Optimal Pricing and Capacity Planning in Operations Management

by

Dehui Tong

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Joseph L. Rotman School of Management
University of Toronto

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-77662-9

Our file Notre référence
ISBN: 978-0-494-77662-9

Abstract

Optimal Pricing and Capacity Planning in Operations Management

Dehui Tong

Doctor of Philosophy

Graduate Department of Joseph L. Rotman School of Management

University of Toronto

2011

Pricing and capacity allocation are two important decisions that a service provider needs to make to maximize service quality and profit. This thesis attempts to address the pricing and capacity planning problems in operations management from the following three aspects.

We first study a capacity planning and short-term demand management problem faced by firms with industrial customers that are insensitive to price incentives when placing orders. Industrial customers usually have downstream commitments that make it too costly to instantaneously adjust their schedule in response to price changes. Rather, they can only react to prices set at some earlier time. We propose a hierarchical planning model where price decisions and capacity allocation decisions must be made at different points of times. Customers first sign a service contract specifying how capacity at different times will be priced. Then, when placing an order, they choose the service time that best meets their needs. We study how to price the capacity so that the customers behave in a way that is consistent with a targeted demand profile at the order period. We further study how to optimally allocate capacity. Our numerical computations show that the model improves the operational revenue substantially.

Second, we explore how a profit maximizing firm is to locate a single facility on a general network, to set its capacity and to decide the price to charge for service. Stochastic demand is generated from nodes of the network. Customers demand is sensitive to both the price and the time they expect to spend on traveling and waiting. Considering the combined effect of location and price on the firm's profit while taking into account the demand elasticity, our model provides managerial insights about how the interactions of these decision variables impact the firm's profit.

Third, we extend this single facility problem to a multiple facility problem. Customers have multiple choices for service. The firm maximizes its profit subject to customers' choice criteria. We propose a system optimization model where customers cooperate with the firm to choose the facility for service and a user equilibrium model where customers choose the facilities that provide the best utility to them. We investigate the properties of the optimal solutions. Heuristic algorithms are developed for the user equilibrium model. Our results show that capacity planning and location decisions are closely related to each other. When customers are highly sensitive to waiting time, separating capacity planning and location decisions could result in a highly suboptimal solution.

Dedication

To my parents Mingkui Tong and Tingyun Xue, whose boundless love and firm belief in me helped me to accomplish my goals

and

To my family, Pengfei Yang and Robert Yang for their love and support.

Acknowledgements

I would like to express my gratitude and indebtedness to my supervisors, dissertation committee, friends and family members who have helped me successfully fulfill this endeavor.

First of all, I am heartily thankful to my supervisors, Professors Oded Berman and Dmitry Krass, who provided me the encouragement, guidance and support throughout my program of studies. I am also deeply grateful to Professors Joseph Milner and Opher Baron for their thoughtful suggestions, insightful directions and tremendous help without which these research could not have been possible. As well, I would like to thank Professor Philipp Afeche for his constructive and valuable comments while serving as my dissertation committee member.

Second, I would like to thank my mother Tingyun Xue and father Mingkui Tong, who instilled me with a love of learning, the courage and confident to always take on new challenges. I am grateful to my mother-in-law Minjie Liu and father-in-law Zhongming Yang, for their warm support and help.

I would like to send all my love and gratefulness to my husband Pengfei, who has the greatest confidence and pride in me and supported me unconditionally; and my son Zihan, who has been so independent and patient, for bearing my frequent and extended absence from home.

My heartfelt thanks to Yanrong Cao and Andrew Ching, for brightening my day during gloomy times.

Last but not least, I must thank Rongbing Huang, Binbin Liu, Seokjin Kim for their companionship and kindness in boosting my spirit.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the past decade, there has been a growing consensus among researchers and practitioners alike that the pricing decisions that induce demand are closely linked with capacity planning decisions. Pricing decisions influence the demand patterns that form the basic inputs to any capacity management, and pricing decisions in turn must ultimately be based on capacity constraints. To improve service quality and to increase profit, integration of pricing and capacity decisions in practice are necessary, as can be observed in many industries. For example, in the airline industry where capacity is fixed well in advance and can be augmented only at a relatively high marginal cost, pricing service (i.e., setting ticket prices) dynamically over time in light of the remaining capacities has been widely applied as a standard practice. In the fashion retail business, it is common to find that retailers set the prices of goods differently according to their colors, with due regard to the stock of the items and the expected demand. In the service industry, the price and capacity decisions often depend on the locations of the facilities.

While researchers and practitioners strive to make better decisions, the need for improvement never ends. The pricing and capacity allocation decision varies from industry to industry, so that there is no one universal solution that fits all situations. Making decisions on price and capacity is a complicated process, with other management decisions that need to be made at the same time, for example, on production quantity, inventory control, and the locations of service facilities.

This thesis studies several aspects of the integration of capacity allocation and pricing decisions in Operations Management. The objective is to study how the integration of price and capacity decisions can affect a firm's profit, and how the joint decisions interact with other managerial decisions. We address the problems from the following three aspects.

We begin in Chapter 2 with the analysis of a demand and capacity management problem that originates from a concrete distributor. The challenge of the concrete distribution planning is that the capacity supply and demand are highly imbalanced. On a daily basis, the system frequently runs out of capacity and thus causes serious delays during peak demand periods, while during off-peak times it is under-utilized. On the other hand, customer demands are highly time-sensitive because concrete is a very perishable product and the delay cost is relatively high.

A general approach to dealing with uncertainty demand under capacity constraints is to adopt some dynamic pricing mechanism so that the arriving orders are priced according to the remaining capacity level and other available market information at the time of order arrival. Price is a variable that can be controlled on a continuous basis. Though dynamic pricing has been adopted to many industries such as airline, hotel, car rental etc., it cannot be applied to firms like the concrete distributor with industrial customers. Offering customer price incentives on a short notice is not effective, since when industrial customers place their orders they have already deployed workers and equipments for the delivery, and thus the cost to change the planned delivery time is high. Instead of this dynamic pricing approach, we propose a variable pricing approach within a hierarchical planing scheme, by pricing service according to capacity usage time to regulate the fluctuation of demand. At the strategic stage, a portion of service fee is set according to capacity usage time. Therefore, customers requesting service during peak times will be charged more, but less during off-peak periods. At the operational stage, customers observe posted prices, arrive and are assigned capacity slots dynamically over a planning horizon. We show that a pricing strategy can induce the customers to cooperate with the firm, such that the demand flow using the capacity over time is consistent with the firm's preference. Our computational experiments show that using the hierarchical planning approach can improve the firms' operational revenue.

While in some problems we can investigate the interaction of pricing and capacity allocation decisions directly, there are other problems where the joint analysis of pricing and capacity must be combined with other management decisions. For example, the parking fee and the size of a parking lot in a downtown area can be intuitively different than those in a rural area. When a downtown parking lot is planned, it is necessary to carefully consider the capacity level (lot size) and the prices to charge with regard to the location to make the project viable and to maximize profit. Obviously, charging a higher price will result in a lower occupancy, but charging too low will make the parking system highly congested and may result in revenue loss. Hence, when a firm considers the location of a new facility, the decisions on pricing and capacity are often made simultaneously. Two examples of these problems involving location decisions are studied in Chapters 3 and 4.

In Chapter 3, we focus on several of the most important strategic decisions for a service provider facing uncertain customer demand, including setting the location of the facility, determining the service capacity, and choosing the price to charge for service. The service provider operates a single facility. Utility-maximizing customers are assumed to reside at the nodes of the network, generating Poisson demand streams. Customer utility is affected by the price, travel distance and the waiting time at the facility selected by the customer. The unique feature of our model is, that we explicitly recognize that the total demand generated by each customer is affected by the degree of congestion at the facilities, which, in turn, is affected by the choices made by other customers. Thus, the distribution of customer flows is guided by the equilibrium: the demand rate at a facility depends on the waiting time incurred, which is by itself a function of the demand rate. The objective is to maximize the total profit of the facilities. We start with a general G/G/1 system and show that an optimal location exists on the nodes of a network. Afterwards, we consider a M/M/1 system and analyze the correlations of the three decision variables. An exact solution procedure is developed for the M/M/1 system with exponential demand elasticities.

Chapter 4 extends the single facility problem to a multiple-facility one, where customers have multiple choices for service. We developed models to optimize jointly the three strategic decisions with an emphasis on customers' behavior: location of the facilities, service capacities,

and pricing for service. When capacities are limited and congestions exist in the system, customers' decisions are inter-related. Two models are developed: a system optimization model and a user equilibrium model. In the system optimization model, the firm can assign customers to facilities to maximize the profit. In the user equilibrium model, customers are self-interested, the firm maximizes its profit subject to the equilibrium behavior of the customers. We formulated the user equilibrium problem as a bi-level programming with equilibrium constraints. We discuss the existence of the customer equilibrium and show that the distribution of the equilibrium flow can be solved as a convex optimization problem via a variational inequality approach. We suggest several heuristic algorithms to solve the problem and present a numerical analysis to study the customer's equilibrium behavior effects on the firm's profit. We show that the price decision is independent of the capacity allocation and location decisions. Our numerical results further demonstrate that location and capacity allocation can be made independently when customers sensitivity to travel distance dominates the waiting time. However, when customers are more sensitive to waiting time than travel distance, capacity allocation and location decisions should be jointly optimized to achieve maximum profit.

Combining the three chapters discussed above, we present practical formulations and provide optimal solutions for pricing and capacity allocation decisions in a dynamic distribution system and strategic location models. Our objective for developing these solution procedures is to use them to gain understanding of the relationship between capacity and prices. Though our study is largely theoretical in nature, we hope our results provide some important insights for real life managerial problems.

Chapter 2

Variable Pricing in a Capacity Constrained Just-In-Time Supply Chain

2.1 Introduction

A central concept of contemporary operations management is the use of pricing to dynamically match supply and demand. To maximize revenue, joint pricing/allocation schemes have been widely used by capacity-constrained service industries such as airlines, hotels, and car rental agencies. Price is often used as a control variable, by lowering the price customers are admitted for service and by raising the price sufficiently high they are turned away.

A fundamental assumption of such dynamic pricing schemes is that price and capacity allocation decisions are made at the same time. Customers are assumed to be willing to change their choices in response to the price when placing their orders. In this paper, however, we study a revenue management problem where price and capacity allocation decisions must be made at different points in time. We consider a firm serving industrial customers that have their own commitments to their workforce and downstream customers. Because the commitments must be made a long time before orders will be placed, the customers have little flexibility to alter their schedules when capacity reservations are made. To reduce congestion, the service provider firm

establishes prices that would encourage such customers to schedule their commitments more evenly over a service period, such as a day. In our approach, the firm determines prices for alternate supply schedules when the overall demand and preferences for capacity schedules are known. Then the customers react to the prices and place orders at a later time. Next, the firm allocates capacity to these orders to maximize its profit. In particular, we study how the firm should price capacity so that customers behave consistently with a target demand profile, and as a result the firm can allocates its production capacity accordingly.

The study is motivated by the demand management problem of a ready-to-mix concrete distributor. On-time delivery is very important in a competitive concrete market that has little product differentiation. The firm's customers sign long-term contracts for capacity usage. Each day, they place demand for capacity that typically peaks at certain times of the day so that there is insufficient capacity to serve all the customers, resulting in delays in service. To address the problem the firm could expanse its capacity, but it would be under-utilized during the off-peak periods. According to common operations management practice, strategic pricing could be used to improve the dispatching schedule. However, there are a number of factors that make the problem difficult:

First, customers often have inflexible schedules when placing their orders. Their cost of changing the delivery times is very high, as normally they have already committed the manpower and equipment for the delivery time. Thus online dynamic pricing may not be effective in moving demand away from the peak.

Second, customers may have multiple orders. A customer may request the usage of the capacity at a set of specific times. Concrete is a special product that requires delivery to be made according to specified schedules. Thus, deciding which customers should be admitted at the operational stage is very difficult when the capacity is limited.

Third, deviation from an agreed delivery schedule can be very costly. The short shelf-life of concrete requires that every step from production to final utilization by the customer is "Just-In-Time." The customer has to set up the construction site to allow offloading to start immediately upon the arrival of the concrete delivery truck. Any delay will incur waiting costs for customers' workers and equipments. Ready-to-mix concrete can even solidify in the truck

if offloading is delayed by a few hours.

In this study, we intend to provide an integrated pricing and admission control framework that applies to problems with similar characteristics. These characteristics include demand fluctuation, perishable capacity, limited flexibility of schedules, and multiple demands from each customer. We use a hierarchical planning approach to manage the imbalance of the capacity and demand. The problem was studied under a monopolistic setting with the consideration of the special feature that pricing decision cannot be made jointly with the admission control. We split the problem into two components: strategic pricing and operations control.

On strategic pricing, we study the ability to smooth the expected demand by pricing the use of capacity over time, i.e., charging customers usage cost according to the time of usage. We model customers as self-interested with their own utilities that depend on the time of service, congestion level of the system, and price. At this stage, we study the following two questions: 1) How should the firm set the prices to induce customers to act such that their demand is consistent with the firm's preferred booking limit; 2) what pricing strategy achieves the maximum profit?

At the operational level, the firm determines whether to accept or reject customers' requested schedules based on the availability of capacity and expected future demand. Customers arrive randomly, observe posted prices, and choose a delivery schedule. With a strategic pricing policy, it is expected that the average demand rate would be more evenly distributed within the workday. We input this more evenly distributed demand to explore the optimal admission strategies.

The organization of this Chapter is as follows. Section 2.2 provides a brief literature review. In Section 2.3 we discuss how incentive pricing may be used to induce the customers to schedule according to the firm's capacity. We discuss how to find the booking limit and incentive prices simultaneously to maximize the firm's profit. In Section 2.4 we analyze the firm's operational problems. We formulate the admission control problem and discuss its structural properties. Problems of realistic size are too big to be exactly solve. Therefore, we provide an upper bound and value function approximation for the operational problem. We show the benefits of the hierarchical planning approach through numerical computations and perform sensitivity

analysis in Section 2.5. Concluding remarks and further research extension are provided in Section 2.6.

2.2 Related Literature

Our work is related to research in several areas. The first stream of literature is from the dynamic pricing area, see e.g., Gallego and van Ryzin (1994, 1997). Price is treated as a regulator of demand for the purpose of admission control. At the time of admission where capacity information is known exactly, a firm can use price to open or block some class of customers to maximize its profit. For a comprehensive review of the literature in dynamic pricing, see McGill and van Ryzin (1999), Elmaghraby and Keskinocak (2003), Bitran and Caldentey (2003), and Talluri and van Ryzin (2005). Dynamic pricing in the above literature has had wide application in the airline, hotel and retailing industries. However, such models cannot be applied to cases where customers are insensitive to price at the time when the orders are placed.

We consider a case where pricing decisions cannot be made together with admission control. Therefore we use a hierarchical planning framework (see e.g., Bitran and Tirupati (1989) for an introduction of hierarchical planning approach) where decisions are made at both strategic and operational levels.

At the strategic level, we consider the problem of analyzing and influencing the behavior of customers at shared resources. This problem has been extensively studied by many different research disciplines such as computer science, economics, and transportation. For example, Mirrlees (1971) studies an optimal income taxation problem faced by government or large organizations with the objective to maximize the social welfare (see, e.g., Boadway et al. (1998) for a broad survey). Works in transportation research are more closely related to our research. Wardrop (1952) defines the traffic equilibrium that later serves as a foundation of many research in studying the toll pricing on a traffic network. Smith (1979) provides a theoretical foundation that guarantees the existence and uniqueness of the traffic equilibrium. Dafermos (1973) addresses the problem of pricing in a multiclass transportation network. Our work is

somewhat similar to Dafermos' approach to find the optimal prices. He focuses on toll pricing that is system optimal, and customers are discriminated by their classes. In contrast, we investigate a pricing policy that can induce customers to follow the firm's target booking limit, and our pricing policy has a simplified structure that is more practical in its application. There are also some works that focus on price inefficiency in a network of shared resources. Koutsoupias and Papadimitriou (1999) are the first to quantify degradation in network performance due to unregulated traffic. Roughgarden and Tardos (2002) later prove that if the congestion cost of each edge is a linear function of its flow, then the total congestion cost of the routes chosen by selfish network users is at most 4/3 times the minimum possible total congestion cost. Cole et al. (2003) further extend the result to price network edges with heterogeneous users. We have no intent to quantify the inefficiency of selfish routing, which has been well studied. Instead, we are interested in finding a pricing strategy that can induce a preferred booking limit and regulate demand flow to help admission control at the operational level in the context of hierarchical planning approach.

Our admission control model at the operational level can be viewed as a particular instance of the general class of network revenue management models that are typically applied to sequential reservations for an airline network, hotel, or car rental service. The network revenue management model studies problems in a stochastic and dynamic environment and answers the questions of whether a new request to use a set of resource links should be accepted or rejected. Bertsimas and Popescu (2003) investigate dynamic policies for allocating scarce inventory to stochastic demand for multiple fare classes in a network environment so as to maximize total expected revenues. They propose and analyze a new algorithm based on approximate dynamic programming. The admission control algorithm in our paper uses a similar approach as the Certainty Equivalent Control algorithm they introduced. The major difference between our work and theirs is that they treat the demand as exogenously given and independent of price, while our demand depends on the price established in the framework of the strategic level. In addition, they consider the extension to overbooking at the last stage of the booking process, while our overbooking decisions are made "on the go".

2.3 Strategic Decisions on Pricing

We consider a firm that needs to allocate capacity to a large number of customers over a short horizon such as a day. The firm has limited capacity per unit time ("slots"), which is used to serve customers with various schedule preferences. Customers are divided into separate classes. Each class is defined by the price they pay per unit of capacity, the utility they receive from service, their feasible set of capacity allocations, and their valuation of service delay. At the strategic level, the problem we consider is how the firm should establish time-dependent variable prices for capacity over the service horizon. Given the prices, we assume that each customer chooses the capacity slots that fit her schedule to maximize the utility. In fact, this corresponds to a two-stage Stackelberg game with complete, but imperfect, information between the firm and customers. The firm is the leader setting prices while anticipating the subsequent behavior of the customers. Each price vector defines a different subgame, and given prices customers play the subgame. The decisions of the customers result in a capacity utilization schedule.

The objective at the strategic level is to establish a price scheme so that the customers' equilibrium demand in the game framework is consistent with the firm's preferred capacity utilization schedule (the target schedule). If the firm does not know which target schedule is most profitable, the optimal schedule and price scheme need to be established at the same time. In this section, we demonstrate how to establish the price scheme and the firm's target schedule. In the next section we discuss the capacity allocation problem at the operational level that uses the decisions of the strategic pricing as an input.

2.3.1 Problem Definition

We use column vectors throughout the paper. Vectors and matrices are in bold format. $\mathbf{a}'$ is the transpose of a vector $\mathbf{a}$. $\mathbf{A} = (\mathbf{a}_1, \cdots, \mathbf{a}_n)$ is a matrix obtained by combining its vector elements, $\mathbf{a}_i, i = 1, \cdots, n$ along the natural dimension.

The main problem we address in this section is as follows. A firm has a finite service horizon. The service horizon is divided into $T \geq 2$ time intervals of equal length that are used to serve a large number of customers with various schedule preferences. Without loss of

Table 2.1: List of Notation

$\mathbf{a}_s^m$	A demand pattern of class-m customers, a $0-1$ vector of length T
$\mathbf{A}^m$	The demand pattern matrix of class-m customers, $\mathbf{A}^m = (\mathbf{a}_1^m, \cdots, \mathbf{a}_{S_m}^m)$
$\mathbf{A}$	The demand pattern matrix for all customers, $\mathbf{A} = (\mathbf{A}^1, \cdots, \mathbf{A}^M)$
c_t	The maximum capacity at service period t
$\mathbf{c}$	The maximum capacity vector, $\mathbf{c} = (c_1, \cdots, c_T)'$
δ	The probability of one arrival in a reservation period
$\mathbf{e}^m$	A unit vector in length S_m
$g(\mathbf{y}, m, i)$	The overflow cost of admitting a type (m, i) arrival when the system state is $\mathbf{y}$
h^m	The rejection cost for a class-m customer
I_m	The indifferent set of class-m customers
K	The total number of reservation periods
κ_t	The unit overflow cost in service period t
$\boldsymbol{\kappa}$	The overflow costs vector, $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_T)'$
l_m	The total number of deliveries from a class-m customer
λ^m	The total number of class-m customers
λ	The total number of all customers
M	The total number of customer classes
$\mathcal{M}$	The set of customer classes, $\mathcal{M} = \{1, \cdots, M\}$
$\bar{p}^m$	The nominal price of a class-m customer
$\bar{\mathbf{p}}^m$	The nominal prices vector for class-m customers, $\bar{\mathbf{p}}^m = (\bar{p}^1, \cdots, \bar{p}^M)'$
$\tilde{p}_t$	The variable price of a unit service at period t, identically applied to all customer classes
$\tilde{\mathbf{p}}$	The variable prices vector, $\tilde{\mathbf{p}} = (\tilde{p}_1, \cdots, \tilde{p}_T)'$
p_s^m	The price of a demand pattern s for a class-m customer
$\mathbf{p}^m$	The price vector for class-m customer, $\mathbf{p}^m = (p_1^m, \cdots, p_{S_m}^m)'$
$\mathbf{P}$	The set of price vectors, $\mathbf{P} = \{\mathbf{p}^1, \cdots, \mathbf{p}^M\}$
ψ_i^m	The probability that a class-m customer chooses a demand pattern i in a reservation period
$r(m, i)$	The operational revenue by admitting a type (m, i) arrival
S_m	The total number of demand patterns of a class-m customer
$\mathcal{S}_m$	The set of demand patterns for class-m customers, $\mathcal{S}_m = \{1, \cdots, S_m\}$
T	The total number of service periods
θ_t	The service level coefficient at period t
$\boldsymbol{\theta}$	The service level coefficient vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_T)'$
u_s^m	The utility of a class-m customer that chooses demand pattern s
$\mathbf{u}^m$	The utility vector for class-m customers, $\mathbf{u}^m = (u_1^m, \cdots, u_{S_m}^m)'$
$\hat{u}^m$	The equilibrium utility of class-m customers
$\hat{\mathbf{u}}^m$	A vector of $\hat{u}^m$ in length S_m
v_s^m	The valuation of a class-m customer for a demand pattern s
$\mathbf{v}^m$	The service valuation vector for a class-m customer, $\mathbf{v}^m = (v_1^m, \cdots, v_{S_m}^m)'$
$\tilde{v}_t^m$	The valuation of a class-m customer for a unit service at period t
$\tilde{\mathbf{v}}^m$	The service valuation vector for a class-m customer, $\tilde{\mathbf{v}}^m = (\tilde{v}_1^m, \cdots, \tilde{v}_T^m)'$
V_k	The expected maximum operational revenue with k reservation periods to go
$\bar{V}_k$	The approximate expected operational revenue with k reservation periods to go
w_t^m	The congestion cost of a class-m customer for a unit service at period t
$\mathbf{w}^m$	The congestion cost vector in each service period for class-m, $\mathbf{w}^m = (w_1^m, \cdots, w_T^m)'$
x_s^m	The number of class-m customers that request demand pattern s
$\mathbf{x}^m$	The demand pattern assignment vector for class-m customers, $\mathbf{x}^m = (x_1^m, \cdots, x_{S_m}^m)'$
$\mathbf{x}$	The assignment vector obtained by stacking its individual vectors, $\mathbf{x} = (\mathbf{x}^1; \cdots; \mathbf{x}^M)$
$\mathbf{X}$	The set of demand pattern assignment vectors, $\mathbf{X} = \{\mathbf{x}^1, \cdots, \mathbf{x}^M\}$
y_t	The total number of customers in service period t
$\mathbf{y}$	A vector indicating the number of customers in each service periods, $\mathbf{y} = (y_1, \cdots, y_T)'$
$\tilde{y}_t$	The firm's target flow at service period t
$\tilde{\mathbf{y}}$	The firm's target flow vector, $\tilde{\mathbf{y}} = (y_1, \cdots, y_T)'$

generality, we fix the length of a time interval at one. Let c_t be the maximum capacity at time t and $\mathbf{c} = (c_1, \cdots, c_T)'$. There are a total of M customer classes. We denote the set of customer classes by $\mathcal{M} = \{1, \cdots, M\}$. A customer class-$m$ has S_m potential demand fulfillment patterns, $\mathbf{a}_s^m, s = 1, \cdots, S_m$. Each pattern $\mathbf{a}_s^m = (a_1^m, \cdots, a_T^m)'$ is a 0-1 vector of length T such that $\sum_{t=1}^{T} a_t^m = l_m$, where l_m is the number of deliveries requested by a class-m customer. Let $\mathbf{A}^m = (\mathbf{a}_1^m, \cdots, \mathbf{a}_{S_m}^m)$ be the $T \times S_m$ matrix of feasible demand fulfillment patterns of class m. Let $\mathcal{S}_m = \{1, \cdots, S_m\}$ be the set of demand pattern indexes. For example, a feasible demand pattern matrix for a type 2 customer and 4 service periods, i.e., $l_m = 2$ and $T = 4$ can be:

$$\mathbf{A}^2 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

Let λ^m be the total number of class-m customers. Let x_s^m be the number of class-m customers that request demand pattern $s \in \mathcal{S}_m$ and $\mathbf{x}^m = (x_1^m, \cdots, x_{S_m}^m)'$. For notational convenience, let $\mathbf{X} = \{\mathbf{x}^1, \cdots, \mathbf{x}^M\}$. Let y_t be the total number of customers requesting service at time t and let $\mathbf{y} = (y_1, \cdots, y_T)'$. Then,

$$\mathbf{y} = \sum_{m \in \mathcal{M}} \mathbf{A}^m \mathbf{x}^m. \tag{2.1}$$

We assume that customers value the service according to the time of delivery and the demand patterns. Let v_s^m be the valuation of a class-m customer for a demand pattern s and $\mathbf{v}^m = (v_1^m, \cdots, v_{S_m}^m)'$. Further we assume that customers incur a congestion cost based on the number of other customers requesting service at the same time, i.e. the congestion cost depends the number of orders in the system $\mathbf{y}$. Though $\mathbf{y}$ is not known in advance, we assume that customers either observe the firm's posted booking limit or learn this through their experience (an equilibrium behavior that will be discussed later) to evaluate their congestion cost. Let $w_t^m(\mathbf{y})$ be the congestion cost of a class-m customer for a unit service at time t and $\mathbf{w}^m(\mathbf{y}) = (w_1^m(\mathbf{y}), \cdots, w_T^m(\mathbf{y}))'$. The congestion cost for a class-m customer that has demand

pattern s thus is $\mathbf{w}^m(\mathbf{y})'\mathbf{a}_s^m$. Let p_s^m be the price class m pays for a demand pattern s and $\mathbf{p}^m = (p_1^m, \cdots, p_{S_m}^m)'$. Let $\mathbf{P} = \{\mathbf{p}^1, \cdots, \mathbf{p}^M\}$. A class-$m$ customer who chooses demand pattern s receives a utility:

$$u_s^m(\mathbf{X}, \mathbf{P}) = v_s^m - p_s^m - \mathbf{w}^m(\mathbf{y})'\mathbf{a}_s^m \qquad \forall m \in \mathcal{M}, \forall s \in \mathcal{S}_m. \tag{2.2}$$

where $\mathbf{y}$ depends on $\mathbf{X}$ through (2.1). Let $\mathbf{u}^m(\mathbf{X}, \mathbf{P}) = (u_1^m(\mathbf{X}, \mathbf{P}), \cdots, u_{S_m}^m(\mathbf{X}, \mathbf{P}))'$.

We use the Wardrop Equilibrium, see Wardrop (1952) and Appendix A to describe the customers' equilibrium behavior. We assume that given prices, customers choose the demand pattern so that in equilibrium, no customer has any incentive to alter her preference. The Wardrop equilibrium principle applies to a situation where there is a very large number of infinitesimal users. This implies that in our case, when a customer unilaterally switches her choice from one schedule to another, there is no measurable change in the utility of other customers. We thus relax the integrality of $\mathbf{x}^m$ requirement in our model. (For a discussion of the relationship of Nash Equilibrium and Wardrop Equilibrium, see Haurie and Marcotte (1985).)

Definition 2.1. Customer Equilibrium. *Given a price $\boldsymbol{P}$, $\boldsymbol{X}^*$ is in equilibrium, if for every class-m customer,*

$$x_s^{m*} > 0 \qquad \textit{implies} \qquad u_s^m(\boldsymbol{X}^*, \boldsymbol{P}) \geq u_t^m(\boldsymbol{X}^*, \boldsymbol{P}) \qquad \forall t \in \mathcal{S}_m. \tag{2.3}$$

If $x_s^m > 0$, there is positive flow for demand pattern s. The definition implies that the equilibrium utility of all positive flow demand patterns for class-m customers is the same. All other unused demand patterns for class-m customers would provide a weakly lower utility. However, several demand patterns may provide the same maximum utility. Let $\hat{u}^m$ be the equilibrium utility of class-m customer induced by $\mathbf{X}^*$ given $\mathbf{P}$. An alternative definition of the customer equilibrium is, given $\mathbf{P}$, $\mathbf{X}^*$ is in equilibrium if and only if for every class $m \in \mathcal{M}$ and

demand pattern $s \in \mathcal{S}_m$:

$$u_s^m(\mathbf{X}^*, \mathbf{P}) \begin{cases} = \hat{u}^m & \text{if } x_s^{m*} > 0, \\ \leq \hat{u}^m & \text{if } x_s^{m*} = 0. \end{cases} \tag{2.4}$$

To ensure the existence of a customer equilibrium, we assume that $\mathbf{w}^m(\mathbf{y})$ has a monotonicity property (see Appendices B for this definition and other relevant comments). We next express the customer equilibrium defined in (2.3) and (2.4) in another form which will be used later as constraints in a mathematical programming formulation in Section 2.3.3. Let $\hat{\mathbf{u}}^m$ be a column vector of $\hat{u}^m$ in length S_m, then (refEquilibrium2) can be written as,

$$\begin{aligned} (\hat{u}^m - u_s^m(\mathbf{X},\mathbf{P}))x_s^m = 0 \quad & \forall s \in \mathcal{S}_m, \forall m \in \mathcal{M}, \\ \hat{\mathbf{u}}^m - \mathbf{u}^m(\mathbf{X},\mathbf{P}) \geq \mathbf{0} \quad & \forall m \in \mathcal{M}. \end{aligned} \tag{2.5}$$

(2.5) requires that for any class-m, the utilities associated with $x_s^m > 0$, $\forall s \in \mathcal{S}_m$, are the same and equal to $\hat{u}^m$. All other demand patterns provide no better utility.

Consider a firm that has a preferred booking limit $\tilde{y}_t$ at each service time t. Let $\tilde{\mathbf{y}} = (\tilde{y}_1, \cdots, \tilde{y}_T)'$ be the *Target Flow*. In general, the equilibrium flow given by (2.1) resulting from self-interested customers will not match the firm's target flow $\tilde{\mathbf{y}}$.

In this section, we consider two problems. First, can we select a price $\mathbf{P}$ so that self-interested customers choose the firm's target flow? Second, how should the firm select a target flow to maximize its profit?

We assume that the price charged to customers is composed of two parts: a class-dependent nominal price and a time-dependent variable price. Let $\bar{p}^m$ be the nominal price for class-m. We use such nominal price to capture the practice of offering class-dependent and loyalty discounts. Let $\tilde{p}_t$ be the variable price charged at time t. Typically $\tilde{p}_t$ will be higher at more popular times. Let $\bar{\mathbf{p}} = (\bar{p}^1, \cdots, \bar{p}^M)'$ be the nominal price vector and $\tilde{\mathbf{p}} = (\tilde{p}_1, \cdots, \tilde{p}_T)'$ be the variable price vector. Then, the price for a class-m who receives demand pattern s is

$$p_s^m = \bar{p}^m l_m + \tilde{\mathbf{p}}' \mathbf{a}_s^m. \tag{2.6}$$

It has been established in the traffic equilibrium literature (Dafermos, 1973) that when multiple classes of customers are offered a proper set of class-dependent prices, the user optimal solution is also system optimal. In our case, we can find a price for each order such that the user optimal solution maximizes the combined utility of the firm and the customers, i.e., the solution is system optimal. Our price follows an additive structure as in (2.6), i.e., variable prices are identically applied to all customer types. (We cannot solve for the prices in the multiplicative case.) In the next two subsections, we show how to find the prices so that customers behave in a way that is consistent with the firm's targeted demand profile, and how to establish the booking limit and pricing policy so as to maximize the profit when the firm does not have a preferred booking limit.

2.3.2 Strategic Pricing for a Specific Target Flow

In this subsection, we find a pricing strategy that can induce customers to choose demand patterns such that the resulting equilibrium demand is consistent with the firm's target flow. Two relevant questions are: 1) Is there a pricing policy that induces a given target flow? 2) How can we find a price mechanism based on (2.6) to induce a specific target flow?

Let $\tilde{\mathbf{y}}$ be the firm's preferred target flow. The target flow should be below the maximum capacity. Further, it is possible that the firm may want to satisfy some service level in its target flow. For example, it may not want to exceed some percentage of the maximum capacity. In general there maybe a constraint

$$(f_t(\tilde{\mathbf{y}}) - \theta_t c_t)^+ \leq \xi_t, \qquad \forall t \in \mathcal{T} \tag{2.7}$$

where $f_t(\tilde{\mathbf{y}})$ is a function $R^T \to R$, $\theta_t \geq 0$ is a utilization coefficient, and $\xi_t \geq 0$.

