Decision Making under Uncertainty: Reliability and Incentive Compatibility

by

Tingting Cui

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

Doctor of Philosophy

in

Engineering- Industrial Engineering and Operations Research

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Zuo-Jun Shen, Chair
Professor J. George Shanthikumar
Professor Carlos Daganzo
Professor Ying-Ju Chen

Spring 2010

Decision Making under Uncertainty: Reliability and Incentive Compatibility

by

Tingting Cui

Abstract

Decision Making under Uncertainty: Reliability and Incentive Compatibility

by

Tingting Cui

Doctor of Philosophy in Engineering- Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun Shen, Chair

This dissertation studies analytical and computational aspects of two types of problems in applied operations research. In the first part of the dissertation, we consider the reliable facility location models in which facilities are subject to unexpected failures. We propose a compact mixed integer programming formulation that is polynomial in size. To compute optimal facility locations that balance the trade-off between normal operation and failure costs, we develop two exact algorithms: one is based on Lagrangian Relaxation, and the other is a hybrid of neighborhood search and cutting plane procedures. To obtain more managerial insights, we further investigate a Continuum Approximation (CA) model that predicts the total system cost without details about facility locations and customer assignments. The CA model is a valuable tool for sensitivity analysis, as well as a fast heuristic for large problem instances.

The second part of the dissertation is dedicated to theoretical and applied mechanism design, which consists of two chapters. In the first chapter, we study a problem of allocating limited capacity of a queueing system to serve several segments of customers, who differ in their willingness to pay and their sensitivity to delays, both of which are their private information. We show that probabilistic admission control, randomized priority rule, and strategic idleness can emerge as optimal solutions in a revenue maximizing mechanism.

In the second chapter of part two, we revisit the optimal auction design problem and propose a robust formulation based on an uncertainty set that characterizes the conservativeness of the bidders' beliefs, with two special cases being the Bayesian and Ex post formulations. Using the network approach, we identify the necessary and sufficient conditions under which the expected revenues achieved by different formulations are identical. Furthermore, we show that in a multiple-object auction, the auctioneer's expected revenue may strictly decrease as the bidders' beliefs become more uncertain.

To my husband Kai.

Contents

List of Figures

List of Tables

Acknowledgements

My study at U.C. Berkeley has been the most enriching experience of my life. I express my gratitude to everyone who has made this journey truly rewarding for me.

I am deeply indebted to my advisor, Professor Zuo-Jun Shen, who has been not just a thesis supervisor but also a cordial friend and a personal mentor. He has taught me to conduct independent research, to deliver effective presentations, and to be a confident teacher. His wisdom and advice have proved invaluable over the course of my doctoral study, and will definitely continue to benefit me in the future as I start my own academic career.

I would also like to thank other members of my committee, Professor George Shanthikumar, Professor Carlos Daganzo, and Professor Ying-Ju Chen. Their time, effort and feedback has been invaluable for me.

I remain forever grateful to my Master's thesis advisor, Professor J. Cole Smith, for introducing me to the world of operations research and encouraging me to pursue doctoral studies at U.C. Berkeley.

Finally, my eternal thanks to my family for their unconditional love, affection, and support. Most of all, to Kai, my husband and best friend, who is always with me in all my joys and sorrows. Words are never enough to express my love for them.

Chapter 1

Outline

This dissertation concerns itself with two diverse topics arise from applied operations research: reliable facility location and multiple dimension mechanism design. For this reason, the report is organized in two parts. In this chapter, we briefly outline the contents of each part.

Part I

The classic uncapacitated fixed charge location problem (UFL) selects facility locations and customer assignments in order to balance the trade-off between initial setup costs and day-to-day transportation costs. However, some of the constructed facilities may become unavailable due to disruptions caused by natural disasters, terrorist attacks or labor strikes. When a facility failure occurs, customers may have to be reassigned from their original facilities to others that require higher transportation costs. We present facility location models that minimize normal construction and transportation costs as well as hedge against facility failures within the system.

In Chapter 2, we proposed an implicit formulation of the reliable facility location problem that are based on level assignment. Unlike scenario based stochastic programs that have exponentially many variables and constraints, our formulation is polynomial in size. The formulation can also be easily linearized as a mixed integer program (MIP), which can be solved using off-the-rack MIP solvers such as ILOG CPLEX. However, due to the large scale nature of the problem, the CPLEX solver is not reliable and usually requires excessive computation time. This motivates our development of a Lagrangian algorithm, which provides decomposition of the customer assignment problem, along with two custom-designed algorithms for the individual assignment problems. Our computation results indicate that the Lagrangian algorithm is efficient for mid-sized problem instances. However, the algorithm performance is not ideal for large test instances.

To further improve computational efficiency, we introduce a group of new algorithms in Chapter 3. All of the algorithms rely on a neighborhood search procedure to improve the upper bound/best available feasible solution, but differ in the way to find lower bounds of the optimal solution. We proposed three different methods to update the lower bounds, listed in increasing order of accuracy: an MIP formulation assuming no facility failures, a linear MIP formulation that assumes fixed failure probabilities, and a Lagrangian relaxation of the original problem with varying failure probabilities. Once all the neighborhood solutions have been evaluated, they are eliminated from future consideration using integer cuts. Our test results indicates that these algorithms provide a good combination of speed and accuracy.

Both Chapter 2 and Chapter 3 address discrete models of the reliable facility location problems, for which only computational results are available and very few insights can be drawn from the optimal solutions. To provide more managerial insights, we develop a Continuum Approximation (CA) model in Chapter 4. The CA model precisely predicts the system costs by focusing on important decisions like the number of facilities and the influential areas, omitting details of facility locations and customer assignments. Since the system cost is approximated as continuous function of the key parameters, the CA model is a handy tool for sensitivity analysis. Furthermore, the continuous results from the CA model can be translated into a discrete facility location design, making the CA model an alternative heuristic approach to the discrete models. Our extensive test results show that the CA model provide approximate solutions that are close to the optimal solutions found by the discrete models, especially for the larger problem instances.

Part II

In Chapter 5, we study a problem setting in which a capacity-constrained server (modeled as an $M/M/1$ queueing system) intends to serve several segments of customers. Customers request the same amount of task; nevertheless, they are heterogeneous in two attributes: their willingness to pay, and their willingness to wait, both of which are privately observed by this customer but unknown to the server. In the absence of the information about the customers' preference, the server faces an *adverse selection* problem and therefore must design an appropriate mechanism: On one hand, the mechanism must induce the customers to reveal their true preferences at their own will; on the other hand, this mechanism must maximize the server's long-run average expected payoff in the presence of capacity constraint and congestion effect.

We show that a well-designed menu of probabilistic admission control along with priority pricing contracts may force customers to reveal their true valuations and at the same time induce customers that are more sensitive to the delay to opt for higher priorities. Thus, the probabilistic admission control allows the server to identify the customers that are willing to pay more for the service (thereby reducing the undesirable congestion) and consequently may enable the server to increase its revenue. We further find that randomized priority rule, probabilistic admission control, and

strategic idleness can emerge as optimal solutions (from the server's perspective). Moreover, even though ex ante the server may exhibit specific preference/ranking over different groups of customers, the server may probabilistically admit more than one group and the preference becomes endogenous.

In Chapter 6, we switch our focus to the robustness of some well-known auction mechanisms. In an attempt to answer the Wilson's doctrine, that criticizes game theory models assuming too much common knowledge amongst the players, we're particularly interested in the revenue difference of optimal mechanisms under the Bayesian Nash (less robust) and the Ex post (more robust) settings. We adopt a novel approach based on a graph theory representation of the incentive compatibility (IC) constraints, to characterize the necessary and sufficient condition under which a fixed allocation rule achieves the same revenue under the Bayesian and the Ex post settings. Based on this result, we show that Bayesian optimal allocation for a single-object auction has a revenue equivalent implementation in the Ex post setting, however, the same result do not apply for multiple-object auctions. We further characterize Ex post optimal allocations for special cases of multiple-object auctions, and verify the revenue difference from the Bayesian optimal solutions.

Chapter 2

Reliable Facility Location: Formulation and Lagrangian Relaxation

2.1 Introduction

The classic uncapacitated fixed charge location problem (UFL) selects facility locations and customer assignments in order to balance the trade-off between initial setup costs and day-to-day transportation costs. However, some of the constructed facilities may become unavailable due to disruptions caused by natural disasters, terrorist attacks or labor strikes. When a facility failure occurs, customers may have to be reassigned from their original facilities to others that require higher transportation costs. In this chapter we present facility location models that minimize normal construction and transportation costs as well as hedge against facility failures within the system.

The reliable location model was first introduced by Snyder and Daskin Snyder and Daskin [2005] to handle facility disruption. Their motivating example is as follows. Consider a supply network that serves 49 cities, consisting of all state capitals of the continental United States and Washington, DC. Demands are proportional to the 1990 state populations and the fixed costs are proportional to the median house prices. The optimal UFL solution for this problem is shown in Figure 2.1. This solution has a fixed cost of \$348,000 and a transportation cost of \$509,000 (at \$0.00001 per mile per unit of demand). However, if the facility in Sacramento, CA failed, customers from

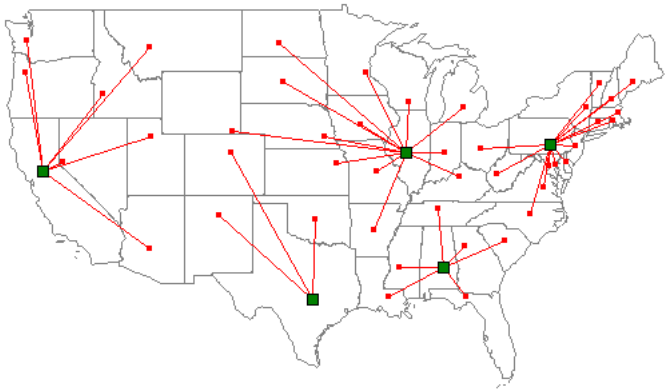


Figure 2.1. UFL solution to 49-data set

the entire west-coast region would have to get service from the facilities in Springfield, IL and Austin TX, which would increase the transportation cost to $1,081,000 (112%). Table 2.1 lists the "failure cost", the transportation cost associated with each facility failure.

Location	Failure Cost	% Increase
Sacramento, CA	1,081,229	112%
Harrisburg, PA	917,332	80%
Springfield, IL	696,947	37%
Montgomery, AL	639,631	26%
Austin, TX	636,858	25%
Transp. cost w/o failures	508,858	0%

Table 2.1. Failure costs of UFL solution

Snyder and Daskin Snyder and Daskin [2005] suggested that locating facilities in the capitals of CA, NY, TX, PA, OH, AL, OR, and IA (Figure 2.2) is a more reliable solution. In this solution, the maximum failure cost is reduced to $500,216, less than the smallest failure cost in Table 2.1. However, three additional facilities are used in this solution resulting in a total location and day-to-day transportation cost of $919,298 - a 7.25% increase from the UFL optimal solution.

Realistically, no company would accept a supply network with high normal operating costs just to hedge against very rare facility disruptions. In order to balance the trade-off between normal operating costs and failure costs, the network structure should depend on how likely the candidate sites may get disrupted, as well as their closeness to the potential customers. In Snyder and Daskin Snyder and Daskin [2005], all facility locations are assumed to have identical failure probabilities, which might not be very representative of practical situations. Let us illustrate how site-dependent failure probabilities impact the choice of facility locations. Specifically, suppose that the facilities are vulnerable to hurricane related disasters. Facilities located in the

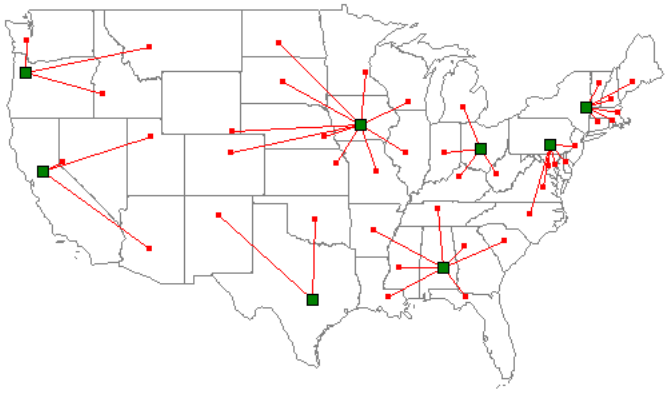


Figure 2.2. A more reliable solution

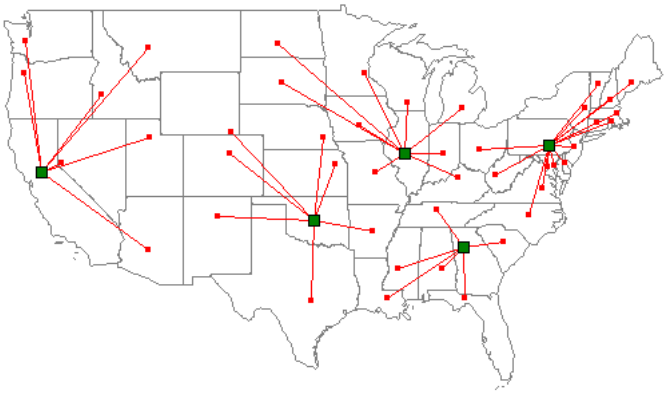


Figure 2.3. A cost efficient solution

Gulf coast area (TX, LA, MS, AL and FL) all have a 10% chance of disruption, while other potential sites have a much lower failure probability of 0.1%. It is cost efficient here to hedge against disruption by locating facilities in the capitals of CA, PA, IL, GA and OK (Figure 2.3). In this solution, the two facilities along the Gulf coast (TX and AL) are moved to adjacent "safer" locations. Although the failure costs of CA and PA are high, we choose not to build "backup" facilities for them because their probability of disruption is so small. The expected failure cost in this solution is about \$4,000, compared to \$130,344 in the UFL optimal solution in Figure 2.1, and the location and day-to-day transportation costs are increased by only 3.6%. Table 2.2 compares the normal operating costs and the expected failure costs of the three solutions.

