac{c}{qv-p}$.

Combining the two cases, we have $\lambda_e^N = \min\{\lambda_0 + wq, \mu - \frac{c}{qv-p}\}$.

Proof of Lemma 11

First note that, according to equation (3.19), any $(q, f^N(q))$ dominates any (q, μ) such that $\mu > f^N(q)$. In words, the curve $\mu = f^N(q)$ dominates the area above this curve in the $\mu - q$ space. Hence, the optimal quality level and capacity level in this benchmark model must be in the area $\mu \leq f^N(q)$.

So if $\hat{\mu} \leq f^N(\hat{q})$, then $q^N = \hat{q}, \mu^N = \hat{\mu}$.

If $\hat{\mu} > f^N(\hat{q})$, because of the concavity of $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu)$, (q^N, μ^N) must be on the curve $\mu = f^N(q)$. Denote the profit function on the curve $\mu = f^N(q)$ as $V^N(q, f^N(q))$.

$$
\begin{aligned}
V^N(q, f^N(q)) &= p(\lambda_0 + wq) - \alpha(q) - \beta(f^N(q)) \\
\frac{dV^N(q, f^N(q))}{dq} &= pw - \alpha'(q) - \beta'(f^N(q))\left(w - \frac{cv}{(qv-p)^2}\right) \\
\frac{d^2V^N(q, f^N(q))}{dq^2} &= -\alpha''(q) - \beta''(f^N(q))(f'^N(q))^2 - \beta'(f^N(q))\frac{2cv^2}{(qv-p)^3}.
\end{aligned}
$$

It is easy to see that $\frac{d^2V^N}{dq^2} > 0$.So $V^N(q, f^N(q))$ is concave in q. By the first order condition, we know that q^N is the unique solution of equation (3.20). $\mu^N = f^N(q^N)$. Again, because of concavity, we know that $\mu^N < \hat{\mu}$.

At $\hat{q}$, $\frac{rcv}{(qv-p)^2} - \alpha'(q) = 0$. Then

$$
\frac{dV^N(q, f^N(q))}{dq}|_{q=\hat{q}} = [p - \beta'(f^N(\hat{q}))]\left(w - \frac{cv}{(\hat{q}v-p)^2}\right)
$$

Since $f^N(\hat{q}) < \hat{\mu}$, then $p - \beta'(f^N(\hat{q})) > 0$. Therefore, if $w \leq \frac{cv}{(\hat{q}v-p)^2}$, $\frac{dV^N(q,f^N(q))}{dq}|_{q=\hat{q}} \leq 0$, that is, $q^N \leq \hat{q}$. If $w > \frac{cv}{(\hat{q}v-p)^2}$, $\frac{dV^N(q,f^N(q))}{dq}|_{q=\hat{q}} > 0$ and $q^N > \hat{q}$.

Proof of Lemma 9

When $\alpha(q) = \frac{k}{(1-q)^l}, k > 0, l \geq 1$ and $\beta(\mu) = m\mu^2, m > 0$,

$$V(p, f(p)) = \frac{p(\lambda_0 + w)}{1-q} - \frac{k}{(1-q)^l} - m\left(\frac{\lambda_0 + w}{1-q} + \frac{c}{qv-p} - w\right)^2 - pw,$$

and

$$\begin{aligned}
\frac{dV(q,f(q))}{dq} &= \frac{p(\lambda_0+w)}{(1-q)^2} - \frac{lk}{(1-q)^{l+1}} \\
&- 2m\left(\frac{\lambda_0 + wq}{1-q} + \frac{c}{qv-p}\right)\left[\frac{\lambda_0+w}{(1-q)^2} - \frac{cv}{(qv-p)^2}\right] \\
&= \frac{1}{(1-q)^2}\left[LH(q) - RH(q)\right], \qquad (5.39)
\end{aligned}$$

where

$$\begin{aligned}
LH(q) &= p(\lambda_0+w) + 2mc\left[\frac{v(\lambda_0+wq)(1-q)}{(qv-p)^2} - \frac{\lambda_0+w}{qv-p} + \frac{cv(1-q)^2}{(qv-p)^3}\right], \\
RH(q) &= \frac{lk}{(1-q)^{l-1}} + \frac{2m(\lambda_0+wq)(\lambda_0+w)}{1-q}.
\end{aligned}$$

1. $\lim_{q\to\frac{p}{v}} LH(q) = \infty$ and $\lim_{q\to\frac{p}{v}} RH(q) < \infty$, so $\lim_{q\to\frac{p}{v}} \frac{dV(q,f(q))}{dq} > 0$;

2. $\lim_{q\to 1} LH(q) < \infty$ and $\lim_{q\to 1} RH(q) = \infty$, so $\lim_{q\to 1} \frac{dV(q,f(q))}{dq} < 0$;

3. It is clear that $RH(q)$ is increasing in q for $l \geq 1$. Moreover, because $\frac{-2v(1-q)(\lambda_0 v+pw)}{(qv-p)^3} < 0$ and $\frac{cv(1-q)^2}{(qv-p)^3}$ is decreasing in q, overall $LH(q)$ is decreasing in q. Therefore $\frac{dV(q,f(q))}{dq}$ is decreasing in q. Together with 1-2, this means that $\frac{dV(q,f(q))}{dq} = 0$ has a unique solution in $(p/v, 1]$. This also means that $V(q, f(q))$ is unimodal in q.

Proof of Proposition 12

(A) These are first order conditions of (3.6) and (3.7), two concave functions.

(B) This follows from (3.8) after we plug in $\alpha(q) = \frac{k}{(1-q)^l}$ and $\beta(\mu) = m\mu^2$.

(C) For all $q \in (0,1)$, define $h_q(l) = \frac{lk}{p(1-q)^{l+1}} - \frac{\lambda_0+w}{(1-q)^2} = \frac{1}{p(1-q)^2}\left[\frac{lk}{(1-q)^{l-1}} - p(\lambda_0+w)\right]$.

- When $l = 1$, $h_q(1) < 0$ due to the assumption $k < p(\lambda_0 + w)$.
- When $l \to \infty$, $h_q(l) \to \infty$.
- $h_q(l)$ is continuous and increasing in l.