Suppose $\tilde{\mathbf{y}}$ is posted as public information for all customers. $\tilde{\mathbf{y}}$ would be a reasonable approximation of the number of orders in each service periods by the end of the reservation periods. Thus, we assume customers would use $\tilde{\mathbf{y}}$ to evaluate their congestion cost. Further, we assume the congestion costs $\mathbf{w}^m(\tilde{\mathbf{y}}), \forall m \in \mathcal{M}$ are known. For notational convenience, let $\mathbf{e}^m$ be a unit vector in length S_m. Given these costs and a price set $\mathbf{P}$, finding a customer equilibrium

$\mathbf{X}^*$ in (2.3) or (2.4) is equivalent to solving the problem,

$$\begin{aligned} \max_{\mathbf{X}} \quad & \sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \mathbf{p}^{m\prime} - \mathbf{w}^m(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m \\ s.t. \quad & \mathbf{e}^{m\prime}\mathbf{x}^m \le \lambda^m \qquad \forall m \in \mathcal{M}, \\ & \mathbf{x}^m \ge \mathbf{0} \qquad \forall m \in \mathcal{M}. \end{aligned} \tag{P1}$$

If the resulting number of orders $\mathbf{y}^* = \sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^{m*}$ from **P1** is the same as the target flow $\tilde{\mathbf{y}}$, then the price $\mathbf{P}$ induces the flow $\tilde{\mathbf{y}}$. However, in general, $\mathbf{y}^*$ generated by the utility-maximizing customers is not consistent with the firm's target flow $\tilde{\mathbf{y}}$. We define,

Definition 2.2. *A target flow $\tilde{\boldsymbol{y}}$ is a feasible target flow if it satisfies the service level constraint in (2.7) and there is an assignment of customers $\boldsymbol{X}$ that satisfies*

$$\begin{aligned} & \sum_{m\in\mathcal{M}} \boldsymbol{A}^m\boldsymbol{x}^m = \tilde{\boldsymbol{y}}, \\ & \boldsymbol{e}^{m\prime}\boldsymbol{x}^m \le \lambda^m \qquad \forall m \in \mathcal{M}, \\ & \boldsymbol{x}^m \ge \boldsymbol{0} \qquad \forall m \in \mathcal{M}. \end{aligned} \tag{2.8}$$

Definition 2.3. *Given a feasible target flow $\tilde{\boldsymbol{y}}$ and a price matrix $\boldsymbol{P}$, the assignment $\boldsymbol{X}^*$ solving **P1** is the* user optimal solution *under $\boldsymbol{P}$. Also, if $\boldsymbol{X}^*$ satisfies*

$$\sum_{m\in\mathcal{M}} \boldsymbol{A}^m\boldsymbol{x}^{m*} = \tilde{\boldsymbol{y}}, \tag{2.9}$$

then $\boldsymbol{X}^*$ is consistent *with target flow $\tilde{\boldsymbol{y}}$, and $\boldsymbol{P}$ is* an incentive optimal price *to target flow $\tilde{\boldsymbol{y}}$.*

The following Propositions 2.1 and 2.2 show that for any feasible target flow, there is a pricing policy $(\bar{\mathbf{p}}, \tilde{\mathbf{p}})$ defining price p_s^m through (2.6) that can induce the target flow. Proposition 2.1 implies that a variable pricing strategy exists for any feasible target flow even when the nominal price $\bar{\mathbf{p}}$ has been predetermined. Proposition 2.2 indicates that when the firm has flexibility to select the nominal price, the nominal price and variable price can be found at the same time.

By including the target flow constraint in **P1** and letting $\tilde{\mathbf{p}} = \mathbf{0}$, define the problem **P2** as:

$$\begin{aligned} \max_{\mathbf{X}} \quad & \sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \bar{p}^m l_m \mathbf{e}^{m\prime} - \mathbf{w}(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m \\ s.t. \quad & \sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m = \tilde{\mathbf{y}}, \qquad\qquad (\mathbf{P2}) \\ & \mathbf{e}^{m\prime}\mathbf{x}^m \le \lambda^m \qquad \forall m \in \mathcal{M}, \\ & \mathbf{x}^m \ge \mathbf{0} \qquad \forall m \in \mathcal{M}. \end{aligned}$$

Proposition 2.1. *Let $\boldsymbol{X}^*$ be the optimal solution of **P2**. Let $\hat{\boldsymbol{\alpha}}$ be the Lagrangian multiplier vector w.r.t the target flow constraint in **P2** at $\boldsymbol{X}^*$ for any given feasible target flow $\tilde{\boldsymbol{y}}$ and nominal price $\bar{\boldsymbol{p}}$. Then, $\tilde{\boldsymbol{p}} = \hat{\boldsymbol{\alpha}}$ is the variable price vector such that $\boldsymbol{X}^*$ is an optimal solution to **P1** that is consistent with $\tilde{\boldsymbol{y}}$.*

Proof. The optimality conditions of **P2** (Bertsekas, 1995) state that $\mathbf{X}^*$ is a global maximum for **P2** if and only if $\mathbf{X}^*$ is feasible and there exists a T dimensional vector $\hat{\boldsymbol{\alpha}}$ such that $\mathbf{X}^*$ is an optimal solution of the following problem:

$$\begin{aligned} \max_{\mathbf{X}} \quad & \sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \bar{p}^m l_m \mathbf{e}^{m\prime} - \mathbf{w}^m(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m - \hat{\boldsymbol{\alpha}}'(\sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m - \tilde{\mathbf{y}}) \qquad (2.10) \\ s.t. \quad & \mathbf{e}^{m\prime}\mathbf{x}^m \le \lambda^m \qquad \forall m \in \mathcal{M}, \\ & \mathbf{x}^m \ge \mathbf{0} \qquad \forall m \in \mathcal{M}. \end{aligned}$$

The objective function of (2.10) can be rewritten as

$$\sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \bar{p}^m l_m \mathbf{e}^{m\prime} - \hat{\boldsymbol{\alpha}}'\mathbf{A}^m - \mathbf{w}^m(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m + \hat{\boldsymbol{\alpha}}'\tilde{\mathbf{y}}. \qquad (2.11)$$

When the firm charges nominal price $\bar{\mathbf{p}}$ and variable price $\tilde{\mathbf{p}} = \hat{\boldsymbol{\alpha}}$, the customer equilibrium

solution is found by substituting $\mathbf{P}$ with $\bar{\mathbf{p}}$ and $\tilde{\mathbf{p}}$ in **P1**. Then,

$$\begin{aligned} \max_{\mathbf{X}} \quad & \sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \bar{p}^m l_m \mathbf{e}^{m\prime} - \hat{\boldsymbol{\alpha}}'\mathbf{A}^m - \mathbf{w}^m(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m \\ s.t. \quad & \mathbf{e}^{m\prime}\mathbf{x}^m \le \lambda^m \qquad \forall m \in \mathcal{M}, \\ & \mathbf{x}^m \ge \mathbf{0} \qquad \forall m \in \mathcal{M}. \end{aligned} \tag{2.12}$$

Since the last term in (2.11) is a constant for any fixed $\hat{\boldsymbol{\alpha}}$ and target flow $\tilde{\mathbf{y}}$, the optimization problems (2.12) and (2.10) are equivalent. Therefore, the optimal solution of (2.10), $\mathbf{X}^*$, is an optimal solution of (2.12), which is the user optimal solution and is consistent with $\tilde{\mathbf{y}}$ under the variable pricing policy $\tilde{\mathbf{p}} = \hat{\boldsymbol{\alpha}}$. □

When the firm has the flexibility to choose the nominal price $\bar{\mathbf{p}}$ together with the variable price $\tilde{\mathbf{p}}$, we have the following results.

Define problem **P3**, by letting $\bar{\mathbf{p}} = 0$ in **P2**, as:

$$\begin{aligned} \max_{\mathbf{X}} \quad & \sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \mathbf{w}(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m \\ s.t. \quad & \sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m = \tilde{\mathbf{y}}, \qquad\qquad (\mathbf{P3}) \\ & \mathbf{e}^{m\prime}\mathbf{x}^m \le \lambda^m \qquad \forall m \in \mathcal{M}, \\ & \mathbf{x}^m \ge \mathbf{0} \qquad \forall m \in \mathcal{M}. \end{aligned}$$

Proposition 2.2. *Let $\boldsymbol{X}^*$ be the optimal solution to **P3**. Let $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ be the Lagrangian multiplier vectors w.r.t the target flow constraints and demand constraints, respectively, at $\boldsymbol{X}^*$. Then given any feasible target flow $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{p}} = \boldsymbol{\alpha}$ and $\bar{\boldsymbol{p}} = (\beta^1/l_1, \cdots, \beta^m/l_m, \cdots, \beta^M/l_M)'$ define an optimal price policy such that $\boldsymbol{X}^*$ is a user optimal solution of **P1** that is consistent with $\tilde{\boldsymbol{y}}$.*

Proof. Observe that **P3** is a linear programming problem, thus $\mathbf{X}^*$ is an optimal solution of **P3** if and only if $\mathbf{X}^*$ is feasible and there exist $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\beta^m \ge 0$, $\beta^m = 0$ for all $m \in \mathcal{M}$

with $\mathbf{e}^{m\prime}\mathbf{x}^{m*} < \lambda^m$, and $\mathbf{X}^*$ is the optimal solution of the following problem,

$$\begin{aligned} \max_{\mathbf{X}} \quad & \sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \mathbf{w}(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m - \boldsymbol{\alpha}'(\sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m - \tilde{\mathbf{y}}) - \sum_{m\in\mathcal{M}} \beta^m(\mathbf{e}^{m\prime}\mathbf{x}^m - \lambda^m) \quad (2.13) \\ s.t. \quad & \mathbf{x}^m \geq \mathbf{0} \qquad \forall m \in \mathcal{M}. \end{aligned}$$

The objective function in (2.13) can be rewritten as,

$$\sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \boldsymbol{\alpha}'\mathbf{A}^m - \beta^m\mathbf{e}^{m\prime} - \mathbf{w}(\tilde{\mathbf{y}})'\mathbf{A}^m)\mathbf{x}^m + \boldsymbol{\alpha}'\tilde{\mathbf{y}} + \sum_{m\in\mathcal{M}} \beta^m\lambda^m. \quad (2.14)$$

Assume that we use $\bar{\mathbf{p}} = (\beta^1/l_1, \cdots, \beta^m/l_m, \cdots, \beta^M/l_M)'$ and $\tilde{\mathbf{p}} = \boldsymbol{\alpha}$ as the nominal price and the variable price, respectively, the first term in (2.14) is the objective function of the user optimal problem defined by **P1** under the proposed pricing policy. The last two terms in (2.14) are constant given fixed multiplier $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and target flow. Thus the user optimal problem **P1** is equivalent to (2.13), i.e., the optimal solution of (2.13) and (2.14) $\mathbf{X}^*$, which is consistent with target flow $\tilde{\mathbf{y}}$, is also the user optimal solution defined in **P1** when using the Lagrangian multipliers as the pricing strategy, i.e., $p_s^m = \beta^m + \boldsymbol{\alpha}'\mathbf{a}_s^m$, $\forall m \in \mathcal{M}, \forall s \in S_m$. □

To find the Lagrangian multipliers, let **P4** be the dual of **P2**.

$$\begin{aligned} \min_{\tilde{\mathbf{p}},\boldsymbol{\beta}} \quad & \tilde{\mathbf{p}}'\tilde{\mathbf{y}} + \sum_{m\in\mathcal{M}} \beta^m\lambda^m \\ s.t. \quad & \tilde{\mathbf{p}}'\mathbf{A}^m + \beta^m\mathbf{e}^{m\prime} \geq \mathbf{v}^{m\prime} - \bar{p}^m l_m \mathbf{e}^{m\prime} - \mathbf{w}^m(\tilde{\mathbf{y}})'\mathbf{A}^m \qquad \forall m \in \mathcal{M}, \qquad (\mathbf{P4}) \\ & \boldsymbol{\beta} \geq \mathbf{0}. \end{aligned}$$

P4's optimal solution provides the Lagrangian multipliers of **P2** (Bertsekas, 1999). Alternatively the KarushKuhnTucker (KKT) conditions of **P2** at $\mathbf{X}^*$ provide the Lagrangian multipliers as given in the following proposition.

Proposition 2.3. *The pricing strategy* $p_s^m = \bar{p}^m l_m + \tilde{\boldsymbol{p}}' \boldsymbol{a}_s^m$ *can induce a feasible target flow* $\tilde{\boldsymbol{y}}$ *if and only if* $\boldsymbol{P}$ *satisfies*

$$v_s^m - \bar{p}^m l_m - \tilde{\boldsymbol{p}}' \boldsymbol{a}_s^m - \boldsymbol{w}^m(\tilde{\boldsymbol{y}})' \boldsymbol{a}_s^m - \beta^m + \pi_s^m = 0, \forall m \in \mathcal{M}, \forall s \in \mathcal{S}_m, \tag{2.15}$$

where $\beta^m \geq 0$, $\beta^m = 0\ \forall m \notin \mathcal{A}_1(\boldsymbol{X}^*)$, $\pi_s^m \geq 0$ *and* $\pi_s^m = 0\ \forall (s,m) \notin \mathcal{A}_2(\boldsymbol{X}^*)$, *where* $\mathcal{A}_1(\boldsymbol{X}^*) = \{m | \boldsymbol{e}^{m\prime} \boldsymbol{x}^{m*} = \lambda^m, m \in \mathcal{M}\}$ *and* $\mathcal{A}_2(\boldsymbol{X}^*) = \{(s,m) | x_s^{m*} = 0, m \in \mathcal{M}, s \in \mathcal{S}_m\}$ *are the set of active constraints at the optimal solution* $\boldsymbol{X}^*$ *of* ***P2****.*

Proof. If (2.15) is satisfied, $\mathbf{X}^*$ is an equilibrium when the pricing strategy $p_s^m = \bar{p}^m l_m + \tilde{\mathbf{p}}' \mathbf{a}_s^m$ is used. This is true because for any $x_s^m > 0$, $\pi_s^m = 0$, (2.15) becomes

$$v_s^m - \bar{p}^m l_m - \tilde{\mathbf{p}}' \mathbf{a}_s^m - \mathbf{w}^m(\tilde{\mathbf{y}})' \mathbf{a}_s^m = \beta^m, \forall m \in \mathcal{M}, \forall s \in S_m, \tag{2.16}$$

where β^m can be interpreted as the equilibrium utility for a class-m customer. For any $x_s^m = 0$, $\pi_s^m \geq 0$, (2.15) becomes

$$v_s^m - \bar{p}^m l_m - \tilde{\mathbf{p}}' \mathbf{a}_s^m - \mathbf{w}^m(\tilde{\mathbf{y}})' \mathbf{a}_s^m - \beta^m = -\pi_s^m \leq 0, \forall m \in \mathcal{M}, \forall s \in S_m, \tag{2.17}$$

which implies that all utilities associated with zero flow must be no greater than the equilibrium utility.

On the other hand, if $p_s^m = \bar{p}^m l_m + \tilde{\mathbf{p}}' \mathbf{a}_s^m$ can induce target flow $\tilde{\mathbf{y}}$, then a user optimal flow of **P1**- $\mathbf{X}^*$ is consistent with $\tilde{\mathbf{y}}$. Thus (2.15) is just the KKT condition for $\mathbf{X}^*$. □

It should be noted, however, that it is possible that the pricing strategy obtained from solving **P4** (or (2.15)) may not be unique. Therefore there may be some degree of freedom in choosing a pricing strategy. Several criteria can be adopted. For example, the firm may be able to choose negative variable prices, which is equivalent to discounting price of the less popular service periods without increasing prices for the popular service periods, or to choose the variable prices that are small as compared to the nominal price, or to add a constant to all the prices and use the same target flow.

2.3.3 Strategic Pricing When Choosing the Target Flow

In some cases, the firm may not have a preferred target flow or may not know which target flow is the best to optimize its profit. We next show how to establish the target flow and pricing strategy with an objective of maximizing the profit.

The firm's problem of finding the target flow and prices that would maximize its profit, denoted by **P5**, is as follows:

$$
\begin{array}{lll}
\max_{\mathbf{P},\tilde{\mathbf{y}},\hat{\mathbf{u}}} & \sum_{m\in\mathcal{M}} \bar{p}^m l_m \mathbf{e}^{m\prime}\mathbf{x}^m + \tilde{\mathbf{p}}'\tilde{\mathbf{y}} & \\
s.t. & (\hat{u}^m - u_s^m(\tilde{\mathbf{y}},\tilde{\mathbf{p}}))x_s^m = 0 \qquad \forall s\in\mathcal{S}_m, \forall m\in\mathcal{M}, & (a) \\
& \hat{\mathbf{u}}^m - \mathbf{u}^m(\tilde{\mathbf{y}},\tilde{\mathbf{p}}) \geq \mathbf{0} \qquad \forall m\in\mathcal{M}, & (b) \\
& \mathbf{e}^{m\prime}\mathbf{x}^m \leq \lambda^m \qquad \forall m\in\mathcal{M}, & (c) \\
& \mathbf{x}^m \geq \mathbf{0} \qquad \forall m\in\mathcal{M}, & (d) \quad (\mathbf{P5}) \\
& \hat{u}^m \geq 0 \qquad \forall m\in\mathcal{M}, & (e) \\
& \sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m = \tilde{\mathbf{y}}, & (f) \\
& \tilde{\mathbf{y}} \leq \boldsymbol{\theta}\cdot\mathbf{c}, & (g) \\
& \mathbf{u}^m = \mathbf{v}^m - \bar{p}^m l_m \mathbf{e}^m - \mathbf{A}^{m\prime}\tilde{\mathbf{p}} - \mathbf{A}^{m\prime}\mathbf{w}^m(\tilde{\mathbf{y}}) \qquad \forall m\in\mathcal{M}. & (h)
\end{array}
$$

In **P5**, (a) and (b) express the equilibrium of demand and price, (e) implies that customers have nonnegative utilities, (f) expresses the target flow constraint, and (h) defines the utility for class-m. (g) is the service quality constraint, we use $f_t(\mathbf{y}) = y_t$ and require $y_t \leq \theta_t c_t$, $\forall t \in \mathcal{T}$ for simplicity in the service level constraint as defined in (2.7). We denoted it as $\mathbf{y} \leq \boldsymbol{\theta}\cdot\mathbf{c}$, where $\cdot$ defined the element-wise multiplicative operation, i.e., we allow different utilization requirements for different service periods.

The firm's problem is a mathematical program with nonlinear complementary constraints. This type of problems is typically difficult to solve. We solve **P5** in two steps. We first find the firm's optimal target flow. Then, the optimal price to induce the optimal target flow can be established by the results in the previous section.

The following proposition shows that the optimal target flow can be formulated as a non-linear optimization problem with linear constraints.

Proposition 2.4. *The optimal target flow of the firm $\tilde{\boldsymbol{y}}^*$ can be found by solving:*

$$\begin{aligned} \max_{\boldsymbol{x},\tilde{\boldsymbol{y}}} \quad & \sum_{m\in\mathcal{M}} (\boldsymbol{v}^{m\prime} - \boldsymbol{w}^m(\tilde{\boldsymbol{y}})'\boldsymbol{A}^m)\boldsymbol{x}^m \\ s.t. \quad & \sum_{m\in\mathcal{M}} \boldsymbol{A}^m\boldsymbol{x}^m = \tilde{\boldsymbol{y}}, \\ & \sum_{s\in S_m} \boldsymbol{e}^{m\prime}\boldsymbol{x}^m \le \lambda^m \qquad \forall m\in\mathcal{M}, \\ & \tilde{\boldsymbol{y}} \le \boldsymbol{\theta}\cdot\boldsymbol{c}, \\ & \boldsymbol{x}^m \ge \boldsymbol{0} \qquad \forall m\in\mathcal{M}. \end{aligned} \tag{2.18}$$

and using $\tilde{\boldsymbol{y}}^*$ as a given target flow, the optimal prices can be established by Proposition 2.2.

Proof. Let $\tilde{\mathbf{y}}^*$ be the optimal solution from (2.18). We next show that the optimal prices obtained from Proposition 2.2 that induces $\tilde{\mathbf{y}}^*$ provide the firm with maximal profit. Let $\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}^*$, $\boldsymbol{\pi}^*$ be the Lagrangian multipliers corresponding to the target flow, demand and positiveness of the demand $\mathbf{x}^m$ constraints at the optimal solution of **P3**, where the target flow is replaced by $\tilde{\mathbf{y}}^*$. From the KKT conditions, $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ and $\boldsymbol{\pi}^*$ must satisfy,

$$v_s^m - \mathbf{w}^m(\tilde{\mathbf{y}}^*)'\mathbf{a}_s^m - \boldsymbol{\alpha}^{*\prime}\mathbf{a}_s^m - \beta^{m*} + \pi_s^{m*} = 0, \forall m\in\mathcal{M}, \forall s\in S_m \tag{2.19}$$

Multiplying (2.19) by x_s^{m*}, and combining with the constraint that $x_s^{m*} * \pi_s^{m*} = 0$, we have

$$v_s^m x_s^{m*} - \mathbf{w}^m(\tilde{\mathbf{y}}^*)'\mathbf{a}_s^m x_s^{m*} - \boldsymbol{\alpha}^{*\prime}\mathbf{a}_s^m x_s^{m*} - \beta^{m*}x_s^{m*} = 0, \forall m\in\mathcal{M}, \forall s\in S_m \tag{2.20}$$

Thus,

$$\sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \mathbf{w}^m(\tilde{\mathbf{y}}^*)'\mathbf{A}^m)\mathbf{x}^{m*} - \boldsymbol{\alpha}^{*\prime}\tilde{\mathbf{y}}^* - \sum_{m\in\mathcal{M}} \beta^{m*}\mathbf{e}^{m\prime}\mathbf{x}^m = 0 \tag{2.21}$$

and

$$\boldsymbol{\alpha}^{*\prime}\tilde{\mathbf{y}}^* + \sum_{m\in\mathcal{M}} \beta^{m*}\mathbf{e}^{m\prime}\mathbf{x}^m = \sum_{m\in\mathcal{M}} (\mathbf{v}^{m\prime} - \mathbf{w}^m(\tilde{\mathbf{y}}^*)'\mathbf{A}^m)\mathbf{x}^{m*} \tag{2.22}$$

The left hand side of (2.22) represents the firm's profit from variable pricing. The right hand side represents customers' maximum utility and is maximized in (2.18). The firm receives revenue $\boldsymbol{\alpha}^{*\prime}\tilde{\mathbf{y}}^*$, when $\tilde{\mathbf{p}} = \boldsymbol{\alpha}^*$, $\bar{p}^m = \beta^{m*}/l_m, \forall m\in\mathcal{M}$ and $\tilde{\mathbf{y}}$ is $\tilde{\mathbf{y}}^*$. That is the $\tilde{\mathbf{y}}^*$ maximize the

customers' utility and these prices transfer all of the utilities to the firm which maximize the firm's profit. □

From the proof of Proposition 2.4, we can see that the firm can adjust prices to take all of the rents and reduce the customers' net utilities to zero. The firm can also charge an arbitrary nominal price $\bar{p}^m < \beta^{m*}/l_m$ and induce the same target flow. However, in such cases, some of the profit will be left to the customers' side.

2.3.4 Examples

Example 1 below shows how customers respond to different variable prices in equilibrium.

Example 1

Consider the following problem: Let the planning horizon be $T = 3$. There are two classes of customers with $l_1 = 1$, $l_2 = 2$, $\lambda^1 = 2$, and $\lambda^2 = 3$, so that the total demand of orders is $2 + 3 * 2 = 8$. Class 1 and 2 customers' valuations of the service are $\mathbf{v}^1 = (8, 12, 6)'$ and $\mathbf{v}^2 = (25, 23)'$. Customers incur congestion cost that depends on the number of customers in the service periods. Assume that $\mathbf{w}^1(\mathbf{y}) = \mathbf{w}^2(\mathbf{y}) = (y_1, \cdots, y_3)'$, where y_t is the number of customers in the service period t, respectively. The firm charges nominal prices for the class 1 and class 2 customer $\bar{p}^1 = 3$ and $\bar{p}^2 = 4$, respectively. Assume the firm wants to induce a target flow $\tilde{\mathbf{y}} = (3, 3, 2)'$.

Assume customers from the two classes use the following demand patterns respectively:

$$\mathbf{A}^1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{A}^2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Since the target flow and nominal price are given, we substitute them into the objective function in **P1**:

$$\max_{\mathbf{x}} (2 - \tilde{p}_1)x_1^1 + (6 - \tilde{p}_2)x_2^1 + (1 - \tilde{p}_3)x_3^1 + (11 - \tilde{p}_1 - \tilde{p}_2)x_1^2 + (10 - \tilde{p}_2 - \tilde{p}_3)x_2^2, \qquad (2.23)$$

let $\mathbf{X}^*$ be the solution and $\mathbf{y}^*$ be the corresponding flow, then we need to check if $\mathbf{y}^* = \tilde{\mathbf{y}}$. The result depends on the value of variable price $\tilde{\mathbf{p}}$.

- Case 1: $\tilde{\mathbf{p}} = (0, 0, 0)'$, i.e., no variable incentive pricing.

  Substitute $\tilde{\mathbf{p}} = (0, 0, 0)'$ into (2.23), the customer optimization problem as in **P1** becomes:

$$\begin{aligned} \max_{\mathbf{x} \geq 0} \quad & 2x_1^1 + 6x_2^1 + x_3^1 + 11x_1^2 + 10x_2^2 \\ s.t. \quad & x_1^1 + x_2^1 + x_3^1 \leq 2, \\ & x_1^2 + x_2^2 \leq 3. \end{aligned}$$

  It can be verified that $\mathbf{x}^{1*} = (0, 2, 0)'$ and $\mathbf{x}^{2*} = (3, 0)'$ solve the above problem. The flow allocated to each period is $\mathbf{y}^* = (3, 5, 0)'$, which is different from the target flow $\tilde{\mathbf{y}} = (3, 3, 2)'$. Therefore, $\mathbf{y}^*$ is not consistent with the target flow $\tilde{\mathbf{y}}$ under price policy $\tilde{\mathbf{p}} = (0, 0, 0)'$, and $\tilde{\mathbf{p}} = (0, 0, 0)'$ is not the incentive optimal price to $\tilde{\mathbf{y}} = (3, 3, 2)'$.

- Case 2: $\tilde{\mathbf{p}} = (1, 5, 0)'$.

  Substitute $\tilde{\mathbf{p}} = (1, 5, 0)'$ into (2.23), the customer optimization problem as in **P1** becomes:

$$\begin{aligned} \max_{\mathbf{x} \geq 0} \quad & x_1^1 + x_2^1 + x_3^1 + 5x_1^2 + 5x_2^2 \\ s.t. \quad & x_1^1 + x_2^1 + x_3^1 \leq 2, \\ & x_1^2 + x_2^2 \leq 3. \end{aligned}$$

  This problem has a solution $\mathbf{x}^{1*} = (2, 0, 0)'$, $\mathbf{x}^{2*} = (1, 2)$. The flow allocated to each period is $\mathbf{y}^* = (3, 3, 2)'$, which is exactly the target flow. Therefore, $\mathbf{y}^*$ is the same as $\tilde{\mathbf{y}}$ under $\tilde{\mathbf{p}} = (1, 5, 0)'$, $\tilde{\mathbf{p}} = (1, 5, 0)'$ is an incentive optimal price to $\tilde{\mathbf{y}} = (3, 3, 2)'$.

Note that $\tilde{\mathbf{p}} = (1, 5, 0)'$ induces multiple solutions, some solutions are consistent with the target flow $\tilde{\mathbf{y}}$ and some are not. For example, $\mathbf{x}^{1*} = (0, 2, 0)'$, $\mathbf{x}^{2*} = (0, 3)'$ is also a solution, but the flow allocated to each period is $\mathbf{y}^* = (0, 5, 3)'$, which is different from $\tilde{\mathbf{y}}$. Another solution is $\mathbf{x}^{1*} = (1, 0, 1)'$, $\mathbf{x}^{2*} = (2, 1)'$, the resulting flow is $\mathbf{y}^* = (3, 3, 2)'$, which is exactly the target flow.

Example 2 shows how to obtain the price that can induce a given target flow.

Example 2

Consider the following problem: Let the planning horizon be $T = 4$. There are two classes of customers with $l_1 = 2$, $l_2 = 3$, $\lambda^1 = 2$, and $\lambda^2 = 2$, so that the total demand of orders is $2*2+2*3 = 10$. Congestion cost $\mathbf{w}^1(\mathbf{y}) = \mathbf{w}^2(\mathbf{y}) = (y_1, 4y_2, 4y_3, y_4)'$, where y_t, $t = 1, \cdots, 4$ are the number of customers in each service period. Class 1 customers' valuations of the service are $\mathbf{v}^1 = (22, 29, 23)'$ and class 2 customers' valuations of the service are $\mathbf{v}^2 = (34, 35)'$. The firm's objective is to find a pricing strategy $(\bar{\mathbf{p}}, \tilde{\mathbf{p}})$ to maximize its profit. Assuming the demand pattern matrices for the two classes of customers are:

$$\mathbf{A}^1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{A}^2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

- If the firm wants to induce a feasible target flow $\tilde{\mathbf{y}} = (2, 3, 3, 2)'$.

  Since the target flow is known and the nominal price is not given, we use the result of Proposition 2.2 to find the optimal price strategy. From Proposition 2.2, the Lagrangian multipliers of the following problem provide the optimal price.

$$\begin{aligned} \max_{\mathbf{x}\geq 0} \quad & 8x_1^1 + 5x_2^1 + 9x_3^1 + 8x_1^2 + 9x_2^2 \qquad (2.24) \\ s.t. \quad & x_1^1 + x_1^2 = 2, \\ & x_1^1 + x_2^1 + x_1^2 + x_2^2 = 3, \\ & x_2^1 + x_3^1 + x_1^2 + x_2^2 = 3, \\ & x_3^1 + x_2^2 = 2, \\ & x_1^1 + x_2^1 + x_3^1 \leq 2, \\ & x_1^2 + x_2^2 \leq 2. \end{aligned}$$

The dual of (2.24) is

$$\begin{aligned} \min_{\mathbf{p}} \quad & 2\tilde{p}_1 + 3\tilde{p}_2 + 3\tilde{p}_3 + 2\tilde{p}_4 + 4\bar{p}^1 + 6\bar{p}^2 \\ s.t. \quad & \tilde{p}_1 + \tilde{p}_2 + 2\bar{p}^1 \geq 8, \\ & \tilde{p}_2 + \tilde{p}_3 + 2\bar{p}^1 \geq 5, \\ & \tilde{p}_3 + \tilde{p}_4 + 2\bar{p}^1 \geq 9, \\ & \tilde{p}_1 + \tilde{p}_2 + \tilde{p}_3 + 3\bar{p}^2 \geq 8, \\ & \tilde{p}_2 + \tilde{p}_3 + \tilde{p}_4 + 3\bar{p}^2 \geq 9. \end{aligned} \tag{2.25}$$

The optimal solution of above problem are $\tilde{\mathbf{p}} = (0, 4, 4, 1)'$ and $\bar{\mathbf{p}} = (2, 0)'$. The firm's strategic profit is 34 under this price strategy.

- If the firm's target flow is unknown.

Since the firm's target flow is not available, we first find the firm's optimal target flow. From Proposition 2.4,

$$\begin{aligned} \max_{\mathbf{x}\geq 0} \quad & (22 - y_1 - 4y_2)x_1^1 + (29 - 4y_2 - 4y_3)x_2^1 + (23 - 4y_3 - y_4)x_3^1 \\ & +(34 - y_1 - 4y_2 - 4y_3)x_1^2 + (35 - 4y_2 - 4y_3 - y_4)x_2^2 \\ s.t. \quad & x_1^1 + x_1^2 = y_1, \\ & x_1^1 + x_2^1 + x_1^2 + x_2^2 = y_2, \\ & x_2^1 + x_3^1 + x_1^2 + x_2^2 = y_3, \\ & x_3^1 + x_2^2 = y_4, \\ & x_1^1 + x_2^1 + x_3^1 \leq 2, \\ & x_1^2 + x_2^2 \leq 2. \end{aligned} \tag{2.26}$$

Solving (2.26), the optimal target flow $\tilde{\mathbf{y}} = (1.24, 1.97, 1.97, 1.74)'$. Substitute this target

flow in (2.26), we have

$$\begin{aligned} \max_{\mathbf{x}\geq 0} \quad & 12.88x_1^1 + 13.24x_2^1 + 13.88x_3^1 + 17x_1^2 + 17.50x_2^2 \qquad (2.27)\\ s.t. \quad & x_1^1 + x_1^2 = 1.24,\\ & x_1^1 + x_2^1 + x_1^2 + x_2^2 = 1.97,\\ & x_2^1 + x_3^1 + x_1^2 + x_2^2 = 1.97,\\ & x_3^1 + x_2^2 = 1.74,\\ & x_1^1 + x_2^1 + x_3^1 \leq 2.\\ & x_1^2 + x_2^2 \leq 2. \end{aligned}$$

To find the optimal prices, we need to find the Lagrangian multipliers of (2.27). The dual of (2.27) is

$$\begin{aligned} \min_{\mathbf{p}} \quad & 1.24\tilde{p}_1 + 1.97\tilde{p}_2 + 1.97\tilde{p}_3 + 1.74\tilde{p}_4 + 4\bar{p}^1 + 6\bar{p}^2 \qquad (2.28)\\ s.t. \quad & \tilde{p}_1 + \tilde{p}_2 + 2\bar{p}^1 \geq 12.88,\\ & \tilde{p}_2 + \tilde{p}_3 + 2\bar{p}^1 \geq 13.24,\\ & \tilde{p}_3 + \tilde{p}_4 + 2\bar{p}^1 \geq 13.88,\\ & \tilde{p}_1 + \tilde{p}_2 + \tilde{p}_3 + 3\bar{p}^2 \geq 17,\\ & \tilde{p}_2 + \tilde{p}_3 + \tilde{p}_4 + 3\bar{p}^2 \geq 17.50. \end{aligned}$$

Solving (2.28), the optimal prices of the firm are $\tilde{\mathbf{p}} = (0, 8.5, 8.5, 0.5)'$ and $\bar{\mathbf{p}} = (2.2, 0)'$. The firm's strategic profit is 43.13 under this price strategy.

2.4 Operational Decision on Admission Control

In this section we study how the firm should allocate capacity when customers try to make reservations. Recall that at the strategic level the firm determined a target capacity vector and time-dependent prices to maximize the profit it receives in the equilibrium solution. However, this is done long before the actual reservations are placed and capacity is consumed. The

operational problem considers how the firm should behave at this later date.

We consider customers arriving over a horizon, the "reservation period", place requests for capacity during some service period. Suppose the reservation period is day 0, customers are to reserve capacity starting on day 1. We model customers as arriving according to a rate λ Poisson process during the reservation period. To simplify modeling, we divide the reservation period into K stages, with at most one arrival occurring during each stage. Thus the probability of one arrival in a stage is

$$\delta = \lambda / K.$$

(This is the common discrete time dynamic formulation.) Customers then request capacity during the service period and the firm accepts or rejects the reservation. We approach the problem through an approximate dynamic programming formulation.