In this chapter, we seek to design supply networks that are both reliable and cost efficient. We minimize the expected transportation costs in both the regular and the failure scenarios (plus the fixed construction costs) to balance the trade-off between

Solution 1			
Location	Failure Cost	Failure Probability	Expected Cost
Sacramento, CA	1,081,229	0.001	1,081
Harrisburg, PA	917,332	0.001	917
Springfield, IL	696,947	0.001	697
Montgomery, AL	639,631	0.1	63,963
Austin, TX	636,858	0.1	63,686
Expected failure cost			130,344
Normal operating cost			857,128
Solution 2			
Location	Failure Cost	Failure Probability	Expected Cost
Sacramento, CA	500,216	0.001	500
Albany, NY	419,087	0.001	419
Austin, TX	476,374	0.1	47,637
Harrisburg, PA	409,383	0.001	409
Columbus, OH	434,172	0.001	434
Montgomery, AL	474,640	0.1	47,464
Salem, OR	389,484	0.001	389
Des Moines, IA	452,305	0.001	452
Expected failure cost			97,706
Normal operating cost			919,298
Solution 3			
Location	Failure Cost	Failure Probability	Expected Cost
Sacramento, CA	1,058,226	0.001	1,058
Harrisburg, PA	908,672	0.001	909
Springfield, IL	681,786	0.001	682
Atlanta, GA	679,022	0.001	679
Oklahoma City, OK	660,985	0.001	661
Expected failure cost			3,989
Normal operating cost			888,009

Table 2.2. Comparisons of the normal operating costs and the expected failure costs

normal and emergency operating costs. The failure of each facility site is assumed to be independent and the probability is taken as a prior. Unlike in Snyder and Daskin Snyder and Daskin [2005], the failure probabilities are allowed to be site-dependent. The facility location decisions and customer assignments are made at the first stage, before any failures occur. Each customer can be assigned to up to $R \geq 1$ facilities to hedge against failures. After any disruptions occur, each customer is served by her closest assigned operating facility; if all her assigned facilities have failed then a penalty cost is charged. We feel that it is reasonable to restrict each customer's facility assignments to a pre-determined subset of all open facilities. In reality a customer may not be able to get service from all facilities due to system compatibility, limited capacity, or simply excessive transportation costs. Our computational results indicate that the choice of R has no significant impact on the network structure of the optimal solutions.

The extensive literature on facility location dates back to its original formulation in 1909 and the Weber problem Weber and Friedrich [1929]. Traditionally, facility location problems are modeled as discrete optimization problems and solved with mathematical programming techniques. Daskin Daskin [1997] and Drezner Drezner [1995] provide good introductions to and surveys of this topic.

Recently, reliability issues in supply chain design are of particular interest. Most of the existing literature focuses on facility congestions from stochastic demand. Daskin Daskin [1982], Daskin et al. [1983], Ball and Lin Ball and Lin [1993], ReVelle and Hogan ReVelle and Hogan [1989], and Batta *et. al.* Batta et al. [1989] all attempted to increase the system availability through redundant coverage.

Focus on system failures due to facility disruptions in supply chain design is gaining attention recently Qi and Shen [2007], Qi et al. [2008]. In the traditional locational analysis literature, Snyder and Daskin Snyder and Daskin [2005] propose an implicit formulation of the stochastic P-median and fixed-charge problems based on level assignments, where the candidate sites are subject to random disruptions with equal probability. Works by Shen et al. [2009] and Berman et al. [2007] relax the assumption of uniform failure probabilities. Shen et al. [2009] formulate the stochastic fixed-charged problem as a nonlinear mixed integer program and provides several heuristic solution algorithms. Berman et al. [2007] focus on an asymptotic property of the problem. They prove that the solution to the stochastic P-median problem coincides with the deterministic problem as the failure probabilities approach zero. They also propose heuristics with bounds on the worst-case performance.

2.2 Formulation

Define I to be the set of customers, indexed by i, and J to be the set of candidate facility locations, indexed by j. For the ease of notation, we also use I and J to indicate the cardinalities of the sets. Each customer $i \in I$ has a demand rate of λ_i. The cost to ship a unit of demand from facility $j \in J$ to customer $i \in I$ is denoted by d_{ij}. Associated with each facility $j \in J$ are the fixed location cost f_j and the probability of failure $0 \leq q_j < 1$. The events of facility disruptions are assumed to be independent.

Each customer is assigned to up to $R \geq 1$ facilities, and can be serviced by these and only these facilities. There is a cost ϕ_i associated with each customer $i \in I$ that represents the penalty cost of not serving the customer per unit of missed demand. This cost may be incurred even if some of her assigned facilities are still online, given that ϕ_i is less than the cost of serving i via any of these facilities. This rule is modeled using an "emergency" facility, indexed by $j = J$, that has fixed cost $f_J = 0$, failure probability $q_J = 0$ and transportation cost $d_{iJ} = \phi_i$ for customer $i \in I$.

The variables used in this model are the location variables (X), the assignment

variables (Y) and the probability variables (P):

$$\begin{aligned} X_j &= \begin{cases} 1, & \text{if a facility } j \text{ is open} \\ 0, & \text{otherwise} \end{cases} \\ Y_{ijr} &= \begin{cases} 1, & \text{if facility } j \text{ is assigned to customer } i \text{ at level } r \\ 0, & \text{otherwise} \end{cases} \\ P_{ijr} &= \text{probability that facility } j \text{ serves customer } i \text{ at level } r. \end{aligned}$$

We employ the modeling techniques introduced by Snyder and Daskin Snyder and Daskin [2005] for assigning customers to facilities at multiple levels. A "level-r" assignment for a customer $i \in I$ will serve her if and only if all of her assigned facilities at levels $0, \cdots, r-1$ have failed. At optimality, each customer $i \in I$ should have exactly R assignments, unless i is assigned to the emergency facility at certain level $s < R$. If a customer i is indeed assigned to exactly R regular facilities at levels $0, \cdots, R-1$, she must also be assigned to the emergency facility J at level R to capture the possibility that all of the R regular facilities may fail. Finally, P_{ijr} is the probability that facility j serves customer i at level r, given her other assigned facilities at levels 0 to $r-1$.

The reliability UFL problem (**RUFL**) is formulated as:

$$\text{(RUFL)} \quad \text{Min} \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{j=0}^{J} \sum_{r=0}^{R} \lambda_i d_{ij} P_{ijr} Y_{ijr} \tag{2.1a}$$

$$\text{s.t.} \sum_{j=0}^{J-1} Y_{ijr} + \sum_{s=0}^{r-1} Y_{iJs} = 1 \quad \forall\, 0 \le i \le I-1,\ 0 \le r \le R \tag{2.1b}$$

$$\sum_{r=0}^{R-1} Y_{ijr} \le X_j \quad \forall\, 0 \le i \le I-1,\ 0 \le j \le J-1 \tag{2.1c}$$

$$\sum_{r=0}^{R} Y_{iJr} = 1 \quad \forall\, 0 \le i \le I-1 \tag{2.1d}$$

$$P_{ij0} = 1 - q_j \quad \forall\, 0 \le i \le I-1,\ 0 \le j \le J \tag{2.1e}$$

$$P_{ijr} = (1-q_j) \sum_{k=0}^{J-1} \frac{q_k}{1-q_k} P_{i,k,r-1} Y_{i,k,r-1} \tag{2.1f}$$

$$\forall\, 0 \le i \le I-1,\ 0 \le j \le J,\ 1 \le r \le R$$

$$X_j,\ Y_{ijr} \in \{0,1\} \quad \forall\, 0 \le i \le I-1,\ 0 \le j \le J,\ 0 \le r \le R. \tag{2.1g}$$

The objective function (2.1a) is the sum of the fixed costs and the expected transportation costs. Constraints (2.1b) enforce that for each customer i and each level r, either i is assigned to a regular facility at level r or she is assigned to the emergency facility J at certain level $s < r$ (taking $\sum_{s=0}^{r-1} Y_{iJs} = 0$ if $r = 0$). Constraints

(2.1c) limit customer assignments to only the open facilities, while constraints (2.1d) require each customer to be assigned to the emergency facility at a certain level. (2.1e)-(2.1f) are the "transitional probability" equations. P_{ijr}, the probability that facility j serves customer i at level r, is just the probability that j remains open if $r = 0$. For $1 \leq r \leq R$, P_{ijr} is equal to $\frac{q_k(1-q_j)}{1-q_k} P_{i,k,r-1}$ given that facility k serves customer i at level $r-1$. Note that constraints (2.1b) imply that $Y_{i,k,r-1}$ can equal 1 for at most one $k \in J$, which guarantees correctness of the transitional probabilities.

Formulation (2.1a)-(2.1g) is nonlinear. However, the only nonlinear terms are $P_{ijr}Y_{ijr},\ 0 \leq i \leq I-1,\ 0 \leq j \leq J,\ 0 \leq r \leq R$, each being a product of a continuous variable and a binary variable. We apply the linearization technique introduced by Sherali and Alameddine Sherali and Alameddine [1992] by replacing each $P_{ijr}Y_{ijr}$ with a new variable W_{ijr}. For each $0 \leq i \leq I-1$, $0 \leq j \leq J$ and $0 \leq r \leq R$ a set of new constraints is added to the formulation to enforce $W_{ijr} = P_{ijr}Y_{ijr}$:

$$W_{ijr} \leq P_{ijr} \tag{2.2a}$$
$$W_{ijr} \leq Y_{ijr} \tag{2.2b}$$
$$W_{ijr} \geq 0 \tag{2.2c}$$
$$W_{ijr} \geq P_{ijr} + Y_{ijr} - 1. \tag{2.2d}$$

The linearized formulation (**LRUFL**) is stated below:

$$\text{(LRUFL)} \quad \text{Min} \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{j=0}^{J} \sum_{r=0}^{R} \lambda_i d_{ij} W_{ijr} \tag{2.3a}$$

$$\text{s.t. } (2.1b) - (2.1d) \tag{2.3b}$$

$$P_{ijr} = (1-q_j) \sum_{k=0}^{J-1} \frac{q_k}{1-q_k} W_{i,k,r-1} \tag{2.3c}$$

$$\forall\ 0 \leq i \leq I-1,\ 0 \leq j \leq J,\ 1 \leq r \leq R$$

$$(2.2a) - (2.2d) \quad \forall\ 0 \leq i \leq I-1,\ 0 \leq j \leq J,\ 1 \leq r \leq R \tag{2.3d}$$

$$X_j,\ Y_{ijr} \in \{0,1\} \quad \forall\ 0 \leq i \leq I-1,\ 0 \leq j \leq J,\ 0 \leq r \leq R. \tag{2.3e}$$

Unlike scenario based stochastic programming problems that have exponentially many variables and constraints, our formulation is compact and polynomial in size. Proposition 1 shows the equivalence of (LRUFL) to the scenario based formulation.

Proposition 1. *If $R = J$, then formulation* (2.1a)-(2.1g) *is equivalent to the stochastic programming formulation that covers all failure scenarios.*

In general, (LRUFL) is not equivalent to the scenario based formulation if $R < J$. However, our computational results show that the choice of R has little impact on

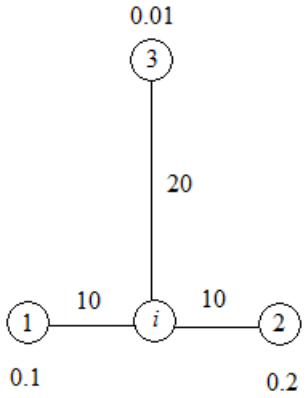


Figure 2.4. Fraction of a supply network

the optimal facility locations. Similar to classic facility location problems, we do not enforce a customer to be served by her closest open facility in our formulation. It is proved in Snyder and Daskin Snyder and Daskin [2005] that the optimal solution always assigns a customer to open facilities level by level in increasing order of distance, given that all facilities are equally likely to fail. The following proposition extends this result to the case where the facility failure probabilities are different across sites.

Proposition 2. *In any optimal solution* $(\mathbf{X}, \mathbf{Y}, \mathbf{P})$ *of (RUFL), if* $Y_{ijr} = 1$ *and* $Y_{ik,r+1} = 1$*, then* $d_{ij} \leq d_{ik}$*, for all* $0 \leq i \leq I-1,\ 0 \leq j \leq J$*, and* $0 \leq r \leq R$*.*

Proposition 2 tells us that for a given subset of facilities assigned to a customer, the optimal assignment levels only depend on the distances from the customer to these facilities. However, if more than R facilities are constructed, it may be sub-optimal to assign each customer to her R closest facilities. As the following example shows, it may be optimal to assign a customer to a facility that is farther away but less likely to fail.

Example 1. *Consider a fraction of a supply network depicted in Figure 2.4. Three facilities are constructed around customer* i*. The distances from* i *to the facilities are* $d_{i1} = d_{i2} = 10$*, and* $d_{i3} = 20$*. The failure probabilities of the three facilities are* $q_1 = q_3 = 0.1$*, and* $q_2 = 0.2$*. The demand rate at* i *is* $\lambda_i = 1$ *and the penalty for not serving a unit of demand is* $\phi_i = 1000$*. Suppose that each customer is only allowed one primary and one back-up facility (*$R = 2$*). If we assign customer* i *to the two closest facilities 1 and 2, then the expected transportation/penalty cost for this customer is 29.8. However, the optimal strategy is to assign* i *to facilities 1 and 3, which reduces the expected transportation cost to 11.98.*

Example 1 implies that even with fixed facility locations, the customer assignment problem is combinatorial and requires more sophisticated solution methods. In this

regard, our model is harder than that in Snyder and Daskin Snyder and Daskin [2005], in which the customer assignment problem with fixed facility locations can be easily solved. We discuss how to decompose the customer assignment problem using Lagrangian relaxation in section 2.3, and how to efficiently solve the individual customer assignment problem in section 2.4.

2.3 Lagrangian Relaxation

The linear mixed-integer program (LRUFL) can be solved using commercial software packages like ILOG CPLEX, but generally such an approach takes an excessively long time even for moderately sized problems. This fact motivates the development of a Lagrangian relaxation algorithm. Relaxing constraints (2.1c) with multipliers μ yields the following objective function:

$$\sum_{j=0}^{J-1}(f_j - \sum_{i=0}^{I-1}\mu_{ij})X_j + \sum_{i=0}^{I-1}\sum_{j=0}^{J}\sum_{r=0}^{R}\lambda_i d_{ij} W_{ijr} + \sum_{i=0}^{I-1}\sum_{j=0}^{J-1}\sum_{r=0}^{R-1}\mu_{ij} Y_{ijr}.$$

For given value of μ, the optimal value of $\mathbf{X}$ can be found easily:

$$X_j = \begin{cases} 1 & \text{if } f_j - \sum_{i=0}^{I-1}\mu_{ij} < 0 \\ 0 & \text{otherwise.} \end{cases}$$

To find the optimal $\mathbf{Y}$, the customer assignment decision, note that the problem is separable in i. For given Lagrangian multipliers μ an individual customer's assignment problem is referred to as the *relaxed subproblem* (**RSP**). The complexity of RSP is demonstrated in Example 1, in which the simple heuristic leads to suboptimal solutions. We discuss efficient algorithms for RSP in Section 2.4.