Thus, for $q = q_{min}$ there exists a threshold $\tilde{l} > 1$ such that $h_{q_{min}}(l) \leq 0$ for $l \leq \tilde{l}$ and $h_{q_{min}}(l) > 0$ for $l > \tilde{l}$. Recall that q_{min} solves $\frac{cv}{(qv-p)^2} = \frac{\lambda_0+w}{(1-q)^2}$ and $\hat{q}$ solves $\frac{cv}{(qv-p)^2} = \frac{lk}{p(1-q)^{l+1}}$. Therefore, $\hat{q} \geq q_{min}$ for $l \leq \tilde{l}$ and $\hat{q} < q_{min}$ for $l > \tilde{l}$.

Proof of Lemma 12

Let $G(q) = pR(q) - \beta(R(q))$. Then

$$G^{''}(q) = pR''(q) - \beta''(R(q))[R'(q)]^2 - \beta'(R(q))R''(q) = [p - \beta'(R(q))]R''(q) - \beta''(R(q))[R'(q)]^2.$$

For $q \geq R^{-1}(\hat{\mu})$, $\beta^{'}(R(q)) \leq \beta^{'}(\hat{\mu}) = p$. So $G^{''}(q) \leq 0$.

Proof of Proposition 13

The feasible area is bounded with a concave set and the optimal solution is not within the area, so the optimal is on the boundary and the optimal point on the boundary is achieved at q_R^*.

Proof of Proposition 14

When $f(q)$ and $R(q)$ intersect twice at q_L and q_H, the feasible region is below both curves. It is marked by the shaded area in Figure 3.33. Since we focus on the non-trivial case of $\mu^* > R(q^*)$, the point (q^*, μ^*) lies in the shaded area between $f(q)$ and $R(q)$ to the right of $(q_H, f(q_H))$ (also denoted by point Y in the figures).

We first show that (q^{**}, μ^{**}) must lie on the boundary of the feasible region indicated by the solid curves in Figure 3.33. The idea and proof are similar to those of Lemma 16: Suppose (q, μ) is in the interior of the feasible region. Since (q^*, μ^*) is outside the feasible

region, so must be $(\hat{q}, \hat{\mu})$. Connect (q, μ) and $(\hat{q}, \hat{\mu})$ by a straight line and denote its intersection with the feasible region boundary by $(q^{'}, \mu^{'}) = \gamma(\hat{q}, \hat{\mu}) + (1-\gamma)(q, \mu)$ for some $\gamma \in (0, 1)$. By concavity we have $V^I(q^{'}, \mu^{'}) \geq \gamma V^I(\hat{q}, \hat{\mu}) + (1-\gamma)V^I(q, \mu)$. Since $(\hat{q}, \hat{\mu})$ maximizes V^I, we must have $V^I(q^{'}, \mu^{'}) \geq V^I(q, \mu)$. Therefore, (q^{**}, μ^{**}) must be on the boundary of the feasible region. More precisely, (q^{**}, μ^{**}) must satisfy $\mu^{**} = f(q^{**})$ if $q_L \leq q^{**} \leq q_H$; and $\mu^{**} = R(q^{**})$ if $q^{**} \leq q_L$ or $q^{**} \geq q_H$.

Next, we show that $q^{**} \geq q_H$ by exclusion. Take any $(q, f(q))$ for $q < q_L$.

Theorem 8 demonstrates that there are two possible cases: 1) $q^* \geq \hat{q} \geq q_{min}$ and 2) $q^* < \hat{q} < q_{min}$. We will study them one by one.

Case 1: $q^* \geq \hat{q} \geq q_{min}$

In this case, both $(\hat{q}, \hat{\mu})$ is to the right of $(q_{min}, f(q_{min})$.

For $l \leq \tilde{l}$, there are three possible situations: a) $(q^*,\ \mu^*) = (\hat{q},\ \hat{\mu})$, $\hat{q} > q_H$ and $\hat{\mu} \leq R(q_H)$ (see Figure 5.1); b) $(q^*,\ \mu^*) = (\hat{q},\ \hat{\mu})$, $\hat{q} > q_H$ and $\hat{\mu} > R(q_H)$ (see Figure 5.2); c) $q^* \geq \hat{q},\ \mu^* = f(q^*),\ q^* > q_H$ (see Figure 5.3).

For case a), first note because of the concavity of function $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu)$, $q_H \leq R^{-1}(\hat{\mu}) \leq \tilde{q}_b \leq \hat{q}$. And the profit function value at any point (q, μ) in the feasible region $\mu \leq R(q)$ such that $q \leq \hat{q}$ is no greater than $V_b(R^{-1}(\mu))$. It is easy to see this for point $(q, \mu) \notin M$ (concavity). For any point $(q, \mu) \in M$, the profit function is $\frac{p(\lambda_0+w)}{1-q} - \alpha(q) - \beta(\mu) - pw$, which is smaller than $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu)$ because $\mu > f(q) = \frac{\lambda_0+w}{1-q} + \frac{c}{qv-p} - w$ for points in set M. And $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu) < V_b(R^{-1}(\mu))$ because of the concavity of $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu)$ and the fact that $\hat{q} > q_H$. That is, for any point $(p, \mu) \in M$, the profit value is also lower than $V_b(R^{-1}(\mu))$. $\tilde{q}_b$ is the maximum of $V_b(R^{-1}(\mu))$ and is not in set M (since $\tilde{q}_b \geq q_H$). So it is the optimal quality level with the budget constraint in this case, i.e., $q_b^* = \tilde{q}_b$.