Customers request capacity to maximize their utilities. For any given class-m and demand pattern s, the utility of s is determined by the equilibrium assignment, $\mathbf{X}^*$, and the price matrix $\mathbf{P}$ through $u_s^m(\mathbf{X}^*, \mathbf{P})$. However, at the operational level, $\mathbf{X}^*$ is not known, but $\tilde{\mathbf{y}}$ and $\mathbf{P}$ are. Because $u_s^m(\mathbf{X}^*, \mathbf{P})$ depends only on the $\tilde{\mathbf{y}}$ vector, we can define the customer's utility(with a slight abuse of notation) as $u_s^m(\tilde{\mathbf{y}}, \mathbf{P})$. Potentially a class-$m$ customer can obtain the maximum utility $\hat{u}^m$ from several demand patterns. In this case, she would be indifferent to choosing any of them. Let $I_m(\tilde{\mathbf{y}}, \mathbf{P}) = \{s | u_s^m(\tilde{\mathbf{y}}, \mathbf{P}) = \hat{u}^m, s \in \mathcal{S}_m\}$ be the indifference set of class-m. Upon arrival, each class-m customer chooses a demand pattern from the indifference set I_m. We assume that in equilibrium customers choose demand pattern $i \in I_m$ according to some probability ψ_i^m. Thus each arrival is characterized as a type (m, i).

The firm then has the choice whether or not to accept the reservation request of an arriving customer. If the request is accepted, the capacity for demand pattern i is allocated to that customer and cannot be allocated to other customers. If the request is rejected, the customer departs. We do not allow a rejected customer to try to reserve a second time by suggesting an alternative demand pattern. We acknowledge this may be a limitation of the model as customers could repeatedly request capacity for alternative demand patterns until one is accepted. Allowing this behavior would then require the firm to determine which of each customer types'

demand patterns to accept. As a result, this will greatly increase the size of the problem which will be beyond the scope that can be addressed through dynamic programming techniques. Moreover, the cost of changing demand patterns for industrial customers is high and therefore this assumption is relevant. This is in fact the main reason why online dynamic pricing is less relevant in this study.

Naturally, there are two definitions of capacity: target flow capacity and maximum capacity. The Target Flow allocation (TFA) model assumes that the maximum number of jobs allocated to a time period is given by the target flow and that no bookings can be made above this limit. The Maximum Capacity Allocation (MCA) model assumes that the limit is given by the maximum capacity $\mathbf{c}$. We assume that there is an additional congestion cost incurred by the firm if an order is booked beyond the target flow. This can be done as long as the total booking is below the maximum capacity.

2.4.1 Model and Formulation

The dynamic program is formulated as follows. The system state is the total number of orders assigned to each service slot, denote as $\mathbf{y} = (y_1, \ldots, y_T)$. The initial state is $\mathbf{y}_K = 0$, where we count time backward from 0.

Recall that $\mathbf{a}_i^m$ is the i^{th} column of the demand pattern matrix $\mathbf{A}^m$ of a class-m customers. Let $u = 1$ if a customer is accepted, 0 otherwise. Let $r(m, i)$ be the revenue associated with admitting a type (m, i) arrival:

$$r(m, i) = \bar{p}^m l_m + \tilde{\mathbf{p}}' \mathbf{a}_i^m.$$

Let h^m be the cost associated with rejecting a class-m customer. Under the MCA model, additional costs are incurred if capacity is reserved beyond the target flow. Let $g(\mathbf{y}, m, i)$ be the overflow cost for booking a type (m, i) customer in state $\mathbf{y}$ and $g(\mathbf{y}, m, i) = 0$ if $\mathbf{y} < \tilde{\mathbf{y}}$. We assume $g(\mathbf{y}, m, i)$ is non-decreasing in $\mathbf{y}$ for all (m, i):

$$g(\mathbf{y}, m, i) \leq g(\mathbf{y} + \mathbf{a}_t^n, m, i), \forall (n, t).$$

For example, let $\mathbf{y}(m,i,1)$ be the allocated capacity if a type (m,i) customer is accepted, $g(\mathbf{y},m,i) = \sum_{t\in\mathcal{T}} \kappa_t \times \min\left[y_t(m,i,1) - y_t, (y_t(m,i,1) - y_t^T)^+\right]$ is a piecewise linear overflow cost function, where κ_t is the cost charged for each unit over the target flow at service period t.

Let $V_k(\mathbf{y})$ be the expected maximum profit with k periods to go. Then the dynamic programming formulation for the MCA model is

$$V_k(\mathbf{y}) = (1-\delta)V_{k-1}(\mathbf{y}) + \delta \sum_{m\in\mathcal{M}} \sum_{i\in I_m} \psi_i^m V_k(\mathbf{y},m,i),$$

where $V_k(\mathbf{y},m,i) = \max\{r(m,i) - g(\mathbf{y},m,i) + V_{k-1}(\mathbf{y}+\mathbf{a}_i^m), -h_m + V_{k-1}(\mathbf{y})\}$ if $\mathbf{y}(m,i,1) \le \mathbf{c}$ and $V_k(\mathbf{y},m,i) = -h_m + V_{k-1}(\mathbf{y})$ otherwise.

Let $V_0(\mathbf{y}) = 0$ for all $\mathbf{y}$ be the boundary condition. The objective is

$$\max V_K(\mathbf{0}).$$

The Target Flow Allocation (TFA) model is the same as the MCA formulation where $\mathbf{c}$ is replaced by $\tilde{\mathbf{y}}$. In this case, there is no overbooking cost and $g(\mathbf{y},m,i)$ is 0.

2.4.2 Structural Properties

We now present several structural properties of the dynamic program. These hold for both the MCA and the TFA formulations.

Lemma 2.1. *The value function $V_k(\boldsymbol{y},m,i)$ is non-increasing as a function of $\boldsymbol{y}$:*

$$V_k(\boldsymbol{y}) \ge V_k(\boldsymbol{y}+\boldsymbol{a}_i^m), \forall (m,i).$$

Proof. Suppose we have an optimal policy $\mathbf{u}^*$ when we start from state $\mathbf{y}+\mathbf{a}_i^m$ at period k. $\mathbf{u}^*$ will always be feasible when we start from $\mathbf{y}$. Since the overflow cost $g(\mathbf{y},m,i)$ increases in $\mathbf{y}$ for all (m,i), the value for the objective function when we start from $\mathbf{y}$ and use $\mathbf{u}^*$ will be higher, but $\mathbf{u}^*$ is not necessarily optimal. Therefore, we must have

$$V_k(\mathbf{y}) \ge V_k(\mathbf{y}+\mathbf{a}_i^m), \forall (m,i). \qquad \square$$

Next, we define the marginal opportunity cost in state $\mathbf{y}$ for admitting (m, i) with k periods to go as

$$\Delta V_k(\mathbf{y}, m, i) = V_{k-1}(\mathbf{y}) - V_{k-1}(\mathbf{y} + \mathbf{a}_i^m).$$

In a standard dynamic pricing formulation, we would want to establish either a capacity threshold or a time threshold for which we would always accept a type-(m, i) arrival given sufficient capacity or time-to-go, and reject otherwise. That is a time threshold, dependent on the capacity, and arrival type would be determined as some time K^* such that

$$K^*(\mathbf{y}, m, i) = \min\{k : \Delta V_k(\mathbf{y}, m, i) \leq r(m, i) - g(\mathbf{y}, m, i) + h_m\}. \tag{2.29}$$

A capacity threshold, dependent on the time to go k and arrival type, is defined as

$$Y^*(k, m, i) = \max\{\mathbf{y} : \Delta V_k(\mathbf{y}, m, i) \geq r(m, i) - g(\mathbf{y}, m, i) + h_m\}, \tag{2.30}$$

where $\max\{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^n\} = (\max(y_1^1, y_1^2, \ldots, y_1^n), \max(y_2^1, y_2^2, \ldots, y_2^n), \ldots)$, i.e., the maximum of the vector elements.

If we can find $K^*(\mathbf{y}, m, i)$ and $Y^*(k, m, i)$, then we could adopt a time or capacity threshold policy for the dynamic program. To establish the threshold policies, we need to show the monotonicity of the marginal cost, i.e., that the following inequality holds:

$$\Delta V_k(\mathbf{y}) \leq \Delta V_k(\mathbf{y} + \mathbf{a}_i^m), \forall (m, i).$$

Unfortunately, because of the capacity constraints and multiple orders, the marginal cost function is not monotone.

Proposition 2.5. *Neither a time threshold policy nor a capacity threshold policy are necessarily optimal. Given any period k and arrival (m, i), there does not exist a threshold vector $Y^*(k, m, i)$ such that if $y_t \geq Y_t^*$ for all $t = 1, \ldots, T$, the arrival should be rejected, and accepted otherwise. Similarly, given state $\boldsymbol{y}$ and an arrival (m, i), there does not exist a time threshold such that if $k \leq K^*(\boldsymbol{y}, m, i)$ the arrival should be rejected, and accepted if $k > K^*(\boldsymbol{y}, m, i)$.*

Proof. We establish the proposition using a counter-example. We show that depending upon the state and arrival type, $\Delta V_k(\mathbf{y}, m, i)$ is not necessarily greater or less than $\Delta V_k(\mathbf{y})$. The time and state thresholds defined in (2.29) and (2.30) do not exist.

Consider the following problem: A firm has 4 service periods $\mathcal{T} = \{1, 2, 3, 4\}$ and 4 reservation periods, $K = 4$. There are three types of customers $\mathcal{M} = \{1, 2, 3\}$. The number of orders per request for each type of customers is $l_1 = 1$, $l_2 = 2$, $l_3 = 3$. The order patterns for each type of customers are expressed in matrixes $\mathbf{A}^1$, $\mathbf{A}^2$, and $\mathbf{A}^3$, respectively:

$$\mathbf{A}^1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{A}^2 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{A}^3 = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

where columns of in the matrxes represent types of arrivals. Thus, we have four types of arrivals, $(1, 1)$, $(2, 1)$, $(2, 2)$ and $(3, 1)$. For example, the first column in $\mathbf{A}^2$ is for arrival $(2, 1)$, i.e., type 2 customer choose their first demand pattern and a unit of capacity in periods 1 and 2 will be reserved upon admission. The probability of arrivals are 0.2, 0.2, 0.2 and 0.3 respectively (there is a 0.1 probability that no customer arrives). Let the profit earned in the current period if the arrival is accepted be $r(1, 1) = 1, r(2, 1) = 5$, and $r(2, 2) = 10, r(3, 1) = 100$. Rejection costs for each class of customers are $h_1 = 0, h_2 = -1$, and $h_3 = 0$. Maximum Capacity is $\mathbf{c} = (2, 1, 2, 2)'$. The expected profit with 2 periods to go given the states $\mathbf{y}$ is given in Table 2.1.

Let the arrival type with 3 period to go be $(1, 1)$, we have

$$\begin{aligned} \Delta V_3((0,0,0,0),1,1) &= V_2(0,0,0,0) - V_2(0,0,1,0) = 3.12 \\ \Delta V_3((1,1,0,0),1,1) &= V_2(1,1,0,0) - V_2(1,1,1,0) = 0.68 \\ \Delta V_3((0,1,1,1),1,1) &= V_3(0,1,1,1) - V_3(0,1,2,1) = 4.12. \end{aligned}$$

We can see that

$$\Delta V_3((0,1,1,1),1,1) \geq \Delta V_3((0,0,0,0),1,1) \geq \Delta V_3((1,1,0,0),1,1).$$

Table 2.2: The expected profits with 2 periods to go

State ($\mathbf{y}$)	Expected Profit $V_2(\mathbf{y})$
0020	0.76
0021	0.76
1110	2.72
2100	3.40
1111	2.72
0022	0.76
0122	-1.40
0011	52.63
0121	-1.40
0111	2.72
0010	52.63
1100	3.40
0000	55.75

Therefore, the monotonicity of $\Delta V_3(\mathbf{y}, 1, 1)$ does not hold. Further, if we check the optimal admission control in period 3, $(1, 1)$ is rejected when the state is $(0, 0, 0, 0)$, accepted when it is $(1, 1, 0, 0)$, and rejected when it is $(0, 1, 1, 1)$. □

2.4.3 An Upper Bound for the Operational Problem

In this subsection, we describe an upper bound for the dynamic programming problems of MCA and TFA models presented in subsection 2.4.1. If the demand realization were known a priori, the firm could optimize the admission control with complete information. Let H_i^m be the number of type (m, i) arrivals over the reservation horizon and let $\mathbf{H}^m = \left(H_1^m, \ldots, H_{|I_m|}^M\right)$ and $\mathbf{H} = (\mathbf{H}^1, \ldots, \mathbf{H}^M)$. Given $\mathbf{H}$, an upper bound is provided by solving the following problem. Recall that x_i^m is the number of class-m customers assigned to the i^{th} demand pattern and let $\mathbf{x}^m = \{x_i^m, \ldots, x_{|I_m|}^m\}$. An upper bound for the operational problem is

$$
\begin{aligned}
V^{\mathrm{UB}}(\mathbf{H}) = \max_{\mathbf{x}} \quad & \sum_{m\in\mathcal{M}} \bar{p}^m l_m \mathbf{e}^{m\prime}\mathbf{x}^m + \tilde{\mathbf{p}}' \sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m - \sum_{m\in\mathcal{M}} h_m \mathbf{e}^{m\prime}(\mathbf{H}^m - \mathbf{x}^m) \\
& -\boldsymbol{\kappa}' \left(\left(\sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m \right) - \tilde{\mathbf{y}} \right)^+ \\
s.t. \quad & \sum_{m\in\mathcal{M}} \mathbf{A}^m\mathbf{x}^m \le \mathbf{c}, \\
& \mathbf{x}^m \le \mathbf{H}^m \qquad \forall m \in \mathcal{M}, \\
& x_i^m \qquad \text{integer}, \forall m \in \mathcal{M}, \forall i \in I_m.
\end{aligned} \tag{2.31}
$$

The problem maximizes the revenue less overflow and rejection cost, subject to the capacity and demand constraints. $V^{\mathrm{UB}}(\mathbf{H})$ is an upper bound on the profit for a sample path $\mathbf{H}$. By taking expectations over all $\mathbf{H}$, $E_{\mathbf{H}}(V^{\mathrm{UB}}(\mathbf{H}))$ would provide an upper bound for the expected profit of the dynamic formulations in subsection 2.4.1. However, it is impossible to calculate this expectation $E_H(V^{\mathrm{UB}}(\mathbf{H}))$, as there are far too many realizations in any reasonably sized problem, and for each realization we must solve an integer programming problem. Therefore, we only estimate the upper bound by using a Monte Carlo simulation. One can also solve (2.31) with the expected arrival rates and use this as the approximation of the cost to go. Next, we will use the upper bound estimated to evaluate the dynamic programming approximation developed.

2.4.4 Value Function Approximation

In this subsection, we present an approximation of the expected value function during the reservation period. The approximation will be used in the development of a heuristic for the operational problem.

Let $\bar{z}_k^m$ be the expected aggregate demand for class-m during the periods $k-1$ to 1. Let $\bar{\mathbf{z}}_k = (z_k^1, \ldots, \bar{z}_k^m)$. Given current state $\mathbf{y}$ and the expected aggregate demand $\bar{\mathbf{z}}_k$, let $\bar{V}_k(\mathbf{y}, \bar{\mathbf{z}}_k)$ be the approximate expected profit to go, approximating $V_k(\mathbf{y})$. For the MCA model, $\bar{V}_k(\mathbf{y}, \bar{\mathbf{z}}_k)$ is given by solving the following integer problem:

$$
\begin{aligned}
\bar{V}_k(\mathbf{y}, \bar{\mathbf{z}}_k) = \max_{\mathbf{x}} \quad & \sum_{m \in \mathcal{M}} \bar{p}^m l_m \mathbf{e}^{m\prime} \mathbf{x}^m + \tilde{\mathbf{p}}' \sum_{m \in \mathcal{M}} \mathbf{A}^m \mathbf{x}^m - \sum_{m \in \mathcal{M}} h_m(\bar{z}_k^m - \mathbf{e}^{m\prime} \mathbf{x}^m) \\
& -\boldsymbol{\kappa}' \left(\left(\sum_{m \in \mathcal{M}} \mathbf{A}^m \right) \mathbf{x}^m + \mathbf{y} - \tilde{\mathbf{y}} \right)^+ \\
s.t. \quad & \sum_{m \in \mathcal{M}} \mathbf{A}^m \mathbf{x}^m \leq \mathbf{c} - \mathbf{y}, \\
& \mathbf{e}^{m\prime} \mathbf{x}^m \leq \bar{z}_k^m \quad \forall m \in \mathcal{M}, \\
& x_i^m \qquad \text{integer} \qquad \forall m \in \mathcal{M}, i \in I_m.
\end{aligned} \tag{2.32}
$$

The value function approximation for the TFA model is obtained by substituting $\tilde{\mathbf{y}}$ in for $\mathbf{c}$ and noting the last term of the objective function is then identically 0. For the MCA model, the third term is non-linear. However, introducing the auxiliary vector $\boldsymbol{\tau}$ (a T dimensional vector), we can rewrite (2.32) as

$$
\begin{aligned}
\bar{V}_k(\mathbf{y}, \bar{\mathbf{z}}_k) = \max_{\mathbf{x}} \quad & \sum_{m \in \mathcal{M}} \bar{p}^m l_m \mathbf{e}^{m\prime} \mathbf{x}^m + \tilde{\mathbf{p}}' \sum_{m \in \mathcal{M}} \mathbf{A}^m \mathbf{x}^m - \boldsymbol{\kappa}'\boldsymbol{\tau} - \sum_{m \in \mathcal{M}} h_m(\bar{z}_k^m - \mathbf{e}^{m\prime} \mathbf{x}^m) \\
s.t. \quad & \sum_{m \in \mathcal{M}} A^m x^m \leq \mathbf{c} - \mathbf{y}, \\
& \mathbf{e}^{m\prime} \mathbf{x}^m \leq \bar{z}_k^m \quad \forall m \in \mathcal{M}, \\
& \boldsymbol{\tau} \geq \sum_{m \in \mathcal{M}} \mathbf{A}^m \mathbf{x}^m + \mathbf{y} - \tilde{\mathbf{y}}, \\
& \boldsymbol{\tau} \geq 0, \\
& x_i^m \qquad \text{integer} \quad \forall m \in \mathcal{M}, i \in I_m.
\end{aligned} \tag{2.33}
$$

The problem (2.33) is still a difficult linear integer program. However, in some cases it would be natural for the potential demand patterns to take deliveries in consecutive time periods. For example, the concrete distributing problem requires that the deliveries have to be made consecutively. In this case, all the 1's in each column of the matrix of demand patterns would be consecutive. Then it is easy to show that the solution to the linear programming relaxation of (2.33) will be integer (Wolsey and Nemhauser, 1998).

We now present a simple value function approximation heuristic for solving the MCA and

TFA models.

Observe $\bar{V}_{k-1}(\mathbf{y}, \bar{\mathbf{z}}_{k-1}) - \bar{V}_{k-1}(\mathbf{y} + \mathbf{a}_i^m, \bar{\mathbf{z}}_{k-1})$ approximates the opportunity cost of accepting a type (m, i) customer in period k. This opportunity cost is compared with the marginal benefit of accepting the customer, $r(m, i) - g(\mathbf{y}, m, i) + h_m$. Doing so provides an approximate dynamic program heuristic. Formally:

Value Function Heuristic (VF-Heuristic)

Step 1 *For each arrival, check if the arrival in period k is feasible for the available capacity:*

If $\boldsymbol{y} + \boldsymbol{a}_i^m \geq \boldsymbol{c}$*, reject the arrival, else go to the next step.*

Step 2 *Compare marginal benefit and opportunity cost:*

Marginal Benefit$= r(m, i) + h_m - g(\mathbf{y}, m, i)$,

Opportunity Cost$= \bar{V}_{k-1}(\mathbf{y}, \bar{\mathbf{z}}_{k-1}) - \bar{V}_{k-1}(\mathbf{y} + \mathbf{a}_i^m, \bar{\mathbf{z}}_{k-1})$, *where* $\bar{V}$ *is obtained by solving (2.33).*

If (Marginal Benefit $\geq$ *Opportunity Cost)*

Accept the arrival.

Else

Reject the arrival.

As an alternative, we will compare the VF-Heuristic to a **Myopic** heuristic, where the firm admits a customer request as long as there is sufficient capacity.

2.4.5 Performance of the Value Function Heuristic

We use Monte Carlo simulation to evaluate the performance of the VF-Heuristic. We compare it to the myopic heuristic.

We test several test cases with different sizes, defined by the number of service periods, T, the number of customer types, M, and the number of reservation periods, K. We consider three groups of computational experiments. For the small-size cases defined by $T = 6, M = 3, 4, 5$ and $K = 5, 6, 7, 8$, we can solve the TFA and MCA dynamic programs through backward induction. For these cases, we can calculate the relative efficiency (RE) of the VF-Heuristic and the Myopic heuristic given by V/V', where V is the profit given by the heuristic and V' is the optimum

given by the backward induction.

We also test medium-size cases with $T = 16, 24, 32; M = 5, 7;$ and $K = 32, 96, 100, 200;$ and large-size cases with $T = 48, 96, M = 9$ and $K = 500$. For these sizes the exact solution cannot be found. We therefore report on the performance of the heuristic by comparing the optimal solution of the heuristic with the upper bound given by (2.31). This is the upper bound efficiency (UE): V/V^{UB} where V^{UB} is the upper bound value.

The difficulty in solving any given instance is related to the ratio of the demand to the capacity (namely, the implied utilization or congestion index of the system). For the TFA model we report the congestion index $CI_y = \frac{\sum_{m\in\mathcal{M}} \lambda_m l_m}{\sum_{t\in\mathcal{T}} \tilde{y}_t}$ with respect to the target flow vector, $\tilde{y}$. For the MCA model we report CI_y and the congestion index CI_c, given with respect to the total capacity: $CI_c = \frac{\sum_{m\in\mathcal{M}} \lambda_m l_m}{\sum_{t\in\mathcal{T}} c_t}$.

Table 2.3 represents the results for the TFA model, comparing the VF-Heuristic to the Myopic heuristic. Table 2.4 presents the same for the MCA model. We observe for the small-size cases that the VF-Heuristic generally performs well compared with the optimum value. Further, we observe that it outperforms the Myopic heuristic (except for trivial differences in two cases). We note that for these small cases the upper bound efficiency is approximately 5% to 10% lower than the relative efficiency.

From Table 2.3 we can make several observations on the performance of the VF-Heuristic. First, it generally outperforms the Myopic heuristic, in most cases by a considerable amount. Second, the optimality gap is for the most part small, on the order of 10% to 20% versus the upper bound. Further, we observe the performance of the VF-Heuristic is mostly unaffected by the congestion index (at least on the ranges that we test). This is in contrast to the myopic heuristic whose performance degrades as the congestion increases. Next, we use the VF-Heuristic to evaluate the hierarchical approach to the pricing and allocation problem.

Table 2.3: Performance of the value function heuristic-TFA

(T, M, K)	CI_y	$RE^{VF-Heur}$	RE^{Myopic}	$UE^{VF-Heur}$	UE^{Myopic}
(6,5,6)	0.70	0.9945	0.9907	0.9497	0.9461
(6,4,6)	0.57	0.9794	0.9626	0.8811	0.8661
(6,3,5)	0.58	0.9972	0.9887	0.9499	0.9418
(6,3,8)	1.27	0.9744	0.8959	0.9305	0.8555
(6,3,7)	1.11	0.9958	0.8735	0.8327	0.7305

(a) Small size cases;

(T, M, K)	CI_y	$UE^{VF-Heur}$	UE^{Myopic}
(16,5,32)	1.00	0.7518	0.5666
(16,5,32)	1.11	0.8461	0.5939
(16,5,32)	1.24	0.6938	0.4047
(16,5,32)	1.42	0.7692	0.5388
(16,5,32)	1.66	0.7084	0.4719
(24,5,100)	1.01	0.8816	0.8767
(24,5,100)	1.12	0.8435	0.8788
(24,5,100)	1.26	0.8203	0.8011
(24,5,100)	1.44	0.8444	0.8317
(24,5,100)	1.68	0.7643	0.6760
(24,5,100)	2.02	0.6472	0.5491
(32,5,96)	1.00	0.8651	0.8600
(32,5,96)	1.11	0.7839	0.7868
(32,5,96)	1.25	0.7630	0.7406
(32,5,96)	1.42	0.6764	0.6623
(32,5,96)	1.66	0.6542	0.6131
(32,5,96)	2.00	0.5845	0.5329
(32,7,200)	1.00	0.9023	0.9043
(32,7,200)	1.11	0.8605	0.8650
(32,7,200)	1.25	0.8324	0.7884
(32,7,200)	1.42	0.8246	0.7828
(32,7,200)	1.66	0.7387	0.6377
(32,7,200)	2.00	0.7187	0.6208

(b) Medium size cases;

(T, M, K)	CI_y	$UE^{VF-Heur}$	UE^{Myopic}
(48,9,500)	1.00	0.9121	0.8743
(48,9,500)	1.11	0.8248	0.6444
(48,9,500)	1.25	0.9035	0.7045
(48,9,500)	1.42	0.9651	0.6756
(48,9,500)	1.66	0.9556	0.4987
(48,9,500)	2.00	0.8312	0.3002
(96,9,500)	1.00	0.9643	0.9227
(96,9,500)	1.11	0.9538	0.8256
(96,9,500)	1.24	0.9570	0.7787
(96,9,500)	1.42	0.9128	0.7204
(96,9,500)	1.66	0.9296	0.7134
(96,9,500)	1.99	0.9051	0.5959

(c) Large size cases.

Table 2.4: Performance of the value function heuristic-MCA

(T, M, K)	CI_y	CI_c	$RE^{VF-Heur}$	RE^{Myopic}	$UE^{VF-Heur}$	UE^{Myopic}
(6,5,6)	0.90	0.70	0.9948	0.9962	0.8975	0.8987
(6,4,6)	0.73	0.57	0.9842	0.9847	0.9244	0.9249
(6,3,5)	0.78	0.58	0.9830	0.9709	0.8390	0.8286
(6,3,8)	1.67	1.27	0.9013	0.8332	0.8851	0.8182
(6,3,7)	1.48	1.11	0.9128	0.8778	0.8929	0.8587

(a) Small size cases;

(T, M, K)	CI_y	CI_c	$UE^{VF-Heur}$	UE^{Myopic}
(16,5,32)	1.00	0.83	0.8427	0.6797
(16,5,32)	1.11	0.83	0.8767	0.6989
(16,5,32)	1.24	0.83	0.8991	0.6853
(16,5,32)	1.42	0.83	0.7916	0.6743
(16,5,32)	1.66	0.83	0.9188	0.789
(24,5,100)	1.01	0.83	0.9525	0.9594
(24,5,100)	1.12	0.83	0.9705	0.9671
(24,5,100)	1.26	0.83	0.9538	0.9264
(24,5,100)	1.44	0.83	0.9854	0.9521
(24,5,100)	1.68	0.83	0.9661	0.9229
(24,5,100)	2.02	0.83	0.9650	0.9487
(32,5,96)	1.00	0.83	0.9375	0.9358
(32,5,96)	1.11	0.83	0.9241	0.9201
(32,5,96)	1.25	0.83	0.9387	0.9367
(32,5,96)	1.42	0.83	0.9202	0.9202
(32,5,96)	1.66	0.83	0.9246	0.9252
(32,5,96)	2.00	0.83	0.9427	0.9405
(32,7,200)	1.00	0.83	0.9386	0.9259
(32,7,200)	1.11	0.83	0.9587	0.9448
(32,7,200)	1.25	0.83	0.9731	0.9227
(32,7,200)	1.42	0.83	0.9660	0.9337
(32,7,200)	1.66	0.83	0.9269	0.9060
(32,7,200)	2.00	0.83	0.9621	0.9339

(b) Medium size cases;

(T, M, K)	CI_y	CI_c	$UE^{VF-Heur}$	UE^{Myopic}
(48,9,500)	1.00	0.83	0.9502	0.9014
(48,9,500)	1.11	0.83	0.9519	0.8238
(48,9,500)	1.25	0.83	0.9374	0.8486
(48,9,500)	1.42	0.83	0.9803	0.8522
(48,9,500)	1.66	0.83	0.9780	0.8251
(48,9,500)	2.00	0.83	0.9745	0.8448
(96,9,500)	1.00	0.83	0.9740	0.9441
(96,9,500)	1.11	0.83	0.9628	0.9140
(96,9,500)	1.24	0.83	0.9872	0.9382
(96,9,500)	1.42	0.83	0.9878	0.9395
(96,9,500)	1.66	0.83	0.9703	0.9264
(96,9,500)	1.99	0.83	0.9723	0.9520

(c) Large size cases.

2.5 Numerical Computations: Hierarchical Planning vs. Non-Hierarchical Planning

In this section we conduct numerical studies to demonstrate the possible benefits that can be obtained through implementing the proposed hierarchical planning approach. We use Monte Carlo simulation for all the numerical computations. We demonstrate the magnitude of the increase that a hierarchical approach can have over a non-hierarchical planning approach. Then, we study the effects of various target flows, customer valuation patterns, and service levels.

2.5.1 Implementation Details

Recall that in our hierarchical planning approach, we first solve the strategic level problem that produces a pricing vector $\tilde{\mathbf{p}}$ and a target flow $\tilde{\mathbf{y}}$. These are then presented to the customers who at the operational level need to determine their demand schedules. We want to compare how well the hierarchical planning approach perform as opposed to the non-hierarchical planning approach. The difference between the hierarchical planning approach (HP) and non-hierarchical planning approach (NHP) is defined by their respective demand generating processes, their schedule selection processes, and the pricing structure used to evaluate their resulting revenues. We next define the processes for both cases.

Demand Generating and Schedule Selection Process

For our numerical testing, we divide the reservation horizon into K periods as described above and assume that the probability of an arrival in a period is given by $\delta = \lambda/K$ and the probability of no arrival is then $1-\delta$. That is, we simulate a discretized Poisson process. Upon arrival customers request one of their demand schedules. Under HP, class-m customers choose from I_m in $\mathcal{S}_m$ according to ψ_i^m. Depending upon the information known, the values of ψ_i^m may vary. Further, these values may be related to behavioral choices. Therefore, we will compare several reasonable values of the ψ_i^m's. First, customers may choose their preferred demand pattern by randomizing uniformly over their indifference set. Second, customers may try to

weight their choices based on the posted target flow. If customers observe the firms acceptance and rejection over a long time, they may plan their schedule to fit the firm's preferred target flow. Further, if the firm post their ideal schedule $\mathbf{X}^*$, the customers may weight their choice by $\mathbf{X}^*$. This could produce an outcome considered the best possible, comparable to using the ideal schedule itself. Thus we suggest the following three possible definitions of ψ_i^m,

$$\psi_i^m = \begin{cases} \frac{1}{|I_m|} & \text{Under uniform weighting;} \\ \frac{\tilde{\mathbf{y}}' \mathbf{a}_i^m}{\sum_{i \in \hat{I}_m} \tilde{\mathbf{y}}' \mathbf{a}_i^m} & \text{Weighted by } \tilde{\mathbf{y}}; \\ \frac{x_i^{m*}}{\sum_{i \in \mathcal{S}_m} x_i^{m*}} & \text{Weighted by } \mathbf{X}^*. \end{cases} \tag{2.34}$$

Under NHP, class-m customers choose schedules from $\mathcal{S}_m$. We assume that customers choose their preferred demand pattern by randomizing uniformly over their feasible demand set, i.e., $\psi_i^m = \frac{1}{|\mathcal{S}_m|}$.

Pricing Structure

The HP and NHP differ in their pricing structures. The HP uses both the nominal prices $\bar{\mathbf{p}}$ and the variable prices $\tilde{\mathbf{p}}$ generated at the strategic level. The NHP only uses nominal prices as these are known prior to solving the strategic level problem. Recall that the time dependent variable prices may have a degree of freedom in their definition. Because we wish to compare the HP and NHP problem solution revenues, we need to define their price structure carefully. In practice, one might expect that the firm would choose $\tilde{\mathbf{p}}$ to maximize its revenue subject to some individual rationality constraints on the customers' choices. However, these prices are unknown. Using these prices would result in revenues that are not comparable with those of the NHP approach. Thus, we choose the variable price $\tilde{\mathbf{p}}$ such that $\tilde{\mathbf{p}}' \tilde{\mathbf{y}} = 0$. The revenue of both approaches are then comparable.

Initially, we assume the customers prefer service delivery in the middle of the service period. That is, their value for service has a peak at $T/2$. We will define this valuation pattern and compare it to two alternatives in Section 2.5.3.

We use the following parameters in our numerical examples. In the demand generating process, we let the probability of an arrival in a reservation period, i.e., λ/K, to be between

0.2 and 0.5. Depending on the value of M, the probability of an arrival for a class-m customer varies and is approximately equal to $\frac{1}{2M}$. (For $M = 5$, the values are 0.08, 0.10, 0.16, 0.07, and 0.09; for $M = 7$, the values are 0.04 0.05, 0.02, 0.02, 0.03, 0.02, and 0.04; for $M = 9$, the values are 0.04, 0.03, 0.06, 0.08, 0.09, 0.09, 0.03, 0.05, and 0.03). In doing so, we are attempting to model an environment with small, medium and large customers. The values of l_m are as follows: for $M = 5$, $l = (2, 3, 5, 8, 10)'$; for $M = 7$, $l = (1, 2, 5, 6, 7, 10, 15)'$; for $M = 9$, $l = (1, 2, 5, 8, 10, 15, 20, 30, 40)'$. We vary the value of T, M, and K in the test cases as shown in the Tables 2.4 to 2.6. We also vary the value of θ from 0.6 to 1, where θ is defined as in **P5** (g) as $\tilde{\mathbf{y}} \leq \theta \mathbf{c}$. That is, we use a scalar θ that expresses the maximum target flow as a percentage of the firm's capacity.