We use standard subgradient optimization technique to update the Lagrangian multipliers μ, as described in Fisher Fisher [2004]. If the Lagrangian process fails to converge in a certain number of iterations, we use branch-and-bound to close the gap. As a benchmark, we tested our algorithm on the same data sets used by Snyder and Daskin Snyder and Daskin [2005]. The computational results are discussed in Section 2.5.

2.4 The Relaxed Subproblem

Below is the MIP formulation of the relaxed subproblem with respect to customer i (RSP_i). For ease of notation, we omit the subscript i in Y_{ijr}, P_{ijr} and W_{ijr}.

$$(\text{RSP}_i) \quad \text{Min } \Phi_i = \sum_{j=0}^{J}\sum_{r=0}^{R} \lambda_i d_{ij} W_{jr} + \sum_{j=0}^{J-1}\sum_{r=0}^{R-1} \mu_{ij} Y_{jr} \tag{2.4a}$$

$$\text{s.t. } \sum_{j=0}^{J-1} Y_{jr} + \sum_{s=0}^{r-1} Y_{Js} = 1 \quad \forall\ 0 \le r \le R \tag{2.4b}$$

$$\sum_{r=0}^{R-1} Y_{jr} \le 1 \quad \forall\ 0 \le j \le J-1 \tag{2.4c}$$

$$\sum_{r=0}^{R} Y_{Jr} = 1 \tag{2.4d}$$

$$P_{j0} = 1 - q_j \quad \forall\ 0 \le j \le J \tag{2.4e}$$

$$P_{jr} = (1-q_j)\sum_{k=0}^{J-1} \frac{q_k}{1-q_k} W_{k,r-1} \quad \forall\ 0 \le j \le J,\ 1 \le r \le R \tag{2.4f}$$

$$Y_{jr} \in \{0,1\} \quad \forall\ 0 \le j \le J,\ 0 \le r \le R \tag{2.4g}$$

$$(2.2\text{a}) - (2.2\text{d}). \tag{2.4h}$$

We propose two methods to solve the relaxed subproblem: one exact algorithm that finds the optimal customer assignment, and one fast approximate algorithm that provides an lower bound.

2.4.1 An Exact Algorithm

Following a similar argument to Proposition 2, given the subset of facilities that serve a certain customer, it is optimal to assign this customer to the facilities level by level in increasing order of the distances. Therefore the objective value of (RSP_i) only depends on the set of facilities that serve customer i. Define $\Phi_i(S)$ to be the minimum cost to serve customer i, using only facilities in S; i.e.,

$$\Phi_i(S) = \text{Min } \sum_{j=0}^{J}\sum_{r=0}^{R} h_i d_{ij} W_{jr} + \sum_{j\in S} \mu_{ij} \tag{2.5a}$$

$$\text{s.t. } (2.4\text{b}) - (2.4\text{g}) \tag{2.5b}$$

$$\sum_{r=0}^{R-1} Y_{jr} = 0 \quad \forall\ j \in \{1, \cdots, J-1\} \setminus S. \tag{2.5c}$$

It is clear that (RSP_i) is equivalent to the following minimization of a set function (MSF_i):

$$
\begin{aligned}
(\text{MSF}_i) \quad & \text{Min } \Phi_i(S) && \text{(2.6a)}\\
& \text{s.t. } S \subseteq \{0, \cdots, J-1\} && \text{(2.6b)}\\
& \quad |S| \leq R. && \text{(2.6c)}
\end{aligned}
$$

We solve (MSF) using a special branch-and-bound algorithm, based on some unique properties of the set function Φ, as described in Proposition 3.

Proposition 3. *The set function Φ_i is supermodular, for all $i = 0, \cdots, I-1$.*

The minimization of a supermodular set function can be solved more efficiently, using the branch-and bound algorithm developed by Goldengorin et al. Goldengorin et al. [1999]. The algorithm keeps track of A and B, the set of facilities that have been forced in or out for each branch-and-bound node. The supermodularity of the objective function allows us to force out a facility if its addition to set A does not reduce the total cost. In an unconstrained problem, it is also possible to force in a facility if its deletion from $\{0, \cdots, J-1\} \setminus B$ increases the total cost. However, since MSF is subject to the cardinality constraint (2.6c), the second option does not apply here.

2.4.2 An Approximate Solution

Although the exact algorithm in Section 2.4.1 takes advantage of special structure of the problem, its worst case complexity is still exponential. In this section we provide a fast approximate algorithm that finds lower bounds for the Lagrangian procedure.

In our approximate solution, we replace the variable probability P_{jr} with fixed numbers. Let $j_0, j_1, \cdots, j_{J-1}$ be an ordering of the facilities such that $q_{j_0} \leq q_{j_1} \leq \cdots \leq q_{J-1}$. Define

$$
\begin{aligned}
\alpha_r &= (1 - q_{j_r}) \prod_{\ell=0}^{r-1} q_{j_\ell} \\
\beta_r &= \prod_{\ell=0}^{r-1} q_{j_\ell}.
\end{aligned}
$$

We define a reformulation of the relaxed subproblem (RRSP) by replacing P_{jr}

with α_r if $0 \leq j \leq J-1$, and replacing P_{Jr} with β_r:

$$(\text{RRSP}_i) \quad \text{Min} \sum_{j=0}^{J-1} \sum_{r=0}^{R-1} (\lambda_i d_{ij} \alpha_r + \mu_{ij}) Y_{jr} + \sum_{r=0}^{R} \lambda_i d_{iJ} \beta_r Y_{Jr} \tag{2.7a}$$

$$\text{s.t.} \sum_{j=0}^{J-1} Y_{jr} + \sum_{s=0}^{r-1} Y_{Js} = 1 \quad \forall\ 0 \leq r \leq R \tag{2.7b}$$

$$\sum_{r=0}^{R-1} Y_{jr} \leq 1 \quad \forall\ 0 \leq j \leq J-1 \tag{2.7c}$$

$$\sum_{r=0}^{R} Y_{Jr} = 1 \tag{2.7d}$$

$$Y_{jr} \in \{0,1\} \quad \forall\ 0 \leq j \leq J,\ 0 \leq r \leq R. \tag{2.7e}$$

The following proposition states that we can solve (RRSP) for a lower bound of (RSP).

Proposition 4. *The (RRSP) formulation* (2.7a)-(2.7e) *yields a lower bound to the relaxed subproblem* (2.4a)-(2.4h).

We note that the (RRSP) formulation (2.7a)-(2.7e) leads to a combinatorial assignment problem, which can be solved in strongly polynomial time using the Hungarian algorithm Kuhn [2005]. In our numerical tests, we use both the exact and the approximate algorithm to get the best combination of speed and accuracy.

Although our compact MIP formulation and the Lagrangian relaxation algorithm are significant improvements over scenario based stochastic programming formulations, the worst case complexity is still exponential, due to the NP-hardness of the underlying problem. Furthermore, because only numerical results are available from the discrete model, very few managerial insights can be drawn from the optimal solutions. In the next section, we overcome these difficulties by introducing the continuum approximation (CA) model.

2.5 Computational Results

The Lagrangian Algorithm was tested on two types of networks - the "real" network based on the US map with 49 or 88 nodes and the "random" network generated on a unit square region with 50 or 100 nodes (the data set was kindly provided by L. Snyder and is available from his website Snyder). The failure probabilities q_j in the real networks are calculated using $q_j = 0.1e^{-D_j/400}$, in which D_j is the great cycle distance (in miles) between location j and New Orleans, LA. In the random networks, q_j

are randomly generated from a uniform distribution between 0 and 0.2. For each data set, we test our algorithm for $R = 2$, 3 and 4. The Lagrangian relaxation/branch-and-bound procedure is executed to a tolerance of 0.5%, or up to 3600 seconds (60 minutes) in CPU time. The algorithm was coded in C++ and tested on an Intel Pentium 4 3.20GHz processor with 1.0 GB RAM under Linux. Parameter values for the Lagrangian relaxation algorithm can be found in Table 2.3, and the algorithm performance is summarized in Table 2.4.

Parameter	Value
Optimal tolerance	0.005
Maximum number of approximate iterations at root node	1000
Maximum number of exact iterations at root node	500
Maximum number of approximate iterations at child nodes	200
Maximum number of exact iterations at child nodes	100
Initial value for μ_{ij}	optimal dual of LP relaxation

Table 2.3. Parameter values for the Lagrangian relaxation

Nodes	R	Root LB	Root UB	Root gap	Overall UB	Overall gap	CPU time
49	2	875,899	880,098	0.479	880,098	< 0.500	6
49	3	870,417	874,423	0.460	874,423	< 0.500	25
49	4	870,125	874,323	0.483	874,323	< 0.500	49
88	2	122,755	123,365	0.497	123,365	< 0.500	244
88	3	121,743	122,348	0.497	173,5400	< 0.500	419
88	4	121,727	122,329	0.494	122,329	< 0.500	925
50	2	6,332.89	6,362.71	0.471	6,362.71	< 0.500	1
50	3	6,336.73	6,362.71	0.410	6,362.71	< 0.500	2
50	4	6,338.15	6,362.71	0.387	6,362.71	< 0.500	2
100	2	11,881.1	11,981.0	0.841	11,970.2	< 0.500	69
100	3	11,853.3	12,127.9	2.317	11,970.0	< 0.500	94
100	4	11,883.6	12,036.1	1.283	11,970.0	< 0.500	120

Table 2.4. LR Algorithm Performance

We notice that the maximum re-assignment level R does not affect the optimal facility locations in all of our test instances, although a higher R in general helps to reduce the optimal cost. Figure 2.5 and 2.6 illustrate the optimal facility locations for the 49-node and the 88-node problem respectively. Table 2.5 and 2.6 list the percentage of covered demand, fixed cost, and failure probability at each optimal facility location in the two problem instances. In both cases, the optimal solutions avoid highly risky areas such as LA and MS. In areas with moderate risk, clusters of facilities are formed to hedge against possible disruptions. In areas with low risk (OR, CA and AZ), facilities are located relatively sparsely.

Our algorithm appears to have performed efficiently on the random test instances. However, the algorithm convergence is slow for some of the real test instances. Due to

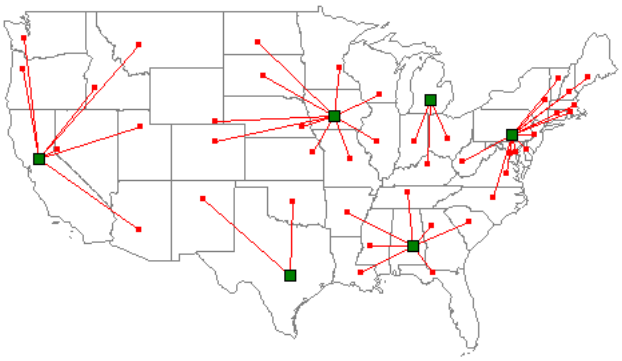


Figure 2.5. Optimal Solution to the 49-Node Problem

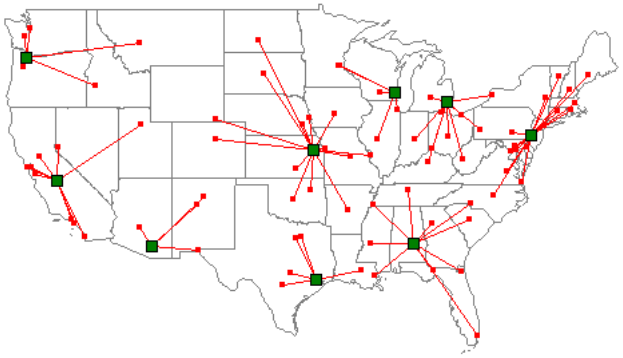


Figure 2.6. Optimal Solution to the 88-Node Problem

the computational complexity of finding exact solutions for the relaxed subproblems (RSP), we can only afford to run the exact algorithm for a very limited number of iterations (100 at each B&B node as compared to 2000 in Snyder and Daskin Snyder and Daskin [2005]), The approximate algorithm for RSP is fast, but the bound it provides can be lax in some circumstances. For fixed-charge location problems, it is generally more efficient to relax the assignment constraint (2.1b) instead of the linking constraint (2.1c). However, in our case relaxing (2.1c) allows us to decompose the customer assignment problem, a key step in the algorithm development. To improve the efficiency of the algorithm, we introduce a new algorithm in Chapter 3 based on neighborhood search and cutting plain procedures.

Location	Demand Covered	Fixed Cost	Failure Probability
Sacramento, CA	19%	115,800	0.001
Austin, TX	9%	72,600	0.043
Harrisburg, PA	29%	38,400	0.012
Lansing, MI	12%	48,400	0.013
Montgomery, AL	17%	62,200	0.053
Des Moines, IA	15%	49,500	0.014

Table 2.5. Optimal Locations for the 49-Node Problem

Location	Demand Covered	Fixed Cost	Failure Probability
Houston, TX	10%	58,000	0.043
Philadelphia, PA	27%	49,400	0.012
Detroit, MI	10%	25,600	0.013
Milwaukee, WI	10%	53,500	0.014
Portland, OR	3%	59,200	0.001
Tucson, AZ	5%	66,800	0.001
Fresno, CA	17%	80,300	0.001
Montgomery, AL	9%	62,200	0.053
Topeka, KS	8%	48,800	0.024

Table 2.6. Optimal Locations for the 88-Node Problem

Chapter 3

Reliable Facility Location: the Search-and-Cut Algorithm

3.1 Introduction

In this chapter, we study the reliable version of the *uncapacitated fixed-charge location problem* (UFLP), although our model can be easily extended to address other facility location problems. The uncertainty of facility failure is specified by individual and independent failure probability inherent at each facility. To hedge against failures within the system, each customer is assigned to multiple facilities, ordered by levels. The facility at the lowest level is a customer's primary facility, which will serve the customer as long as it remains operational. A facility at a higher level is the customer's backup facility, and it only serves the customer when all facilities at lower levels have failed. In the rare event that all facilities assigned to a customer have failed, a penalty cost is incurred, which can be taken as the loss of goodwill, or the cost to serve the customer at a competitor's facility. Our goal is to minimize the sum of fixed location costs, the expected transportation costs at all levels, and the expected penalty costs. This problem will be referred to as the *reliable uncapacitated fixed-charge location problem* (RUFLP).

This chapter can be viewed as an extension of Chapter 2, where we do not limit the customer assignment levels. Using a combination of neighborhood search and cutting plane process, our search-and-cut algorithm that outperform the Lagrangian algorithm in Chapter 2 in both execution time and solution quality, especially for larger R. The new algorithm also works as a heuristic, which can solve extremely large problem instances.