For case b), we can just consider the boundary line formed by curve $\mu = R(q)$ with the segment $q_L \leq q \leq q_H$ replaced by $\mu = f(q)$ because of the concavity and $\hat{q} > q_H$. Also note that the profit function on curve $f(q)$, $V(q, f(q))$, has a single mode, denoted by q_V. And q_V is in between $\bar{q}(\hat{\mu})$ and $\hat{q}$. Proposition 12 shows that $V(q, f(q))$ is unimodal under Lemma 5. Although the analysis is for the case where $\hat{\mu} > f(\hat{q})$, the property is not restricted to that case. Because of the concavity of $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu)$, the profit value at $(\bar{q}(\hat{\mu}), \hat{\mu})$ is

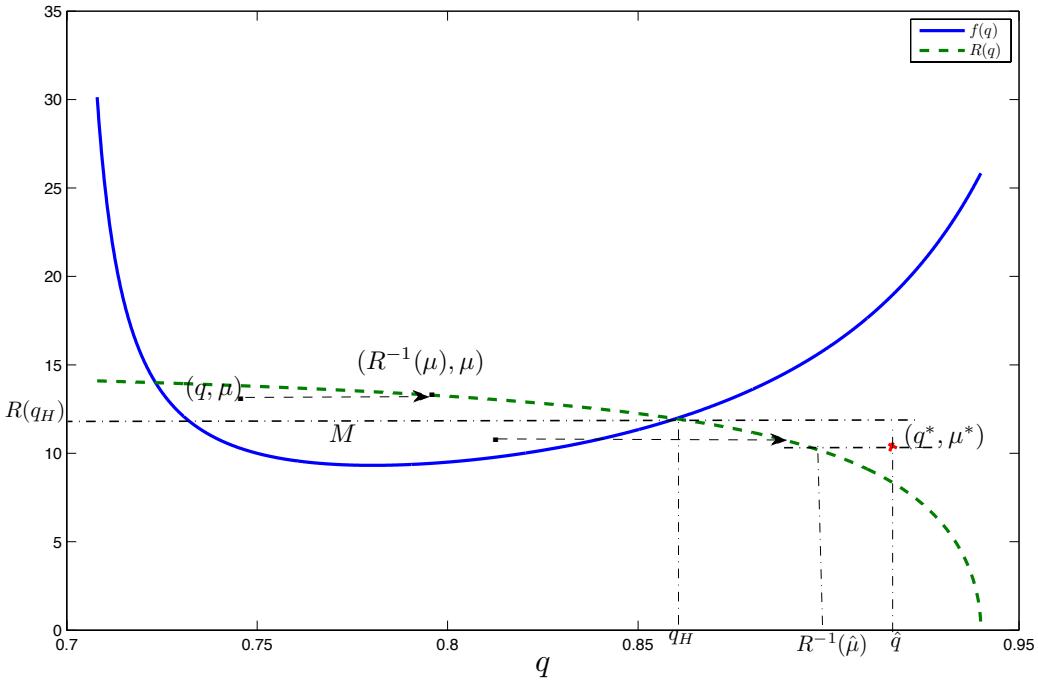


Figure 5.1: Case a)

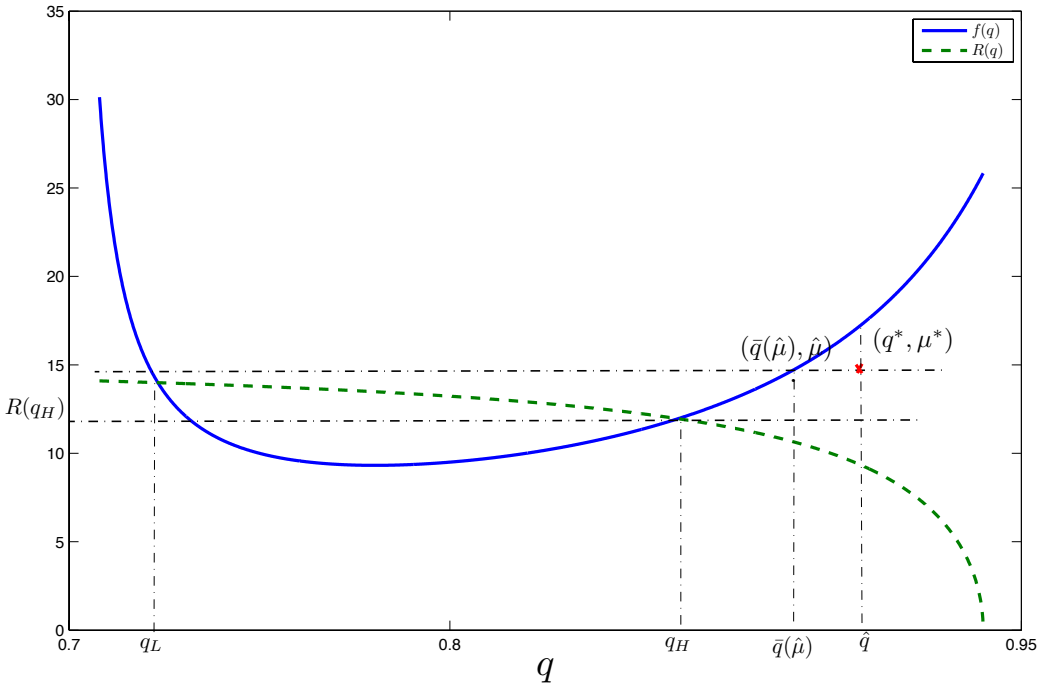


Figure 5.2: Case b)