The nominal price is as follows: $\bar{\mathbf{p}} = (20, 20, 18, 18, 16)'$ for $M = 5$; $\bar{\mathbf{p}} = (20, 20, 20, 18, 18, 16, 16)'$ for $M = 7$; $\bar{\mathbf{p}} = (20, 20, 20, 20, 20, 18, 18, 16, 16)'$ for $M = 9$. Thus customers with medium volume receive 10% discount, and large volume receive 20% discount, off the base price of 20 per order. We set the rejection penalty cost to be 5 per order, i.e., $h_m = 5 * l_m$. The congestion cost is assumed to be linear $w_t^m(y_t) = w_t^m * y_t$, where w_t^m is a value dependent on the size of the problem. w_t^m is between 1 and 2 for $T = 16$ and 32; w_t^m is between 0.2 and 0.5 for $T = 48$ and 96. Further, we consider both the TFA and MCA models for HP. For the MCA model, we assume linear overflow cost κ_t, where the value of κ_t varies between 1 and 5. We assume that the target flow is unknown and is obtained through **P5**.

We present the *operational revenue* (average revenue over the sample paths) and the *strategic efficiency*. The latter is the ratio of the operational revenue to the optimal strategic revenue. This defines how well the HP performs relative to the base case.

2.5.2 Overall Performance

In this subsection, we compare the performance of the HP and NHP for both the TFA and MCA models.

We consider four test cases (T, M, K): (16,5,100), (32,7,200), (48,9,500) and (96,9,500) and vary the service level from 0.7 to 0.9. In all cases, we assume ψ_i^m's are uniformly weighted over I_m.

Table 2.5 presents the operational revenues and the strategic efficiencies for the TFA and MCA models. Comparing the results, we observe that the MCA model, as expected produces higher revenue than the TFA model. In general the operational revenue ratio is higher in the TFA case. We observe that for the most part, as the service level decreases, i.e., the system becomes more congested due to the reduce of capacity, the relative performance of the HP over the NHP increases. Finally, we observe that the strategic efficiency increases as the size of the problem increases. The strategic efficiency in MCA model are greater than 1 in some cases, because overbooking is not allowed at the strategic level. When calculating the strategic revenue, the capacity assignment is limited to the target flow. If overbooking above the target flow is allowed, more operational revenue might be generated. Based on the observations in Table 2.5, we conclude that the HP improves the operational revenue on average by 16%.

Table 2.5: Hierarchical Planning vs. No Planning

(T, M, K)	Model	θ	SE^{HP}	SE^{NHP}	OR^{HP}	OR^{NHP}	$\frac{OR^{HP}}{OR^{NHP}}$
(16,5,100)	TFA	0.9	0.7238	0.6625	3480	3185	1.09
		0.8	0.7265	0.6319	3359	2922	1.14
		0.7	0.7308	0.6060	3136	2600	1.20
	MCA	0.9	0.8725	0.8119	4195	3903	1.07
		0.8	1.0069	0.8371	4656	3871	1.20
		0.7	1.0855	0.8920	4658	3827	1.21
(32,7,200)	TFA	0.9	0.7139	0.5347	3361	2517	1.33
		0.8	0.5934	0.5249	2793	2471	1.13
		0.7	0.5771	0.4937	2488	2129	1.16
	MCA	0.9	0.7955	0.7416	3745	3491	1.07
		0.8	0.8168	0.7447	3845	3506	1.09
		0.7	0.9290	0.8068	4006	3479	1.15
(48,9,500)	TFA	0.9	0.8705	0.7325	53755	45234	1.18
		0.8	0.8176	0.6838	48141	40267	1.19
		0.7	0.9023	0.6594	47451	34679	1.36
	MCA	0.9	0.9619	0.8283	59401	51150	1.16
		0.8	1.0126	0.8565	59627	50433	1.18
		0.7	1.2429	0.9423	65364	49557	1.31
(96,9,500)	TFA	0.9	0.7801	0.7486	124194	119175	1.04
		0.8	0.8478	0.6889	127251	103407	1.23
		0.7	0.7793	0.6560	103695	87293	1.18
	MCA	0.9	0.8768	0.8440	139574	134356	1.03
		0.8	1.0088	0.8801	151426	132106	1.14
		0.7	1.1252	0.9809	149717	130517	1.14

2.5.3 The Effect of Service Level

The service level is one of the factors that has impact on the optimal target flow decision of the firm and thus ultimately affects the operational revenues (see constraint (g) in **P5**). However, the effect of service level is not immediately obvious. In the MCA model, a target flow planned according to a high service level may allow many customers to be scheduled at the same time at the strategic level. However, at the operational level, such a target flow may not perform well.

We observe that the two figures below have quite different shapes. In Figure 2.1, the operational revenue increases as the service level increases. This is because in the TFA model, the target flow is a hard constraint. Any flow above the target flow is not accepted. Therefore, as the service level increases, there is more capacity and thus the operational revenue increases. Note that in some cases (not shown), there is some fluctuation in the HP operational revenue as θ increases. Note also that the operational revenue in the HP case tends to follow the NHP case. However, as can be observed in Figure 2.1, the operational revenue for the HP drops when $\theta = 100\%$. We explain this by noting that the target flow for the high service level allocates capacity away from the peak. Because customers under the NHP do not choose their delivery schedules based on the target flow, they may not receive capacity and the operational revenue will suffer. In comparison, we observe in Figure 2.2 that the operational revenues for the HP case decreases as the service level increases (the maximum operational revenue occurs at $\theta = 65\%$). This is a significant result. Recall that the strategic revenue must increase as the service level increases. The decrease in the operational revenue expresses the inefficiency of the HP approach for the MCA model. If θ is considered an artificial variable put in place by the firm to ensure the target flow spread out, then what we observe is that the firm can benefit from such a restriction at the operational level. This means that at the strategic level, reserving some safety capacity is beneficial.

Figure 2.1: The effect of service level on the operational revenue under the TFA case

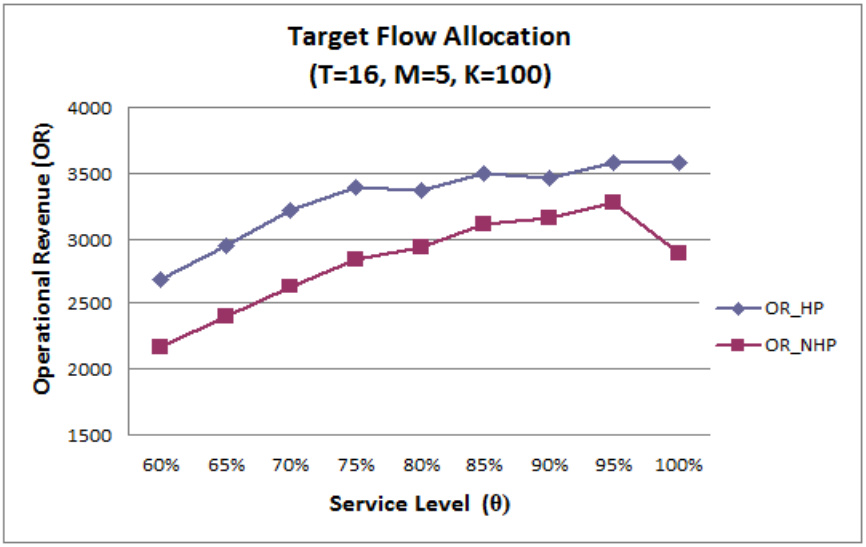
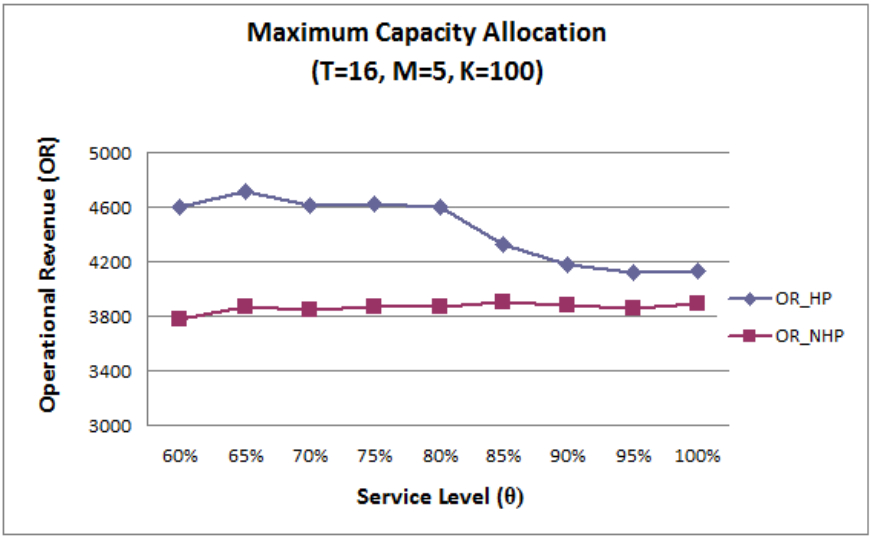


Figure 2.2: The effect of service level on the operational revenue under the MCA case



2.5.4 The Effect of Customer Valuation Patterns

The customers' valuations of service may vary with the service delivery time. Our previous analysis assumes that customers have higher valuation for the middle of the service period. Customers in other settings, however, may have different valuation patterns. In this subsection,

we study how changes in customer valuation patterns affect the HP and NHP performance. We consider three different valuation patterns: Peak, Uniform and Decreasing. We generate customer valuation patterns as follows, for each class-m, we choose v^m from a uniform random variable distributed between 50 and 120 and use it as a base valuation. Let $\tilde{v}_t^m$ be the valuation of a class-m customer for a unit service at service period t and $\tilde{\mathbf{v}}^m = (\tilde{v}_1^m, \cdots, \tilde{v}_T^m)'$. We determine $\tilde{v}_t^m$ through the following functions,

$$\tilde{v}_t^m = \begin{cases} 0.8v^m + \frac{0.8v^m}{T}t, \forall t \leq \frac{T}{2}, 1.6v^m - \frac{0.8v^m}{T}t, \forall t > \frac{T}{2} & \text{Peak;} \\ v^m, \forall t = 1, \cdots, T & \text{Uniform;} \\ 1.2v^m - \frac{0.4v^m}{T}t, \forall t = 1, \cdots, T & \text{Decreasing.} \end{cases}$$

Observe that the peak and decreasing patterns allow a maximum 20% deviations from the base valuation. Then, the the valuation of a class-m customer for a demand pattern i can be determined by $v_i^m = \tilde{\mathbf{v}}^{m'}\mathbf{a}_i^m$.

Tables 2.6 presents the results for the TFA and MCA models, respectively. We observe that the valuation pattern has little consistent effect on the HP vs. the NHP.

Table 2.6: Sensitivity to the valuation patterns

(T, M, K)	Model	Valuation Pattern	SE^{HP}	SE^{NHP}	OR^{HP}	OR^{NHP}	$\frac{OR^{HP}}{OR^{NHP}}$
(16,5,100)	TFA	Peak	0.7406	0.6206	3425	2870	1.19
		Uniform	0.6901	0.5969	3318	2870	1.15
		Decreasing	0.7483	0.5977	3593	2870	1.25
	MCA	Peak	0.9803	0.8352	4534	3862	1.17
		Uniform	0.8390	0.8034	4034	3862	1.04
		Decreasing	0.8583	0.8044	4121	3862	1.06
(32,7,200)	TFA	Peak	0.5855	0.4838	2756	2278	1.21
		Uniform	0.5650	0.4838	2660	2278	1.16
		Decreasing	0.5937	0.4838	2795	2278	1.22
	MCA	Peak	0.8170	0.7434	3846	3500	1.09
		Uniform	0.7595	0.7434	3575	3500	1.02
		Decreasing	0.7873	0.7434	3706	3500	1.05
(48,9,500)	TFA	Peak	0.7982	0.6755	47001	39774	1.18
		Uniform	0.8007	0.6755	47151	39774	1.18
		Decreasing	0.6869	0.6755	40444	39774	1.01
	MCA	Peak	1.0009	0.8548	58934	50334	1.17
		Uniform	1.0118	0.8548	59576	50334	1.18
		Decreasing	0.9068	0.8548	53396	50334	1.06

2.5.5 The Effect of Information Disclosed

Tables 2.7 shows the operational revenue of the HP and NHP under the three weighting scenarios for the TFA and MCA models, respectively. We use "U", "Y" and " X" to denote the un-weighted, weighted by target flow $\tilde{\mathbf{y}}$, and weighted by the optimal assignment $\mathbf{X}^*$ cases, respectively. The tests are conducted under various demand patterns. In Table 2.7, we observe that the $\frac{OR^{HP}}{OR^{NHP}}(X)$ and $SE^{HP}(X)$ are consistently higher than that of the weighted by target flow and unweighted cases for both the TFA and the MCA models. We observe that neither the unweighted nor the weighted by target flow cases dominate one another. However if the customers know the flow in the equilibrium solution and choose their demand patterns according to this flow, then the firm can achieve a higher operational revenue. Thus it is not enough for the customers to simply choose based on what they perceive to be the firm's preference through the expressed target flow. They also have to know how the firm prefers to allocate the target flow.

Table 2.7: Sensitivity to the information disclosed

(T, M, K)	Model	Valuation Pattern	SE^{HP} (U)	SE^{HP} (Y)	SE^{HP} (X)	$\frac{OR^{HP}}{OR^{NHP}}$ (U)	$\frac{OR^{HP}}{OR^{NHP}}$ (Y)	$\frac{OR^{HP}}{OR^{NHP}}$ (X)
(16,5,100)	TFA	Peak	0.7406	0.7561	0.7606	1.19	1.21	1.22
		Uniform	0.6901	0.6889	0.7974	1.15	1.15	1.33
		Decreasing	0.7483	0.7487	0.7918	1.25	1.25	1.32
	MCA	Peak	0.9803	0.9609	0.9812	1.17	1.15	1.17
		Uniform	0.8390	0.8525	0.9498	1.04	1.06	1.18
		Decreasing	0.8583	0.8975	0.9104	1.06	1.11	1.13
(32,7,200)	TFA	Peak	0.5855	0.6150	0.6432	1.21	1.27	1.32
		Uniform	0.5650	0.6017	0.6338	1.16	1.24	1.30
		Decreasing	0.5937	0.5952	0.6542	1.22	1.23	1.35
	MCA	Peak	0.8170	0.8281	0.8761	1.09	1.11	1.17
		Uniform	0.7595	0.7815	0.8758	1.02	1.05	1.17
		Decreasing	0.7873	0.7849	0.8696	1.05	1.05	1.16
(48,9,500)	TFA	Peak	0.7982	0.7842	0.9023	1.18	1.16	1.33
		Uniform	0.8007	0.8343	0.8373	1.18	1.23	1.23
		Decreasing	0.6869	0.7266	0.8516	1.01	1.07	1.26
	MCA	Peak	1.0009	0.9521	1.1087	1.17	1.11	1.29
		Uniform	1.0118	0.9920	1.0011	1.18	1.16	1.17
		Decreasing	0.9068	0.8977	1.0340	1.06	1.05	1.20

2.6 Conclusions

In this study, we consider a special class of demand management and capacity planning problems where customers of several classes are unable to change their schedule instantaneously in response to price incentives. In particular, we study a hierarchical planning model to consider the special requirement that price decisions and capacity allocation decisions must be made at different points of time.

The model is composed of a strategic level and an operational level. At the strategic level, we attempt to shed light on the following questions: (1) What target flow should a firm choose to maximize its profit? (2) If the target flow is given, what price should the firm charge their customers to induce the target flow? We first show that a simple price strategy where the same price difference is imposed on all customer classes can induce any feasible target flow. Then we establish the procedures of finding an optimal target flow and the price strategy to induce the target flow. The planning at this strategic level helps to better match the supply and the demand at the operational level. At the operational level, we treat the regulated deterministic demand at the strategic level as stochastic arrivals. Customers arrive randomly with a non-flexible schedule. We study the structural properties of the admission system and propose heuristic algorithms to allocate the capacities.

Our numerical results demonstrate that by pricing the capacity to induce a preferred demand profile, the hierarchical planning approach can effectively balance capacity and demand and hence substantially improves the system performance. By comparing with the naive non-hierarchical planning approach, we make the following observations:

1. The hierarchical planning model improves the operational revenue and upper bound efficiency significantly. The operational revenue improves approximately up to 20% by our experiments.

2. The hierarchical planning provides higher operational revenues for various customer valuation patterns.

There are several extensions to the above models that can be considered. In this study, customers are assumed to have perfect information regarding the other customers when responding

to the variable prices through an equilibrium behavior. A more general setting of imperfect information might be considered in the future. As a second extension, elastic demand can be considered. This is a very important extension because the demand in some real life scenarios would be sensitive to prices. Finally, the numerically analysis is limited by data availability. The numerical computation is conducted by Monte Carlo simulation due to the lack of data. The results could be validated using actual data.

Chapter 3

Pricing, Capacity Planning and Location on a Single-Facility Network

3.1 Introduction

The price charged for service, the location of the facilities, and the processing capacity of the facilities are among the most important decisions a service provider needs to make in a competitive market. Many service companies are starting to consider the price and time (travel time and service waiting time) factors when designing their service network. For example, Walmart neighborhood market stores usually average about 42,000 square feet in size, and are typically located in the highly populated areas, while Sam's Club stores (which tend to charge lower prices) have an average size of approximately 133,000 square feet, and are usually located in the suburban areas (Walmart, 2009). Many similar cases can be found, for example, the price and the number of gas pumps vary at different gas stations. Six Flags, a chain of amusement parks, prices their seasonal passes differently at different locations. A 2009 seasonal pass at Discovery Kingdom park in Vallejo, CA costs $49.99, while the same type of ticket at Magic Mountain Park in Los Angels, CA sells for $59.99. A parking lot is more expensive and crowded in a downtown area than in a suburban area. A Starbucks coffee in a convention center costs

more than the one in a residential community a block away, because business customers are willing to pay a premium to save time on traveling. Macdonald and Nelson (1991) found that fixed market basket of goods costs about 4 percent less in suburban locations than in central city stores. Motivated by an integrative, customer-focused decision making, our objective is to find out how to make the capacity, pricing and location decisions strategically to maximize profit when customers are sensitive to both price and time.

The price, location and capacity decisions are closely related to each other. To maximize profit, the interactions among these decisions must be well understood. The nature of these interactions, however, are not immediately obvious. Larger capacity may not necessarily lead to a higher profit. More capacity on one hand will tend to reduce the waiting time, but on the other hand may attract more customers that could increase the congestion and thus lengthen the waiting time. In addition, a higher price possibly comes with a self-adjusted reduced demand, but a reduced demand implies lower average waiting times if all other parameters remain fixed, and demand can go up again as a result.

Most facility location models ignore these interactions entirely and simply assume that customers choose the closest facility available for service. From a customer's perspective, however, the choice of visiting facilities may not be based just on easy access (proximity), but also on the price set by the firm and the service quality, often measured in terms of the average waiting time.

The objective of this study is to explore the interplay of three strategic decisions: the location of service facilities, the price charged for service, and the available service capacity. In the current Chapter, we start by analyzing a single facility network. The analysis to the multi-location setting is extended in the following Chapter. We consider the following setting. A profit maximizing firm is to locate a single facility on a general network, to select the capacity and to determine the price to charge for service. Stochastic demand is generated from nodes of the network. Customers are both price and time sensitive. The basic research questions we attempt to address include: (1) What is the optimal capacity and price for a given location? (2) How does the optimal price depend on location of the capacity? (3) How sensitive are location, price and capacity decisions to the elasticity of the customers with respect to these factors.

The paper is organized as follows. Section 3.2 contains the literature review. In Section 3.3 we present our model and assumptions. Properties of the optimal solutions are discussed in Section 3.4 and Section 3.5. In Section 3.6, we analyze a special case of our model and examine the impact of several key factors on the firm's profit through an example. Finally in Section 3.7 we provide some concluding remarks.

3.2 Literature Review

Many researchers have considered customer response to congestion and price when designing a service network. Two streams of literature in location models are related to this topic: spatial pricing models and facility location and congestion models. Spacial pricing models consider the joint decision of locating facilities and pricing service when customers are sensitive to the price and travel distance. Wagner and Falkson (1975) were the first to study the pricing and location problem on a network facing price sensitive demands. Later work, for example, Hansen et al. (1981), investigates a plant location problem under uniform delivered pricing, where the firm decides where and how many facilities to locate, and what uniform price to charge. In a generalized version of the uncapacitated facility location problem, Hanjoul et al. (1990) consider a firm whose objective is to maximize profits by choosing its price and the number, locations, sizes and market areas of its plants. Other relevant papers include Logendran and Terrell (1991) and So (2000). Berman et al. (2007) give a review in spatial pricing models.

Facility location and congestion models focus on location problems in the presence of stochastic demand and congestion. This topic has been extensively studied and it continues to be an active area (See Berman and Krass (2002) for an overview). The most recent works include Aboolian et al. (2008) and Elhedhli (2006), who study a service system design problem seeking to locate a set of service facilities with sufficient capacities and to assign stochastic customer demands to each facility, so as to minimize the fixed costs (of opening facilities and acquiring service capacities) and variable costs (for access and waiting). Berman and Kaplan (1987) examine a problem of locating a set of facilities on a network where the demand for service at the facility consists of a decay function of the average time customers spent in the

system. Berman and Drezner (2005) extends the model in Berman and Kaplan (1987) and locate more than one facility on the network and consider a service level constraint.

These research has investigated the interactions among various subsets of the price, location, and capacity space. This is practical when customers have little simultaneous information on price and waiting times. Traditionally, the pricing and capacity decisions were made separately by the marketing and operations functional areas within the firms. With the development of information technology, customers have access to more and more information.

To the best of our knowledge, the only paper that considers the price, location and capacity simultaneously is Dobson and Stavrulaki (2007), who study a monopolist selling a single product to time-sensitive customers located on a line segment. Our work is different from theirs in two ways. First, instead of locating on a line segment, we analyze the problem on a network setting, where a facility can be placed anywhere on the network. Second, we assume that customer utility functions reflect the trade-off preferences among price, location and waiting time. The customers respond the firm's decisions by adjusting their demand, which increases with their utilities. In Dobson and Stavrulaki (2007)'s work, customers react the firm's decision by ordering or not, based on their reservation prices.

3.3 A General Model

We consider a profit maximizing firm who offers a single service at a price p and serves customers located on a general network $V(N, L)$, where N is a set of nodes and L is a set of links. d_{ix} denotes the shortest distance from node i to a facility at $x \in V$. Customer demand occurs at the nodes of the network and is sensitive to the traveling distance d, the price paid for service p and the expected waiting time w. We model the facility service process as a general queueing system such as a $G/G/1$ system, in which demand arrives with a mean arrival rate λ, and is processed with a mean service rate μ. We assume that the coefficients of variation of inter-arrival time and service time are exogenously determined, denoted as ν_a and ν_s, respectively. Customers experience an average delay of $w(\mu, \lambda)$, which depends on the capacity μ and the total demand λ arriving at the facility. The firm incurs a cost $C \cdot \mu$ per unit time, where $p > C > 0$. The

firm's objective is to locate the facility in V, select a uniform price p and determine the capacity μ so as to maximize its profit.

There is a maximum demand associated with each node $i \in N$, denoted as λ_i^{max}. The effective demand generated at node i that travels to facility located at $x \in V$ is assumed to be a general distributed random variable with mean λ_{ix}, which decays with respect to the travel distance to the facility d_{ix}, the price paid for the service p and the expected waiting time w. Let $F(d) \in [0,1]$, where $F'(d) \leq 0$ is the decay function associated with travel distance d, let $G(p) \in [0,1]$, where $G'(p) \leq 0$ is the decay function associated with the price p, and let $H(w) \in [0,1]$, where $H'(w) \leq 0$ is the decay function associated with the expected waiting time at the service facility w. Assuming that the demand decays from the three factors are multiplicative, λ_{ix} then satisfies

$$\lambda_{ix} = \lambda_i^{max} F(d_{ix}) G(p) H(w), \qquad \forall i \in N, \forall x \in V. \tag{3.1}$$

Therefore, the total demand rate arriving at the facility located at $x \in V$ is,

$$\lambda = \sum_{i=1}^{N} \lambda_i^{max} F(d_{ix}) G(p) H(w). \tag{3.2}$$

The single facility problem formulation is

$$\begin{aligned} \max_{x \in V, p \in \mathcal{P}, \mu \in \mathcal{U}} \quad & R(x,p,\mu) = p \cdot \lambda - C\mu \\ s.t. \quad & \lambda = \sum_{i=1}^{N} \lambda_i^{max} F(d_{ix}) G(p) H(w(\mu,\lambda)). \end{aligned} \tag{3.3}$$

In this formulation, the total demand rate (3.2) is a function of waiting time and waiting time in turn is a function of the total demand rate. The feedback loop control is a special feature of our model, which represents the interactions among the price, the travel distance and the expected waiting time. The decisions of price and capacity are intertwined with each other: providing more capacities could attract more customers, but more customers may cause an increased service waiting time, which leads to a reduced demand. Similarly, charging a

higher price may reduce demand, which then implies a decrease in waiting time, which may cause demand to increase. It becomes even more complex when location factor is also taken into consideration, as demand also varies with differences in travel distances. To maximize the profit, the firm must understand the complex interplay between these factors and carefully calibrate the decisions of location, price and capacity.

3.4 Optimal Location of the Facility

We first investigate the problem of locating the facility when the price charged for service and the capacity of the facility are given. The potential locations of the facility are on the whole network, thus is an infinite set. We will show in this section that the optimal location of the facility can be reduced to a finite set.

Consider a link $(a, b) \in L$ with a length of l_{ab} on the network. Let x be a point on link (a, b), i.e., $x \in (a, b)$. With a bit abuse of notation, let $d_{ax} = x$. We define x_i as a *break point* on the link if x_i satisfies,

$$x_i + d_{ia} = l_{ab} - x_i + d_{ib}, \qquad \text{for some } i \in N.$$

That is, x_i is a point of location where demands at node i is indifferent when traveling to x_i through node a or node b. Further, we define the set of break points on the link as

$$B_{ab} = \cup_{\forall i \in N}\{x_i \in (a, b)\} \cup \{0, l_{ab}\}.$$

We sort the break points in B_{ab} in ascending order. Denote $\hat{x}$ and $\tilde{x}$ as two arbitrarily consecutive break points on link (a, b). We call an interval between any two consecutive break points a *primary region*. Assuming that $x \in (\hat{x}, \tilde{x})$, we have the following propositions that shows the optimal location exists on the nodes of the network under a mild assumption, namely that $F(d)$ is convex.

Proposition 3.1. *If $F(d_{ix})$ is convex, then $A(x) \doteq \sum_{i \in N} \lambda_i^{max} F(d_{ix})$ is convex on any primary region.*

Proof. Let N_a be the set of nodes where the shortest distance to x is through a. Let N_b be the set of nodes where the shortest distance to x is through b.

$A(x)$ can be expressed as

$$A(x) = \sum_{i \in N_a} \lambda_i^{max} F(x + d_{ia}) + \sum_{i \in N_b} \lambda_i^{max} F(l_{ab} + d_{ib} - x).$$

Thus, we have the first and second order conditions

$$A'(x) = \sum_{i \in N_a} \lambda_i^{max} F'(x + d_{ia}) - \sum_{i \in N_b} \lambda_i^{max} F'(l_{ab} + d_{ib} - x), \tag{3.4}$$

$$A''(x) = \sum_{i \in N_a} \lambda_i^{max} F''(x + d_{ia}) + \sum_{i \in N_b} \lambda_i^{max} F''(l_{ab} + d_{ib} - x). \tag{3.5}$$

Since $F''(x) \geq 0$, we must have $A''(x) \geq 0$ by (3.5). Therefore, $A(x)$ is a convex function of x on any primary region. □

Proposition 3.2. *If $F(d_{ix})$ is convex, then $A(x)$ is convex on any link.*

Proof. Since $A(x)$ is convex on any primary region, it suffices to show that for any break point $\hat{x}$ on (a, b), $\lim_{\epsilon \to 0} A'(\hat{x} - \epsilon) \leq \lim_{\epsilon \to 0} A'(\hat{x} + \epsilon)$. Suppose $\hat{x}$ is the breakpoint with respect to a node i_0 only. So that at $\hat{x} - \epsilon$ $i_0 \in N_a$ and at $\hat{x} + \epsilon$ $i_0 \in N_b$

From equation (3.4) and (3.5), we have

$$\begin{aligned} A'(\hat{x} - \epsilon) &= \sum_{i \in N_a} \lambda_i^{max} F'(\hat{x} - \epsilon + d_{ia}) - \sum_{i \in N_b} \lambda_i^{max} F'(l_{ab} + d_{ib} - (\hat{x} - \epsilon)), \\ A'(\hat{x} + \epsilon) &= \sum_{i \in N_a} \lambda_i^{max} F'(\hat{x} + \epsilon + d_{ia}) - \sum_{i \in N_b} \lambda_i^{max} F'(l_{ab} + d_{ib} - (\hat{x} + \epsilon)) \\ &\quad -\lambda_{i_0}^{max} F'(\hat{x} + \epsilon + d_{ia}) - \lambda_{i_0}^{max} F'(l_{ab} + d_{ib} - (\hat{x} + \epsilon)). \end{aligned}$$

As $\epsilon \to 0$, the first two items of $A'(\hat{x} - \epsilon)$ and $A'(\hat{x} + \epsilon)$ are the same. Since $F'(x) \leq 0$, the negatives of the third and the fourth items of $A'(\hat{x} + \epsilon)$ are nonnegative. Therefore, we have

$$\lim_{\epsilon \to 0} A'(x - \epsilon) \leq \lim_{\epsilon \to 0} A'(x + \epsilon).$$ □

Thus, $A(x)$ is convex on the entire link (a, b).

Proposition 3.3. *If $F(d)$ is convex, given p and μ, the optimal location exists in N.*

Proof. The first order condition of (3.2) ($\lambda = A(x)G(p)H(w)$) w.r.t. x gives

$$\lambda'(x) = A'(x)G(p)H(w) + A(x)G(p)H'(w)w'(\lambda)\lambda'(x).$$

Then, we have

$$\lambda'(x) = \frac{A'(x)G(p)H(w)}{1 - A(x)G(p)H'(w)w'(\lambda)}. \tag{3.6}$$

We know that $w'(\lambda) \geq 0$ and $H'(w) \leq 0$. The denominator in (3.6) thus is positive as the second term $A(x)G(p)H'(w)w'(\lambda) \leq 0$. The numerator in (3.6) shows that $\lambda'(x)$ and $A'(x)$ have the same sign. Since $A(x)$ is convex on any link, $A'(x)$ is non-decreasing on link (a, b). Consider three cases: (1) If $A'(0) \geq 0$, then $\lambda'(x) \geq 0, \forall x \in [0, l_{ab}]$, the optimal location is at node b; (2) If $A'(0) \leq 0$ and $A'(l_{ab}) \leq 0$, then $\lambda'(x) \leq 0, \forall x \in [0, l_{ab}]$, and the optimal location is at node a; and (3) If $A'(0) \leq 0$ and $A'(l_{ab}) \geq 0$, then $\lambda(x)$ is non-increasing until $\lambda'(x) = A'(x) = 0$ for some $x \in [0, l_{ab}]$, and $\lambda(x)$ is non-decreases as x goes toward node b and the optimal location is either at node a or at node b.

Applying the above argument to the whole network, we conclude that one of the two nodes always gets higher arrival rate than that of the link. For fixed p and μ, the optimal location of the facility exists in N. □

3.5 Optimal Price and Capacity Assignment Given a Facility Location

We have reduced the search for the optimal location to a finite set. We next study the problem of setting the price and capacity to maximize the profit when the facility location is known. Once the profit maximizing problem at a fixed facility location has been addressed, the best location can be determined by comparing profits across all possible locations. We first analyze the problem assuming that the facility is operating as a G/G/1 queueing system. Further, we simplify it to a M/M/1 system to gain more insights.

3.5.1 G/G/1 System

In a G/G/1 system, the expected waiting time, which includes the queueing time plus service time, can be approximated by (Hopp and Spearman, 2000),

$$w(\mu,\lambda) \approx \frac{\nu\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu}, \qquad \mu > \lambda, \tag{3.7}$$

where $\nu = \frac{\nu_a+\nu_s}{2}$ is the average of the squared coefficient of variations of the arrival process and service process, denoted as ν_a and ν_s, respectively.

We start from studying the first order conditions of the firm's profit (3.3) w.r.t. p and μ with location x fixed, which are

$$\frac{\partial R(p,\mu)}{\partial p} = \lambda + p * \frac{d\lambda}{dp} = 0, \tag{3.8}$$

$$\frac{\partial R(p,\mu)}{\partial \mu} = p * \frac{d\lambda}{d\mu} - C = 0. \tag{3.9}$$

To find $\frac{d\lambda}{dp}$, we hold the capacity of the facility μ and location x fixed. From (3.2) and (3.7), we have the derivatives w.r.t. p,

$$\frac{dw}{dp} = \frac{\nu\frac{d\lambda}{dp}}{(\mu-\lambda)^2}, \tag{3.10}$$

$$\frac{d\lambda}{dp} = A(x)G'(p)H(w) + A(x)G(p)H'(w)\frac{dw}{dp}. \tag{3.11}$$

Therefore from (3.10) and (3.11), we have

$$\frac{d\lambda}{dp} = \frac{A(x)G'(p)H(w)}{1-\frac{\nu A(x)G(p)H'(w)}{(\mu-\lambda)^2}}. \tag{3.12}$$

To find $\frac{d\lambda}{d\mu}$, we hold location x and price p fixed, from (3.2) and (3.7) we have the derivatives w.r.t. μ,

$$\frac{dw}{d\mu} = -\frac{1}{\mu^2} - \frac{\nu\lambda(2\mu-\lambda)}{\mu^2(\mu-\lambda)^2} + \frac{\nu\frac{d\lambda}{d\mu}}{(\mu-\lambda)^2}, \tag{3.13}$$

$$\frac{d\lambda}{d\mu} = A(x)G(p)H'(w)\frac{dw}{d\mu}. \tag{3.14}$$

Therefore from (3.13) and (3.14), we have

$$\frac{d\lambda}{d\mu} = \frac{A(x)G(p)H'(w)(-\frac{1}{\mu^2} - \frac{\nu\lambda(2\mu-\lambda)}{\mu^2(\mu-\lambda)^2})}{1 - \frac{\nu A(x)G(p)H'(w)}{(\mu-\lambda)^2}}. \tag{3.15}$$

Substitute (3.12) and (3.15) into (3.8) and (3.9) , we have

$$G(p)(1 - \frac{\nu A(x)G(p)H'(w)}{(\mu-\lambda)^2}) + pG'(p) = 0, \tag{3.16}$$

$$A(x)G(p)H'(w)(-\frac{p}{\mu^2} - \frac{\nu\lambda(2\mu-\lambda)p}{\mu^2(\mu-\lambda)^2} + \frac{\nu C}{(\mu-\lambda)^2}) - C = 0. \tag{3.17}$$

Proposition 3.4. *The effective demand decreases in price p, with location x and capacity μ fixed.*

Proof. The results follows from (3.12), where the numerator is less than or equal to 0 and the denominator is positive because of the non-increasing property of $H(w)$. □

Proposition 3.5. *The effective demand increases in capacity μ, with location x and price p fixed.*

Proof. The results follows from (3.15). Since $H'(w) \leq 0$ and $\mu > \lambda$ and $-\frac{1}{\mu^2} - \frac{\nu\lambda(2\mu-\lambda)1}{\mu^2(\mu-\lambda)^2} \leq 0$, the numerator of (3.15) is positive. The denominator is positive as well because of the non-increasing property of $H(w)$. Therefore $\frac{d\lambda}{d\mu} \geq 0$. □

Proposition 3.6. *The optimal price and capacity of the facility together with the equilibrium arrival rate and waiting time satisfy (3.2),* (3.7), (3.16) *and* (3.17)

Proposition 3.6 offers the necessary conditions of the optimal price and capacity, the equilibrium demand rate and waiting time. However, the system is highly nonlinear. The following proposition shows that when (x, p, μ) are given, there is always a unique equilibrium that defines the arrival rate and waiting time.