The remainder of the chapter is organized as follows. In Section 3.2, we introduce

the formulation of the RUFLP problem. Section 3.3 is dedicated to solution algorithms for RUFLP, with Section 3.3.1 introducing a linear reformulation that provides a lower bound, Section 3.3.2 discussing an exact algorithm, and Section 3.3.3 proposing an approximate solution. Numerical experiment design and computational results are discussed in Section 3.4.

3.2 Formulation

Let $N(|N| = n)$ be the set of customer demand aggregation points and $M(|M| = m)$ be the set of candidate locations for the facilities. We denote demand rate at node i as λ_i for each $i \in N$, at the fixed cost to locate a facility at node j as f_j for each $j \in M$. Let d_{ij} be the unit cost to serve demand from node i at a facility located at node j. We model facility disruptions as independent events, happening at location $j \in M$ with probability $0 \leq q_j < 1$.

To hedge against disruption risk, each customer is assigned to up to $R > 1$ facilities, and can be served by these and only these facilities. A penalty cost ϕ_i is incurred for each unit of unmet demand due to facility failures, which can be taken as the lost of good will, or the cost to serve the customers at a competitor's facility.

We will use $S \subset M$ to denote the set of facilities selected. If a facility is located at site j, we call it facility j.

Define $S[i] = \{k \mid \phi_i \geq d_{ik}, k \in S\}$ to be the set of facilities in S whose unit shipping cost to customers at $i \in N$ is lower than the unit penalty cost. Define $i[r] \in S$ $(r = 1, 2, .., |S[i]|)$ to be the facility that serves customers at $i \in N$. at level r. Assume $q_{0[i]} = 1$ for $i \in N$. Define $C_i(S)$ to be the sum of shipping cost and lost sales cost of one unit of demand to customers at $i \in N$. Then, it can be verified that

$$C_i(S) = \left(\sum_{r=1}^{|S[i]|} \left(\prod_{t=0}^{r-1} q_{i[r]} \right) \left(1 - q_{i[r]}\right) d_{i,i[r]} + \left(\prod_{t=0}^{|S[i]|} q_{i[t]} \right) \phi_i \right). \tag{3.1}$$

Define $F(S)$ to be the total cost of shipping and fixed facilty given location set S, such that

$$F(S) = \sum_{j \in S} f_j + \sum_{i \in N} \lambda_i C_i(S). \tag{3.2}$$

The Reliable Uncapacitated Fixed-Charge Location problem (RUFL) is formulated:

$$\min_{S \subseteq M} \{F(S)\}.$$

Define $E[i] = \{j \mid \phi_i \geq d_{ij}, j \in M\}$ to be the set of facilities in M which their unit shipping cost to customers at $i \in N$ is lower than the unit lost sales cost ϕ_i. Let

$e_i = |E[i]|$. Denote x_j to be a binary variable which is one if we open a facility at $j \in M$ and zero otherwise. Denote y_{ijr} to be a binary variable which is one if facility $j \in M$ is assigned to customers at $i \in N$ at level $r = 1, .., e_i$ and zero otherwise. Define P_{ijr} to be the probability that facility $j \in M$ is assigned to customers at $i \in N$ at level $r = 1, .., e_i$. Then $P_{ij1} = 1 - q_j$ and $P_{ijr} = (1 - q_j) \sum_{k \in M} \frac{q_k}{1-q_k} P_{i,k,r-1} y_{i,k,r-1}$ for $i \in N$, $j \in E[i]$, $2 \leq r \leq e_i$. Given the above defintitions the reliable uncapacitated facility location problem $RUFL$ is formulated as:

$$
\begin{aligned}
\text{(RUFL)} \quad \text{Minimize } Z = & \sum_{j \in M} f_j x_j + \sum_{i \in N} \sum_{j \in E[i]} \sum_{r=1}^{e_i} \lambda_i d_{ij} P_{ijr} y_{ijr} \\
\text{s.t.} & \sum_{r=1}^{e_i} y_{ijr} \leq x_j \ \forall i \in N, \ j \in E[i], \\
& \sum_{j \in E[i]} y_{ijr} = 1 \ \forall i \in N, \ 1 \leq r \leq e_i, \\
& P_{ij1} = 1 - q_j, \\
& P_{ijr} = (1 - q_j) \sum_{k \in M} \frac{q_k}{1 - q_k} P_{i,k,r-1} y_{i,k,r-1} \ \forall i \in N, \ j \in E[i], \ 2 \leq r \leq e_i, \\
& x_j \in \{0, \ 1\} \ \forall j \in M, \\
& y_{ij} \in \{0, \ 1\} \ \forall i \in N, \ j \in E[i].
\end{aligned}
$$

Note that $RUFL$ is a nonlinear mixed integer program which is large-scale in nature. Unlike in Chapter 2, R, the number of facilities assigned to a customer, is not limited to a certain fixed value and customers are able to be assigned at all open facilities with a shipping cost less than unit slaes lost cost, which is in fact more realistic. When R is fixed the above model can be rewritten by replacing $max\{R, e_i\}$ with e_i for $i \in N$.

Here we develop a solution approach which finds an optimal solution more efficiently regardless of whether R is fixed or not. To do this we first simplify the customer assignment assumption such that customers are assigned to open facilities level by level in an increasing order of shipping cost. This might not be the optimal customer assignment, but later we show that even with this assumptions our solution approach is frequently able to come up with better solutions with less amount of time compared to the lagrangian relaxation algorithm in Chapter 2. We also develop an efficient approximate approach which is capable of solving large problem instances.

3.3 Algorithms for $RUFL$

Our exact and approximate solution approaches is based on obtaining efficient lower bounds for $RUFL$. We describe the lower bound in Section 3.1. The exact

approach presented in 3.2 is based on finding succesive improved lower bounds and their corresponding upper bounds. In Section 3.3, we prsent heuristics which are in fact partial solutions to exact approach.

3.3.1 A Lower Bound for $RUFL$

Let $q_{[1]} \leq q_{[2]} \leq ... \leq q_{[m-1]} \leq q_{[m]}$ be an ordering of failure probabilities. Define $P_r = \left(\prod_{t=1}^{r-1} q_{[r]}\right)\left(1 - q_{[r]}\right)$. By replacing P_{ijr} with fixed failure probabilities P_r $\forall i \in N, j \in E[i]$ in $RUFL$ will result the following mixed integer program which we call $RMIP$:

$$\text{(RMIP)} \quad \text{Minimize} Z = \sum_{j \in M} f_j x_j + \sum_{i \in N} \sum_{j \in E[i]} \sum_{r=1}^{e_i} \lambda_i d_{ij} P_r y_{ijr} \tag{3.3}$$

$$\text{s.t.} \sum_{r=1}^{e_i} y_{ijr} \leq x_j \ \forall i \in N,\ j \in E[i], \tag{3.4}$$

$$\sum_{j \in E[i]} y_{ijr} = 1 \ \forall i \in N,\ 1 \leq r \leq e_i, \tag{3.5}$$

$$x_j \in \{0,\ 1\} \ \forall j \in M, \tag{3.6}$$

$$y_{ijr} \in \{0,\ 1\} \ \forall i \in N,\ j \in E[i]. \tag{3.7}$$

Theorem 1. *Define Z^*_{RMIP} to be the optimal value for the objective function of RMIP. Also let S^*_{RUFL} and Z^*_{RUFL} be an optimal location set and the optimal value of the objective function of RUFL, respectively. Z^*_{RMIP} is a lower bound for Z^*_{RUFL}, i.e.*

$$Z^*_{RMIP} \leq Z^*_{RUFL} = \sum_{j \in S^*_{RUFL}} f_j + \sum_{i \in N} \lambda_i C_i(S^*_{RUFL}). \tag{3.8}$$

Proof. First, we introduce an equivalent formulation of (RUFL) by "splitting" the decision variables:

$$y_{jr} = \begin{cases} 1 & \text{if the level } r \text{ facility is in the same distance as facility } j \\ 0 & \text{otherwise.} \end{cases}$$

$$z_{jr} = \begin{cases} 1 & \text{if the level } r \text{ facility has the same failure probability as facility } j \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that RUFL is equivalent to the following problem:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{j\in M} f_j x_j + \sum_{i\in N}\sum_{j\in E[i]}\sum_{r=1}^{e_i} \lambda_i d_{ij} P_{ijr} y_{ijr} \\
\text{s.t.} \quad & \sum_{r=1}^{e_i} y_{ijr} \le x_j \ \forall i \in N,\ j \in E[i], && (3.9)\\
& \sum_{j\in E[i]} y_{ijr} = 1 \ \forall i \in N,\ 1 \le r \le e_i, && (3.10)\\
& P_{ij1} = 1 - q_j, && (3.11)\\
& P_{ijr} = (1-q_j)\sum_{k\in M} \frac{q_k}{1-q_k} P_{i,k,r-1} z_{i,k,r-1} \\
& \forall i \in N,\ j \in E[i],\ 2 \le r \le e_i, && (3.12)\\
& x_j \in \{0,\ 1\}\ \forall j \in M, && (3.13)\\
& y_{ij}, z_{ij} \in \{0,\ 1\}\ \forall i \in N, j \in E[i], && (3.14)\\
& y_{ij} = z_{ij}\ \forall i \in N,\ j \in E[i]. && (3.15)
\end{aligned}$$

If we remove the last constraint (3.15), the customer is allowed to choose an arbitrary combination of transportation cost and failure probability. We call this relaxed problem RELAX:

$$\begin{aligned}
\text{(RELAX)} \quad \text{Minimize} \quad & \sum_{j\in M} f_j x_j + \sum_{i\in N}\sum_{j\in E[i]}\sum_{r=1}^{e_i} \lambda_i d_{ij} P_{ijr} y_{ijr} \\
\text{s.t.} \quad & (3.9) - (3.14).
\end{aligned}$$

Next, we show that the RELAX is equivalent to formulation RMIP, based on the following claim. **Claim 1**. *There exists an optimal solution* $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{P}^*)$ *to RELAX, such that if* $z^*_{ijr} = 1$, $z^*_{ik,r+1} = 1$ *for some* $i \in N$, *then* $q_j \le q_k$. To prove Claim 1, let $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{P})$ be an optimal solution to formulation RELAX, such that $z_{ijr} = 1$, $z_{ik,r+1} = 1$ for some $i \in N$, and $q_j > q_k$. Let u and v be the facilities assigned to customer i at level r and $r+1$, *i.e.* $y_{iur} = 1$ and $y_{iv,r+1} = 1$. We construct a new solution $(\mathbf{x}', \mathbf{y}', \mathbf{z}', \mathbf{P}')$ as follows:

$$\begin{aligned}
\mathbf{x}' &= \mathbf{x};\\
\mathbf{y}' &= \mathbf{y};\\
z'_{h\ell s} &= \begin{cases} 1 & \text{if } h=i, \ell=k,\ s=r \text{ or } h=i,\ell=j,\ s=r+1,\\ 0 & \text{if } h=i,\ell=j,\ s=r \text{ or } h=i,\ell=k,\ s=r+1,\\ z_{h\ell s} & \text{otherwise;}\end{cases}\\
P'_{h\ell s} &= \begin{cases} \frac{1-q_k}{1-q_j} P_{jr} & \text{if } h=i,\ell=k,\ s=r,\\ \frac{q_k(1-q_j)}{1-q_k} P'_{k,s-1} = q_k P_{jr} & \text{if } h=i,\ell=j,\ s=r+1,\\ 0 & \text{if } h=i,\ell=j,\ s=r \text{ or } h=i,\ell=k,\ s=r+1,\\ P_{h\ell s} & \text{otherwise.}\end{cases}
\end{aligned}$$

By construction, $(\mathbf{x}', \mathbf{y}', \mathbf{z}', \mathbf{P}')$ is a feasible solution to RELAX. Define $G(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{P})$ to be the objective value of RELAX associated with solution $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{P})$. The following assertion holds:

$$\begin{aligned} G(\mathbf{x}', \mathbf{y}', \mathbf{z}', \mathbf{P}') - G(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{P}) &= \lambda_i(P'_{ikr}d_{iu} + P'_{ij,r+1}d_{iv} - P_{ijr}d_{iu} - P_{ik,r+1}d_{iv}) \\ &= \lambda_i[d_{iu}(P'_{ikr} - P_{ijr}) + d_{iv}(P'_{ij,r+1} - P_{ik,r+1})] \\ &= \lambda_i\{d_{iu}[\frac{1-q_k}{1-q_j}P_{ijr} - P_{ijr}] - d_{iv}(q_kP_{ijr} - \frac{q_j(1-q_k)}{1-q_j}P_{ijr})\} \\ &= \frac{q_j - q_k}{1-q_j}\lambda_i P_{ijr}(d_{iu} - d_{iv}) \\ &\geq 0, \end{aligned}$$

where the last inequality follows from the fact that $d_{iu} \leq d_{iv}$ (see Proposition 4 in **?**). This implies that if optimal solution does not satisfy the monotonic condition in Claim 1, we can always construct an alternative optimal solution by swapping j and k, which completes the proof of Claim 1. Following from Claim 1, it is straight forward that $\mathbf{P}^*$, the optimal service probability to RELAX satisfies

$$\mathbf{P}^*_{ijr} = P_r, \ \forall i \in N, j \in E[i], 1 \leq r \leq e_i,$$

which implies that $Z^*_{RMIP} = G(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{P}^*) \leq Z^*_{RUFL}$, where the inequality follows from the fact that RELAX is a relaxation of RUFL. □

We note that RMIP is equivalent to its relaxation without the integrality constraints on $\mathbf{y}$. This makes RMIP easy to solve.

3.3.2 An Exact Approach for $RUFL$

We note that any feasible location vector $\mathbf{x}$ including the one produced by solving $RMIP$ generates a feasible solution to $RUFL$. This is achieved by first defining the assignment vector $\mathbf{y}(\mathbf{x})$ using the suumption that customers are assigned to open facilities level by level in an increasing order of shipping cost. Denote $S_\mathbf{x}$ as the set of facility locations under vector $\mathbf{x}$, then the resulting value of $Z_{RUFL}(\mathbf{x}$, $\mathbf{y}(\mathbf{x})) = \sum_{j \in S_\mathbf{x}} f_j + \sum_{i \in N} \lambda_i C_i(S_\mathbf{x})$ provides an upper bound for $RUFL$.

At each step we find an improved lower bound by solving an improved $RMIP$. An improved $RMIP$ is $RMIP$ with additional "cuts" which eliminates the pre-visited vectors from the feasible region (at the first step $RMIP$ is solved without any cuts). The location set found solving $RMIP$ is used as a starting point in a neighbourhood search to find an improved upper bound using a descent appraoch. For each location set in the neighbourhood we find its corresponding sum of shipping and fixed costs assuming customers are assigned to open facilities level by level in an increasing order of shipping cost. The neighborhood of a location set $S_\mathbf{x}$ and the descent approach is defined as follows.