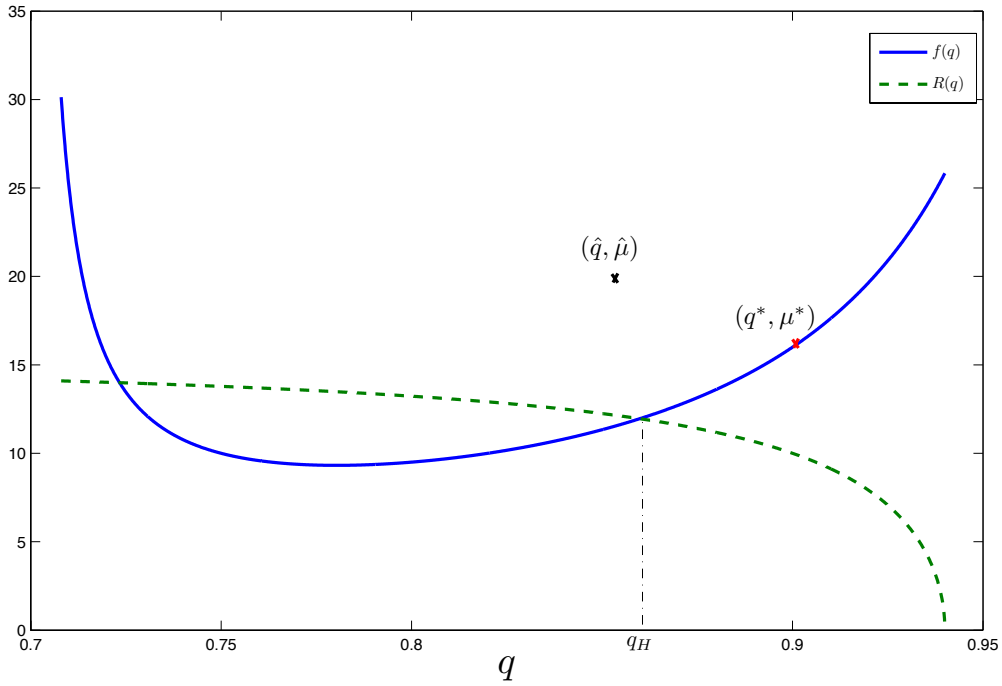


Figure 5.3: Case c)

greater than the value at $(q_{min}, f(q_{min}))$ since $\bar{q}(\hat{\mu}) > q_{min}$. That is, $V(q, f(q))$ is increasing during $(q_{min}, \bar{q}(\hat{\mu}))$. $V(q, f(q))$ is decreasing for $q \geq \hat{q}$. So q_V must be in between $\bar{q}(\hat{\mu})$ and $\hat{q}$. Because $q_H < \bar{q}(\hat{\mu})$, then $(q_H, f(q_H))$ dominates any point (q, μ) such that $q_L \leq q \leq q_H$ and $\mu = f(q)$. Because $\hat{q} > q_L$ ($\hat{q} > q_{min}$ and $q_L < q_{min}$), any point $(q, R(q))$ such that $q \leq q_L$ is dominated by $(\underline{q}(R(q)), R(q))$ and then is further dominated by $(q_H, f(q_H))$. Therefore, $(q_H, f(q_H))$ dominates any point (q, μ) such that $q \leq q_H$ on the boundary line. And $V_b(q)$ is concave, hence $q_b^* = \max\{q_H, \tilde{q}_b\}$.

For case c), similarly as in case b), any point $(q, R(q))$ such that $q \leq q_L$ is dominated by $(\underline{q}(R(q)), R(q))$. And $V(q, f(q))$ is increasing till $q^* > q_H$. So $(q_H, f(q_H))$ dominates any point (q, μ) such that $q \leq q_H$. Therefore, $q_b^* = \max\{q_H, \tilde{q}_b\}$.

Now we study the case in which $l > \tilde{l}$ and $R(q)$ does intersect with $f(q)$ in the $\mu - q$ space. Note that in this case, $\hat{q} < q_{min}$. There are four possible situations: i) $(q^*, \mu^*) = (\hat{q}, \hat{\mu})$ and $\hat{q} < q_L$; ii) $q^* \leq \hat{q}$, $\mu^* = f(q^*)$ and $\hat{q} \leq q_L$; iii) $q^* < q_L$, $\mu^* = f(q^*)$ and $q_L < \hat{q} < q_H$, $\hat{\mu} > R(q_L)$; iv) $q^* \leq \hat{q}$, $\mu^* = f(q^*)$ and $q^* > q_H$. Case iv) can happen only when $q_H < q_{min}$. For case i), because of the concavity of $V_b(q)$ and function $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu)$, we know that $q_b^* = \tilde{q}_b$ and $\tilde{q}_b \leq \hat{q}$. For case ii), again because of the concavity of $V_b(q)$, $\tilde{q}_b$ must be no greater than q_L and $q_b^* = \tilde{q}_b$. For case iii), since $V(q, f(q))$ is unimodal and $q^* < q_L$, then $V(q, f(q))$ is decreasing for $q \geq q_L$. Because of the concavity of function $p\mu - \frac{pc}{qv-p} - \alpha(q) - \beta(\mu)$, we know that $(q_L, R(q_L))$ dominates any (q, μ) such that $q \geq q_L$ in the feasible region. And $\tilde{q}_b$ must be no greater than $\hat{p}$. Because $V_b(q)$ is concave, then $q_b^* = \min\{q_L, \tilde{q}_b\}$. Case iv) is actually similar as case c) when $l \leq \tilde{l}$. So in this case, $q_b^* = \max\{q_H, \tilde{q}_b\}$.

Combining all the cases analyzed above, we get Proposition 14.

5.3 Proofs of Chapter 4

Proof of Lemma 14

The probability of a customer existing in the system, leaving after a given time t is

$$\mathrm{P}\{\psi > t\} = \sum_{n=0}^{n=\infty} r^n \frac{(\theta t)^n e^{-\theta t}}{n!} = e^{-\theta(1-r)t} \tag{5.40}$$

Given this probability the distribution can be derived as $f(\psi) = \theta(1-r)e^{-\theta(1-r)\psi}$. Then we can calculate the number of purchases made by a given customer who enters the system at time τ can be calculated as

$$\begin{aligned} \mathrm{E}(x^N|\tau) &= 1 + r\theta(T-\tau)\mathrm{P}\{\psi > (T-\tau)\} + r\int_0^{T-\tau} \theta\psi f(\psi)\mathrm{d}\psi \\ &= 1 + \frac{r}{1-r}(1 - e^{-\theta(1-r)(T-\tau)}) \end{aligned} \tag{5.41}$$

As a result the total number of purchases by the new customers who arrive during the period $[0, T]$,

$$\begin{aligned} \mathrm{E}(X^N) &= \int_0^T \{1 + \frac{r}{1-r}(1 - e^{-\theta(1-r)(T-\tau)})\}\lambda\mathrm{d}\tau = \frac{\lambda T}{1-r} + \frac{r\lambda}{\theta(1-r)^2} \\ &- \frac{r\lambda}{\theta(1-r)^2}e^{-\theta(1-r)T} \end{aligned} \tag{5.42}$$

Proof of Theorem 10

$$\mathrm{E}(X) = \mathrm{E}(X^E) + \mathrm{E}(X^N) = \frac{\lambda T}{1-r} + \frac{2r\lambda}{\theta(1-r)^2}\left(1 - e^{-\theta(1-r)T}\right). \tag{5.43}$$

Proof of Theorem 11

As quality goes to 1, or capacity goes to infinity profit goes to $-\infty$. Also for $q = \frac{p}{V}$ and $\mu = 0$, profit is increasing in the capcity and quality levels. This implies that the solution is on the interior and on a point where first order conditions are zero.