Proposition 3.7. *For any (x, p, μ),* (3.2) *and* (3.7) *defines a unique equilibrium arrival rate λ^* and waiting time w^*.*

Proof. From (3.2), arrival rate λ decreases in w. From (3.7), λ increases with w. Therefore there exists one unique arrival rate and waiting time given any (x, p, μ) by the continuity of function λ and w. □

3.5.2 M/M/1 System

In M/M/1 system, the squared coefficients of variation $\nu_a = 1$ and $\nu_s = 1$. Expression (3.7) is exact and is simplified as

$$w = \frac{1}{\mu - \lambda}, \qquad \mu > \lambda. \tag{3.18}$$

Alternatively, the relationship between capacity and waiting time is $\mu = \lambda + 1/w$. There is one to one correspondence between the waiting time and the capacity in both a G/G/1 and a M/M/1 system. However, in a M/M/1 system, we found that using the waiting time as decision variable is more convenient. The problem using w as decision variable becomes

$$\begin{aligned} \max_{x \in V, p \in \mathcal{P}, w} \quad & R(x, p, w) = p \cdot \lambda - C\mu \\ s.t. \quad & \lambda = A(x)G(p)H(w), \\ & \mu = \lambda + 1/w. \end{aligned}$$

Using w and p as decision variables, the firm's profit can be expressed as

$$R(x, p, w) = (p - C)A(x)G(p)H(w) - C/w.$$

The first order condition of the revenue function w.r.t. p and w with location x fixed are,

$$\frac{\partial R(p, w)}{\partial p} = A(x)H(w)(G(p) + (p - C)G'(p)) = 0, \tag{3.19}$$

$$\frac{\partial R(p, w)}{\partial w} = (p - C)A(x)G(p)H'(w) + C/w^2 = 0. \tag{3.20}$$

The next two propositions give necessary conditions for the optimal price and capacity with location fixed.

Proposition 3.8. *The price decision is independent of the location and capacity decisions. The optimal price is a solution of*

$$G(p) + (p - C)G'(p) = 0. \tag{3.21}$$

Proposition 3.9. *If location x and price p are known, the optimal equilibrium waiting time w^* satisfies*

$$A(x)G(p)H'(w^*){w^*}^2 = C/(C - p). \tag{3.22}$$

The equilibrium arrival rate to the facility and the optimal capacity of the facility hence satisfy

$$\lambda^* = A(x)G(p)H(w^*), \tag{3.23}$$

$$\mu^* = \lambda^* + 1/w^*. \tag{3.24}$$

The proofs of Proposition 3.8 and 3.9 follow directly from the first order conditions in (3.19) and (3.20). To find the sufficient condition, we need to know the the curvature of the objective value. The curvature of the profit function depends on the specific form of the decay function. For example, if the price and/or waiting time delay function are linear (the second derivatives vanish), it can be verified that the profit function is concave w.r.t. p and/or w. Later we will show that the profit function is unimodal when the decay functions are exponential.

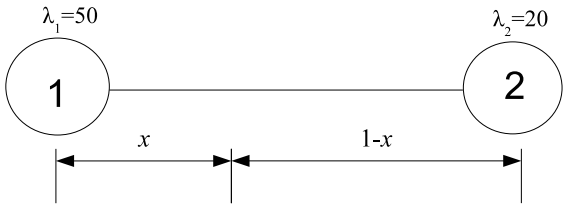
3.5.3 An Example

The pricing decision is intertwined with the location and capacity decisions in a $G/G/1$ system. However, the pricing decision is independent of the location and capacity decisions in a $M/M/1$ system. We now use a small example to examine the inter-relations of the three decision variables and investigate the independence of the pricing decision.

Consider a 2-node network as shown in Figure 3.1. The demand rates at each node are $\lambda_1^{max} = 50$ and $\lambda_2^{max} = 20$. We assume exponential decay functions, $F(d) = \exp(-\alpha d)$, $G(p) = \exp(-\beta p)$, and $H(w) = \exp(-\gamma w)$, where α, β and γ are the demand elasticities to price, travel distance and waiting time respectively. The unit capacity cost is $C = 0.5$. A firm is about to locate a facility on the network, select the price to charge for service and choose an appropriate

capacity level.

Figure 3.1: A 2-node network -single facility



We study the relationship of two of the three decision variables when the third one is fixed.

1. Location and Price.

   Recall that the optimal price in M/M/1 satisfies (3.21) and is independent of the location decision. However, we cannot derive the same conclusion for a G/G/1 system, in which the optimal price satisfies (3.16). Comparing (3.16) and (3.21), the dependency of the price and location seems related to the squared coefficient of variation and customers' elasticities. Figure 3.2 plots the firm's profit w.r.t. price when $\alpha = 0.4$. The locations of the facility are at node 1, a point that is 0.2 distance away from node 1, and node 2. The price elasticity is set at two levels, $\beta = 0.2$ and $\beta = 1$. The squared coefficient of variations are set at $\nu = 1$ and $\nu = 10$. $\nu = 1$ is the special case of M/M/1, $\nu = 10$ represents a system with a high variation of interarrival and service times. Figure 3.2 shows the four plots of the combinations $(\beta = 0.2, \nu = 1)$, $(\beta = 0.2, \nu = 10)$, $(\beta = 1, \nu = 1)$ and $(\beta = 1, \nu = 10)$. The value of the optimal price and the profit are shown as a pair of numbers on the graph. We can see that the optimal prices of all the three locations are the same in each subplot.

   The dependency of price and location could also relate to customers' sensitivity to dis-

tance. When customers are more sensitive to the travel distance, price might vary more w.r.t. location. Figure 3.3 is a similar plot to Figure 3.2 with higher distance elasticity coefficient ($\alpha = 2$). In Figure 3.3, we can see that when the price elasticity and the squared coefficient of variation are high ($\beta = 1, \nu = 10$), the optimal price varies with the location, but the optimal prices are still close to each other.

Figure 3.4 shows the relationship of the demand with respect to the price. The overall demand is decreasing w.r.t. price. The higher the squared coefficient of variation the lower the demand with all others being fixed. The higher the price elasticity the lower the demand with all others fixed. The maximum profit is obtained when the marginal benefit of increasing the price equals the marginal cost due to the decreases of the demand.

In summary, location decisions impact the magnitude of the demand and the profit obtained. However pricing decision is relatively independent of the location decision except for the cases exhibiting extremely high variation of the demand inter-arrival and service times ($\nu > 10$) and the demands are highly sensitive to the travel distance ($\alpha > 2$).

Figure 3.2: Profit vs. price with capacity fixed ($\alpha = 0.4$)

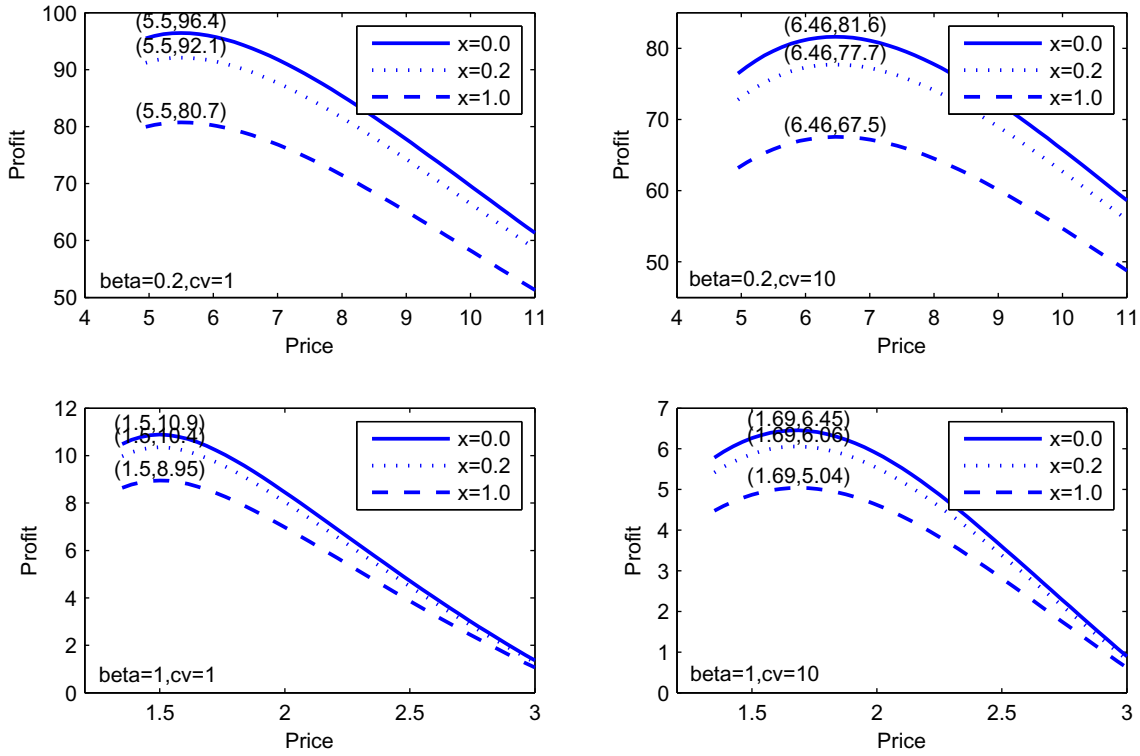


Figure 3.3: Profit vs. price with capacity fixed ($\alpha = 2$)

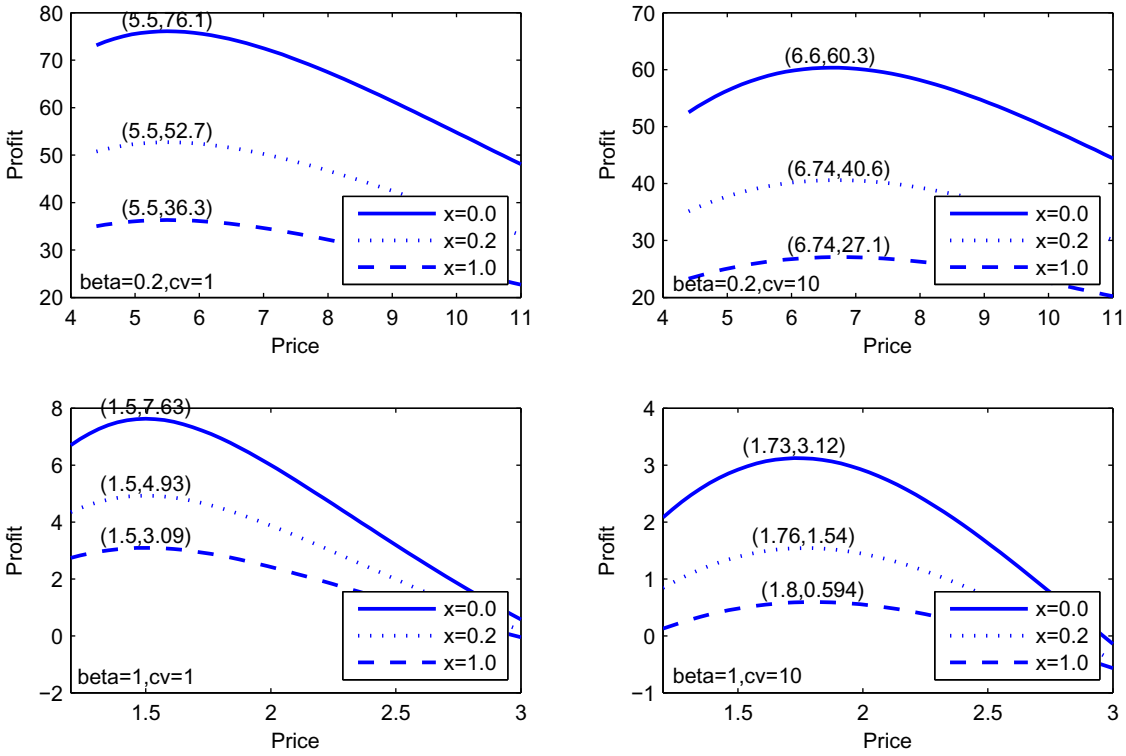
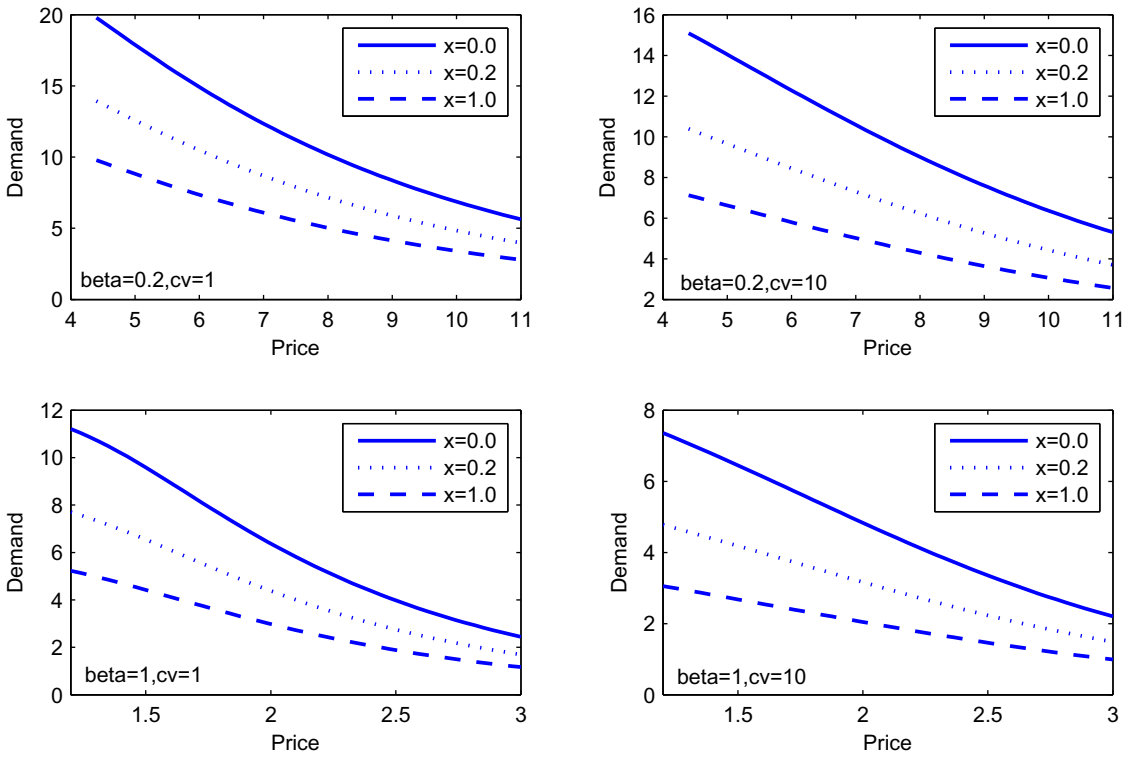


Figure 3.4: Demand vs. price with capacity fixed ($\alpha = 2$)



2. Location and Capacity.

   Figures 3.5 and 3.6 show the capacity impact on the total demand and the profit for various combinations of waiting time elasticity and squared coefficient of variation. The waiting time elasticity is set at two levels $\gamma = 0.1$ and $\gamma = 1$. The squared coefficient of variations are set to $\nu = 1$ and $\nu = 10$. The four plots are for the four combinations $(\gamma = 0.1, \nu = 1)$, $(\gamma = 0.1, \nu = 10)$,$(\gamma = 1, \nu = 1)$ and $(\gamma = 1, \nu = 10)$. Figure 3.5 shows that the demand increases w.r.t. the capacity. The higher the waiting time elasticity and/or the coefficient of variations, the more sensitive the demand is to the capacity. Comparing the curvature of the plot $(\gamma = 0.1, \nu = 1)$ with the other three, we can see that the curvature in the plot $(\gamma = 0.1, \nu = 1)$ is flatter than others. The maximum profit is obtained when the marginal profit improvement from increasing the demand equals the marginal cost of increasing the capacity.

   Figure 3.6 shows that the location and the capacity decisions are correlated with each other. From the four plots, we can see that capacity decisions must be made together with the location decisions to achieve maximum profit.

Figure 3.5: Demand vs. capacity with price fixed ($\alpha = 0.4, \beta = 0.4$)

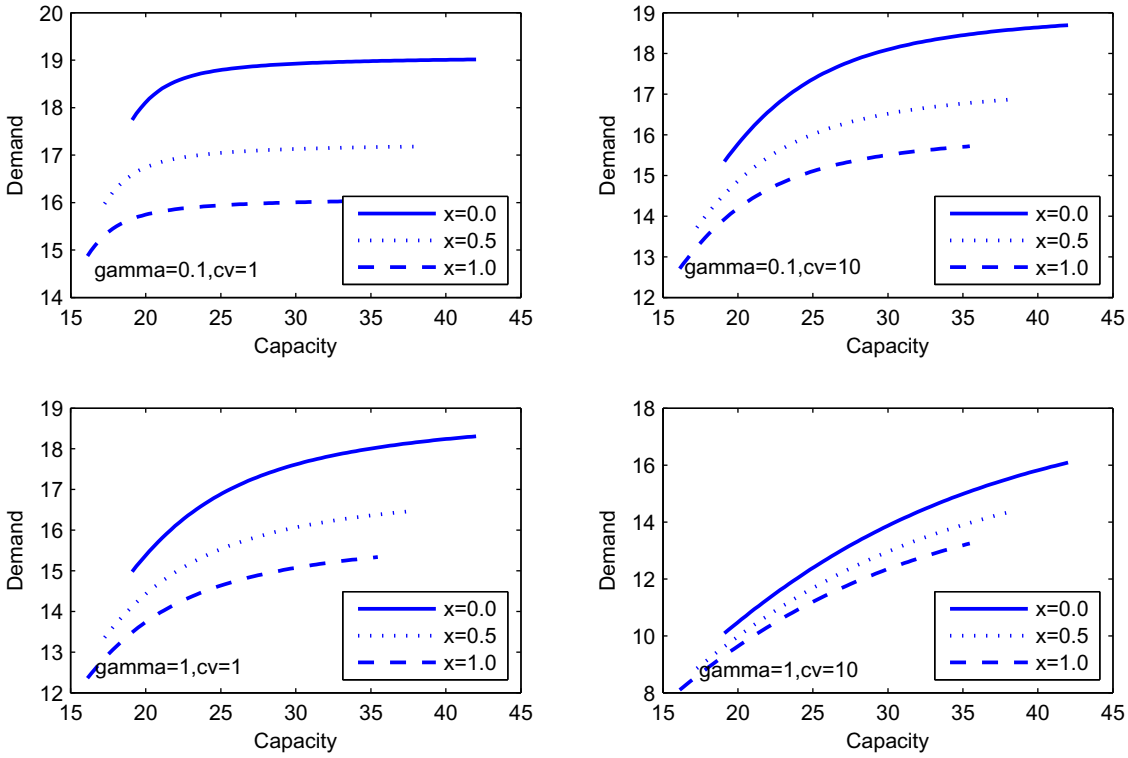
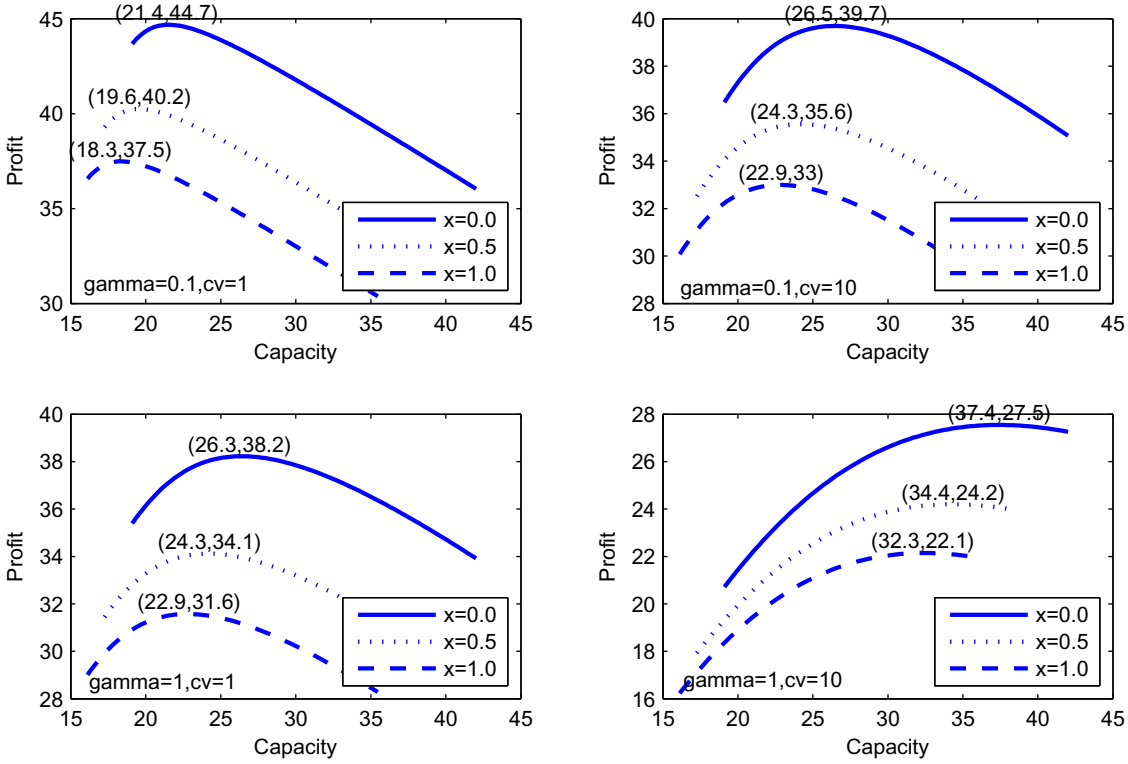


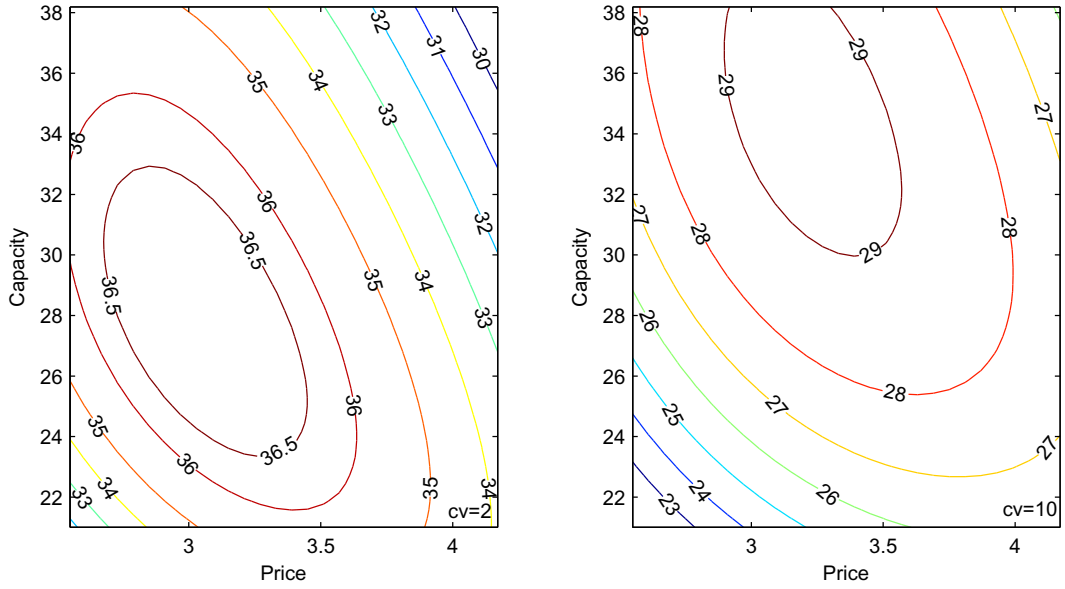
Figure 3.6: Profit vs. capacity with price Fixed ($\alpha = 0.4, \beta = 0.4$)



3. Price and Capacity.

   Figure 3.7 shows the profit contour w.r.t. price and capacity when the facility is fixed at node 1. The demand elasticities are $\alpha = 0.4$, $\beta = 0.4$, and $\gamma = 0.8$. The squared coefficient of variation is set to $\nu = 2$ and $\nu = 10$. We can see that the price and the capacity decisions need to be considered jointly to achieve the maximum profit. Trying to find an optimal solution with either one of them arbitrarily fixed could result in a suboptimal solution.

Figure 3.7: Profit contour vs. price and capacity at node 1



3.6 M/M/1 System with Exponential Decay Functions

Propositions 3.8 and 3.9 provide the necessary conditions of the optimal price and capacity, however the solution that satisfies the conditions may not be sufficient to provide optimality, due to two reasons: (1) the curvature of the revenue is unknown; and (2) the solution may not be unique depending upon the specific form of the decay functions. To simplify the problem and gain more insight of the interplay among the three decision variables, we use two assumptions for the remainder of this study: *(a) The facility operates as a $M/M/1$ queueing system;* and *(b) The demand decay functions are exponential.*

Let $F(d) = \exp(-\alpha d)$, $G(p) = \exp(-\beta p)$, and $H(w) = \exp(-\gamma w)$, where α, β and γ are demand elasticities w.r.t the travel distance, the price paid for service and the waiting time respectively. The formulation of the problem can be expressed as

$$\begin{aligned} \max_{x\in V, p\in\mathcal{P}, \mu\in\mathcal{U}} \quad & R(x,p,\mu) = p\cdot\lambda - C\cdot\mu \\ s.t. \quad & \lambda = \sum_{i=1}^{N} \lambda_i \exp(-\alpha d_{ix} - \beta p - \gamma w), \\ & w = \frac{1}{\mu - \lambda}. \end{aligned}$$

Proposition 3.10. *Given x, $w(p,\mu)$ is a decreasing convex function w.r.t. (p,μ).*

Proposition 3.11. *Given x, $\lambda(p,\mu)$ decreases in p, increases in μ and concave in (p,μ).*

Proof of Propositions 3.10 and 3.11. Substitute the demand decay functions $F(d) = \exp(-\alpha d)$, $G(p) = \exp(-\beta p)$, and $H(w) = \exp(-\gamma w)$ into (3.10), (3.12), (3.13), and (3.15). Let $\nu = 1$, we have

$$\frac{d\lambda}{dp} = \frac{-\beta\lambda}{1+\gamma w^2\lambda}, \tag{3.25}$$

$$\frac{dw}{dp} = \frac{-w^2\beta\lambda}{1+\gamma w^2\lambda}, \tag{3.26}$$

$$\frac{d\lambda}{d\mu} = \frac{\gamma w^2\lambda}{1+\gamma w^2\lambda}, \tag{3.27}$$

$$\frac{dw}{d\mu} = \frac{-w^2}{1+\gamma w^2\lambda}. \tag{3.28}$$

The second derivative of the waiting time gives

$$\begin{aligned} \frac{d^2w}{d\mu^2} &= \frac{\gamma^2 w^6\lambda + 2w^3}{(1+\gamma w^2\lambda)^3}, \\ \frac{d^2w}{dp^2} &= \frac{\beta^2 w^2\lambda(1+2w\lambda)}{(1+\gamma w^2\lambda)^3}, \\ \frac{d^2w}{d\mu dp} &= \frac{\beta w^3\lambda(2-\gamma w)}{(1+\gamma w^2\lambda)^3}. \end{aligned}$$

The determinant of the Hessian matrix of $w(p,\mu)$ is

$$\frac{2\beta^2 w^5 \lambda}{1+\gamma w^2 \lambda} \geq 0.$$

Therefore $w(p,\mu)$ is convex w.r.t. (p,μ).

The second derivative of the demand rate gives

$$\frac{d^2\lambda}{d\mu^2} = \frac{\gamma w^3 \lambda(w\gamma - 2)}{(1+\gamma w^2\lambda)^3}, \tag{3.29}$$

$$\frac{d^2\lambda}{dp^2} = \frac{\beta^2\lambda - 2\beta^2\gamma w^3\lambda^3}{(1+w^2 f_\gamma)^3}, \tag{3.30}$$

$$\frac{d^2\lambda}{d\mu dp} = -\frac{\beta\gamma w^2\lambda(2w\lambda+1)}{(1+\gamma w^2\lambda)^3}. \tag{3.31}$$

The determinant of the Hessian matrix of $\lambda(p,\mu)$ is

$$-\frac{2\beta^2\gamma w^3\lambda^2}{1+\gamma w^2\lambda} \leq 0.$$

Therefore $\lambda(p,\mu)$ is concave w.r.t. (p,μ). □

Proposition 3.12. *There exists a unique optimal price p^* and capacity μ^* given any x such that*

$$p^* = \frac{1}{\beta} + C, \tag{3.32}$$

$$\mu^* = \frac{2\gamma L(x) + \beta\gamma C}{4L(x)^2}, \tag{3.33}$$

$$w^* = \frac{2}{\gamma} L(x), \tag{3.34}$$

$$\lambda^* = \frac{\beta\gamma C}{4L(x)^2}, \tag{3.35}$$

where

$$L(x) = -LambertW\left(-\frac{1}{2}\left(\frac{\beta\gamma C\exp(\beta p^*)}{\sum_{i=1}^n \lambda_i \exp(-\alpha d(i,x))}\right)^{\frac{1}{2}}\right).$$

Proof. Using (3.3) (Recall that $\lambda = A(x)G(p)H(w)$) and (3.26), the first order of the revenue

w.r.t. p and gives

$$\frac{\partial R(p,\mu)}{\partial p} = \lambda + p\frac{d\lambda}{dp} = \lambda\left(1 - p\left(\beta + \gamma\frac{dw}{dp}\right)\right) = \lambda\big(1 - \frac{\beta p}{1+\gamma w^2\lambda}\big) = 0. \tag{3.36}$$

Using (3.3) and (3.27), the first order of the revenue w.r.t. μ gives

$$\frac{\partial R(p,\mu)}{\partial \mu} = p\frac{d\lambda}{d\mu} - C = \frac{\gamma w^2\lambda p}{1+\gamma w^2\lambda} - C = 0. \tag{3.37}$$

Solving (3.36) and (3.37), we have $p^* = \frac{1}{\beta} + C$ and the optimal capacity μ^* satisfies

$$w^2\lambda = \beta C/\gamma, \tag{3.38}$$

Thus,

$$w^2\sum_{i=1}^{N}\lambda_i \exp\big(-\alpha d_{ix} - \beta p^* - \gamma w\big) = \beta C/\gamma, \tag{3.39}$$

which can be rewritten as

$$w^2\exp(-\gamma w) = \frac{\beta C\exp(\beta p^*)}{\gamma\sum_{i=1}^{N}\lambda_i\exp(-\alpha d_{ix})}. \tag{3.40}$$

Further we rewrite (3.40) as,

$$(-\frac{1}{2}\gamma w)^2\exp(-\frac{1}{2}\gamma w)^2 = \frac{\beta\gamma C\exp(\beta p^*)}{4\sum_{i=1}^{N}\lambda_i\exp(-\alpha d_{ix})}. \tag{3.41}$$

Since $w > 0$ and $-\frac{1}{2}\gamma w < 0$, (3.41) is equivalent to,

$$(-\frac{1}{2}\gamma w)\exp(-\frac{1}{2}\gamma w) = -\big(\frac{\beta\gamma C\exp(\beta p^*)}{4\sum_{i=1}^{N}\lambda_i\exp(-\alpha d_{ix})}\big)^{\frac{1}{2}}, \tag{3.42}$$

which is a form of LambertW function and has a solution on its principle branch, i.e, $\frac{1}{2}\gamma w < 1$,

$$w^* = -\frac{2}{\gamma}L(x) \tag{3.43}$$

where

$$L(x) = -LambertW\bigg(-\frac{1}{2}\bigg(\frac{\beta\gamma C\exp(\beta p^*)}{\sum_{i=1}^{n}\lambda_i\exp(-\alpha d(i,x))}\bigg)^{\frac{1}{2}}\bigg).$$

Substitute w^* into (3.18) and (3.38), then

$$\mu^* = \frac{2\gamma L(x) + \beta\gamma C}{4L(x)^2},$$

and

$$\lambda^* = \frac{\beta\gamma C}{4L(x)^2}.$$

Next, we show that the first order solution p^* and μ^* is indeed an optimal solution.

Claim 1, $R(p,\mu)$ is unimodal in p, the unique stationary point gives the maximum solution.