Define $N_k(S)$, the *distance-k neighborhood* of $S \subseteq M$ as

$$N_k(S) = \{S' \subseteq M : \ |S - S'| + |S' - S| \leq k\};$$

i.e. S' is in the distance-k neighborhood of S if the number of non-overlapping elements in the two sets does not exceed k.

Once the neighborhood is well defined, the descent algorithm is straightforward: use the solution to $RMIP$ a starting subset $S_{\mathbf{x}}$; evaluate the change in the value of the objective function for all the subsets in the neighborhood; if an improved subset exists in the neighborhood, move the search to the best vector in the neighborhood. Repeat the process with the new subset until no improved vector exists in the neighborhood. The last subset the solution. Denote $S_{\overline{\mathbf{x}}}$ as the solution subset to the descent approach, then if the resulting vlaue of the objective function $\sum_{j \in S_{\overline{\mathbf{x}}}} f_j + \sum_{i \in N} \lambda_i C_i(S_{\overline{\mathbf{x}}})$ is less than the current upper bound then the solution to the descent appraoch $\overline{\mathbf{x}}$ and $\mathbf{y}(\overline{\mathbf{x}})$ is our new and improved upper bound.

To complete the step, for each starting location set in the descent appraoch we inroduce a "cut" to $RMIP$ to eliminate all of the feasible vectors which arein its neighbourhood and have already been examined. Denote $S_{\widehat{\mathbf{x}}}$ to be an starting location set in the descent approach, then the following constarint will ensure that all location vectors in the neighborhood of $\widehat{\mathbf{x}}$ (and have already been examined) are infeasible:

$$\sum_{j \in S_{\widehat{\mathbf{x}}}} x_j - \sum_{j \in M - S_{\widehat{\mathbf{x}}}} x_j \leq \ |S_{\widehat{\mathbf{x}}}| - k - 1. \tag{3.16}$$

We note that that (3.16) does not make any location vector infeasible unless they are in the neighborhood of $\widehat{\mathbf{x}}$. The addition of this cut to the improved RMIP will help to improve the lower bound in the next steps. The procedure terminates when the gap between the current lower bound and upper bound is within a tolerable limit.

Now consider Problem $RMIP(l)$ as follows:

$$\min \ Z_{RMIP(l)} = \sum_{j \in M} f_j x_j + \sum_{i \in N} \sum_{j \in E[i]} \sum_{r=1}^{e_i} \lambda_i d_{ij} P_r y_{ijr} \qquad (RMIP(l))$$

$$s.t. \qquad (3.4) - (3.7),$$

$$\sum_{j \in S} x_j - \sum_{j \in M - S} x_j \leq \ |S| - k - 1, \text{ for } S \in A_{RMIP(r)} \text{ and } r = 1, 2, .., l - 1. \tag{3.17}$$

Denote $\mathbf{x}^*_{RMIP(r)}$ as the optimal location vector for $RMIP(r)$, and $S_{\mathbf{x}^*_{RMIP(r)}}$ as the set of facility locations under vector $\mathbf{x}^*_{RMIP(r)}$, then $A_{RMIP(r)}$ in (3.17) is the set of all the starting subsets in the descent approach with $S_{\mathbf{x}^*_{RMIP(r)}}$ as the original starting subset. Therefore, to solve $RMIP(l)$, we need to solve $RMIP(r)$ and the descent approach with $S_{\mathbf{x}^*_{RMIP(r)}}$ as the starting subset for $r = 1, 2, .., l - 1$. Note that

$RMIP(1)$ does not include constraint (3.17) and is the original $RMIP$ by definition. Note that constraints (3.17) ensures that all of the subsets that have already been examined to become infeasible without enforcing infeasibility to the subsets which have not been examined before.

Denote $\mathbf{x}(r)$ and $\mathbf{y}(\mathbf{x}(r))$ to be respectively the location and customer assignment solution vectors for the best found solution after solving $RMIP(r)$ and performing descent approach with $S_{\mathbf{x}^*_{RMIP(r)}}$ as the original starting subset. Note that $Z_{RUFL}(\mathbf{x}(r)\ ,\ \mathbf{y}(\mathbf{x}(r)))$ is the resulting value of the objective function. Define $UB(r)$ $= \min\{UB(r-1), Z_{RUFL}(\mathbf{x}(r)\ ,\ \mathbf{y}(\mathbf{x}(r)))\}$ as the improved upper bound after solving $RMIP(r)$ and performing descent approach with $S_{\mathbf{x}^*_{RMIP(r)}}$ as the original starting subset. Denote $\mathbf{x}_{UB(r)}$ and $\mathbf{y}(\mathbf{x}_{UB(r)})$ to be the corresponding location and customer assignment solution vectors of the upper bound found $UB(r)$, respectively, such that $UB(r) = Z_{RUFL}(\mathbf{x}_{UB(r)}\ ,\ \mathbf{y}(\mathbf{x}_{UB(r)})) = \sum_{j\in S_{\mathbf{x}_{UB(r)}}} f_j + \sum_{i\in N} \lambda_i C_i(S_{\mathbf{x}_{UB(r)}})$. We note that $UB(r)$, is non-increasing in r such that $UB(1) \geq UB(2) \geq ... \geq UB(r-1) \geq UB(r)$. From formulation of $RMIP(l)$ we conclude that the optimal value of the objective function in $RMIP(r)$, is non-decreasing in r such that $Z^*_{RMIP(1)} \leq Z^*_{RMIP(2)} \leq ... \leq Z^*_{RMIP(r-1)} \leq Z^*_{RMIP(r)}$.

Assume a specified tolerence level $\epsilon \geq 0$, then the exact approach based on consequent improvements in lower and upper bounds is described as follows:

The Search-and-Cut Algorithm

Step 0: Set $l = 1$, $S_{\mathbf{x}^*} = \{\}$, and $Upper\ Bound = \infty$.

Step 1: Solve $RMIP(l)$ and find $\mathbf{x}^*_{RMIP(l)}$and $\mathbf{y}(\mathbf{x}^*_{RMIP(l)})$ the optimal location and customer assignment solution vectors in $RMIP(l)$, and $S_{\mathbf{x}^*_{RMIP(l)}}$ as the set of facility locations under vector $\mathbf{x}^*_{RMIP(l)}$, and set $Z^*_{RMIP(l)} = \sum_{j\in S_{\mathbf{x}^*_{RMIP(l)}}} f_j + \sum_{i\in N} \lambda_i C_i(S_{\mathbf{x}^*_{RMIP(l)}})$. Set $Lower\ Bound = Z^*_{RMIP(l)}$. If $\frac{Upper\ Bound - Lower\ Bound}{Lower\ Bound} < \epsilon$ then go to Step 4. Otherwise, go to step 2.

Step 2: Perform the descent approach with $S_{\mathbf{x}^*_{RMIP(r)}}$ as the original starting subset and find $A_{RMIP(l)}$−the set of all the starting subsets in the descent approach; $\mathbf{x}(l)$, $\mathbf{y}(\mathbf{x}(l))$−the location and customer assignment solution vectors of the best found solution; and $Z_{RUFL}(\mathbf{x}(l)\ ,\ \mathbf{y}(\mathbf{x}(l))) = \sum_{j\in S_{\mathbf{x}(l)}} f_j + \sum_{i\in N} \lambda_i C_i(S_{\mathbf{x}(l)})$−the objective function value.

Step 3: If $Upper\ Bound > Z_{RUFL}(\mathbf{x}(l)\ ,\ \mathbf{y}(\mathbf{x}(l)))$, then set $Upper\ Bound = Z_{RUFL}(\mathbf{x}(l), \mathbf{y}(\mathbf{x}(l)))$, $\mathbf{x}^* = \mathbf{x}(l)$, $\mathbf{y}(\mathbf{x}^*) = \mathbf{y}(\mathbf{x}(l))$ and $S_{\mathbf{x}^*} = S_{\mathbf{x}(l)}$. Set $l = l + 1$ and go to Step 1.

Step 4: Stop. Optimal location set is $S_{\mathbf{x}^*}$, optimal location vector is $\mathbf{x}^*$, optimal customer assignment vector is $\mathbf{y}(\mathbf{x}^*)$, and the optimal objective function value is $Z^{\mathcal{A}} = Upper\ Bound$.

3.3.3 Approximate Approaches for $RUFL$

Next we will present heuristic $H(l)$ which is an approximate solution approach based on the solving $RMIP(l)$ and performing descent approach with $S_{\mathbf{x}^*_{RMIP(l)}}$ as the original starting subset.

Heuristic $H(l)$

Step 0: Set $r = 1$, $S_{H(0)} = \{\}$, and $UB(0) = \infty$.

Step 1: If $r > l$ go to Step 2, otherwise go to Step 1-1.

Step 1-1: Solve $RMIP(r)$ and find $S_{\mathbf{x}^*_{RMIP(r)}}$. Set $\overline{S} = S_{\mathbf{x}^*_{RMIP(r)}}$, $\widehat{S} = \overline{S}$, $S_{H(r)} = \overline{S}$, and $A_{RMIP(r)} = \{\overline{S}\}$. Find $C_i(\widehat{S})$ for $i \in N$ using (3.1) and $F(\widehat{S})$ using (3.2).

Step 1-2: Evaluate all the subsets in the distance-k neighborhood of $\overline{S}$, and find $\hat{S}$, the subset with the minimum objective value.

Step 1-3: If $F(\widehat{S}) < F(\overline{S})$ then $\overline{S} = \widehat{S}$, $S_{H(r)} = \widehat{S}$, $A_{RMIP(r)} = A_{RMIP(r)} \cup \{\widehat{S}\}$ and go back to Step 1-1. Otherwise go to Step 1-4.

Step 1-4: Set $UB(r) = \min\{F(S_{H(r)}), UB(r-1)\}$. If $UB(r) = UB(r-1)$ then $S_{H(r)} = S_{H(r-1)}$. Set $r = r + 1$ and repeat Step 1.

Step 2: Stop. Output location set is $S_{H(l)}$, and objective function value is $UB(l)$.

We note that for different values of l we may end up with differnt solutions. Also it is clear that in order to solve heuristic $H(l)$, we must succesively solve heuristics $H(1)$, $H(2)$, .., $H(l-2)$, and $H(l-1)$.

3.4 Computational Results

We tested our exact and heuristic algorithms on two different types of data. The exact algorithm is tested on four data sets with 50, 75, 100, and 150 nodes. These data sets are based on 1990 census data, with each node representing one of the 50, 75, 100 or 150 largest cities in the U.S. Demands λ_i are set to the city population divided by 10^4, while the fixed cost f_j is set to the median home value in the city. The transportation cost d_{ij} is calculated based on the great circle distance between node i and j.

In all four data sets, the set of facilities M is equal to the set of customers N (each demand point is a potential facility site). Penalty cost ϕ_i is set to 10,000 for each customer, and the failure probabilities q_j are calculated using $q_j = \beta + 0.1\alpha e^{-d_j/400}$, where $\beta = 0.01$, and d_j is the great cycle distance (in miles) between node j and New

Orleans, LA. For each data set, we fix $\alpha = 1.0$, and vary the maximum assignment level R from 3 to 10; then set at $R = M$. We also fix $R = 4$ and vary α from 1.05 to 1.45 at 0.05 increment.

The heuristic algorithm $H(1)$ based on distance-2 neighborhood is tested on data sets based on dense networks with up to 600 nodes (the data set was kindly provided by J.E. Beasley and is available from his website Beasley (2009)). The transportation cost d_{ij} is calculated based on the shortest path distance between node i and j. Demands and fixed costs are randomly generated from uniform distributions between 10 and 110, and between 1,000 and 11,000, respectively. Penalty cost is set to be 1,000 for each customer, and failure probabilities are randomly generated from a uniform distribution between 0.01 and 0.11.

Our algorithms are coded in C++ and tested on an Intel Pentium 4 3.20GHz processor with 1.0 GB RAM under Linux. The neighborhood search-cutting plane procedure is executed to a precision of 0.005, or up to 3600 seconds in CPU time in the exact algorithm, and executed for a single iteration in the heuristic algorithm. For each "real" test instance, we report the computational times of three different search-and-cut algorithms, based on distance-k neighborhood, where $k \in \{1, 2, 3\}$.As a comparison to our exact algorithm, we also test the Lagrangian Relaxation algorithm of Cui et al. (2009). To simplify presentation, from now on we will refer the Lagrangian Relaxation algorithm of Cui et al. (2009) as the "LR method". The test results are summarized in Tables 3.1 - 3.4 for the exact algorithm, and in Table 3.5 for the heuristic algorithm.