Proof of Theorem 12

If both r and λ are dependent on q and μ then the optimal value can be calculated using the first order conditions.

$$\begin{aligned}
\frac{\partial \Pi}{\partial q} &= p\frac{\partial \lambda}{\partial q}\underbrace{\left\{\frac{T}{1-r}+\frac{2r}{\theta(1-r)^2}(1-e^{-\theta(1-r)T})\right\}}_{A} \\
&+ p\lambda\frac{\partial r}{\partial q}\underbrace{\left\{\frac{T}{(1-r)^2}+\frac{2(1+r)}{\theta(1-r)^3}(1-e^{-\theta(1-r)T})-\frac{2rT}{(1-r)^2}e^{-\theta(1-r)T}\right\}}_{B} \\
&- \alpha^{'}(q)=0, \qquad (5.44) \\
\frac{\partial \Pi}{\partial \mu} &= p\frac{\partial \lambda}{\partial \mu}\underbrace{\left\{\frac{T}{1-r}+\frac{2r}{\theta(1-r)^2}(1-e^{-\theta(1-r)T})\right\}}_{A} \\
&+ p\lambda\frac{\partial r}{\partial \mu}\underbrace{\left\{\frac{T}{(1-r)^2}+\frac{2(1+r)}{\theta(1-r)^3}(1-e^{-\theta(1-r)T})-\frac{2rT}{(1-r)^2}e^{-\theta(1-r)T}\right\}}_{B} \\
&- \beta^{'}(\mu)=0. \qquad (5.45)
\end{aligned}$$

Part a)

For this part only the retention rate is considered to be endogenous in the parameters. Hence,

$$\begin{aligned}
\frac{\partial \Pi_r}{\partial q} &= p\lambda\frac{\partial r}{\partial q}\underbrace{\left\{\frac{T}{(1-r)^2}+\frac{2(1+r)}{\theta(1-r)^3}(1-e^{-\theta(1-r)T})-\frac{2rT}{(1-r)^2}e^{-\theta(1-r)T}\right\}}_{B} \\
&- \alpha^{'}(q)=0, \qquad (5.46) \\
\frac{\partial \Pi_r}{\partial \mu} &= p\lambda\frac{\partial r}{\partial \mu}\underbrace{\left\{\frac{T}{(1-r)^2}+\frac{2(1+r)}{\theta(1-r)^3}(1-e^{-\theta(1-r)T})-\frac{2rT}{(1-r)^2}e^{-\theta(1-r)T}\right\}}_{B} \\
&- \beta^{'}(\mu)=0. \qquad (5.47)
\end{aligned}$$

This implies that $\frac{\partial \Pi}{\partial q}(q^r,\mu^r)=p\frac{\partial \lambda}{\partial q}>0$, hence $q^r<q*$. The conditions follow from the fact that B is increasing in r. The proof for μ is also similar.

Part b)

The proof is similar to part a.

Proof of Proposition 16

The optimal value when there are no new arrivals can be calculated as below.

$$\begin{aligned}\frac{\partial \Pi^E}{\partial q} &= p\frac{\partial \lambda}{\partial q}\underbrace{\left\{\frac{r}{\theta(1-r)^2}(1-e^{-\theta(1-r)T})\right\}}_{C} + p\lambda\frac{\partial r}{\partial q}\underbrace{\left\{\frac{(1+r)}{\theta(1-r)^3}(1-e^{-\theta(1-r)T}) - \frac{2rT}{(1-r)^2}e^{-\theta(1-r)T}\right\}}_{D} \\ &- \alpha^{'}(q) = 0, \qquad (5.48)\\ \frac{\partial \Pi^E}{\partial \mu} &= p\frac{\partial \lambda}{\partial \mu}\underbrace{\left\{\frac{r}{\theta(1-r)^2}(1-e^{-\theta(1-r)T})\right\}}_{C} + p\lambda\frac{\partial r}{\partial \mu}\underbrace{\left\{\frac{(1+r)}{\theta(1-r)^3}(1-e^{-\theta(1-r)T}) - \frac{2rT}{(1-r)^2}e^{-\theta(1-r)T}\right\}}_{D} \\ &- \beta^{'}(\mu) = 0. \qquad (5.49)\end{aligned}$$

As C and D are both less that A and B in (5.44), it is clear that $\frac{\partial \Pi}{\partial q}(q^E, \mu^E) > 0$. Hence $q^E < q^*$. The conditions stem from the fact that $A > 2C$ and $B > 2D$. Proof for μ is similar.

Proof of Proposition 17

The condition mentioned in the proposition is the solution to the following;

$$\frac{\partial \Pi_d}{\partial p_d} = \lambda_d T + p\frac{\partial \lambda_d}{\partial p_d}\{Ap - (p-p_d)T\} = 0, \qquad (5.50)$$

where A is defined in (5.44).