Let $\hat{p}$ be the stationary point that satisfies $\frac{\partial R(p,\mu)}{\partial p} = 0$. If $p < \hat{p}$, by the convexity of w (see proof of Proposition 3.11), $\frac{dw}{dp}|_{p=p} \leq \frac{dw}{dp}|_{p=\hat{p}} < 0$. Thus from (3.36), we have $1 - p*(\beta + \gamma\frac{dw}{dp}|_{p=p}) > 1 - \hat{p}*(\beta + \gamma\frac{dw}{dp}|_{p=\hat{p}}) = 0$. Hence $R(p,\mu)$ is increasing in p when $p < \hat{p}$. If $p > \hat{p}$, by the convexity of $w, \frac{dw}{dp}|_{p=p} \geq \frac{dw}{dp}|_{p=\hat{p}}$, thusm $1-p*(\beta+\gamma\frac{dw}{dp}|_{p=p}) < 1-\hat{p}*(\beta+\gamma\frac{dw}{dp}|_{p=\hat{p}}) = 0$. Hence $R(p,\mu)$ is decreasing when $p > \hat{p}$. Therefore, $R(p,\mu)$ has only one stationary point in p where the optimal revenue is obtained.

Claim 2, $R(p,\mu)$ has one or two stationary point in μ, μ^* is the larger root that satisfies $\frac{\partial R(p,\mu)}{\partial \mu} = 0$.

The second derivative of the revenue gives

$$\frac{\partial^2 R(p,\mu)}{\partial \mu^2} \quad = \quad p\frac{d^2\lambda}{d\mu^2}.$$

Therefore, the curvature of the revenue depends on $\frac{d^2\lambda}{d\mu^2}$, (3.29). Let $w^*(p,\mu)$ be the equilibrium waiting time given (p,μ). Since $w^*(p,\mu)$ decreases w.r.t. μ, we start from a sufficiently small capacity $\mu_0 \geq 0$ such that $w^*(p,\mu_0) > 2/\gamma$ and therefore $\frac{d^2\lambda}{d\mu^2}|_{\mu=\mu_0} \geq 0$. As $\mu \geq \mu_0$ increases, the waiting time $w^*(p,\mu)$ decreases to some point $\hat{\mu}$ such that $w^*(p,\hat{\mu}) = 2/\gamma$ or $\frac{d^2\lambda}{d\mu^2}|_{\mu=\hat{\mu}} = 0$. Therefore $R(p,\mu)$ is convex in $[\mu_0,\hat{\mu}]$. When $\mu \geq \hat{\mu}$, $w^*(p,\hat{\mu}) < 2/\gamma$ or $\frac{d^2\lambda}{d\mu^2}|_{\mu=\hat{\mu}} \leq 0$. Therefore $R(p,\mu)$ is concave in $[\hat{\mu},\infty]$. The optimal μ is either μ_0 or some μ such that $\frac{\partial R(p,\mu)}{\partial \mu} = 0$. Since

$\lim_{\mu \to 0} R(p, \mu) = 0$, optimal μ is the one that satisfies $\frac{\partial R(p,\mu)}{\partial \mu} = 0$. Therefore optimal μ is the larger root that satisfies $\frac{\partial R(p,\mu)}{\partial \mu} = 0$. Therefore, μ^* in (3.33) is the larger stationary point and the optimal capacity. □

Proposition 3.12 provides many interesting insights. (3.32) shows that the optimal price is independent of the location decision, and the optimal price decreases with β, i.e., optimal price is lower when the demand is more sensitive to price. The optimal capacity in (3.33) is closely related with location through $L(x)$ and increases with γ. That is, the optimal capacity gets bigger as the demand gets more sensitive to the waiting time.

Let $\rho = \lambda/\mu$ be the system utilization rate. Proposition 3.13 shows the utilization rate at the optimal price and capacity.

Proposition 3.13. *When the price and capacity are optimal, the utilization rate $\rho^* = \lambda^*/\mu^*$ of the system satisfies,*

$$\rho^* = \frac{\beta C}{2L(x) + \beta C}, \tag{3.44}$$

Proof. It follows directly from the ratio of (3.35) and (3.33). □

Proposition 3.13 shows that the utilization rate ρ^* decreases as α or γ increases (L(x) increases in α and γ).

From Propositions 3.3 and 3.12, it is straightforward to use the following algorithm to solve the single facility problem:

Algorithm - Single facility

Step1: *Obtain $p^* = \frac{1}{\beta} + C$.*

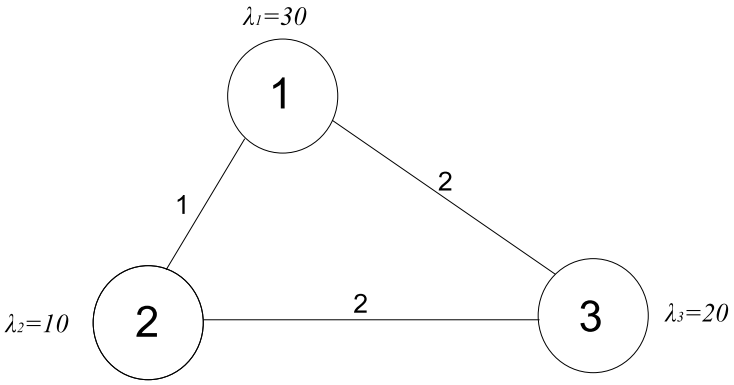
Step2: *For each $x \in N$, calculate w^*, μ^*.*

Step3: *Evaluate the revenue function and choose the location that gives the best revenue.*

3.6.1 An Example

Consider a 3-node network as shown in Figure 3.8 . The demand rates at each node are $\lambda_1 = 30$, $\lambda_2 = 10$, and $\lambda_3 = 20$. The demand elasticities to price, travel distance and waiting time are $\alpha = 0.1$, $\beta = 0.2$, and $\gamma = 0.3$, respectively. The unit capacity cost is $C = 3$. A firm is about to locate a facility on the network, select the price to charge for service and choose the appropriate capacity level.

Figure 3.8: A 3-node network - single facility



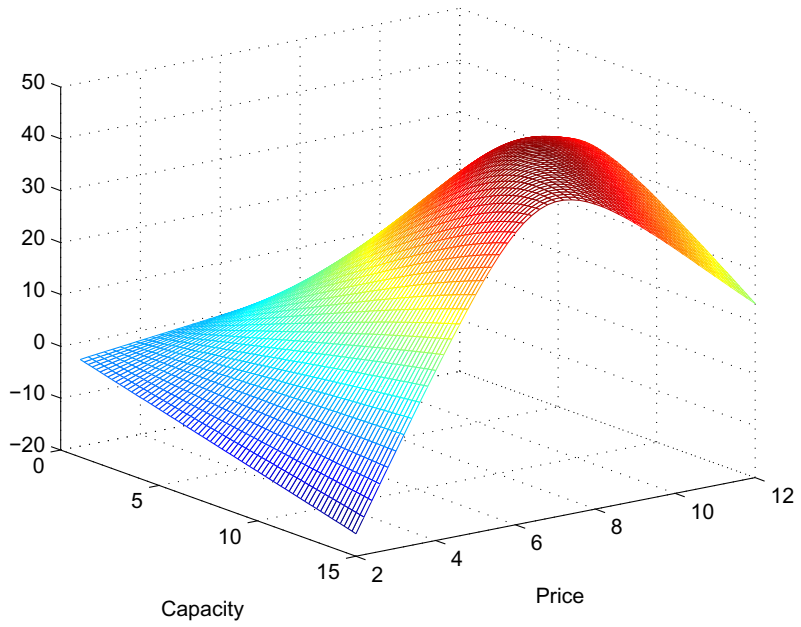
Note that the price is independent of the location and the capacity decisions. Proposition 3.12 tells us the optimal price ($p^* = 1/\beta + C = 8$) and the optimal capacities and revenues at each node, which are summarized in Table 3.1.

We observe that the optimal location is at node 1, which gives a profit of 43.74. In Figure 3.9, we plot the profit surface with respect to prices and capacities when the facility is at node 1. We can see that the profit is a single mode function of prices and capacities.

Table 3.1: Optimal capacity, price and profit at various locations

Node	Optimal Price	Optimal Capacity	Equilibrium Demand	Equilibrium Waiting Time	Profit
1	8	11.9795	9.7694	0.4525	42.2166
2	8	11.5788	9.4097	0.4610	40.5414
3	8	11.4164	9.2642	0.4646	39.8644

Figure 3.9: The profit surface at node 1



3.7 Conclusions

We considered how a firm should integrate the three decisions of locating a single facility, charging the price for service and determining the capacity of the facility to maximize profit, taking into account that customer demand is influenced by the firm's decisions. We studied the single facility model on a general network, where the demand arrival process and the service times are stochastic and demand rate is a decreasing function of the price, the travel distance and the expected waiting time at the facility. We established the existence of equilibrium demand and waiting time. We showed that the optimal location exists on the nodes of the network so that the choice of locations is reduced to a finite set. If service and arrival processes can be represented by the M/M/1 system, the optimal price is independent of the location and capacity decisions. Furthermore, when the demand is exponential with respect to the three decision variables, the optimal capacity can be expressed in a closed form for a fixed facility location. These results provide the following managerial insights:

1. Pricing decisions can be made independently of location decisions in a broad range of demand types. According to our numerical examples, when demand is not very sensitive to travel distances, and the average variance of the demand inter-arrival time and service time is not extremely large, the optimal price is insensitive to the location of the facility.

2. Capacity decisions however must be made with the consideration of the location of the facility. Our numerical examples showed that the optimal capacities vary greatly across different locations of the facilities with all other parameters fixed.

3. When the location of the facility is known, pricing and capacity decisions need to be made jointly to achieve maximum profit.

We acknowledge that the model formulated here does not account for all factors that may influence strategic decisions of a firm. To provide better understand the interactions of demand flow with the capacity and price, we limited our analysis to a single facility case. Extension to a multiple facility case is analyzed in the next Chapter. In addition, in our model, the firm behaves as a Monopolist, the competition is considered implicitly through the elasticity of the

demand. Future studies may consider competition between multiple firms or locating facilities on a network with pre-existing facilities.

Chapter 4

Pricing, Capacity Planning and Location on a Multiple-Facility Network

4.1 Introduction

Determining the locations of facilities, the service capacities, and choosing the price to charge for service are the three most important strategic decisions a service provider needs to make. Traditionally, pricing, location planning and capacity allocation are made separately. To maximize profit and improve service quality, a joint pricing, location, and capacity allocation scheme is necessary.

There is a growing consensus among researchers and practitioners alike that the three strategic decisions should be integrated. However, it is not immediately clear how the decisions should be combined and under what conditions an integrated decision is superior to separated decisions. To answer these questions, we need to understand customers' behaviors in response to the strategic decisions. A typical assumption in location literature is that customers patron the closest facility for service and their decisions of where to shop are independent of each other. The assumption works well for many situations. However, it may not be sufficient enough to characterize customers' true behaviors. As in our setting, a customer may decide to visit a

particular facility because that facility offers faster service and better price, not necessarily the closest travel distance. Customers from the same node may find it beneficial to visit different facilities to avoid congestion. Moreover, due to the presence of congestion, a customer's choice of service depends not only on his own decision but also on other customers' preferences as well.

In this study, we focus on a customer-oriented approach to design a multiple facility service network. In contrast to many studies, customers choose service not by finding the closest facility but by the facility providing the best utility. When facing multiple facilities, customers choose where to shop and adjust the total demand according to their utilities. They evaluate the service according to the price to pay and the time to spend and then choose the one that is most favorable.

The multiple-facility model to be presented here is partly an extension of the single facility model in Chapter 3. Solving the problem for multiple facilities is much more difficult than for the single facility case where customers patron the only available facility for service and the firm does not need to worry about customers switching to other facilities. When there are multiple facilities, customers have multiple choices of where to shop. The final choice is subject to customers' decision rules, for example, choosing the closest facility. We incorporate this customer choice behavior in our model, assuming all customers behave strategically so that the service system will be finally in an equilibrium state such that no customers has an incentive to alter her choice.

The organization of this chapter is as follows. Section 4.2 is the literature review. Section 4.3 provides some background information and introduces the assumptions and notations of the models. Section 4.4 presents a system optimization model where customers can be dictated to achieve the maximum profit of the system. Section 4.5 presents a user optimization model, where customers' equilibrium behavior is captured. Section 4.6 analyzes the properties of the optimal solutions and the insights developed in this section will be carried forward to develop the heuristic algorithms in Section 4.7. Section 4.8 shows the numerical results. The last section 4.9 is the conclusion and managerial insights.

4.2 Literature Review

Locating facilities with consideration of customer behaviors has received considerable attention in both the Economics and Operations Management literature. We review only the literature directly related to our study. Eiselt et al. (1993) and Drezner and Hamacher (2002) provide comprehensive reviews of the broader literature.

This work falls in the general category of Location Problems with Stochastic Demand and Congestion(LPSDC). Berman and Krass (2002) give an overview of LPSDC problems. Originated from coverage-type location models, LPSDC has been a very active research area. For example, Marianonov and Serra (1998) study a location-allocation problem for a congested system, where customer demands are assumed to be generated by a Poisson process, the distribution of service time is exponential, each facility act as a $M/M/1/k$ queueing system with finite capacity k. The objective is to locate m facilities to capture as much demand as possible. In contrast to Marianonov and Serra (1998) assuming customers can be assigned to any open facility within the coverage radius, Berman et al. (2006) study a location model with stochastic demand and congestion, where each customer will tend to patronize the closest open facility. They considered two potential source of lost demand, demand lost due to insufficient coverage and demand lost due to congestion. The objective is to find the minimum number of facilities, and their locations, so that the amount of demand loss from either source does not exceed certain pre-set levels. The most recent works include Aboolian et al. (2008) and Aboolian et al. (2009).

None of above studies, however, incorporates the consumer choice equilibrium into their models. When there are congestions in the system, customers decisions are correlated with each other and over time they learn to distribute themselves more "evenly", so that the system is in an equilibrium state, where no customers has any incentive to change her choice. One feature of our model is to capture this type of customer equilibrium behaviors.

Another area of literature closely related to our work is the location models with elastic demands. In Huff (1964), customers split their demands between several facilities with the frequency of a visit to a facility increasing with the attractiveness of the facility and decreasing

with the travel distance. Berman and Kaplan (1987) consider the time that customers spend at the service facilities and presented an algorithm to find the supply and demand equilibrium. Our work is similar to Berman and Kaplan in defining the supply and demand equilibrium, but we also add one more dimension of price and consider the customer choice equilibrium.

The customers choice equilibrium in this work are modeled using traffic equilibrium which has traditionally been developed for transportation planning and has been penetrated in recent years to other scientific fields such as telecommunication, power management and health care etc. For example, Zhang et al. (2009) study a health care facility network design problem with congestion. In discussing a traffic equilibrium model, Wardrop (1952) introduces the Wardrop equilibrium in which customers traveling between the same origin and destination have the same utility. Further, Dafermos (1980) uses the theory of variational equalities to establish the existence of a traffic equilibrium and devises an algorithm for computing an equilibrium. Aashtiani and Magnanti (1981) present a traffic equilibrium model considering the congestion effect, and discuss existence and uniqueness of the equilibria.

Berman and Drezner (2005) maybe the closest work related to our study, where location of congested facilities with distance sensitive demand is discussed. However, whereas in Berman and Drezner demand generated at a node is distance dependent and service level constraint is applied, our demand depends on customer's utility, which is a exponentially decay function of distance, price and waiting time. In Berman and Drezner's model, customers from a node have to choose only one facility for service. In our model, customers can choose any facility for service as long as it is justified by their utilities. Customers at a node need not visit the same facility, and they can split the demand to different facilities. In equilibrium, nobody has any incentive to alter her choice.

4.3 Assumptions and Backgrounds

Consider a general network $V(N, L)$, where N is the set of nodes and L is the set of links. A service provider is about to locate a fixed number of facilities on V, and in the mean time to decide the prices to charge for service and set up the capacities. We assume that facilities

must be located at nodes. Let $M \subseteq N$ be the set of potential locations of the facilities. Let $S \subseteq M$ be the set of locations of the facilities decided. We use $|S|$ to denote the total number of facilities to be located. Let $p = \{p_j, j \in S\}$ and $\mu = \{\mu_j, j \in S\}$ be the set of prices charged for service and the capacity levels to be set up, respectively.

The demands are generated from the nodes of the network. Customers are sensitive to the travel distances, prices for service and the expected waiting times at the facilities. Assume there is a facility at j, the posted price for service is p_j and the expected waiting time at the facility is w_j, then the disutility of a customer from node i to visit a facility at node j for service is,

$$u_{ij} = \alpha d_{ij} + \beta p_j + \gamma w_j, \qquad \forall i \in N, \forall j \in S.$$

We allow fractional flows, i.e., customers from the same node may choose multiple facilities. Let λ_i be the maximum demand at node i, and y_{ij} be the proportion of the maximum demand at node i to visit facility j, $\sum_{j \in S} y_{ij} = 1, \forall i \in N$, then the actual demand rate from node i to facility j is defined as an exponential decay function of the disutility u_{ij} as follows,

$$v_{ij} = \lambda_i y_{ij} \exp(-\alpha d_{ij} - \beta p_j - \gamma w_j), \qquad \forall i \in N, \forall j \in S. \tag{4.1}$$

Alternatively, we may say that the actual demand rates must satisfy the following flow conservation constraint,

$$\sum_{j \in S} v_{ij} \exp(\alpha d_{ij} + \beta p_j + \gamma w_j) = \lambda_i, \qquad \forall i \in N, \forall j \in S. \tag{4.2}$$

We will use the actual demand flow v_{ij} or the proportion of demand y_{ij} interchangeably as variables indicating customer demand distribution throughout this chapter.

We assume the demands arrive according to a Poisson process and the service rate follows an exponential distribution. We assume only one server at each facility, so that the service process acts as a $M/M/1$ queueing system. Thus, the expected system waiting time for a facility at node j is

$$w_j = \frac{1}{\mu_j - \sum_{i \in N} v_{ij}}, \qquad \forall j \in S. \tag{4.3}$$

The firm's profit then is:

$$\max R(S,p,\mu) = \sum_{j\in S}(p_j \sum_{i\in N} v_{ij} - C\cdot\mu_j), \tag{4.4}$$

where v_{ij} is subject to customers' decision rules.

Note that the final flow in (4.1) depends on customers' decision rules. If we assume that customers cooperate with the firm to obtain system optimal distribution of flow (e.g., if customers' assignment to facility can be centrally enforced), we obtain a *system optimization* model. Alternatively, if customers behave as non-cooperating players, then we obtain a traffic equilibrium where nobody has any incentive to deviate unilaterally. This results in a *user equilibrium model*. We discuss the models in the following two sections respectively.

4.4 System Optimization Model

In the system optimization model, we assume that customers cooperate with the firm (either willingly or by the firm's enforced assignment to facilities) to maximize its profit. Note that demands are still affected by the travel distances and the prices through the utility function. Let x_j be the binary indicator variable equal to 1 if a facility is located at j and 0 otherwise. The model can now be formulated as a Mixed Integer Nonlinear Programming as follows,

$$\begin{aligned}
\max_{p,\mu,x} \quad & \sum_{j\in M}\left(p_j \sum_{i\in N} v_{ij} - C\cdot\mu_j x_j\right) \\
s.t. \quad & \sum_{i\in N} v_{ij} \le \mu_j \qquad \forall j\in M, \\
& \sum_{j\in M} v_{ij}\exp(\alpha d_{ij}+\beta p_j+\gamma w_j) = \lambda_i \qquad \forall i\in N, \qquad (\mathbf{P1}) \\
& v_{ij} \le K x_j \qquad \forall i\in N, \forall j\in M, \\
& \sum_{j\in M} x_j = |S|, \\
& x_j = 0,1 \qquad \forall j\in M, \\
& v_{ij} \ge 0 \qquad \forall i\in N, \forall j\in M,
\end{aligned}$$

where w_j is expressed using the decision variables as follows:

$$w_j = \frac{1}{\mu_j - \sum_{i \in N} v_{ij}} \qquad \forall j \in M.$$

The first constraint in **P1** ensures that enough capacity is provided to satisfy demand. The second constraint is a flow conservation constraint as defined in 4.2. The third constraint is to ensure that customers visit only the nodes where a facility is located. K is a sufficiently large number so that there is no flow constraint when $x_j = 1$. The fourth constraint is to limit the total number of facilities to be $|S|$. **P1** can be solved with standard non-linear programming algorithms (although the non-linearity in the constraints presents computational challenges).

4.5 User Equilibrium Model

In the user equilibrium model, a customers chooses the facilities that maximize her utility. The system is in equilibrium if no customer will deviate from the current choice to seek for a better utility. As such, customers may not visit the closest facility and customers from a node may not visit the facility at the same node.

To model this customers' choice equilibrium, we use the classical traffic equilibrium model (Nagurney, 1999), which shows how users on a congested transportation network choose travel paths to minimize travel cost from origins to destinations. If the firm has decided the locations, prices and capacities, we define our customer choice equilibrium using the similar definition to traffic equilibrium as,

Definition 4.1. *Customer Choice Equilibrium* $(\boldsymbol{v}^*, \boldsymbol{u}^*)$ *is in an equilibrium if, once established, no customer has any incentive to alter her choice. This state is characterized by the following conditions, which must hold for every node* $i \in N$ *and every node-facility* (i, j) *pair:*

$$u_{ij}(\boldsymbol{v}^*) \begin{cases} = u_i^* & \text{if } v_{ij}^* > 0; \\ \geq u_i^* & \text{if } v_{ij}^* = 0. \end{cases} \qquad \forall i \in N, \forall j \in S. \tag{4.5}$$

The consumer choice equilibrium can be formulated as a generalized nonlinear complemen-

tary problem:

$$
\begin{aligned}
&(u_{ij}(\mathbf{v}) - u_i)v_{ij} = 0 \qquad \forall i \in N, \forall j \in S, && (a)\\
&u_{ij}(\mathbf{v}) - u_i \geq 0 \qquad \forall i \in N, \forall j \in S, && (b)\\
&\textstyle\sum_{j\in M} v_{ij} - D_i(\mathbf{u}) = 0 \qquad \forall i \in N, && (c) \quad \textbf{(P2)}\\
&v_{ij} \geq 0 \qquad \forall i \in N, \forall j \in S, && (d)\\
&u_i \geq 0 \qquad \forall i \in N. && (e)
\end{aligned}
$$

In this formulation:

v_{ij}	is the flow between node i and facility j;
$\mathbf{v}$	is the vector of$(v_{11} \cdots v_{\|N\|1}, v_{12} \cdots v_{\|N\|2}, \cdots\cdots, v_{1\|S\|} \cdots v_{\|N\|\|S\|})'$;
u_i	is an accesibility variable, the lowest disutility for Node-Facility pair (i, j);
$\mathbf{u}$	is the vector of $(u_1, \cdots, u_{\|N\|})'$;
$D_i(\mathbf{u})$	is the demand rate originated from node i, $D_i(\mathbf{u}) = \lambda_i \exp(-u_i)$;
$u_{ij}(\mathbf{v})$	is the disutility function if node i choose facility j for service $u_{ij} = \alpha d_{ij} + \beta p_j + \gamma w_j$ where $w_j = 1/(\mu_j - \sum_{i\in N} v_{ij})$

(a) and (b) in Problem **P2** model the customer choice equilibrium law requiring that for any node-facility pair (i, j), the disutility for all choices of service with positive flow $v_{ij} > 0$, is the same and equal to u_i. u_i is no more than the disutility for any node-facility pair that has no flow, i.e., $v_{ij} = 0$. Constraint (c) requires that the total flow originated from node i equals the total demand $D_i(\mathbf{u})$, which in turn depends upon the congestion in the facility through utility variable $\mathbf{u}$.

Theorem 4.1. ***P2*** *has at least one equilibrium solution.*

Proof. The proof follows directly from Theorem 5.4 of Aashtiani and Magnanti (1981). $u_{ij}(\mathbf{v})$ is a positive continuous function for all node-facility pair (i, j) and $D_i(\mathbf{u})$ is a continuous function that is bounded from above. Therefore **P2** has a solution. □

Let x_j be the binary indicator variable equals to 1 if a facility is located at j and 0 otherwise. The firm's problem is to maximize its revenue while customers act optimally to maximize their

utilities,

$$
\begin{aligned}
\max_{p,\mu,x} \quad & \textstyle\sum_{j\in M}\left(p_j \sum_{i\in N} v_{ij} - C x_j \mu_j\right) \\
s.t. \quad & \textstyle\sum_{i\in N} v_{ij} \le \mu_j \qquad \forall j \in M, \\
& v_{ij} \le K x_j, \qquad \forall i \in N, j \in M, \\
& (u_{ij} - u_i) v_{ij} = 0 \qquad \forall i \in N, j \in M, \\
& u_{ij} - u_i \ge 0 \qquad \forall i \in N, j \in M, \\
& \textstyle\sum_{j\in M} v_{ij} - \lambda_i \exp(-u_i) = 0 \qquad \forall i \in N, \qquad \textbf{(P3)} \\
& v_{ij} \ge 0 \qquad \forall i \in N, j \in M, \\
& u_i \ge 0 \qquad \forall i \in N, \\
& \textstyle\sum_{j\in M} x_j = |S|, \\
& x_j = 0, 1 \qquad \forall i \in N, j \in M, \\
& u_{ij} = \alpha d_{ij} + \beta p_j + \frac{\gamma}{\mu_j - \sum_{i\in N} v_{ij}} + K(1 - x_j) \qquad \forall i \in N, j \in M.
\end{aligned}
$$

where K is a sufficiently large constant. The problem is a Mathematical Problem with Equilibrium Constraints (MPEC), which is in general very difficult to solve.

4.6 Properties of Optimal Solutions

In this section, we study the properties of the optimal solutions in Problem **P3**. The locations of the facilities are assumed to be known. We focus on finding the optimal price and capacity allocation, and how the customers would respond to the firm's decisions.

4.6.1 Optimal Price and Capacity Allocation

If we know the locations of the facilities and how customers distribute their flows (i.e., either $\{v_{ij}, i \in N, j \in S\}$ or $\{y_{ij}, i \in N, j \in S\}$ are known), we have the following results that characterize the optimal price and capacity.

Proposition 4.1. *The optimal price is,*

$$p_j^* = \frac{1}{\beta} + C, \forall j \in S, \tag{4.6}$$

and the optimal capacity of a facility at node j satisfies

$$\mu_j^* = \frac{2\gamma L_j(y) + \beta\gamma C}{4L_j(y)^2}, \forall j \in S, \tag{4.7}$$

where

$$L_j(y) = -LambertW\Big(-\frac{1}{2}\Big(\frac{\beta\gamma C \exp(\beta p_j^*)}{\sum_{i=1}^n \lambda_i y_{ij} \exp(-\alpha d_{ij})}\Big)^{\frac{1}{2}}\Big), \forall j \in S. \tag{4.8}$$

Proof. As a service facility is operating as a M/M/1 queue, from (4.1) and (4.3), we have,

$$\mu_j = \sum_{i \in N} \lambda_i y_{ij} \exp(-\alpha d_{ij} - \beta p_j - \gamma w_j) + 1/w_j. \tag{4.9}$$

The right-hand-side of the above equation is a decreasing function of w_j, so there is one corresponding capacity for any waiting time. Thus if we can find the optimal waiting time, we will know the optimal capacity. We next use w_j as decision variables. The firm's profit function thus can be expressed as,

$$R(p, w) = \sum_{j \in \mathcal{S}} \big((p_j - C) \sum_{i \in N} \lambda_i y_{ij} \exp(-\alpha d_{ij} - \beta p_j - \gamma w_j) - C/w_j\big).$$

The first order conditions of the profit function w.r.t. p_j and w_j gives

$$p_j = \frac{1}{\beta} + C, \forall j \in S, \tag{4.10}$$

$$-\gamma(p_j - C) \sum_{i \in N} \lambda_i y_{ij} \exp(-\alpha d_{ij} - \beta p_j) \exp(-\gamma w_j) + C/w_j^2 = 0, \forall j \in S. \tag{4.11}$$

The latter can be rewritten as,

$$-\Big(\frac{\gamma C}{4(p_j - C)\sum_{i=1}^n \lambda_i y_{ij} \exp(-\alpha d_{ij} - \beta p_j)}\Big)^{\frac{1}{2}} = -\frac{1}{2}\gamma w_j^* \exp(-\frac{1}{2}\gamma w_j^*). \tag{4.12}$$

which is an LambertW function that has a real solution when $\frac{\gamma}{2} w_j < 1$. Thus the optimal waiting time is,

$$w_j^* = -\frac{2}{\gamma} LambertW\Big(-\Big(\frac{\gamma C}{4(p_j - C)\sum_{i=1}^n \lambda_i y_{ij} \exp(-\alpha d_{ij} - \beta p_j)}\Big)^{\frac{1}{2}}\Big). \tag{4.13}$$

Substitute the optimal price (4.10) into (4.13), then

$$w_j^* = \frac{2}{\gamma} L_j(y), \forall j \in S,$$

where

$$L_j(y) \doteq -LambertW\left(-\left(\frac{\beta\gamma C \exp(1+\beta C)}{4\sum_{i=1}^{n} \lambda_i y_{ij} \exp(-\alpha d_{ij})} \right)^{\frac{1}{2}} \right).$$

Substitute w_j^* into (4.9), we have

$$\mu_j^* = \frac{2\gamma L_j(y) + \beta\gamma C}{4L_j(y)^2}, \forall j \in S,$$
$$\lambda_j^* = \frac{\beta\gamma C}{4L_j(y)^2}, \forall j \in S.$$

Therefore, the optimal revenue is,

$$R^* = \sum_{j \in S} \frac{\gamma C(1 - 2L_j(y))}{4Lj(y)^2}.$$ □

4.6.2 Customer Equilibrium Flow

To maximize its profit, the service provider needs to consider how customers distribute their flows. If the locations of the facilities, the prices and the capacities are all known, customers will distribute their demands to maximize their utilities. The equilibrium demand at each facility locations depends on how the customers choose the facilities for service. In this section, we will show that the customer choice equilibrium can be solved by a convex optimization problem.

Theorem 4.2. *A flow $\boldsymbol{v}^*$ is in equilibrium if and only if it satisfies the following variational inequality problem:*

$$F(\boldsymbol{v}^*)^T(\boldsymbol{v} - \boldsymbol{v}^*) \geq 0, v \geq 0, \tag{4.14}$$

where $F(\boldsymbol{v}) = (F_{11}(\boldsymbol{v}) \cdots F_{|N|1}(\boldsymbol{v}), F_{12}(\boldsymbol{v}) \cdots F_{|N|2}(\boldsymbol{v}), \cdots, F_{1|S|}(\boldsymbol{v}) \cdots F_{|N||S|}(\boldsymbol{v}))^T$ *and*

$$F_{ij}(\boldsymbol{v}^*) = \alpha d_{ij} + \beta p_j + \frac{\gamma}{\mu_j - \sum_{i \in N} v_{ij}^*} + \ln \sum_{j \in M} v_{ij}^* - \ln \lambda_i, \forall i \in N, \forall j \in S. \tag{4.15}$$

Proof. We first show that if $\mathbf{v}^*$ is in equilibrium, i.e., $\mathbf{v}^*$ satisfies (4.5), then it also satisfies (4.14). Note from (4.5) for a fixed allocation pair (i, j),

$$(u_{ij}(\mathbf{v}^*) - u_i^*) * (v_{ij} - v_{ij}^*) \geq 0. \tag{4.16}$$

From **P2** (c), the demand flow conservation requires that

$$u_i^* = \ln \lambda_i - \ln \sum_{j \in S} v_{ij}^*. \tag{4.17}$$

We also have,

$$u_{ij}(\mathbf{v}^*) = \alpha d_{ij} + \beta p_j + \frac{\gamma}{\mu_j - \sum_{i \in N} v_{ij}^*}. \tag{4.18}$$

Substitute (4.17) and (4.18) into (4.16), we have

$$(\alpha d_{ij} + \beta p_j + \frac{\gamma}{\mu_j - \sum_{i \in N} v_{ij}^*} + \ln \sum_{j \in S} v_{ij}^* - \ln \lambda_i) * (v_{ij} - v_{ij}^*) \geq 0, \forall v_{ij} \geq 0, \forall i \in N, j \in S. \tag{4.19}$$

Summing over all demand nodes and facility locations, then

$$\sum_{i \in N} \sum_{j \in S} (\alpha d_{ij} + \beta p_j + \frac{\gamma}{\mu_j - \sum_{i \in N} v_{ij}^*} + \ln \sum_{j \in S} v_{ij}^* - \ln \lambda_i) * (v_{ij} - v_{ij}^*) \geq 0, \tag{4.20}$$

which in vector notation, gives (4.14).