Table 3.1. Performance of Exact Algorithms- 50 Nodes

Nodes	R	alpha	SnC UB	LR UB	SnC-1 Time	SnC-2 Time	SnC-3 Time	LR Time
50	3	1	1,021,060	1,020,980	14	9	6	23
50	4	1	1,020,540	1,020,540	26	16	14	54
50	5	1	1,020,520	1,020,520	28	16	12	84
50	6	1	1,020,520	1,020,520	38	26	16	180
50	7	1	1,020,520	1,020,520	40	27	15	273
50	8	1	1,020,520	1,020,520	44	34	25	626
50	9	1	1,020,520	1,020,520	43	33	20	908
50	10	1	1,020,520	1,020,520	47	31	20	1250
50	50	1	1,020,520	-	56	38	26	-
50	4	1.05	1,021,410	1,023,590	30	16	14	67
50	4	1.1	1,022,280	1,026,650	39	18	14	121
50	4	1.15	1,023,160	1,029,710	56	20	15	136
50	4	1.2	1,024,030	1,032,640	90	30	22	133
50	4	1.25	1,024,910	1,035,590	115	36	27	256
50	4	1.3	1,025,790	1,038,550	174	47	34	208
50	4	1.35	1,026,670	1,041,520	194	51	35	488
50	4	1.4	1,027,540	1,044,510	240	61	41	313
50	4	1.45	1,028,370	1,047,510	290	77	40	474

Table 3.2. Performance of Exact Algorithms- 75 Nodes

Nodes	R	alpha	SnC UB	LR UB	SnC-1 Time	SnC-2 Time	SnC-3 Time	LR Time
75	3	1	1,149,130	1,149,070	27	17	28	195
75	4	1	1,148,590	1,148,590	38	26	26	273
75	5	1	1,148,580	1,148,580	52	29	35	382
75	6	1	1,148,580	1,148,580	95	53	52	540
75	7	1	1,148,580	1,148,580	101	73	67	708
75	8	1	1,148,580	1,148,580	124	89	67	2098
75	9	1	1,148,580	1,148,580	134	81	74	2382
75	10	1	1,148,580	1,148,580	135	81	76	2444
75	75	1	1,148,580	-	177	100	95	-
75	4	1.05	1,149,600	1,152,670	48	31	35	229
75	4	1.1	1,150,600	1,156,310	60	31	35	254
75	4	1.15	1,151,610	1,160,000	78	39	40	366
75	4	1.2	1,152,610	1,163,720	107	52	40	621
75	4	1.25	1,153,620	1,167,500	138	68	50	824
75	4	1.3	1,154,630	1,171,320	181	84	56	974
75	4	1.35	1,155,640	1,175,190	223	99	73	1518
75	4	1.4	1,156,660	1,179,110	322	128	91	1915
75	4	1.45	1,157,670	1,183,090	437	147	105	2216

Table 3.3. Performance of Exact Algorithms- 100 Nodes

Nodes	R	alpha	SnC UB	LR UB	SnC-1 Time	SnC-2 Time	SnC-3 Time	LR Time
100	3	1	1,254,080	1,253,910	393	194	255	3612
100	4	1	1,253,010	1,253,010	1662	488	358	3672
100	5	1	1,252,990	1,252,990	1968	655	547	3707
100	6	1	1,252,990	1,254,560	3568	1103	691	3621
100	7	1	1,252,990	1,252,990	3631	1142	671	3745
100	8	1	1,252,990	1,253,460	3658	1315	717	3785
100	9	1	1,252,990	1,254,310	3643	1265	849	3660
100	10	1	1,252,990	1,253,460	3662	1261	885	3740
100	100	1	1,252,990	-	3656	1476	1066	-
100	4	1.05	1,254,040	1,256,560	2668	738	525	3612
100	4	1.1	1,255,060	1,260,120	3361	854	635	3615
100	4	1.15	1,256,060	1,263,280	3630	1207	692	3664
100	4	1.2	1,256,920	1,267,070	3636	1500	863	3618
100	4	1.25	1,257,780	1,269,310	3632	1848	1233	3632
100	4	1.3	1,258,640	1,274,070	3635	2301	1329	3749
100	4	1.35	1,259,500	1,275,410	3633	3106	1607	3633
100	4	1.4	1,260,370	1,277,030	3628	3633	1759	3613
100	4	1.45	1,261,240	1,281,280	3635	3641	1903	3610

Table 3.4. Performance of Exact Algorithms- 150 Nodes

Nodes	R	alpha	SnC UB	LR UB	SnC-1 Time	SnC-2 Time	SnC-3 Time	LR Time
150	3	1	1,363,780	1,371,000	451	215	828	3706
150	4	1	1,362,630	1,369,790	1315	781	1034	4128
150	5	1	1,362,600	-	2265	1180	1463	-
150	6	1	1,362,600	-	3661	2292	2041	-
150	7	1	1,362,600	-	3620	2294	2550	-
150	8	1	1,362,600	-	3697	2583	2696	-
150	9	1	1,362,600	-	3608	2722	2729	-
150	10	1	1,362,600	-	3624	2845	2904	-
150	150	1	-	-	-	-	-	-
150	4	1.05	1,363,630	1,368,280	1919	1044	1109	3610
150	4	1.1	1,364,640	1,369,450	2727	1328	1240	4110
150	4	1.15	1,365,640	1,372,880	3671	1731	1474	4257
150	4	1.2	1,366,650	1,382,480	3629	2650	1809	4076
150	4	1.25	1,367,660	1,385,280	3663	3148	2422	4330
150	4	1.3	1,368,670	1,383,290	3670	3629	3319	4036
150	4	1.35	1,369,680	1,392,260	3673	3638	3618	4124
150	4	1.4	1,370,690	1,395,780	3606	3636	3619	4754
150	4	1.45	1,371,710	1,400,550	3679	3624	3621	3730

Both this chapter and the previous one address discrete models of the reliable facility location problems, for which only computational results are available and very few insights can be drawn from the optimal solutions. To provide more managerial insights, we develop a Continuum Approximation (CA) model in Chapter 4. The CA model precisely predicts the system costs by focusing on important decisions like the number of facilities and the influential areas, omitting details of facility locations and customer assignments. Since the system cost is approximated as continuous function of the key parameters, the CA model is a handy tool for sensitivity analysis. Furthermore, the continuous results from the CA model can be translated into a discrete facility location design, making the CA model an alternative heuristic approach to the discrete models.

Table 3.5. Performance of Heuristic Algorithm

Index	Nodes	LB	UB	Gap	CPU Time
1	100	56005	59104	0.055	1
2	100	54119	57506	0.063	0
3	100	57061	60381	0.058	1
4	100	59160	62091	0.050	1
5	100	51212	54949	0.073	1
6	200	79185	82525	0.042	6
7	200	82317	85734	0.042	5
8	200	82458	85961	0.042	5
9	200	79602	82822	0.040	5
10	200	75862	79467	0.048	9
11	300	99306	102391	0.031	14
12	300	97632	101517	0.040	14
13	300	100576	103351	0.028	21
14	300	100142	103767	0.036	16
15	300	97933	101408	0.035	17
16	400	109138	112694	0.033	34
17	400	107013	109946	0.027	231
18	400	115546	118974	0.030	99
19	400	111601	115729	0.037	294
20	400	109719	113261	0.032	177
21	500	117043	121380	0.037	238
22	500	122883	126744	0.031	77
23	500	124377	128559	0.034	1821
24	500	119575	123328	0.031	351
25	500	122169	125422	0.027	90
26	600	130941	137923	0.053	970
27	600	130296	137564	0.056	1602
28	600	126871	129713	0.022	1047
29	600	130393	134152	0.029	318
30	600	138171	141906	0.027	279

Chapter 4

Reliable Facility Location: a Continuum Approximation Approach

4.1 Introduction

Most of the discrete location models are NP-hard and thus it is difficult to obtain good solutions for large problem instances within a limited time frame. This fact motivates research on the continuum approximation (CA) method as an alternative to solving large-scale facility location problems. Building on the earlier work in Newell Newell [1971, 1973] and Daganzo Daganzo [1984a,b], Daganzo and Newell Daganzo and Newell [1986] propose a CA approach for the traditional facility location problem. While conditions are slowly-varying, the cost of serving the demand near a facility location is formulated as a function of a continuous facility density (number of facilities per unit area) that can be efficiently optimized in a point-wise way. Note that the inverse of facility density is the influence area size (area per facility). The optimization yields the desired facility density and influence area size near each candidate location, which informs the design of discrete facility locations. It is shown in various contexts that the CA approach gives good approximate solutions to large-scale logistics problems by focusing on key physical issues such as the facility size and demand distribution Hall [1984, 1986, 1989], Campbell [1993a,b], Daganzo and Erera [1999], Dasci and Verter [2001]. See Langevin et al. Langevin et al. [1996] and Daganzo Daganzo [2005] for reviews of the CA model. Ouyang and Daganzo Ouyang and Daganzo [2006] and Ouyang Ouyang [2007] propose methods to efficiently transform output from the CA model into discrete design strategies. The former reference also analytically validates the CA method for the traditional facility location problem. Recently, Lim et al. Lim et al. propose a reliability CA model for facility location problems with uniform customer density. For simplification, a specific type of failure-proof facility is assumed to exist; a customer is always re-assigned to a failure-proof

facility after its nearest regular facility fails, regardless of other (and nearer) regular facilities. We relax these rather strong assumptions in our work.

The planar version of the reliable facility location problem is defined over a large set of customers in the continuous metric space $\mathcal{S} \subseteq \mathbb{R}^2$, where the demand rate λ, fixed cost f, failure probability q and the penalty cost ϕ are continuous functions of the location $x \in S$. All these spatial attributes are assumed to vary continuously and slowly in x. Suppose that the cost units are set so that the transportation cost for serving a unit demand at x by a facility at x_j is equal to the distance measured by the Euclidean metric, $\|x - x_j\|$. In addition, we assume that $\phi(x) \geq \max\{\|x - x_j\| : \forall x_j \in S\}$, for all $x \in S$. Under such assumption, a customer shall always be assigned to R facilities if available.

Given any solution with $n > 0$ facilities located at $\mathbf{x} = \{x_1, \cdots x_n\}$, the demand at $x \in S$ could potentially be served by a subset of facilities or not be served at all. We denote the customer assignment plan by $\mathbf{y} = \{(y_1(x), ..., y_R(x)) : \forall x \in \mathcal{S}\}$, where $y_k(x)$ is the index of the facility assigned as the k-th choice to the customer at x. For any given design $\mathbf{x}, \mathbf{y}$, we use $\bar{P}(x|\mathbf{x}, \mathbf{y})$ to denote the probability that the demand at x is not served, while $P(x, x_j|\mathbf{x}, \mathbf{y})$ is the probability that this demand is served by facility j. These probabilities depend on the set of facility distances, $\{\|x - x_j\| : \ j = 1, \cdots, n\}$, the maximum reassignment level R, and the facility failure scenarios, but they must sum up to 1; i.e.,

$$\bar{P}(x|\mathbf{x}, \mathbf{y}) + \sum_{j=1}^{n} P(x, x_j|\mathbf{x}, \mathbf{y}) = 1, \ \forall x \in \mathcal{S}. \tag{4.1}$$

We will derive these probability functions in the next section.

The total expected cost includes three components: fixed facility charges, expected transportation costs for served demand, and expected penalty costs for unserved demand. The optimization problem can now be formulated as follows:

$$\min_{\mathbf{x},\mathbf{y}} \ \sum_{j=1}^{n} f(x_j) + \int_{x \in \mathcal{S}} \left[\phi(x)\bar{P}(x|\mathbf{x}, \mathbf{y}) + \sum_{j=1}^{n} \|x - x_j\| P(x, x_j|\mathbf{x}, \mathbf{y}) \right] \lambda(x) dx. \tag{4.2}$$

In (4.2), the first term is the total fixed facility charges. The integral term is the total expected cost for serving (or not serving) all customer demand in $\mathcal{S}$. The first part of the integrand corresponds to the scenario where the customer at x is not served, incurring a penalty cost of $\phi(x)$. The second part is the expected transportation distance for the customer at x to obtain service.

4.2 Infinite Homogeneous Plane

We first consider the case where $\mathcal{S} = \mathbb{R}^2$, and all parameters, λ, ϕ, q, f, are constant everywhere. We will first identify optimal results for this simpler case and then use them as building blocks to design solution methods for more general cases.

Throughout this section, we focus on the non-trivial case where $q < 1$. Obviously, on a homogeneous plane, given any set of locations $\mathbf{x}$ and any failure scenario, a customer should always go to the nearest "available" facility. Otherwise we could reduce the cost by simply switching this customer over to a closer facility. Snyder & Daskin Snyder and Daskin [2005] used a similar argument to show that each customer should go to a facility only if all nearer facilities have failed. Thus, any design $\mathbf{x}$ (subject to failure) determines the assignment of customer demand.

From the perspective of a generic facility j, it will serve every customer on the 2-d plane with a certain probability (depending on its failure probability and that of other facilities). The whole area $\mathcal{S}$ is partitioned into non-overlapping subareas $\mathcal{R}_{j0}, \mathcal{R}_{j1}, \mathcal{R}_{j2}, \cdots$, such that $\mathcal{R}_{jk}, \forall k$, contains the subset of customers for whom facility j is the $(k+1)^{\text{th}}$ nearest facility. With this definition, for every j there is a non-overlapping partition if we ignore the boundaries of these subareas,

$$\bigcup_k \mathcal{R}_{jk} = \mathcal{S}, \text{ and } \mathcal{R}_{jk} \bigcap \mathcal{R}_{jk'} = \varnothing, \forall k, k'.$$

Since every customer will always go to the nearest available facility, the customer at $x \in \mathcal{R}_{jk}$ will go to facility j only after all of its k "nearest" facilities have failed, and if $k+1 \leq R$. Facility j will serve customers at x with the following service probability:

$$P(x, x_j|\mathbf{x}, \mathbf{y}) = (1-q)q^k, \text{ if } x \in \mathcal{R}_{jk}, \tag{4.3}$$

which decreases with k.

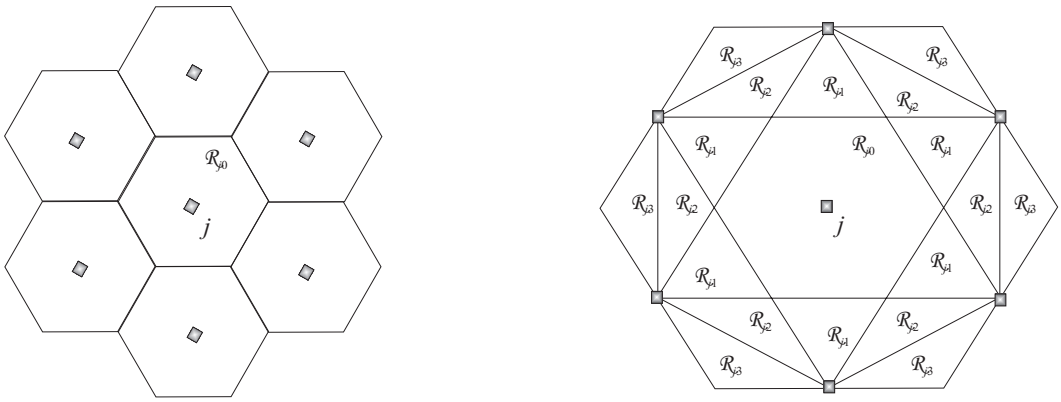
Particularly, the *initial service area* $\mathcal{R}_{j0}$ denotes the subarea of $\mathcal{S}$ served by facility j before any failure; i.e., $\mathcal{R}_{j0} := \{x : \|x - x_j\| \leq \|x - x_i\|, \forall i\} \subseteq \mathcal{S}$. Further denoting the set of initial service areas by $\mathbf{R} := \{\mathcal{R}_{10}, \mathcal{R}_{20}, \ldots, \mathcal{R}_{n0}\}$, they should form another area partition (ignoring boundaries):

$$\bigcup_j \mathcal{R}_{j0} = \mathcal{S} \text{ and } \mathcal{R}_{i0} \bigcap \mathcal{R}_{j0} = \varnothing, \forall i, j.$$

Proposition 5 shows that the optimal facility design on a homogeneous plane has the following special structure.