Proof of Theorem 13

$$\frac{\partial \Pi_d}{\partial q} = \frac{\partial \lambda_d}{\partial q}\{pA - (p-p_d)T\} + p\lambda\frac{\partial r}{\partial q}B = 0, \qquad (5.51)$$

where A and B are defined in (5.44). Now it is clear that if the marketing decision is beneficial we have $\Pi_d(q_d, \mu_d) - \Pi(q_d, \mu_d) \geq 0$ which implies $(\lambda_d - \lambda)B \geq (p - p_d)\lambda_d T$.

$$\begin{aligned}\frac{\partial \Pi_d}{\partial q} - \frac{\partial \Pi}{\partial q} &= pA\left\{\frac{\partial \lambda_d}{\partial q} - \frac{\partial \lambda}{\partial q}\right\} - (p-p_d)T\frac{\partial \lambda_d}{\partial q} \\ &= \frac{\partial \lambda_d}{\partial q}\frac{1}{\lambda_d^2}\left\{pA(\lambda_d - \lambda)(\lambda_d + \lambda) - (p-p_d)T\lambda_d^2\right\} \\ &\geq \frac{\partial \lambda_d}{\partial q}\frac{1}{\lambda_d^2}\left\{(p-p_d)T\lambda_d\lambda\right\} > 0 \qquad (5.52)\end{aligned}$$

which implies $q_d < q$. The proof for μ is similar.

Proof of Proposition 18

The condition mentioned in the proposition is the solution to the following;

$$\frac{\partial \Pi_l}{\partial I_d} = p\lambda \frac{\partial r_l}{\partial I_l} B - g^{'}(I_l) = 0. \tag{5.53}$$

where B is defined in (5.44)

Proof of Theorem 14

$$\frac{\partial \Pi_l}{\partial q} = p\frac{\partial \lambda}{\partial q} A + p\lambda \frac{\partial r_l}{\partial q} B - \alpha^{'}(q) = 0, \tag{5.54}$$

$$\frac{\partial \Pi_l}{\partial \mu} = p\frac{\partial \lambda}{\partial \mu} A + p\lambda \frac{\partial r_l}{\partial \mu} B - \beta^{'}(\mu) = 0, \tag{5.55}$$

where A and B are defined in (5.44). B is increasing in r. Now $\frac{\partial r_l}{\partial q} > \frac{\partial r}{\partial q}$, hence $q_l < q^*$. But for μ $\frac{\partial r_l}{\partial \mu} < \frac{\partial r}{\partial \mu}$, so the solution is dependent on the problem parameters.

Proof of Lemma 15

$$\begin{aligned}
\frac{\partial \lambda_i}{\partial q_j} &= \frac{-\delta V \Lambda \exp^{-\delta(V(q_i-q_j)-c_w(\frac{1}{\mu_i}-\frac{1}{\mu_j}))}}{\left(1+\exp^{-\delta(V(q_i-q_j)-c_w(\frac{1}{\mu_i}-\frac{1}{\mu_j}))}\right)^2} < 0, \\
\frac{\partial \lambda_i}{\partial \mu_j} &= \frac{-\delta c_w \Lambda \exp^{-\delta(V(q_i-q_j)-c_w(\frac{1}{\mu_i}-\frac{1}{\mu_j}))}}{\mu^2\left(1+\exp^{-\delta(V(q_i-q_j)-c_w(\frac{1}{\mu_i}-\frac{1}{\mu_j}))}\right)^2} < 0, \\
\frac{\partial r_i}{\partial q_j} &= -\frac{q_i V}{c_w}\mu_i \exp^{-\frac{\mu_i}{\mu_j}-\frac{V(1-q_j)+I}{c_w}\mu_i} < 0, \\
\frac{\partial r_i}{\partial \mu_j} &= -\frac{q_i}{\mu_j^2}\mu_i e^{-\frac{\mu_i}{\mu_j}-\frac{V(1-q_j)+I}{c_w}\mu_i} < 0.
\end{aligned}$$

Proof of Theorem 15

$$\frac{\partial \Pi^D}{\partial q_i} = p\frac{\partial \lambda_i}{\partial q_i} A + p\lambda_i \frac{\partial r_i}{\partial q_i} B - \alpha^{'}(q_i) = 0, \tag{5.56}$$

$$\frac{\partial \Pi^D}{\partial \mu_i} = p\frac{\partial \lambda_i}{\partial \mu_i} A + p\lambda_i \frac{\partial r_i}{\partial \mu_i} B - \beta^{'}(\mu_i) = 0, \tag{5.57}$$

where A and B are defined in (5.44). Now it is clear that $r^D(q) < r^M(q)$, $\frac{\partial r^D}{\partial q} < \frac{\partial r^M}{\partial q}$. Combined with the fact that the new customer arrival stays the same, implies that $q^D > q^M$. The proof for μ is similar.

BIBLIOGRAPHY

[1] P. Afèche, M. Araghi, and O. Baron. Customer retention, acquisition, and service quality for a call center: optimal promotions, priorities, and staffing. *Working Paper*, University of Toronto, 2013.

[2] O. Z. Akşin, F. de Véricourt, and F. Karaesmen. Call center outsourcing contract analysis and choice. *Management Science*, 54(2):354–368, 2008.

[3] O.Z. Akşin and P.T. Harker. To sell or not to sell: Determining the tradeoffs between service and sales in retail banking phone centers. *Journal of Service Research*, 2(1):19–33, 1999.

[4] G. Allon and A. Federgruen. Outsourcing service processes to a common service provider under price and time competition. Working paper, Kellogg School of Management, 2008.

[5] Gad Allon and Awi Federgruen. Competition in service industries. *Oper*, 55(1):37–55, January-February 2007.

[6] K. S. Anand, M. Fazel Paç, and S. Veeraraghavan. Quality-speed conundrum: Tradeoffs in customer-intensive services. *Management Science*, 57(1):40–56, 2011.

[7] Krishnan S. Anand, M. Fazl Pa, and Senthil Veeraraghavan. Quality-speed conundrum: Tradeoffs in customer-intensive services. *Management Science*, 57(1):40–56, January 2011.

[8] E.W. Anderson and M.W. Sullivan. The antecedents and consequences of customer satisfaction for firms. *Marketing Science*, 12(2):125–143, 1993.

[9] S. Baiman, P. E. Fischer, and M. V. Rajan. Information, contracting, and quality costs. *Management Science*, 46(6):776–789, June 2000.

[10] Baldridge. The malcolm baldrige national quality improvement act of 1987 - public law 100-107. Available at `http://www.quality.nist.gov/PDF_files/Improvement_Act.pdf`., 1987.