We next show that if $\mathbf{v}^*$ satisfies (4.14) then it is in equilibrium, i.e., (4.5) is satisfied. Since (4.19) and (4.14) are equivalent, we use (4.19) here for ease of explanation. Consider pair (k, l), let $v_{ij} = v_{ij}^*, \forall ij \neq kl$ in (4.19), then (4.19) is simplified as:

$$(\alpha d_{kl} + \beta p_l + \frac{\gamma}{\mu_l - \sum_{k \in N} v_{kl}^*} + \ln \sum_{l \in S} v_{kl}^* - \ln \lambda_k) * (v_{kl} - v_{kl}^*) \geq 0, \forall k \in N, l \in S, \tag{4.21}$$

from which (4.5) follows and consequently for every (k, l) pair. □

Theorem 4.3. *The Jacobian matrix $\nabla \boldsymbol{F}(\boldsymbol{v})$ is symmetric and positive semi-definite.*

Proof. The Jacobian matrix can be represented as $\nabla \mathbf{F}(\mathbf{v}) = \mathbf{A} + \mathbf{B}$, where $\mathbf{A}$ and $\mathbf{B}$ are $|N||S| \times$

$|N||S|$ matrices in the form of

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{A}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{A}_{|S|} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{B}_1 & \cdots & \mathbf{B}_1 \\ \mathbf{B}_1 & \mathbf{B}_1 & \cdots & \mathbf{B}_1 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_1 & \mathbf{B}_1 & \cdots & \mathbf{B}_1 \end{pmatrix},$$

where $\mathbf{A}_j$ is a $|N| \times |N|$ matrix with all elements equal to $\frac{\gamma}{(\mu_j - \sum_{i \in N} v_{ij})^2}$, and $\mathbf{B}_1$ is a $|N| \times |N|$ diagonal matrix as

$$\mathbf{B}_1 = \begin{pmatrix} \frac{1}{\sum_{j \in S} v_{1j}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sum_{j \in S} v_{2j}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sum_{j \in S} v_{Nj}} \end{pmatrix}$$

For non-zero vector $\mathbf{z} \in R^{|N||S|}$, $\mathbf{z} = (z_{11} \cdots z_{|N|1}, z_{12} \cdots z_{|N|2}, \cdots\cdots, z_{1|S|} \cdots z_{|N||S|})$, we have

$$\begin{aligned} \mathbf{z}^T \mathbf{A} \mathbf{z} &= \sum_{j \in S} \frac{\gamma}{(\mu_j - \sum_{i \in N} v_{ij})^2} (\sum_{i \in N} z_{ij})^2 \geq 0, \\ \mathbf{z}^T \mathbf{B} \mathbf{z} &= \sum_{i \in N} \frac{1}{\sum_{j \in S} v_{ij}} (\sum_{j \in S} z_{ij})^2 \geq 0, \end{aligned}$$

which implies $\mathbf{A}$ and $\mathbf{B}$ are symmetric and positive semi-definite, and hence $\nabla \mathbf{F}(\mathbf{v})$ is symmetric and positive semi-definite. □

For a general variational inequality problem $VI(\mathbf{F}(\mathbf{v}), \mathcal{K})$, where $\mathcal{K}$ is a feasible set of v, if $\mathbf{F}(\mathbf{v})$ is continuously differentiable on $\mathcal{K}$ and the Jacobian matrix $\nabla \mathbf{F}(\mathbf{v})$ is symmetric and positive semi-definite, then there is a real valued convex function $f : \mathcal{K} \mapsto R$, $\nabla f(\mathbf{v}) = \mathbf{F}(\mathbf{v})$ with $\mathbf{v}^*$ the solution of $VI(\mathbf{F}(\mathbf{v}), \mathcal{K})$ is also the solution of the mathematical programming problem:

$$\begin{aligned} \min \quad & f(\mathbf{v}) \\ s.t. \quad & \mathbf{v} \in \mathcal{K}. \end{aligned}$$

Note that $\nabla \mathbf{F(v)}$ is symmetric and positive semi-definite. The variational inequality problem (4.14) thus can be formulated as the following convex optimization problem,

$$\begin{array}{lll} \min_{\mathbf{v}} & f(\mathbf{v}) & \\ s.t. & \sum_{i \in N} v_{ij} \leq \mu_j & \forall j \in S, \\ & \sum_{j \in S} v_{ij} \leq \lambda_i & \forall i \in N, \quad \textbf{(P5)} \\ & v_{ij} \geq 0 & \forall i \in N, j \in S, \end{array}$$

where

$$f(\mathbf{v}) = -\sum_{j \in S} \gamma \ln(\mu_j - \sum_{i \in N} v_{ij}) + \sum_{i \in N} \sum_{j \in S} (\alpha d_{ij} + \beta p_j - \ln \lambda_i - 1 + \ln \sum_{j \in S} v_{ij}) v_{ij}. \tag{4.22}$$

4.7 Solving the Problem

In this section, we develop several heuristic algorithms to solve our problem. Recall that the price decision is independent of location and capacity allocation decisions. Therefore, we focus on the latter two decisions. The problem is decomposed into two subproblems: location and capacity allocation. That is, we first construct a procedure to search for the set of facility locations, then we consider how much capacity to allocate to the facilities. The location and capacity allocation algorithms are solved iteratively to find the best solution.

4.7.1 Location Algorithms

Assuming the firm's profit can be evaluated (by the capacity allocation algorithms discussed in the next section), we discuss three algorithms to find the best set of locations: the Descent Algorithm, the Greedy Algorithm, and the Variable Neighborhood Search Algorithm. Note that these location algorithms have been well developed and widely used in location literature (see e.g., Berman and Huang (2007)).

Descent Algorithm

This algorithm starts from an arbitrary solution set S and search for the best "neighborhood" of S with a better set of locations. The procedure is repeated iteratively until no better neighborhood can be found. Let $R(S)$ be the firm's optimal revenue if the location set is S. Define the neighborhood of S by adding a node in N/S to S and removing another node from S. The cardinality of such a neighborhood is $|S|(|N| - |S|)$. The procedure of the Descent Algorithm is as follows,

Step 0 *Random locate S facilities on the network*

Step 1 *Check all the subsets S' in the neighborhood of S and calculate $R(S')$. Find the maximum of the revenues R_{max} and subset S'_{max}.*

Step 2 *If $R_{max} \geq R(S)$ change S to S'_{max} and go to step 2.*

Step 3 *Otherwise, the Descent algorithm terminates.*

Greedy Algorithm

This algorithm starts from an empty set of locations and add one facility at a time until $|S|$ facilities are reached. At each iteration, it chooses the facility whose addition cause the greatest improvement in revenue. The procedure is as follows,

Step 0 *$k = 0$, select an initial facility $j^k = j^0$ that gives the best revenue, $S^k = S^0 = \{j^0\}$.*

Step 1 *Check all the facilities $j \in N/S^{k-1}$ and calculate $R(S^{k-1} \cup j)$. Find the maximum of these values R_{max} and j^k_{max} and let $S^k = S^{k-1} \cup j^k_{max}$.*

Step 2 *$k = k + 1$, go to step 2, terminate the algorithm if k=—S—.*

Variable Neighborhood Search Heuristic

Variable Neighborhood Search (VNS) is a metaheuristic that guides a local heuristic search procedure to explore the solution space beyond local optimality (Mladenovic and Hansen, 1997). It systematically changes neighborhoods within a local search algorithm. First, a function

$\rho(S_1, S_2)$ is defined as the number of different locations in S_1 and S_2, where $|S_1| = |S_2|$. For example, if $S_1 = 1, 2, 3$ and $S_2 = 2, 3, 4$, then $\rho(S_1, S_2) = 1$. For every $k = 1, 2 \cdots, M$, define the kth neighborhood of S as

$$N_k(S) = \{Y : \rho(S, Y) = k\}.$$

Comparing to local search algorithms, which are usually 1-opt procedures, VNS does not stop at the $N_1(S)$. Suppose S is the current solution, VNS randomly chooses a solution, say S' from $N_1(S)$ and runs a local search (e.g., descent algorithm) in $N_1(S')$. If there is no improvement, VNS randomly looks for another starting point from $N_2(S)$ and runs a new local search, and so on. If VNS finds an improved solution, this solution will be used as the new current solution. A new iteration will start until some stopping condition is met. The procedure can be stated as follows:

Step 0 *Find an initial solution S. Set the current best solution $S^* = S$ and $k = 1$.*

Step 1 *Randomly choose a solution S' from $N_k(S^*)$.*

Step 2 *Call a local search algorithm based on S'.*

Step 3 *If $k \leq M$ and the returned solution from step 2, S'' is not better than S^*, set $k = k+1$ and go back to step 1 unless the iteration limit is reached. If $k = |S|$ and the returned solution of step 2, S'' is not better than S^*, set $k = 1$ and go back to step 1 unless the iteration limit is reached. Otherwise set $S^* = S''$ and $k = 1$ and go back to step 1.*

From the procedures of the three algorithms, we can see that the Greedy algorithm needs less computation effort than the Descent Algorithm. Variable Neighborhood Search Algorithms has an advantage of finding global optimal locations. We will compare the performance of these three algorithms later in the numerical experiments.

4.7.2 Capacity Allocation Algorithms

In this section, we present two algorithms to solve the lower level capacity allocation problem when the locations of the facilities are known. Recall that customers choose the facilities for

service to maximum their utilities. A customer may not choose the closest facility for service, but choose a facility with longer travel distance to avoid congestion. Customers at the same node may split their flows to different facilities as long as they can obtain the same utility. We use the properties of the optimal solutions in Section 4.6 to develop the algorithms: User Equilibrium Heuristic 1 (UEH1)and User Equilibrium Heuristic 2 (UEH2). UEH1 attempts to find the optimal capacity by considering customers' equilibrium behavior iteratively. UEH2 assumes that customers visit the closest assignment and allocate capacity according to this assignment.

User Equilibrium Heuristic 1

Use one of the location algorithms, for any fixed location S, let $p_j = 1/\beta + C, \forall j \in S$, run the following algorithms:

Step 0 *Initialize: Assume closest visit. Let* N_j *be the set of nodes that visit facility* j*. Set* $k = 0$ *and let*

$$\mu_j^0 = \frac{2\gamma L_j^0(x) + \beta\gamma C}{4L_j^0(x)^2}, \forall j \in S,$$

where

$$L_j^0(x) = -LambertW\bigg(-\frac{1}{2}\bigg(\frac{\beta\gamma C\exp(1+\beta C)}{\sum_{i\in N_j}\lambda_i\exp(-\alpha d_{ij})}\bigg)^{\frac{1}{2}}\bigg), \forall j \in S.$$

Step 1 *Given* μ^k*. Calculate the resulting equilibrium flow* v^k *by P*5

$$\min_{\mathbf{v}} -\sum_{j\in S}\gamma\ln(\mu_j^k - \sum_{i\in N} v_{ij}) + \sum_{i\in N}\sum_{j\in S}(\alpha d_{ij} + C - \ln\lambda_i + \ln(\sum_{j\in S} v_{ij}))v_{ij}$$

$$\begin{aligned} s.t.\quad & \textstyle\sum_{i\in N} v_{ij} \le \mu_j, \forall j \in S, \\ & \textstyle\sum_{j\in S} v_{ij} \le \lambda_i, \forall i \in N, \\ & v_{ij} \ge 0, \forall i \in N, j \in S. \end{aligned}$$

Evaluate waiting time and the original demand distribution,

$$w_j^k = \frac{1}{u_j^k - \sum_{i \in N} v_{ij}^k},$$
$$y_{ij}^k = \frac{v_{ij}^k}{\lambda_i \exp(-\alpha d_{ij} - \beta p^* - \gamma w_j^k)}.$$

Step 2 Given v_{ij}^k, w_j^k, calculate

$$\mu_j^{k+1} = \frac{2\gamma L_j^{k+1}(S) + \beta\gamma C}{4L_j^{k+1}(S)^2}, \forall j \in S,$$

where

$$L_j^{k+1}(S) = -LambertW\bigg(-\frac{1}{2}\bigg(\frac{\beta\gamma C \exp(1+\beta C)}{\sum_{i \in N_j} y_{ij}^k \lambda_i \exp(-\alpha d_{ij})}\bigg)^{\frac{1}{2}}\bigg), \forall j \in S.$$

Step 3 *Repeat Step 1 and Step 2. Stop if* $\|\mu^{k+1} - \mu^k\| \leq \epsilon$.

User Equilibrium Heuristic 2

Use one of the location algorithms, for any fixed location S, let $p_j = 1/\beta + C, \forall j \in S$, run the following algorithms:

Step 0 *Initialize: Assume closest visit. Let* N_j *be the set of nodes that visit facility* j*. Set* $k = 0$ *and let*

$$\mu_j^* = \frac{2\gamma L_j^*(S) + \beta\gamma C}{4L_j^*(S)^2}, \forall j \in S,$$

where

$$L_j^*(S) = -LambertW\bigg(-\frac{1}{2}\bigg(\frac{\beta\gamma C \exp(1+\beta C)}{\sum_{i \in N_j} \lambda_i \exp(-\alpha d_{ij})}\bigg)^{\frac{1}{2}}\bigg), \forall j \in S.$$

Step 1 *Given* μ^**. Calculate the resulting equilibrium flow* $\boldsymbol{v}^*$ *by P*5*, then stop.*

4.7.3 An Uncapacitated Facility Location Problem (UFLP) Formulation

Uncapacitated Facility Location Problem (UFLP) is one of the most commonly used location models. In a classical UFLP, facilities are placed among M possible sites with the objective of minimizing the total travel distance for satisfying all demand at N given locations (see e.g., Mirchandani and Francis (1990)). In this section we present an approach using the UFLP formulation. We assume that locations are made first through the UFLP formulation and the capacity allocation decisions are made after the locations of the facility are determined. The UFLP formulation of our location problem can be stated as,

$$\begin{aligned} \max_{x,y} \quad & \sum_{i\in N}\sum_{j\in M} p^* * y_{ij} \exp(-\alpha d_{ij} - \beta p^*) \\ s.t. \quad & \sum_{j\in M} x_j = |S|, \qquad\qquad (\mathbf{P6}) \\ & \sum_{j\in M} y_{ij} = 1 \qquad \forall i \in N, \\ & y_{ij} \le x_j \qquad \forall i \in N, j \in M, \\ & x_j, y_{ij} = 0, 1 \qquad \forall i \in N, j \in M. \end{aligned}$$

Then, we can use the capacity allocation algorithms UEH1 and UEH2 in Section 4.7.2 to obtain the optimal capacities. We denote the algorithms as UFLP1 (UFLP+UEH1) and UFLP2 (UFLP+UEH2) respectively.

Note that it can be verified that **P6** provides an integer solution and customers are assigned to the closest facility.

4.7.4 An Upper Bound and Lower Bound of Customer Flows

We present an upper bound and a lower bound of customer flows by considering two extreme conditions. If there are sufficient supply of capacities such that there is no need to wait for service, i.e., customers' waiting time is zero, the maximal total demand would be,

$$\Lambda_{UB} = \sum_{j\in M} \lambda_i \exp(-\alpha d_{ij} - \beta p_j). \tag{4.23}$$

On the other hand, if the congestion is so high such that the systems cannot be in equilibrium, the minimal total demand would be,

$$\Lambda_{LB} = \sum_{j \in M} \lambda_i \exp(-\alpha d_{ij} - \beta p_j - 2), \tag{4.24}$$

where $\gamma * w > 2$ as in (4.12).

4.8 An Example: Is Visiting the Closest Facility Optimal?

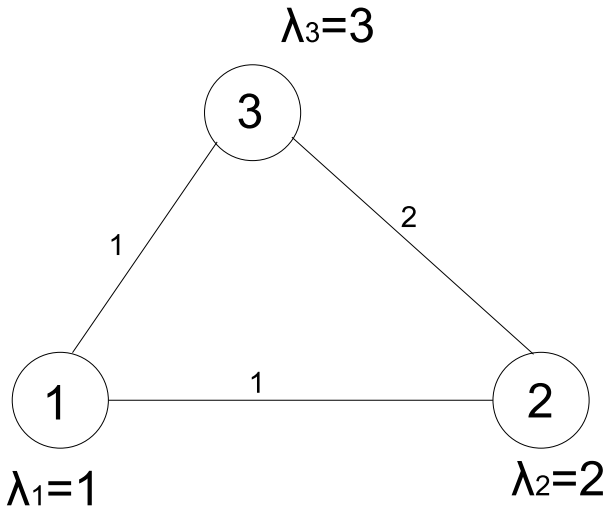
We assume that customers' behavior when visiting their closest facilities for service is quite reasonable and practical when there is no congestion and an uniform price is used. For example, the UFLP formulation in **P6** gives an optimal solution that customers visit their closest facilities. When there is congestion in the system, however, the answer is not immediately obvious. Customers' choice of where to go for service is no longer independent due to the presence of congestions. In this section, we use an example to gain the insight and show that visiting the closest facility actually is not necessarily optimal for both the System Optimization model and the User Equilibrium model.

Consider a firm operating on a 3-node network as shown in Figure 4.1. The travel distances between nodes are $d_{12} = 1$, $d_{31} = 1$, and $d_{32} = 2$. The maximum demand rates at each node are $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$. The demands'elasticity to price and congestion time is $\beta = 0.2$ and $\gamma = 1$ respectively. The demands' elasticity to travel distance varies for analytical purpose. The unit capacity cost is $C = 0.5$. We assume that the firm has decided to locate one facility at node 1 and another one at node 2. Let $y_{31}, 1 - y_{31}$ be the proportion of flows at node 3 that go to node 1 and node 2 respectively. We next find optimal flows y_{31}, and $1 - y_{31}$ and show its impact on the firms's revenue.

The optimal price is 5.5. Given y_{31}, and $1 - y_{31}$, the firm's optimal capacity and revenue, for both System Optimization and User Optimization, are as follows,

$$\mu_j^* = \frac{2L_j(y) + 0.1}{4L_j(y)^2}, \qquad j = 1, 2,$$

Figure 4.1: A 3-node network - multiple facility



and

$$R^* = \sum_{j=1}^{2} \frac{0.5(1 - 2L_j(y))}{4L_j(y)^2},$$

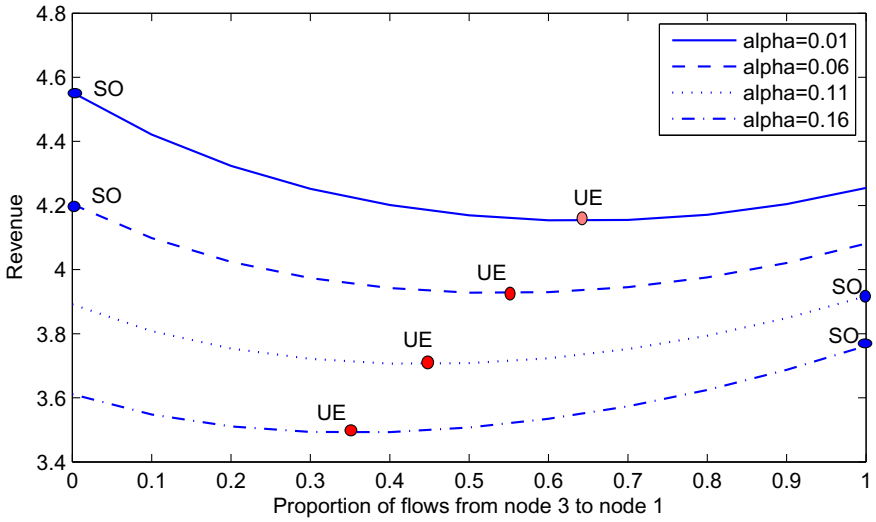
where

$$\begin{aligned} L_1^*(y) &= -LambertW\left(-\frac{1}{2}\left(\frac{0.1\exp(1.1)}{\lambda_1 + \lambda_3 y_{31}\exp(-\alpha)}\right)^{\frac{1}{2}}\right), \\ L_2^*(y) &= -LambertW\left(-\frac{1}{2}\left(\frac{0.1\exp(1.1)}{\lambda_2 + \lambda_3(1 - y_{31})\exp(-2\alpha)}\right)^{\frac{1}{2}}\right). \end{aligned}$$

System Optimization We plot the firm's revenue with respect to flow y_{31} in Figure 4.2 under various travel distance elasticities. We use SO to denote the system optimal solutions in the figure. We can see from Figure 4.2 that $y_{31}^* = 0$ at $\alpha = 0.01$ and 0.06, and $y_{31}^* = 1$ at $\alpha = 0.11$ and 0.16. Obviously, assigning customers to the closest facility is not necessarily optimal. Depending upon customers' elasticity to the travel distance and waiting time, assigning customers at node 3 to their closest facility at node 1 provides higher revenue when the elasticity to demand is relatively large, ($\alpha = 0.11, 0.16$). When the waiting time factor dominates, i.e.

customers are more sensitive to the waiting time than travel distance ($\alpha = 0.01, 0.06$), pooling customers together to the higher demand node 2 provides higher profit.

Figure 4.2: Profit vs. flow distribution



User Equilibrium When customers have the flexibility to choose where to go for service, they will try to maximize their profit. In this case, all customers at node 3, no matter which facility to visit, have the same utility in equilibrium. Therefore,

$$\alpha * d_{31} + \gamma w_1^* = \alpha * d_{32} + \gamma w_2^*,$$

where

$$w_j^* = \frac{2}{\gamma} L_j^*(y_{31}).$$

Solving the above equation at $\alpha = 0.01, 0.06, 0.11$ and 0.16, we have $y_{31} = 0.6455, 0.5406, 0.4414$ and 0.3520 respectively (shown in Figure 4.2 as UEs).

As expected, We can see that the revenues of the user equilibrium solutions are lower than the revenues from the the system optimal solutions. It is also interesting to see that in the User Equilibrium model, customers split their flows to node 1 and node 2.

Note that the user equilibrium flow is not unique given fixed locations and capacities. In this example, when $\alpha = 0$, any flow that results in the same total flow at node 1 and node 2 is a user equilibrium flow. For example ($y_{31} = 2, y_{32} = 1$, and $y_{12} = y_{21} = 0$) and

($y_{31} = 1, y_{32} = 2, y_{12} = 0$, and $y_{21} = 1$) will both be user optimal if we allocate the same capacity at the two nodes.

4.9 Computational Experiments

To test the heuristic algorithms proposed and investigate their managerial implications, we conduct extensive experiments under different demand elasticity scenarios and problems sizes. The number of nodes was set to 10, 20, 30, 50, 200, and 300. The number of facilities was set to 2, 3, 4, 5, 8, and 10. All runs are performed on a Pentium 4 PC equipped with 1GHZ processor and 1GB RAM.

All procedures were coded in C++ and MATLAB. The network used in the experiments were generated randomly. The length of each link was generated over the interval $(0, 1)$ uniformly. All demand weights were generated randomly over the interval $(10, 50)$. For all problem instances, we ensured that no two instances shared a common random seed.

Two set of experiments are conducted: capacity allocation with fixed location problems and location and capacity allocation problems. In the instances of capacity allocation problems, we use KNITRO solver as a benchmark to evaluate the quality of our heuristic solutions. KNITRO is one of the commercial softwares that are capable of solving nonlinear complimentary optimization problems.

The following two performance measures were used in the experiments.

1. Relative error (RE): $(R^* - R)/R^*$, where R is the objective function value generated by the heuristics and R^* is the optimal objective value from KNITRO.

2. Heuristic gap (HG): $(R_{max} - R)/R_{max}$, where R_{max} is the best solution obtained among the heuristics.

4.9.1 Capacity Allocation with Fixed Locations

Tables 4.1 to 4.3 show the REs from the capacity allocation algorithms with the locations of the facilities being fixed, under different network sizes and customer elasticities. Two set of customer elasticities are used: $(\alpha = 0.1, \gamma = 0.3)$ is for the cases that customers' sensitivities

to travel distance and waiting time is similar; and ($\alpha = 0.001, \gamma = 0.3$) is for the cases that customers are extremely sensitive to waiting time. The optimal solution of KNITRO is obtained by starting from the best solution obtained from UEH1 and UEH2.

We have the following observations from our computational experiments:

1. When customers are not very sensitive to waiting time, or are more sensitive to travel distances than waiting times, both UEH1 and UEH2 provide good solutions that are close to the results from KNITRO.

2. When customers are very sensitive to waiting time, UEH1 provides a better solution than UEH2. For example, in Table 4.1(b), the REs of UEH1 are all quite small, but the REs of UEH2 range from 0.1279 to 0.4572.

3. When the number of facilities on the network increases, REs of UEH2 tend to increase.

4. The running times of UEH1 and UEH2 is much shorter than that of KNITRO. However, the running time of UEH1 is longer than UEH2.

These observations provide important managerial insights. Recall that UEH1 finds the optimal capacity by iteratively taking into account customers equilibrium behaviors, and UEH2 finds the optimal capacity by assuming that customers visit closest facility for service. When customers are more sensitive to travel distance than waiting time, they tend to choose the closest facility for service. Thus UEH2, which is based on closest assignment, provides solutions as good as UEH1. However, when customers are highly sensitive to waiting time, assuming that they visit the closest facilities is apparently not sufficient to reflect their true behaviors. UEH1 thus performs much better than UEH2.

We also observe that KNITRO gives good solutions within a reasonable time when the locations of the facilities are fixed and the network is small ($|N| \leq 50$). When the network is relatively large ($|N| \geq 100$), the running time of KNITRO increases greatly, to the extend of 2000 seconds by our experiences.

Table 4.1: Capacity allocation with fixed locations for $|N| = 10, 20, 30$

$\|N\|$	$\|S\|$	RE^{UEH1}	$Time^{UEH1}$	RE^{UEH2}	$Time^{UEH2}$	$Time^{KNITRO}$
10	2	0.0000	0.01	0.0000	0.02	1.03
10	3	0.0018	0.03	0.0021	0.06	0.17
10	4	0.0009	0.01	0.0000	0.02	0.23
10	5	0.0031	0.73	0.0123	0.17	0.32
20	2	0.0000	0.02	0.0000	0.07	0.20
20	3	0.0006	0.06	0.0007	0.08	0.28
20	4	0.0001	0.31	0.0001	0.18	0.81
20	5	0.0005	0.10	0.0007	0.09	1.19
30	2	0.0002	0.76	0.0006	0.23	1.05
30	3	0.0000	0.03	0.0000	0.05	0.56
30	4	0.0000	1.75	0.0004	0.20	1.03
30	5	0.0001	0.29	0.0001	0.20	1.76

(a) $\alpha = 0.1, \gamma = 0.3$;

$\|N\|$	$\|S\|$	RE^{UEH1}	$Time^{UEH1}$	RE^{UEH2}	$Time^{UEH2}$	$Time^{KNITRO}$
10	2	0.0066	0.38	0.1279	0.00	0.37
10	3	0.0091	1.05	0.2086	0.02	0.88
10	4	0.0099	0.64	0.3137	0.03	0.05
10	5	0.0099	0.20	0.3423	0.02	0.04
20	2	0.0013	0.11	0.1133	0.03	1.74
20	3	0.0060	1.05	0.2449	0.03	0.97
20	4	0.0006	2.92	0.3199	0.05	1.08
20	5	0.0001	59.77	0.4447	0.05	0.66
30	2	0.0006	1.48	0.1057	0.02	1.78
30	3	0.0006	1.42	0.2193	0.05	1.53
30	4	0.0196	2.73	0.3347	0.05	0.67
30	5	0.0008	7.19	0.4572	0.06	0.58

(b) $\alpha = 0.001, \gamma = 0.3$.

Table 4.2: Capacity allocation with fixed locations for $|N| = 50, 80, 100$

$\|N\|$	$\|S\|$	RE^{UEH1}	$Time^{UEH1}$	RE^{UEH2}	$Time^{UEH2}$	$Time^{KNITRO}$
50	2	0.0000	0.07	0.0000	0.14	4.18
50	3	0.0000	0.38	0.0000	0.20	4.03
50	4	0.0000	3.37	0.0007	0.30	4.18
50	5	0.0000	5.66	0.0001	0.29	6.04
50	8	0.0000	4.84	0.0003	0.53	9.54
50	10	0.0001	1.80	0.0000	0.67	44.56
80	2	0.0000	2.21	0.0001	0.22	3.48
80	3	0.0000	2.04	0.0003	0.31	8.18
80	4	0.0000	4.22	0.0003	0.36	11.99
80	5	0.0000	6.83	0.0008	0.44	19.46
80	8	0.0000	6.62	0.0002	0.74	51.24
80	10	0.0008	27.47	0.0023	1.07	117.70
100	2	0.0000	0.11	0.0000	0.09	4.94
100	3	0.0000	4.62	0.0001	0.37	17.95
100	4	0.0000	2.06	0.0001	0.45	44.28
100	5	0.0001	4.30	0.0006	0.57	36.02
100	8	0.0001	10.51	0.0004	0.89	115.82
100	10	0.0000	31.47	0.0009	1.42	254.77

(a) $\alpha = 0.1, \gamma = 0.3$;

$\|N\|$	$\|S\|$	RE^{UEH1}	$Time^{UEH1}$	RE^{UEH2}	$Time^{UEH2}$	$Time^{KNITRO}$
50	2	0.0003	0.27	0.1280	0.05	4.94
50	3	0.0001	0.30	0.2295	0.06	4.27
50	4	0.0006	3.52	0.3426	0.09	4.60
50	5	0.0007	0.58	0.4756	0.11	6.16
50	8	0.0003	19.23	0.8223	0.30	11.26
50	10	0.0002	2.31	0.9245	0.38	50.80
80	2	0.0006	0.39	0.1397	0.08	3.55
80	3	0.0003	0.45	0.2177	0.09	9.49
80	4	0.0001	0.72	0.3451	0.14	13.19
80	5	0.0002	0.89	0.3747	0.17	22.58
80	8	0.0006	6.34	0.4068	0.49	52.26
80	10	0.0000	3.05	0.4088	0.59	138.88
100	2	0.0005	0.39	0.1328	0.08	5.73
100	3	0.0010	0.69	0.2187	0.13	21.18
100	4	0.0002	0.72	0.3259	0.17	48.71
100	5	0.0008	1.11	0.3643	0.20	36.74
100	8	0.0010	15.75	0.4093	0.59	134.35
100	10	0.0006	6.84	0.4267	0.73	280.25

(b) $\alpha = 0.001, \gamma = 0.3$.

Table 4.3: Capacity allocation with fixed locations for $|N| = 200, 300$

$\|N\|$	$\|S\|$	RE^{UEH1}	$Time^{UEH1}$	RE^{UEH2}	$Time^{UEH2}$	$Time^{KNITRO}$
200	2	0.0000	1.27	0.0000	0.34	94.25
200	3	0.0000	1.78	0.0000	0.44	121.41
200	4	0.0001	8.70	0.0002	0.59	407.52
200	5	0.0000	12.81	0.0001	0.72	1483.50
200	8	0.0001	10.14	0.0002	1.22	400.37
200	10	0.0005	13.66	0.0000	1.45	40.28
300	2	0.0000	1.53	0.0000	0.45	276.41
300	3	0.0002	3.25	0.0001	0.61	564.12
300	4	0.0002	7.36	0.0002	0.80	281.02
300	5	0.0001	7.84	0.0000	0.98	69.24
300	8	0.0003	36.80	0.0000	1.66	119.97
300	10	0.0003	48.44	0.0000	2.20	168.52

(a) $\alpha = 0.1, \gamma = 0.3$;

$\|N\|$	$\|S\|$	RE^{UEH1}	$Time^{UEH1}$	RE^{UEH2}	$Time^{UEH2}$	$Time^{KNITRO}$
200	2	0.0007	0.77	0.1191	0.17	111.22
200	3	0.0003	1.11	0.2581	0.27	143.26
200	4	0.0010	1.45	0.3405	0.34	431.97
200	5	0.0001	1.78	0.3688	0.42	1602.18
200	8	0.0002	9.81	0.3948	1.17	456.42
200	10	0.0001	7.58	0.4062	1.47	43.50
300	2	0.0008	1.27	0.1375	0.31	298.52
300	3	0.0005	1.81	0.2521	0.45	597.97
300	4	0.0009	2.36	0.2537	0.58	309.12
300	5	0.0001	2.99	0.3686	0.72	70.62
300	8	0.0003	20.61	0.4890	2.00	134.37
300	10	0.0001	71.70	0.5513	2.50	171.89

(b) $\alpha = 0.001, \gamma = 0.3$.

4.9.2 Location and Capacity Allocation

In this section, we show the numerical works of finding the joint optimal locations and capacity allocations by using the algorithms proposed (recall that price decision is independent of the location and capacity decisions).

Tables 4.4, 4.5, and 4.6 show the results for problems under various network sizes and customer elasticities. Table 4.4 is for some small networks. The REs are obtained by comparing the solutions of our algorithms to the solution obtained from KNITRO. Since KNITRO does not work for the discrete variables, we enumerate all sets of locations to obtain the optimal revenue. Tables 4.5 and 4.6 are for medium to large networks. We use HG as a measurement of the

solution quality in these cases, as enumerating locations from KNITRO is practically impossible. We test two demand elasticity scenarios for all cases: $(\alpha = 0.1, \beta = 0.3)$, where customers are sensitive to travel distances and waiting time in approximately the same magnitude, and $(\alpha = 0.001, \beta = 0.3)$, where customers are more sensitive to waiting time than travel distances.

We observe that in all cases, DA+UEH2, GA+UEH2, UFLP1, and UFLP2 run faster than DA+UEH1, GA+UEH1, and VNS. For a network of 30 nodes and 5 facilities, the completion takes just a few seconds for DA+UEH2, GA+UEH2, UFLP1, and UFLP2. It takes a few minutes for DA+UEH1, GA+UEH1, and VNS. The computation time increases greatly for DA+UEH1 and VNS algorithm as the network size increases. For example, in Table 4.6 (a), for a network of 200 nodes and 8 facilities, it takes DA+UEH1 and VNS over 8 hours to complete. The time to complete DA+UEH2 and GA+UEH2 is within an hour.

In addition to the running time, we also have the following observation about the solution quality of these algorithms:

1. The location algorithms combined with UEH1 (DA+UEH1, GA+UEH1) generally provide better solutions than the location algorithms combined with UEH2 (DA+UEH2, GA+UEH2). This patterns can be found in all the tables.

2. For cases with $(\alpha = 0.1, \beta = 0.3)$, all algorithms provide good solutions, though some of them take a longer time than others.

3. For cases with $(\alpha = 0.001, \beta = 0.3)$, the algorithms using UEH1 generate much better solutions than the algorithms using UEH2. For example, in Table 4.5 (b), UGs obtained from DA+UEH2 and GA+UEH2 range from 1% to 12%, while UGs are less than 0.1% in Table 4.5 (a).

4. UFLP1 and UFLP2 give solutions as good as all other algorithms when $(\alpha = 0.1, \beta = 0.3)$. However, for cases with $(\alpha = 0.001, \beta = 0.3)$, even UFLP1 does not provide good solutions for some cases, for example in Table 4.4 (b), for cases $(|N| = 10, |S| = 3)$, $(|N| = 20, |S| = 4)$, and $(|N| = 30, |S| = 5)$, the UGs of UFLP1 are 24%, 16% and 14% respectively. As expected, UFLP2 does not provide good solution for all cases with $(\alpha = 0.001, \beta = 0.3)$.

These observations provide us important guidance in how to decide the price, location and capacity decisions when designing a service network. Our algorithms can be classified into two categories: (1) DA+UEH1, GA+UEH1, VNS make combined location and capacity allocation decisions with consideration of customers' reaction; and (2) DA+UEH2 and GA+UEH2 though jointly optimize location and capacity, ignore customers' reaction when allocating capacities.

UFLP1 considers customer reaction when allocation capacities, however the location decisions was made by assuming that customers visit the closest facility. UFLP2 determines the location and capacity separately and assume that customers visit closest capacities for both decisions.