Proposition 5. *In an infinite homogeneous Euclidean plane, the optimal initial service areas should form a regular hexagon tessellation of the plane, while the facilities are at the centroids of the initial service areas; see Figure 4.1(a).*

With Proposition 5, we can estimate the exact optimal cost incurred by one facility on an infinite homogeneous plane. First of all, the probability that a particular facility serves a customer diminishes approximately exponentially with the distance between them. This is because the number of facilities closer to the customer (i.e., k), is approximately proportional to the square of the distance, while the service probability in (4.3) decreases exponentially with k. From the facility's perspective, the number of



(a) (b)

Figure 4.1. Regular hexagon tessellation in an infinite homogeneous 2-d Euclidean plane: (a) Initial service areas; (b) Service subarea partition for facility j.

available customers grows only polynomially with the distance. Hence, the expected service cost incurred to one facility on an infinite homogenous plane is bounded from above even when $R \to \infty$.

The regular hexagonal tessellation design in Figure 4.1(a) obviously leads to the service subarea partition in Figure 4.1(b). An arbitrary facility j has an initial service area size $A := |\mathcal{R}_{j0}|$ and may fail with a probability of q. Note that on the infinite plane, $n \to \infty$ in general. For this facility to serve customers that only go to R nearest facilities, we define the following useful term:

$$L := \int_{x \in \mathcal{S}} \|x - x_j\| P(x, x_j | \mathbf{x}, \mathbf{y}) dx = \sum_{k=0}^{R-1} \int_{x \in \mathcal{R}_{jk}} \|x - x_j\| (1-q) q^k dx,$$

where the second equality holds from (4.3). The average traveled distance for a customer to get service at the facility is then $L/(RA)$, and the total expected service cost for the facility to serve all its potential customers on the two-dimensional plane is hence λL. Certainly, $L < \infty$ (since $q < 1$) and its value should only depend on three factors, A, R and q.

By dimensional analysis and the Buckingham-Π Theorem Johnson [1944], the dimensionless quantities, $L/A^{\frac{3}{2}}$, R, and q, must be interdependent; *i.e.*, there must exist a unique function G such that

$$L/A^{\frac{3}{2}} = G(R, q). \tag{4.4}$$

Obviously, G(R,q) can be interpreted as the total expected service cost for a facility to serve all its potential customers when $\lambda = 1$ and $A = 1$. The exact functional form of G is unknown; however it only depends on the distance metric and can be estimated by a simulation.

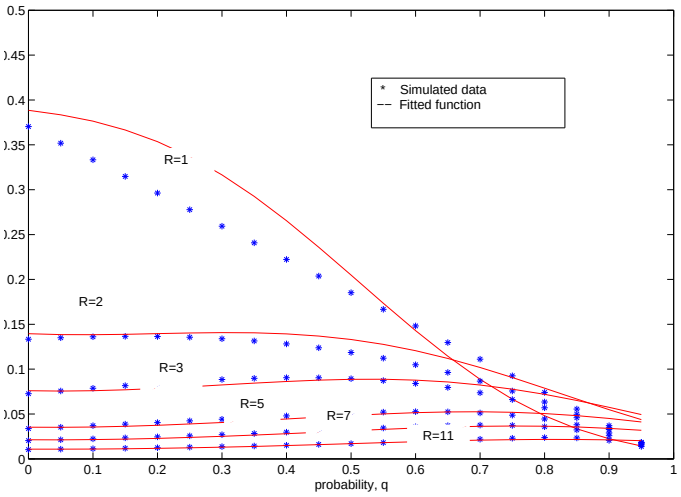


Figure 4.2. Simulated and fitted $L/(RA)^{3/2}$ for Euclidean metric.

For example, we hypothesize that $\ln(L/A^{3/2})$ can be approximated by a linear function of a list of polynomial terms of R and q. For the Euclidean metric, least squares regression with the simulated data in Figure 4.2 (with $1 \leq R \leq 11, 0 \leq q \leq 0.95$) yields

$$G(R,q) \approx \exp(-0.930 - 0.223q + 4.133q^2 - 2.906q^3 - 1.542\pi q^2/R), \tag{4.5}$$

The R-square value for the above regression equals 0.96, indicating a very good fit, especially for $R \geq 2$ and $q \leq 0.5$ (the realistic range of parameters for the reliability problem). In the numerical example, we will use (4.5) to approximate $G(R,q)$. It should be noted, however, that (4.5) is by no means the only way to estimate $G(R,q)$; rather, it is a plausible and simple choice. The CA approach presented in this paper can still be applied with any alternatives of (4.5).

Then, from (4.5)

$$L = G(R,q)A^{3/2}.$$

Note that one facility is built in correspondence to the customers in an area of size A. Intuitively, the optimal size of the initial service area can be obtained by minimizing the average cost per unit area; i.e.,

$$\min_A \{f/A + \lambda L/A | A > 0\}.$$

More detail on employing these results to solve general homogeneous or heterogeneous problems is presented in the following sections.

4.3 Heterogeneous Plane

In realistic cases, we allow the parameters λ, ϕ, q, f to be slowly varying functions of the location x in a bounded area $\mathcal{S}$. Instead of looking for $\mathbf{x}, \mathbf{y}$ (and $\mathbf{R}$) directly, we proposed to use the CA method to look for a continuous function, $A(x) \in \mathbb{R}_+, x \in \mathcal{S}$, that approximates the initial service area size of a facility near x. i.e., $A(x) \approx |\mathcal{R}_{i0}|$ if $x \in \mathcal{R}_{i0}$. We assume that $\mathcal{S}$ is far larger than $A(x)$; i.e. approximately 'infinite'. When all parameters, $f(x), \lambda(x), q(x)$ etc., are *approximately constant* over a region comparable to the size of several influence areas, the influence area size $A(x)$ should also be approximately constant on that scale. We show below that possible demand assignment and the associated service probabilities can be approximated by simple functions.

In a heterogeneous area $\mathcal{S}$, the objective function in (4.2) can be rewritten as:

$$\min_{\mathbf{x},\mathbf{y}} \sum_{j=1}^{n} f(x_j) + \int_{x\in\mathcal{S}} \phi(x)\bar{P}(x|\mathbf{x},\mathbf{y})\lambda(x)dx + \sum_{j=1}^{n} \int_{x\in\mathcal{S}} \|x - x_j\| P(x, x_j|\mathbf{x},\mathbf{y})\lambda(x)dx. \tag{4.6}$$

We will now rewrite (4.6) in terms of the new decision function $A(x)$ using results from Section 4.2.

A facility near location x serves an area of approximate size $A(x)$. We consider large-scale cases where $|\mathcal{S}| \gg A(x), \forall x \in \mathcal{S}$, and as such, $n \geq R$. The demand at x shall not be served if and only if all R closest facilities are nonfunctional simultaneously. Since failures are independent of each other, the probability for this to happen is approximately

$$\bar{P}(x|\mathbf{x},\mathbf{y}) \approx [q(x)]^R. \tag{4.7}$$

For $R \geq 1$, we define an expected "service" cost incurred for facility j as the summation of the fixed charge and the expected transportation costs:

$$\begin{aligned} C_j &:= f(x_j) + \int_{x\in\mathcal{S}} \|x - x_j\| P(x, x_j|\mathbf{x},\mathbf{y})\lambda(x)dx \\ &\approx f(x_j) + \lambda(x_j)L(x_j). \end{aligned}$$

Cost C_j corresponds to facility j which covers an approximate area of size $A(x_j)$. The cost per unit area near x_j, based on (4.4), is

$$\frac{C_j}{A(x_j)} \approx \frac{f(x_j)}{A(x_j)} + \lambda(x_j)G(R, q(x_j))\sqrt{A(x_j)}. \tag{4.8}$$

Substituting expressions (4.7) and (4.8) into (4.6), it is clear that the minimization problem can be approximated by finding the optimal function $A(x) \in [0, \infty)$ that minimizes the following integral:

$$\min_{A(x)} \int_{x\in\mathcal{S}} z(A(x), x)dx, \tag{4.9}$$

where $z(A(x), x)$ is the cost of serving a unit area near x when the influence area size is approximately $A(x)$:

$$z(A(x), x) := \frac{f(x)}{A(x)} + \phi(x)\lambda(x)[q(x)]^R + \lambda(x)G(R, q(x))\sqrt{A(x)}. \tag{4.10}$$

Note that (4.9) can be optimized by minimizing $z(A(x), x)$ over $A(x)$ at every $x \in \mathcal{S}$. In the rest of this subsection, we omit the argument x and use the notation $z(A)$ for simplicity. Formula (4.10) can then be expressed in the following closed form:

$$z(A) := \frac{f}{A} + \phi\lambda q^R + \lambda G(R, q)\sqrt{A}. \tag{4.11}$$

4.4 Feasible Discrete Location Design

Formula (4.9) yields an estimate of the total system cost without providing a discrete facility design. However, the optimal initial service area sizes, $A^*(x), \forall x \in \mathcal{S}$, can be used as guidelines to obtain feasible discrete location designs.

The optimal number of initial service areas, n^*, is approximately given by

$$n^* := \int_{\mathcal{S}} [A^*(x)]^{-1} dx.$$

The disk model by Ouyang and Daganzo Ouyang and Daganzo [2006] searches for a set of n non-overlapping disks, each having a round shape (i.e., approximating hexagons) and a proper size, that cover most of $\mathcal{S}$. A disk centered at x will have size $\alpha A^*(x)$, where the scaling parameter α is slightly smaller than 1 to ensure that the round disks can jointly cover most of $\mathcal{S}$ without leaving the region.

The disks move within $\mathcal{S}$ in search of a non-overlapping distribution pattern. To automate the sliding procedure, repulsive forces acting on the centers of the disks are imposed on any overlapping disks and on any disks that lie outside of $\mathcal{S}$. The disks then move under these forces in small steps, and the disk sizes and forces are updated simultaneously. Ouyang and Daganzo Ouyang and Daganzo [2006] and Ouyang Ouyang [2007] provide detailed discussions on how to choose step sizes, how to introduce necessary random perturbations, and how to decrease α incrementally until all forces vanish (i.e., when a desired non-overlapping pattern is found). Then, the disk centers will be used as the facility locations and the customer demands will be assigned accordingly. This procedure will give a near-optimal feasible solution to the planar problem.

4.5 Computational Results

To test the performance of the CA approach, we consider a $[0, 1] \times [0, 1]$ unit square, where customer demands are distributed according to a density function $\lambda(x)$.

A facility built at location x incurs a cost of $f(x)$ and may fail with probability $q(x)$. As a benchmark, we also construct and solve analogous discrete test instances by partitioning the unit square into $7 \times 7 = 49$ identical square cells; the center of each cell represents a candidate facility location as well as the consolidation point of the customer demand from that cell.

4.5.1 CA as a Heuristic Solution

We group our test instances into two categories: the homogeneous case and the heterogeneous case. In the homogeneous case, all system parameters are constant over space; i.e., $\lambda(x) = \lambda, f(x) = f$, and $q(x) = q \ \forall x$. We generate 16 test instances with key parameters taking values from $q \in \{0.05, 0.10, 0.15, 0.20\}$, $\lambda \in \{50000, 100000, 150000, 500000\}$. The fixed cost is $f = 1000$ for all 16 instances.

In the heterogeneous case, we let the key parameters be continuous functions that can vary across space, defined as follows:

$$\lambda(x) = \lambda(1 + \Delta_\lambda \cos(\pi x_{[2]})), \ f(x) = fe^{-\|x\|}, \ q(x) = q[1 + \Delta_q \cos(\pi \|x\|)], \forall x,$$

where $\|x\|$ is the Euclidean distance from x to the origin, and $x_{[2]}$ is the second coordinate of x. Note that q and λ control the average magnitude of failure probabilities and demand densities, while Δ_q and Δ_λ control the variability of these parameters. We generate 20 test instances in total, with q and Δ_q drawn from $q \in \{0.1, 0.2\}$ and $\Delta_q \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and Δ_λ taking values from $\Delta_\lambda \in \{0.0, 1.0\}$. The average demand density is set to be $\lambda = 100000$ and the average fixed cost is set to $f = 1000$ for all 20 instances.

The penalty cost is fixed at $\phi(x) = \sqrt{2}$, and the reassignment level is set to $R = 2$ for all 36 test instances in both categories.

For each test instance, we implement the CA model through the following procedure:

(i) Compute the continuous solution $A^*(x)$ (4.9), the optimal number of facilities n^*_{CA}, and the predicted total cost Z_{CA} (without discrete facility locations);

(ii) Use the disk model described in Section 4.4 to translate $A^*(x)$ into a feasible planar solution (i.e., facility can be anywhere in the unit square) and compute the planar cost Z^P_{CA}; and

(iii) Round the planar facility locations to the nearest cell centers, and compute Z^C_{CA}, the exact total system cost for the CA solution under continuous customer demand.

For comparison, we solve the discrete version of the problem as follows:

(i) Apply the LR algorithm to obtain the optimal number of facilities n^*_{LR} and the total system cost Z^D_{LR}. The superscript 'D' here stands for discrete customer demand;

(ii) Compute the cost for the LR solution to serve continuous customer demand Z^C_{LR}, where the superscript 'C' stands for continuous customer demand. While doing this, we simply enforce that each customer goes to the nearest existing facility; and

(iii) Compute Z^D_{CA}, the cost for the CA solution under aggregated customer demand..

Particularly, we are interested in the percentage differences between the CA model cost and the LR model cost, for continuous and discrete customer demand respectively; i.e., $\varepsilon^C = \frac{Z^C_{CA} - Z^C_{LR}}{Z^C_{LR}}$ and $\varepsilon^D = \frac{Z^D_{CA} - Z^D_{LR}}{Z^D_{LR}}$. These results are summarized in Table 4.1 for the homogeneous instances and in Table 4.2 for the heterogeneous instances.