[11] S. R. Bhaskaran and V. Krishnan. Effort, revenue, and cost sharing mechanisms for collaborative new product developement. *Management Science*, 55(7):1152–1169, July 2009.

[12] S. Bhattacharyya and F. Lafontaine. Double-sided moral hazard and the nature of share contracts. *The RAND Journal of Economics*, 26(4):761–781, Winter 1995.

[13] W.R. Bishop. The value equation: Building customer loyalty five ways. *Progressive Grocer*, 63:19–20, March 1984.

[14] S. A. Blackwell, S. L Szeinbach, J. H. Barnes, D. W. Garner, and V. Bush. The antecedents of customer loyalty: An empirical investigation of the role of personal and situational aspects on repurchase decisions. *Journal of Service Research*, 1(4):362–375, 1999.

[15] S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning large call centers. *Operations Research*, 52:17–34, 2004.

[16] G. P. Cachon. *Handbooks in Operations Research and Management Science: Supply Chain Management*. Elsevier, 2003.

[17] G. P. Cachon and M. A. Lariviere. Supply chain coordination with revenue-sharing contracts: Strengths and limitations. *Management Science*, 51(1):30–44, January 2005.

[18] C. J. Corbett, G. A. DeCroix, and A. Y. Ha. Optimal shared-savings contracts in supply chains: Linear contracts and double moral hazard. *European Journal of Operational Research*, 163:653667, 2005.

[19] L.J.M. Coulthard. Measuring service quality - a review and critique of research using servqual. *International Journal of Market Research*, 46(4):479–497, 2004.

[20] J. Dana. Competition in price and availability when availability is unobservable. *RAND Journal of Economics*, 32(3):497–513, 2001.

[21] J. D. Dana. Competition in price and availability when availability is unobservable. *The RAND Journal of Economics*, 32(3):497–513, 2001.

[22] F. de Véricourt and Y. Zhou. A routing problem for call centers with customer callbacks after service failure. *Operations Research*, 53:968–981, 2005.

[23] F. de Véricourt and Y.-P. Zhou. A routing problem for call centers with customer callbacks after service failure. *Operations Research*, 53:968–981, 2005.

[24] Zvi Drezner and Horst W Hamacher. *Facility location: applications and theory*. Springer, 2004.

[25] The Economist. Survey: From wild west to wal-mart. *The Economist*, 371:21, April 17 2004.

[26] Peter S. Fader, Bruce G. S. Hardie, and Ka Lok Lee. "counting your customers" the easy way: An alternative to the pareto/nbd model. *Marketing Science*, 24(2):275–284, 2005.

[27] G.J. Fitzsimons. Consumer response to stockouts. *Journal of Consumer Research*, 27:249–266, September 2000.

[28] N. Gans. Customer loyalty and supplier quality competition. *Management Science*, 48(2):208–221, 2002.

[29] N. Gans and Y. Zhou. Call routing schemes for a call-center outsourcing. *Manufacturing Service Oper. Management*, 9(1):33–50, 2007.

[30] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc, 1998.

[31] S. Gupta and D.R. Lehmann. *Models of customer value. Handbook of Marketing Decision Models*. Springer Science, 2008.

[32] Sunil Gupta, Dominique Hanssens, Bruce G. S. Hardie, William Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. Modeling customer lifetime value. *Journal of Service Research*, 9(2):139–155, November 2006.

[33] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[34] J. M. Hall and E. L. Proteus. Customer service competition in capacitated systems. *Manufacturing Service Oper. Management*, 2:144–165, 2000.

[35] J.M. Hall and E.L. Porteus. Customer service competition in capacitated systems. *Manufacturing & Service Operations Management*, 2:144–165, 2000.

[36] S. Hasija, E. J. Pinker, and R. A. Shumsky. Staffing and routing in a two-tier call center. *International Journal of Operational Research*, 1(1/2):8–29, 2005.

[37] S. Hasija, E. J. Pinker, and R. A. Shumsly. Call center outsourcing contracts under information asymmetry. *Management Science*, 54(3):793–807, 2008.

[38] R. Hassin and M. Haviv. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, 2003.

[39] T. Ho, Y. Park, and Y. Zhou. Incorporating satisfaction into customer value analysis: Optimal investment in life-time value. *Marketing Science*, 25:260–277, 2006.

[40] T. Ho, Y.-H. Park, and Y.-P. Zhou. Incorporating satisfaction into customer value analysis: Optimal investment in life-time value. *Marketing Science*, 25:260–277, 2006.

[41] B. Holstrom. Moral hazard in teams. *The Bell Journal of Economics*, 13(2):324–340, Autumn 1982.

[42] J Wesley Hutchinson. Discrete attribute models of brand switching. *Marketing Science*, 5(4):350–371, 1986.

[43] E. Inarra, A. Mauleon, and V. Vannetelbosch. Efficient structure of provision for emergency public services. *Louvain Economic Review*, 65(1):47–62, 1999.

[44] Abel P. Jeuland. Brand choice inertia as one aspect of the notion of brand loyalty. *Management Science*, 25(7):671–682, July 1979.

[45] W.A. Kamakura, S.N. Ramaswami, and R.K. Srivastave. Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing*, 8:329–349, 1991.

[46] Shohreh A. Kaynama and Christine I. Black. A proposal to assess the service quality of online travel agencies: An exploratory study. *Journal of Professional Services Marketing*, 21(1):63–88, 2000.

[47] J. W. Kim and S. C. Park. Outsourcing strategy in two-stage call centers. *Computers and Operations Research*, 37:790–805, 2010.

[48] S. K. Kim and S. Wang. Linear contracts and the double moral-hazard. *Journal of Economic Theory*, 82(2):342–378, October 1998.

[49] L. Kleinrock. *Queueing Systems, Computer Applications*, volume 2. John Wiley & Sons, 1975.

[50] V. Kostami and S. Rajagopalan. Speed quality tradeoffs in a dynamic model. Working paper, University of Southern California, 2010.