Therefore, in a network where customers are not very sensitive to the waiting time, the assumption of visiting the closest facility seems reasonable, so all the algorithms produce good quality solutions. However, when customers are very sensitive to the waiting time, they may redistribute their flows across facilities to avoid congestions. In this case the algorithms based on closest assignment is not sufficient to obtain a good solution.

Table 4.4: Location and capacity allocation for $|N| = 10, 20, 30$

		$DA+UEH1$		$DA+UEH2$		$GA+UEH1$		$GA+UEH2$		$UFLP1$		$UFLP2$			VNS
$\|N\|$	$\|S\|$	UG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$
10	2	0.0000	0.19	0.0000	0.08	0.0000	17.75	0.0000	0.65	0.0000	2.80	0.0000	1.64	0.0000	1.45
10	3	0.0016	23.32	0.0090	0.34	0.0016	50.08	0.0016	0.33	0.0000	0.75	0.0000	0.16	0.0016	49.35
10	4	0.0000	66.65	0.0020	0.52	0.0230	97.51	0.0230	0.37	0.0001	0.46	0.0020	0.11	0.0000	40990.40
10	5	0.0000	77.04	0.0123	0.77	0.1100	129.65	0.1100	0.62	0.0000	0.67	0.0166	0.11	0.0000	219.77
20	2	0.0000	5.42	0.0000	0.68	0.0000	26.35	0.0000	0.56	0.0000	0.85	0.0000	0.36	0.0000	19.40
20	3	0.0058	104.76	0.0000	1.31	0.0000	182.55	0.0000	1.26	0.0000	1.78	0.0000	0.82	0.0000	347.44
20	4	0.0000	1526.16	0.0113	2.02	0.0000	383.14	0.0000	1.83	0.0000	2.31	0.0000	1.07	0.0000	819.31
20	5	0.0000	1519.29	0.0000	4.49	0.0168	504.82	0.0168	4.07	0.0000	0.96	0.0000	0.27	0.0000	2921.67
30	2	0.0000	1.41	0.0000	1.30	0.0000	113.97	0.0000	1.16	0.0000	2.35	0.0000	1.12	0.0000	52.25
30	3	0.0000	640.56	0.0073	3.20	0.0073	236.48	0.0073	2.86	0.0000	46.23	0.0000	16.68	0.0000	890.38
30	4	0.0000	1616.71	0.0180	8.74	0.0177	512.58	0.0177	7.97	0.0001	34.58	0.0001	13.47	0.0000	3924.21
30	5	0.0082	4980.12	0.0006	95.02	0.0006	951.01	0.0006	7.61	0.0002	71.25	0.0004	29.61	0.0001	8998.37

(a) $\alpha = 0.1, \gamma = 0.3$;

		$DA+UEH1$		$DA+UEH2$		$GA+UEH1$		$GA+UEH2$		$UFLP1$		$UFLP2$			VNS
$\|N\|$	$\|S\|$	UG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$
10	2	0.0000	3.49	0.1247	1.08	0.0000	0.72	0.1247	0.29	0.0000	0.08	0.1250	0.02	0.0018	29.26
10	3	0.0000	9.57	0.1324	1.84	0.0000	1.61	0.0960	0.53	0.2409	0.09	0.1913	0.04	0.0078	52.85
10	4	0.0024	9.14	0.2281	3.95	0.0000	4.56	0.2257	1.17	0.0024	0.28	0.3451	0.05	0.0024	80.78
10	5	0.0000	12.15	0.3263	3.33	0.0000	3.91	0.3205	1.42	0.0000	0.20	0.3813	0.06	0.0000	108.72
20	2	0.0035	20.53	0.0501	3.90	0.0182	2.82	0.0501	1.19	0.0241	0.21	0.0536	0.05	0.0000	126.01
20	3	0.0000	40.29	0.0963	11.51	0.0000	8.74	0.0963	2.58	0.0081	0.49	0.1595	0.10	0.0029	243.98
20	4	0.0000	127.92	0.0888	18.77	0.0000	38.64	0.1351	4.30	0.1636	0.59	0.2029	0.09	0.0005	529.07
20	5	0.0000	237.28	0.1817	35.04	0.0000	26.97	0.1864	6.55	0.0102	1.71	0.2812	0.12	0.0028	741.20
30	2	0.0000	56.18	0.0270	13.13	0.0000	11.94	0.0270	2.45	0.0008	0.8	0.0270	0.07	0.0080	273.63
30	3	0.0000	228.83	0.0611	25.77	0.0000	53.43	0.0544	5.48	0.0290	1.95	0.0855	0.11	0.0000	650.07
30	4	0.0032	425.03	0.1005	56.56	0.0000	64.65	0.1076	9.23	0.0032	1.36	0.1323	0.14	0.0023	1245.97
30	5	0.0053	538.96	0.1269	89.34	0.0000	107.25	0.1253	13.71	0.1426	2.47	0.1894	0.19	0.0098	1751.74

(b) $\alpha = 0.001, \gamma = 0.3$.

Table 4.5: Location and capacity allocation for $|N| = 50$

		$DA+UEH1$		$DA+UEH2$		$GA+UEH1$		$GA+UEH2$		$UFLP1$		$UFLP2$		VNS	
$\|N\|$	$\|S\|$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$
50	2	0.0001	9.83	0.0001	1.47	0.0023	20.99	0.0023	1.20	0.0000	657.06	0.0000	537.82	0.0001	7200.50
50	3	0.0003	470.39	0.0003	4.52	0.0003	98.03	0.0003	2.95	0.0000	793.23	0.0000	611.06	0.0000	7200.36
50	4	0.0000	1156.25	0.0001	8.09	0.0000	28.72	0.0001	4.41	0.0000	797.70	0.0000	696.82	0.0000	7201.34
50	5	0.0036	4342.07	0.0039	23.01	0.0009	361.33	0.0015	7.50	0.0000	1459.12	0.0000	1017.59	0.0003	7204.17
50	8	0.0000	9947.34	0.0000	50.67	0.0089	969.29	0.0089	14.06	0.0000	438.59	0.0000	131.24	0.0000	7208.98
50	10	0.0035	6371.45	0.0043	63.38	0.0033	914.95	0.0044	20.56	0.0000	1235.4	0.0000	3.74	0.0000	7232.75

(a) $\alpha = 0.1, \gamma = 0.3$;

		$DA+UEH1$		$DA+UEH2$		$GA+UEH1$		$GA+UEH2$		$UFLP1$		$UFLP2$		VNS	
$\|N\|$	$\|S\|$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$
50	2	0.0005	270.91	0.0154	14.45	0.0000	126.34	0.0116	2.53	0.0126	485.03	0.0271	797.38	0.0028	322.08
50	3	0.0000	354.91	0.0282	40.63	0.0000	79.58	0.0266	5.78	0.0040	0.80	0.0385	585.13	0.0015	819.49
50	4	0.0000	424.22	0.0608	51.94	0.0000	94.55	0.0601	9.92	0.0000	0.58	0.0834	1070.20	0.0134	1257.22
50	5	0.0018	825.66	0.1031	180.33	0.0000	153.63	0.0944	15.11	0.0018	2.20	0.1308	133.53	0.0586	2526.97
50	8	0.0000	3740.24	0.1203	416.64	0.0080	699.76	0.1212	36.69	0.1027	3.23	0.1864	976.21	0.0812	7961.20
50	10	0.0000	8927.73	0.1035	102.36	0.0069	1003.34	0.0782	38.34	0.1112	0.98	0.1806	780.68	0.0960	8965.89

(b) $\alpha = 0.001, \gamma = 0.3$.

Table 4.6: Location and capacity allocation for $|N| = 200, 300$

		$DA+UEH1$		$DA+UEH2$		$GA+UEH1$		$GA+UEH2$		VNS	
$\|N\|$	$\|S\|$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	RE	$Time$
200	2	0.0092	147.06	0.0051	6.28	0.0092	148.52	0.0092	32.00	0.0000	7201.43
200	3	0.0008	438.03	0.0008	31.20	0.0020	356.09	0.0021	65.22	0.0000	10822.80
200	4	0.0000	2962.72	0.0001	121.38	0.0010	664.95	0.0010	103.17	0.0000	14422.20
200	5	0.0011	18816.30	0.0011	268.34	0.0000	1206.52	0.0000	143.73	0.0011	18125.90
200	8	0.0000	68197.60	0.0001	948.84	0.0021	3769.50	0.0023	306.59	0.0015	29062.60
200	10	0.0000	201564.00	0.0003	1341.94	0.0034	5117.44	0.0035	430.06	0.0040	36281.20
300	2	0.0000	88.52	0.0000	19.61	0.0000	381.81	0.0000	76.75	0.0000	7201.92
300	3	0.0000	966.16	0.0000	63.64	0.0117	892.50	0.0118	157.59	0.0000	10851.20
300	5	0.0000	13344.60	0.0001	326.39	0.0041	1631.52	0.0042	254.50	0.0000	14480.00
300	8	0.0000	28640.70	0.0000	648.02	0.0071	2556.78	0.0072	357.55	0.0038	18244.60

(a) $\alpha = 0.1, \gamma = 0.3$;

		$DA+UEH1$		$DA+UEH2$		$GA+UEH1$		$GA+UEH2$		VNS	
$\|N\|$	$\|S\|$	HG	$Time$	HG	$Time$	HG	$Time$	HG	$Time$	RE	$Time$
200	2	0.0000	990.08	0.1015	222.39	0.0000	229.54	0.1015	41.72	0.0000	4865.55
200	3	0.0000	2782.47	0.2136	637.03	0.0000	526.61	0.2136	96.00	0.0000	12800.00
200	4	0.0000	5812.59	0.3273	1371.09	0.0000	895.97	0.3273	167.42	0.0000	26179.20
200	5	0.0000	10769.70	0.4422	3376. 4	0.0000	1351.78	0.4422	253.33	0.0000	46467.80
200	8	0.0000	46425.20	0.4926	20173.20	0.0000	5503.48	0.4926	877.06	0.0000	48900.90
200	10	0.0000	264751.00	0.5434	35984.02	0.0000	6187.46	0.5434	1430.26	0.0000	96283.56
300	2	0.0000	1011.93	0.1029	234.85	0.0000	324.65	0.1029	210.14	0.0000	4999.45
300	3	0.0000	2809.29	0.2147	654.24	0.0000	612.89	0.2147	195.86	0.0000	12882.34
300	5	0.0000	5849.97	0.3279	1385.55	0.0000	899.89	0.3279	156.19	0.0000	26271.90
300	8	0.0000	10802.44	0.4421	3401.20	0.0000	1425.41	0.4421	174.02	0.0000	46527.72

(b) $\alpha = 0.001, \gamma = 0.3$.

4.9.3 Sensitivity Analysis

In this section we study the effect of customers elasticities to travel distance and waiting time on the performance of the heuristic algorithms. We vary customers' elasticity parameters α and γ with all other network parameters being fixed and compare the heuristic gaps among the algorithms. Tables 4.7 and 4.8 show the results of a 10 node 3 facility network. In Table 4.7, customers are roughly equally sensitive to the travel distance and waiting time; In Table 4.8, customers are more sensitive to waiting time than travel distance.

We have the following observations from Tables 4.7 and 4.8:

1. When customers are more sensitive to travel distances than waiting times, i.e., the ratio α to γ is large, all algorithms work well. As shown in Table 4.7, the algorithms give almost the same solution for most cases and the maximal HG is just 9.38%.

2. When customers are more sensitive to waiting times than travel distances, i.e., the ratio α to γ is small, DA+UEH1, VNS and UFLP1 work better than DA+UEH2, GA+UEH2 and UFLP2. Table 4.8 shows that the HG of GA+UHE2 ranges from 0.00% to 27.42% and it can be as high as 34.60% for UFLP2.

These observations are consistent with the numerical experiments in the previous section.

Table 4.7: Sensitivity to customers' elasticity -HG (%)

(α, γ)	$DA+UEH1$	$DA+UEH2$	$GA+UEH1$	$GA+UEH2$	VNS	$UFLP1$	$UFLP2$
$(0.1, 0.1)$	0.84	0.84	0.09	0.11	0.09	0.00	0.00
$(0.1, 0.3)$	1.02	1.02	2.18	0.00	0.00	0.00	0.00
$(0.1, 0.5)$	0.12	0.14	0.12	0.14	0.12	0.00	0.00
$(0.1, 0.8)$	0.00	0.00	1.07	1.07	0.00	0.00	0.00
$(0.3, 0.1)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(0.3, 0.3)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(0.3, 0.5)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(0.3, 0.8)$	0.00	0.00	0.20	3.79	0.00	0.00	0.00
$(0.5, 0.1)$	0.57	0.57	9.38	0.57	0.57	0.00	0.00
$(0.5, 0.3)$	2.11	1.17	2.11	1.17	2.11	0.00	0.00
$(0.5, 0.5)$	0.00	0.00	1.05	0.00	0.00	0.00	0.23
$(0.5, 0.8)$	0.00	0.00	0.00	0.00	0.00	0.00	0.34
$(0.8, 0.1)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(0.8, 0.3)$	7.82	7.82	7.82	7.82	7.82	0.00	0.00
$(0.8, 0.5)$	4.73	4.73	4.73	8.69	4.73	0.00	0.00
$(0.8, 0.8)$	5.91	8.08	5.90	8.08	0.00	0.00	0.00

Table 4.8: Sensitivity to customers' elasticity -HG (%)

(α, γ)	$DA+UEH1$	$DA+UEH2$	$GA+UEH1$	$GA+UEH2$	VNS	$UFLP1$	$UFLP2$
$(0.001, 0.1)$	0.00	0.00	0.00	0.00	0.00	0.36	0.01
$(0.001, 0.5)$	0.06	0.00	0.06	0.00	0.30	0.30	0.00
$(0.001, 1)$	0.00	2.00	0.00	2.00	0.00	0.00	2.00
$(0.001, 5)$	0.00	14.08	0.00	14.08	0.00	0.00	14.07
$(0.001, 10)$	0.00	22.40	35.47	27.42	0.00	0.02	30.30
$(0.01, 0.1)$	0.00	0.00	0.05	0.05	0.00	0.00	0.00
$(0.01, 0.5)$	0.27	0.00	0.45	0.00	0.45	0.27	0.12
$(0.01, 1)$	0.00	0.96	0.62	1.20	0.00	0.62	1.88
$(0.01, 5)$	0.00	8.90	0.00	8.90	0.00	0.00	13.20
$(0.01, 10)$	0.00	20.82	0.00	21.79	0.29	0.00	25.09
$(0.1, 0.1)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(0.1, 0.5)$	0.00	0.02	0.00	0.02	0.00	0.00	0.02
$(0.1, 1)$	0.25	0.00	0.67	0.67	0.25	0.67	0.67
$(0.1, 5)$	0.00	6.89	0.00	6.89	0.00	0.00	7.16
$(0.1, 10)$	0.00	18.07	32.56	19.05	0.39	9.45	19.08
$(0.5, 0.1)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(0.5, 0.5)$	0.00	1.02	0.00	1.02	0.00	0.00	1.02
$(0.5, 1)$	0.00	0.00	0.00	0.00	0.00	2.02	2.02
$(0.5, 5)$	0.00	3.75	0.00	3.75	2.80	2.80	3.75
$(0.5, 10)$	0.00	12.24	0.00	12.24	0.00	0.00	34.60

4.10 Conclusions and Future Research

We studied the problem of designing a multi-facility service network in the presence of congestions and demand elasticities. The objective is to find the locations of the facilities, to decide the price to charge for service and the capacity level to allocate so that the service firm's revenue is maximized. Two models are proposed: a system optimization model and a user equilibrium model. In the system optimization model, customers cooperate with the firm to maximize the firm's revenue. In the user equilibrium model, customers establish equilibrium flow to maximize their own utilities.

We first show through an example that the customers may not visit their closest facilities for service in both the system optimization model and the user equilibrium model, though the system optimization model generates higher revenue. Our study further focus on solving the user optimization model. The properties of the optimal solutions are analyzed. We show that a uniform pricing strategy is optimal and is independent of the location and capacity decisions. When the locations of the facilities, the price charged for service and the capacities are known, we show that customers equilibrium flow problems can be solved by a convex optimization problem via the variational inequality approach. We developed several algorithms to solve the user equilibrium model.

Our numerical experiments suggest that customers' elasticities play a key role in the location and capacity allocation decisions. When the customers value the proximity of the service as much important as the waiting time for service, the price, location and capacity decisions can be made separately. When the customers are very sensitive to the waiting time, however, the separate decision making is not sufficient enough. As indicated by our computational experiments, capacity allocation and location decisions should be jointly optimized in this case.

Several extensions to our modeling frame work are worth further investigate. First, our model and results are based on a specific demand elasticity function and linear utility function. So, it would be interesting to see whether the results can be extended to other demand elasticity functions and more general utility functions. Second, the demand elasticities are assumed to be homogeneous to all customers. One can relax this assumption and consider cases that customers

residing at different nodes have different elasticities and see if the results are different from our observations. Finally, the competition is considered implicitly by the elasticity of the demand in our model. Considering pre-existing competitors on the network and model the loss of demand explicitly to these competitors would also be an interesting topic.

Appendices

Appendix A

Wardrop Equilibrium and Nash Equilibrium

The customer equilibrium considered in this research is a type of Wardrop equilibrium. In this section, we briefly discuss the difference between the Wardrop and Nash equilibria. We refer interested readers to Haurie and Marcotte (1985) for a detailed discussion.

The Wardrop equilibrium was first introduced in Wardrop (1952). Wardrop's first principle for traffic equilibra states: "*The journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route.*" It is equivalent to the definition of customer equilibrium. In our context, we assume there are a very large number of infinitesimal customers and customers are in Wardrop equilibrium if no customer can increase his utility by switching from his current demand pattern to another one. The effect of one individual customer, when unilaterally changes his choice from a particular demand pattern to another one, is infinitesimal.

Nash equilibrium was defined in Nash (1951). In terms of network flows, a flow pattern is in Nash equilibrium if no individual customers on the network can change to a less costly route.

The difference between a Wardrop equilibrium and a Nash equilibrium is that when the players in a Nash game are discrete and finite in number, a Nash equilibrium can be achieved without the costs of all used routes being equal, while in a Wardrop equilibrium the costs of all used routes must be equal. Wardrop's equilibrium represents a limiting case of an infinite

number of infinitesimal players in the Nash equilibrium.

Since the number of customers in our study is large, we use a Wardrop equilibrium, which treats individual user contributions to the costs as infinitesimal.

Appendix B

Existence and Uniqueness of Customer Equilibrium Flow

In this section, we discuss the existence and uniqueness of the equilibrium using variational inequality method. We refer readers to Nagurney (1999) for the variational inequality problem in details. Let $\mathbf{x} = (\mathbf{x}^1; \cdots; \mathbf{x}^M)$ be the column vector of global customer-demand pattern assignment vector with cardinality of $\sum_{m \in \mathcal{M}} \mathcal{S}_m$. We begin with some fundamental definitions.

Definition B.1. *(Variational Inequality Problem) The finite-dimensional variational inequality problem $VI(F, \mathcal{K})$, is to determine a vector $\boldsymbol{x}^* \in \mathcal{K} \subset R^n$, such that*

$$\langle F(\boldsymbol{x}^*)', \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0, \forall \boldsymbol{x} \in \mathcal{K},$$

where F is a given continuous function from $\mathcal{K}$ to R^n, $\mathcal{K}$ is a given closed convex set, and $\langle \cdot, \cdot \rangle$ denotes the inner product in R^n, where R^n in the n-dimensional Euclidean space. For example, $\langle (1,2), (3,4) \rangle = 11$.

Definition B.2. *(Monotonicity) $F(\boldsymbol{x})$ is monotone on $\mathcal{K}$ if*

$$\langle (F(\boldsymbol{x}^1) - F(\boldsymbol{x}^2))', \boldsymbol{x}^1 - \boldsymbol{x}^2 \rangle \geq 0, \forall \boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathcal{K}.$$

For example, any additive, increasing or decreasing function $F(\mathbf{x})$ is monotone.

Definition B.3. *(Strict Monotonicity) $F(\boldsymbol{x})$ is strictly monotone on $\mathcal{K}$ if*

$$\langle (F(\boldsymbol{x}^1) - F(\boldsymbol{x}^2))', \boldsymbol{x}^1 - \boldsymbol{x}^2 \rangle > 0, \forall \boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathcal{K}, \boldsymbol{x}^1 \neq \boldsymbol{x}^2.$$

Definition B.4. *(Strong Monotonicity) $F(\boldsymbol{x})$ is strongly monotone on $\mathcal{K}$ if for some $\boldsymbol{\alpha} > 0$*

$$\langle (F(\boldsymbol{x}^1) - F(\boldsymbol{x}^2))', \boldsymbol{x}^1 - \boldsymbol{x}^2 \rangle \geq \alpha \parallel \boldsymbol{x}^1 - \boldsymbol{x}^2 \parallel^2, \forall \boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathcal{K}.$$

Let $\mathbf{u}^m = (u_1^m, \cdots, u_{S_m}^m)$ be the column vector of utilities for class-m customers with cardinality of S_m. Let $\mathbf{u} = (\mathbf{u}^1; \cdots; \mathbf{u}^M)$ be the column vector of global customer utilities stacked by its natural dimensions with cardinality of $\sum_{m \in \mathcal{M}} S_m$. Then we can formulate (2.4) as a variational inequality problem (Smith, 1979). Using similar proof, we have

Theorem B.1. *Smith (1979) $\boldsymbol{x}^* \in \mathcal{K}$ is a customer equilibrium flow if and only if it solves the following variational inequality problem,*

$$\langle \boldsymbol{u}(\boldsymbol{x}^*)', \boldsymbol{x} - \boldsymbol{x}^* \rangle \leq 0, \forall \boldsymbol{x} \in \mathcal{K} \tag{B.1}$$

Proof. Let $\mathbf{x}^*$ satisfies the condition (2.4) for a Wardrop equilibrium, and let $\mathbf{u}(\mathbf{x}^*)$ be the customer utilities determined by $\mathbf{x}^*$. Regard the customers utilities as fixed. Because $\mathbf{x}^*$ satisfies (2.4) and so only best utilities path are used, total utilities will be reduced by any change of path. Therefore any other path $\mathbf{x} \in \mathcal{K}$ has total utilities at most as great as $\mathbf{x}^*$ which uses best utility pathes, it then follows that the total utilities in vector form:

$$\mathbf{u}(\mathbf{x}^*)' \cdot \mathbf{x} \leq \mathbf{u}(\mathbf{x}^*)' \cdot \mathbf{x}^*, \forall \mathbf{x} \in \mathcal{K} \tag{B.2}$$

Conversely, suppose that (2.4) is not satisfied. Then there is a x_s^m such that

$$x_s^m > 0 \text{ and } u_s^m(\mathbf{x}^*) < u_t^m(\mathbf{x}^*) \tag{B.3}$$

Moving the flow x_s^m along path s to the better utility path t will improve the total utility by

$u_t^m(\mathbf{x}^*)x_s^m - u_s^m(\mathbf{x}^*)x^{*m}_s > 0$. Thus if the resulting flow is $\mathbf{x}$,

$$\mathbf{u}(\mathbf{x}^*)' \cdot \mathbf{x} > \mathbf{u}(\mathbf{x}^*)' \cdot \mathbf{x}^*, \forall \mathbf{x} \in \mathcal{K} \tag{B.4}$$

in which case (B.4) or equivalently (B.3) is not satisfied. We have shown that if (2.4) is satisfied then (B.3) is satisfied. We have also shown that if (2.4) is not satisfied then (B.3) is not satisfied. Therefore (2.4) and (B.3) are equivalent. □

The following two propositions follow directly from the classical variational inequality theorem in (Nagurney, 1999):

Proposition B.1. *If $-\boldsymbol{u}(\boldsymbol{x})$ is strictly monotone, then the equilibrium is unique, if one exists.*

Proposition B.2. *If $-\boldsymbol{u}(\boldsymbol{x})$ is strongly monotone, then there exists a unique equilibrium.*

Let $\mathbf{z}^m = \mathbf{A}^m\mathbf{x}^m$, where z_t^m denotes the number of orders from class-m customers at time t. Thus $\mathbf{y} = \sum_{m\in\mathcal{M}} \mathbf{z}^m$. Let $\mathbf{z} = (\mathbf{z}^1; \cdots ; \mathbf{z}^M)$. Let $\mathbf{w}^m$ be a T dimensional vector of congestion cost for class-m customers, where w_t^m denotes the congestion cost from class-m customers at time t. Let $\mathbf{w} = (\mathbf{w}^1; \cdots ; \mathbf{w}^M)$.

Proposition B.3. *$-\boldsymbol{u}(\boldsymbol{x})$ is monotone(strictly monotone,strongly monotone) if and only if $\boldsymbol{w}(\boldsymbol{z})$is monotone(strictly monotone,strongly monotone).*

Proof. Suppose that $\mathbf{x} \in \mathcal{K}$, $\hat{\mathbf{x}} \in \mathcal{K}$ and $\mathbf{x} \neq \hat{\mathbf{x}}$, to show that $-\mathbf{u}(\mathbf{x})$ is monotone, we need to show that

$$\langle (\mathbf{u}(\mathbf{x}) - \mathbf{u}(\hat{\mathbf{x}}))', \mathbf{x} - \hat{\mathbf{x}} \rangle \leq 0.$$

Let δ_{ts}^m be the (t,s) element of matrix $\mathbf{A}^m$. $u_s^m(\mathbf{x})$ can be written as

$$u_s^m(\mathbf{x}) = v_s^m - \bar{p}^m l_m - \sum_{t\in\mathcal{T}} \tilde{p}_t \delta_{ts}^m - \sum_{t\in\mathcal{T}} w_t^m(\mathbf{y})\delta_{ts}^m.$$

The inner product can be expressed as

$$
\begin{aligned}
& \sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}_m}(u_s^m(\mathbf{x})-u_s^m(\hat{\mathbf{x}}))(x_s^m-\hat{x}_s^m) \\
= & \sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}_m}\Bigg(\big(v_s^m-\bar{p}^m l_m-\sum_{t\in\mathcal{T}}\tilde{p}_t\delta_{ts}^m-\sum_{t\in\mathcal{T}}w_t^m(\mathbf{y})\delta_{ts}^m\big) \\
& -\big(v_s^m-\bar{p}^m l_m-\sum_{t\in\mathcal{T}}\tilde{p}_t\delta_{ts}^m-\sum_{t\in\mathcal{T}}w_t^m(\hat{\mathbf{y}})\delta_{ts}^m\big)\Bigg)(x_s^m-\hat{x}_s^m) \\
= & \sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}_m}\Bigg(\sum_{t\in\mathcal{T}}(-w_t^m(\mathbf{y})+w_t^m(\hat{\mathbf{y}}))\delta_{ts}^m\Bigg)(x_s^m-\hat{x}_s^m) \\
= & \sum_{m\in\mathcal{M}}\sum_{t\in\mathcal{T}}(-w_t^m(\mathbf{y})+w_t^m(\hat{\mathbf{y}}))\sum_{s\in\mathcal{S}_m}\delta_{ts}^m(x_s^m-\hat{x}_s^m) \\
= & \sum_{m\in\mathcal{M}}\sum_{t\in\mathcal{T}}(-w_t^m(\mathbf{y})+w_t^m(\hat{\mathbf{y}}))(z_t^m-\hat{z}_t^m).
\end{aligned}
$$

Therefore we can see that $-\mathbf{u}(\mathbf{x})$ is monotone if and only if $\mathbf{w}(\mathbf{z})$ is monotone. We can show the result of strictly monotone and strongly monotone using similar reasoning. The proof is complete. □

Note that from Propositions B.1 to B.3 we can see that while the existence and uniqueness of the equilibrium depends on the customer's delay cost, it is independent of price, and this is true even for the heterogeneous delay cost cases.

Bibliography

Aashtiani, H.Z., T.L. Magnanti. 1981. Equilibria on a congested transportation network. *SIAM Journal of Algebraic and Discrete Methods* **2**(3) 213–226.

Aboolian, R., O. Berman, Z. Drezner. 2008. Location and allocation of service units on a congested network. *IIE Transactions* **40**(4) 422–433.

Aboolian, R., Y. Sun, G.J. Koehler. 2009. A location-allocation problem for a web services provider in a competitive market. *European Journal of Operational Research* **194**(1) 64–77.

Berman, O., Z. Drezner. 2005. Location of congested capacitated facilities with distance-sensitive demand. *IIE Transactions* **38** 213–221.

Berman, O., R. Huang. 2007. The minisum multipurpose trip location problem on networks. *Transportation Science* **41**(4) 500–515.

Berman, O., E.H. Kaplan. 1987. Facility location and capacity planning with delay-dependent demand. *International Journal of Production Research* **25**(12) 1773–1780.

Berman, O., D. Krass. 2002. Facility location problems with stochastic demands and congestion. *Location analysis: Application and theory* 329–371.

Berman, O., D. Krass, B. Liu. 2007. A monopolist's pricing and location problem on networks. *University of Toronto Working Paper* .

Berman, O., D. Krass, J. Wang. 2006. Locating service facilities to reduce lost demand. *IIE Transactions* **38**(11) 933–946.

Bertsekas, D.P. 1999. *Nonlinear programming*. Athena Scientific, Belmont, Mass.

Bertsimas, D., I. Popescu. 2003. Revenue management in a dynamic network environment. *Transportation Science* **37**(3) 257–277.

Bitran, G.R., R. Caldentey. 2003. An overview of pricing models for revenue management. *Manufacturing and Service Operations Management* **5**(3) 203–229.

Bitran, G.R., D. Tirupati. 1989. Hierarchical production planning. *MIT Sloan School of Management Working Paper 3017-89-MS* .

Boadway, R., M. Marchand, M. Vigneault. 1998. The consequences of overlapping tax bases for redistribution and public spending in a federation. *Journal of public economics* **68**(3) 453–478.

Cole, R., Y. Dodis, T. Roughgarden. 2003. Pricing network edges for heterogeneous selfish users. *In Proceedings of the 35th Annual ACM Symposium on the Theory of Computing* .

Dafermos, S. 1980. Traffic equilibrium and variational inequalities. *Transportation Science* **14**(1) 42–54.

Dafermos, S.C. 1973. Toll patterns for multiclass-user transportation networks. *Transportation Science* **7**(3) 211–223.

Dobson, G., E. Stavrulaki. 2007. Simultaneous price, location, and capacity decisions on a line of time-sensitive customers. *Naval Research Logistics* **54**(1) 1–10.

Drezner, Z., H. Hamacher. 2002. *Facility Location: Applications and Theory*. Springer.

Eiselt, H.A., G. Laporte, J.-F. Thisse. 1993. Competitive location models: A framework and bibliography. *Transportation Science* **27**(1) 44–54.

Elhedhli, S. 2006. Service system design with immobile servers, stochastic demand, and congestion. *Manufacturing and Service Operations Management* **8**(1) 92–97.

Elmaghraby, W.J., P. Keskinocak. 2003. Dynamic pricing: Research overview, current practice and future directions. *Management Science* **49**(10) 1287–1305.

Gallego, G., G. van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science* **40**(8) 999–1020.

Gallego, G., G. van Ryzin. 1997. A multirpodcuct dynamic pricing problem and its applications to network yield management. *Operations Research* **45**(1) 24–41.

Hanjoul, P., P. Hansen, D. Peeters, J.F. Thisse. 1990. Uncapacitated plant location under alternative spatial price policies. *Management Science* **36**(1) 41–57.

Hansen, P., J.F. Thisse, P. Hanjoul. 1981. Simple plant location under uniform delivered pricing. *European Journal of Operational Research* **6**(2) 94–103.

Haurie, A., P. Marcotte. 1985. On the relationship between nash-cournot and wardrop equilibria. *Networks* **15**(3) 295–308.

Hopp, W., M. Spearman. 2000. *Factory Physics*. McGraw-Hill/Irwin.

Huff, D.L. 1964. Defining and estimating a trade area. *Journal of Marketing* **28** 34–38.

Koutsoupias, E., C. Papadimitriou. 1999. Worst-case equilibria. *In Proceedings of the 16th Annual Symposium on Theoretical Aspects of Computer Science* 404–413.

Logendran, R., M.P. Terrell. 1991. Capacitated plant location-allocation problems with price sensitive stochastic demands. *Logistics and Transportation Review* **27**(1) 33–53.

Macdonald, J.M., P. Nelson. 1991. Do the poor still pay more? food price variations in large metropolitan areas. *Journal of Urban Economics* **30**(1) 344–359.

Marianonov, V., D. Serra. 1998. Probabilistic maximal covering location-allocation for congested system. *Journal of Regional Science* **38**(3) 401–424.

McGill, J., G. van Ryzin. 1999. Revenue management: Research overview and prospects. *Transportation Science* **33**(2) 233–256.

Mirchandani, P.B., R.L. Francis. 1990. *Discrete Location Theory*. Wiley-Interscience, New York.

Mirrlees, J. 1971. An exploration in the theory of optimum income taxation. *Review of Economic Studies* **38**(114) 175–208.

Mladenovic, N.P., P. Hansen. 1997. Variable neighborhood search. *Computers and Operations Research* **24**(11) 1097–1100.

Nagurney, A. 1999. *Network Economics: A Variational Inequality Approach*. Kluwer Academic Publishers, Boston.

Nash, J. 1951. Non-cooperative games. *The Annals of Mathematics* **54**(2) 286–295.

Roughgarden, T., E. Tardos. 2002. How bad is selfish routing? *Journal of the ACM* **49**(2) 236–259.

Smith, M.J. 1979. Existence, uniqueness and stability of traffic equilibria. *Transportation Research Part B-Methodological* **13**(4) 295–304.

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

Talluri, K., G. van Ryzin. 2005. *The Theory and Practice of Revenue Management*. Springer.

Wagner, J.L., L.M. Falkson. 1975. Optimal nodal location of public facilities with price-sensitive demand. *Geographical Analysis* **7** 69–83.

Walmart. 2009. Walmart 2009 annual report .

Wardrop, J.G. 1952. Some theoretical aspects of road research. *Proceedings of the Institute of Civil Engineers Part II* 325–378.

Wolsey, L.A., G.L. Nemhauser. 1998. *Integer and Combinatorial Optimization*. Wiley-Interscience.

Zhang, Y., O. Berman, V. Verter. 2009. Incorporating congestion in preventive healthcare facility network design. *European Journal of Operational Research* **198**(3) 922–935.