[51] V. Kostami and S. Rajagopalan. Speed quality tradeoffs in a dynamic model. Technical report, University of Southern California, 2010.

[52] K. Krebsbach. Banks' new mantra: Cross sell like crazy. *Bank Investment Consultant*, 10(8):22–28, 2002.

[53] H. Lee, E. J. Pinker, and R. Shumsky. Outsourcing a two-level service process. *Management Science*, 58(8):1569–1584, 2012.

[54] Rober P. Leone, Vithala R. Rao, Kevin Lane Keller, Anita Man Luo, Leigh McAlister, and Rajendra Srivastava. Linking brand equity to customer equity. *Journal of Service Research*, 9(2):125–138, November 2006.

[55] R. McDougall. *How Much Is Poor Customer Service Costing Your Business?* Upstream Works Software, January 2012.

[56] V. Mehrotra, K. Ross, G. Ryder, and Y.-P. Zhou. Routing to manage resolution and waiting time in call centers with heterogeneous servers. Technical report, University of Washington, Seattle, `http://faculty.washington.edu/yongpin`, 2010.

[57] P. Naor. The regulatoin of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

[58] Non-news is good news. *The Economist*, June 9 2012.

[59] Linda K. Nozick and Mark A. Turnquist. Inventory, transportation, service quality and the location of distribution centers. *European Journal of Operational Research*, 129(2):362 – 371, 2001.

[60] B. Ovchinnikov, A., Boulu, and P.E. Pfeifer. Revenue management with lifetime value considerations. *Working Paper*, University of Virginia, 2013.

[61] A. Parasuraman, V.A. Zeithaml, and L.L. Berry. Servqual: A multiple-item scale for measuring customer perceptions of service quality. *Journal of Retailing*, 64(1):12–40, 1988.

[62] P.E. Pfeifer and A. Ovchinnikov. A note on willingness to spend and customer lifetime value for firms with limited capacity. *Interactive Marketing*, 25(3):178–189, 2011.

[63] R. S. Randhawa and S. Kumar. Usage restriction and subscription services: operational benefits with rational customers. *Manufacturing and Service Oper. Management*, 10(3):429–447, 2008.

[64] F.F. Reichheld. *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value*. Harvard Business School Press, 2001.

[65] F.F. Reichheld and W.E. Sasser Jr. Zero defections: Quality comes to services. *Harvard Business Review*, September-October:105–111, 1990.

[66] W. Reinartz and R. Venkatesan. *Decision models for customer relationship management (CRM)*. Springer Science, 2008.

[67] Werner J. Reinartz, Jacquelyn S. Thomas, and V. Kumar. Balancing acquisition and retention resources to maximize customer profitability. *Journal of Marketing*, 69(1):63–79, January 2005.

[68] J. Ren and Y.-P. Zhou. Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2), 2008.

[69] Z. J. Ren and F. Zhang. Service outsourcing: Capacity, quality, and correlated costs. Working paper, http://apps.olin.wustl.edu/faculty/zhang/Zhang-Journal/qualityoutsourcing.pdf, 2009.

[70] Z. J. Ren and Y. Zhou. Call center outsourcing: Coordinating staffing level and service quality. *Management Science*, 54(2):369–383, February 2008.

[71] G. Roels. The economics of joint production in services. Working paper, UCLA Anderson School of Management, October 2012.

[72] G. Roels, U. S. Karmarkar, and S. Carr. Contracting for collaborative services. *Management Science*, 56(5):849–863, 2010.

[73] R.T. Rust and T.S. Chung. Marketing models of service and relationships. *Marketing Science*, 25(6):560–580, 2006.

[74] T. Rust, K.N. Lemon, and V.A. Zeithmal. Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing*, 68(1):109–127, 2004.

[75] R. Saouma. Optimal second-stage outsourcing. *Management Science*, 54(6):1147–1159, June 2008.

[76] David C. Schmittlein, Donald G. Morrison, and Richard Colombo. Counting your customers: Who are they and what will they do next? *Management Science*, 33(1):1–24, January 1987.

[77] David A. Schweidel, Peter S. Fader, and Eric T. Bradlow. A bivariate timing model of customer acquisition and retention. *Marketing Science*, 27(5):829–843, September-October 2008.

[78] P.B. Seybold and R.T. Marshak. *Customers.com: How to Create a Profitable Business Strategy for the Internet and Beyond*. Times Business, 1998.

[79] R. J. Shapiro. *Futurecast: How Superpowers, Populations, and Globalization Will Change The Way That You Live and Work*. Macmillan, 2008.

[80] K. T. Talluri and G. J. van Ryzin. *The Theory and Practice of Revenue Management*. Springer, 2005.

[81] Jacquelyn S. Thomas. A methodology for linking customer acquisition to customer retention. *Journal of Marketing Research*, 38(2):262–268, May 2001.

[82] Donald M. Topkis. *Supermodularity and Complementarity*. Princeton University Press, 1998.

[83] Andy A. Tsay and Narendra Agrawal. Channel dynamics under price and service competition. *Manufacturing & Service Operations Management*, 2(4):372–391, Fall 2000.

[84] S. Walker. Wachovia extends cross-selling efforts to build on success. *Bond Buyer*, 344, 2003.

[85] W. Whitt. Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2):114–161, 1993.

[86] W. Whitt. Partitioning customers into service groups. *Management Science*, 45(11):1579–1592, 1999.

[87] R. W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, 1989.

[88] M. Xue and J. M. Field. Service coproduction with information stickiness and incomplete contracts: Implications for consulting services design. *Production and Operations Management*, 17(3):357–372, 2008.

[89] Y. Yu, S. Benjaafar, and Y. Gerchak. Capacity sharing and cost allocation among independent firms in the presence of congestion. Working paper, 2009. `http://isye.umn.edu/faculty/pdf/ybg-04-05-09.pdf`.

[90] V.A. Zeithaml. Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence. *Journal of Marketing*, 52:2–22, July 1988.

[91] Y. Zhou and Z. J. Ren. Service outsourcing. In *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, 2010. `http://onlinelibrary.wiley.com/book/10.1002/9780470400531`.