Table 4.1. CA cost estimate, feasible solutions, and LR solutions for the homogeneous cases.

q	$\lambda(10^4)$	Z_{CA}	Z^P_{CA}	Z^C_{CA}	Z^D_{CA}	n^*_{CA}	Z^C_{LR}	Z^D_{LR}	n^*_{LR}	ε^C (%)	ε^D (%)
0.05	5	13908.5	14687.2	14694.2	14694.1	5	14521.9	14281.1	5	1.19	2.89
0.10	5	14430.9	15546.3	15608.2	15600.1	5	15705.8	15134.1	5	-0.62	3.08
0.15	5	15345.4	16666.9	16777.4	16762.3	5	16865.8	16251.5	5	-0.52	3.14
0.20	5	16632.0	18047.2	18201.9	18180.6	5	18318.1	17633.4	5	-0.63	3.10
0.05	10	22151.2	23270.2	23397.5	22895.4	7	23484.7	22607.4	7	-0.37	1.27
0.10	10	23199.4	24797.6	24951.4	24470.8	7	25104.2	24281.5	8	-0.61	0.78
0.15	10	25015.7	26886.8	27063.9	26605.3	7	27192.0	24281.5	8	-0.47	0.74
0.20	10	27568.7	29538.5	29734.8	29298.8	7	30093.5	28954.8	9	-1.19	1.19
0.05	15	28704.7	30085.7	31002.8	30538.9	10	30963.3	29460.2	10	0.13	3.66
0.10	15	29309.9	32162.6	33004.7	32702.9	10	33292.4	31807.3	10	-0.86	2.82
0.15	15	30667.5	35484.1	36263.4	35790.7	10	36339.7	34988.7	11	-0.21	2.29
0.20	15	32880.6	39222.6	39877.9	39546.5	10	40234.9	38945.2	10	-0.89	1.54
0.05	50	65504.6	67197.0	71305.8	65586.0	21	79225.5	54164.0	49	-10.00	21.09
0.10	50	70771.4	73815.2	77062.8	72551.7	21	85395.6	62506.1	49	-9.76	16.07
0.15	50	79752.2	83703.2	87206.1	83885.3	21	94838.5	74026.1	49	-8.05	13.32
0.20	50	92354.9	96457.2	100348.9	95861.2	21	107554.2	88724.3	49	-6.70	8.04

Our test results show that the CA method is a promising tool for finding near optimal solutions. Even under the discrete demand distribution (i.e., considering the continuous demand distribution as an approximation), the optimality gap is below 4% in most test instances. Particularly, even when the demand distribution is significantly variable across space ($\lambda(x)$ varying from 0 to 2λ), the gaps is mostly within $4-7\%$. It should also be noted that most often ε^C is negative, indicating that the CA model is more accurate for systems with continuous demand (i.e., considering the discrete demand as an approximation).

We note that under very high demand density $\lambda = 500,000$, the discrepancy between the CA and the LR solutions is more significant in terms of both the optimal number of facilities and the minimum total cost. This discovery is not surprising. With high demand density, the consolidation of customer demand to the 49 cell centers implies significant costs differences. Intuitively, the discrete model does not

Table 4.2. CA cost estimate, feasible solution, and LR solution for the heterogeneous case.

q	Δ_q	Δ_λ	Z_{CA}	Z_{CA}^P	Z_{CA}^C	Z_{CA}^D	n_{CA}^*	Z_{LR}^C	Z_{LR}^D	n_{LR}^*	ε^C (%)	ε^D (%)
0.1	0.1	0.0	18235	19061.8	20027.3	19563	12	19995.4	18971.1	15	0.16	3.12
0.1	0.2	0.0	18115.3	18891.9	19868.8	19405.5	12	19817.4	18726.6	14	0.26	3.63
0.1	0.3	0.0	18012.8	18753.3	19721.1	19245.2	12	19875.7	18631.9	15	-0.78	3.29
0.1	0.4	0.0	17927.5	18734.3	19943.9	19464.7	12	19493.9	18366.3	15	2.31	5.98
0.1	0.5	0.0	17859.4	18499.1	19500.8	18807.9	12	19453.6	18349.6	14	0.24	2.5
0.2	0.1	0.0	22158.7	23489.3	24376.4	23827.9	12	24046.2	23074.8	14	1.37	3.26
0.2	0.2	0.0	21668.7	22726.7	23573.6	23126.6	12	23747.5	22665.4	15	-0.73	2.03
0.2	0.3	0.0	21243.6	22190.3	23061.7	22621.5	12	23231.2	22076.7	17	-0.73	2.47
0.2	0.4	0.0	20884	21693.4	22593.4	22044.2	12	22540	21426.4	16	0.24	2.88
0.2	0.5	0.0	20590.4	21368.8	22139	21462.3	12	22266.6	20978.2	14	-0.57	2.31
0.1	0.1	1.0	16667.8	18337.7	19286	18573	11	19362.9	17670	14	-0.4	5.11
0.1	0.2	1.0	16646.7	18241.1	19137.8	18505.9	11	19203.7	17529.8	14	-0.34	5.57
0.1	0.3	1.0	16634.4	18136.5	19050.5	18411.5	11	19314.8	17435.3	16	-1.37	5.6
0.1	0.4	1.0	16630.7	18105.1	18971.4	18354	11	18843	17293.5	13	0.68	6.13
0.1	0.5	1.0	16635.4	18198.9	19278.5	18541.6	11	18852.6	17177.7	13	2.26	7.94
0.2	0.1	1.0	18864.4	22717.6	23525.2	22932.2	11	23326.4	22029.6	13	0.85	4.1
0.2	0.2	1.0	18740	22387.6	23258.8	22548.1	11	23270.6	21770.8	13	-0.05	3.57
0.2	0.3	1.0	18659.9	22169.5	22810.1	22284.3	11	23027.1	21369.5	14	-0.94	4.28
0.2	0.4	1.0	18623.1	22075.7	23253	22157.3	11	23075.8	21174.1	16	0.77	4.64
0.2	0.5	1.0	18628.6	21768.2	22486.7	21936.2	11	22686.9	20731.1	14	-0.88	5.81

involve the spatial distribution of customers within these cells. This omission tends to overestimate the marginal benefit of building an additional facility (i.e., in terms of reducing the customers' transportation cost). Hence, when the customer demand is extremely dense, the discrete model would tend to have more facilities in the optimal solution. We anticipate that if we have more than 49 aggregation nodes, the discrepancies would actually reduce.

To verify our hypothesis, we further divide the unit square to 10×10 identical cells and aggregate demand to the 100 cell centers. We test all 4 instances with high demand density $\lambda = 500,000$ in the 100-node network. The results are listed in Table 4.3.

Table 4.3. CA cost estimate, feasible solutions, and LR solutions in the 100-node network.

q	$\lambda(10^4)$	Z_{CA}	Z_{CA}^P	Z_{CA}^C	Z_{CA}^D	n_{CA}^*	Z_{LR}^C	Z_{LR}^D	n_{LR}^*	ε^C (%)	ε^D (%)
0.05	50	65504.6	66973.7	70003.0	68263.3	21	69216.9	66173.5	22	1.14	3.16
0.10	50	70771.4	73570.7	76324.5	74491.9	21	76836.8	73528.5	24	-0.67	1.31
0.15	50	79752.2	83438.5	85746.0	84207.7	21	85840.0	83114.6	23	-0.11	1.32
0.20	50	92354.9	96398.9	98269.2	97045.5	22	98628.9	95837.4	25	-0.36	1.26

The results in Table 4.3 indicate that the CA solutions are more consistent with the optimal solutions with more demand aggregation nodes. In general, the CA model should work at its best with a large number of candidate locations. In this sense, the LR and the CA methods can serve as complements of each other.

4.5.2 CA for Sensitivity Analysis

The system cost predicted by the CA model is continuous in all parameters, and is thus a useful tool for sensitivity analysis. In this section, we demonstrate how to use CA to study the impact of the key parameters on the structure of the optimal system design. In particular, we are interested in knowing how the degree of demand aggregation affects the system cost. In other words, all other things being equal, is it preferable to have evenly distributed demand or aggregated demand?

The CA model suggests that the total cost is determined by (4.9). It is easy to verify that $Z(A(x), x)$ is modular in A and that the point-wise optimal initial service area can be determined by

$$A^* = (\frac{2f}{\lambda G(R,q)})^{\frac{2}{3}}.$$

Plugging A^* back in (4.10) gives us the cost "density" near point x

$$z(x) \equiv z(A^*(x), x) = (2^{-\frac{2}{3}} + 2^{\frac{1}{3}}) f(x)^{\frac{1}{3}} \lambda(x)^{\frac{2}{3}} G^{\frac{2}{3}}(R, q(x)) + \phi(x)\lambda(x)q(x)^R. \quad (4.12)$$

Clearly, $z(x)$ is concave in λ. From Jensen's inequality, we know that the total cost decreases as the degree of demand aggregation increases.

To verify our findings, we designed numerical tests using the LR algorithm. The key parameters are determined by

$$\lambda(x) = \lambda(1 + \Delta_\lambda \cos(\pi x_{[2]})),\ f(x) = 1000,\ q(x) = 0.2, \phi(x) = \sqrt{2}\ \forall x.$$

We generated 30 test instances, 10 each for three different levels of average demand λ at 50000, 100000 or 150000. The demand variation Δ_λ ranges from 0 to 0.9. Like the previous tests, we aggregate demand to 49 discrete points. Each test instance is solved by the LR algorithm, and then the percentage change in the optimal cost is calculated, using the case $\Delta_\lambda = 0$ as the benchmark. The test results are illustrated in Figure 4.3.

Clearly, the test results from the discrete model verify the predictions made by the CA model, with the total cost decreasing by up to 7% as the demand variation increases from 0 to 0.9. This result implies that it is beneficial to aggregate demand. In reality, this principle is commonly implemented through the use of warehouses and distribution centers which serve as points for demand aggregation.

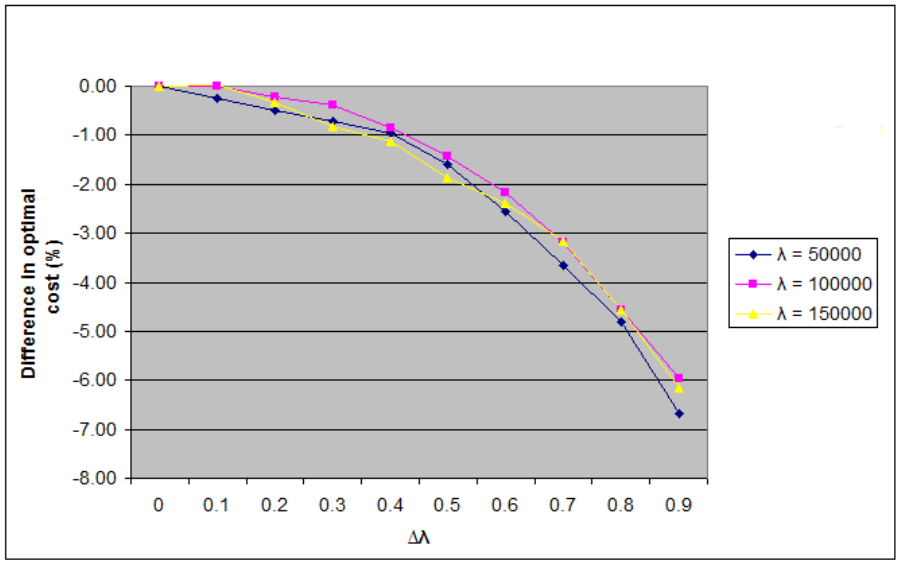


Figure 4.3. Optimal Cost v.s. demand variability

Chapter 5

Multiple-Dimension Mechanism Design in a Queueing System

5.1 Introduction

Consider a capacity-constrained service provider (server) facing customers with different values for service and sensitivities to delays, both of which are their private information. What kind of pricing, scheduling, and admission control policy should the server follow in order to maximize her expected revenue? This question is faced by many firms in the production and service industries, including manufacturing, telecommunication and transportation. A common strategy adopted by these servers is to segment the customers by providing different classes of services. For example, many make-to-order manufacturers charge the customers based on the delivery dates, and transportation firms like Fedex and UPS offer a range of service classes from ground shipping to same day delivery. By offering the option to pay more for faster services, the server may extract more revenue from market segmentation. However, since the server only has aggregate information about the customer attributes but cannot tell apart individual customers, all customers can choose among all service classes in a self-interested way. This gives rise to the *incentive compatibility* issue, which the server must take into account in designing the revenue maximizing admission and scheduling policies.

In this chapter, we provide a unified framework to study the aforementioned revenue maximization problem in the presence of asymmetric information regarding the customers' preferences. Our work is motivated by the recent papers of Afeche [2006], Yahalom et al. [2005], and Katta and Sethuraman [2005]. Afeche [2006] adopts the mechanism design approach to evaluate the incentive compatible *priority pricing* problem in a queueing system where customers' valuations are drawn from a continuous distribution but their delay sensitivities can take only two values. He shows that the revenue-maximizing priority rule does not conform with the celebrated $c\mu$ rule: it may require strategically inserted idleness, randomized priorities, or even reversed $c\mu$

order. Yahalom et al. [2005] allow general distributions over valuation and convex delay cost; this implies that higher moments of delay may be influential in their context. Katta and Sethuraman [2005] impose perfect correlation between valuation and delay sensitivity. Thus, they are able to recast it as a (standard) single-dimensional adverse selection problem; consequently, they provide an efficient algorithm to characterize the optimal mechanism and find that pooling multiple types of customers into the same priority class emerges as an optimal solution.

Despite the insightful elaboration on the incentive issues and the managerial implications that arise from those non-conventional queueing disciplines in the aforementioned work, an important feature of all these mechanisms is that *admission control is made through the design of priority pricing*. For example, in Katta and Sethuraman [2005] and Yahalom et al. [2005], only the priority classes (and the corresponding prices) are specified in the contracts. In Afeche [2006], the contract does specify the admission control. However, he focuses exclusively on the case with deterministic admission control. In other words, a customer is either admitted for sure or discarded entirely depending on the priority pricing scheme (the detailed discussions are deferred to Section 5.2). Thus, all the mechanisms in his model primarily use the priority classes as the sole screening tool to differentiate among customers.

In this paper, we argue that a previously ignored admission control policy plays a significant role in mitigating the information asymmetry between the server and the customers. Specifically, we show that a well-designed menu of admission control along with priority pricing contracts may force customers to reveal their true valuations; at the same time, this menu also induces the customers that are more sensitive to the delay to opt for higher priorities. The intuition is as follows. The customers with high valuations have higher opportunity costs when they do not get the services. Hence, if a *probabilistic* admission control policy is used (with different probability of rejecting customers), customers with high valuations may be willing to pay more for a better chance of getting admitted. Thus, the probabilistic admission control allows the server to choose the right customers to serve (thereby reducing the undesirable congestion) and consequently may enable the server to receive more revenue from those customers.

We illustrate our idea in a stylized model in which both the valuation and the delay sensitivity can take only two values – high or low. This allows us to classify the customers into four groups (types): $\{LH, HH, LL, HL\}$, where the first component specifies whether the valuation is high (H) or low (L), and the second component depicts whether the customer is highly sensitive to the delay (H in this case) or not (L). While not attempting to be all inclusive and the most general possible, this four-type model allows us to derive concrete managerial implications. Specifically, we show that the server may partially admit (through probabilistic admission) more than one customer types, although ex ante one type is more favorable than the other. Moreover, the server may assign different/randomized priorities for customers with same delay sensitivity but different valuations for service. Finally, the optimal contracts may

require strategically inserted idleness to ensure incentive compatibility, which echoes the results of Afeche [2006].

Admission control has long been recognized as a valuable tool to balance the throughput and congestion trade-off in queueing systems; see Stidham [2002] for a good survey. The probabilistic admission policy we propose is widely adopted by connection admission control (CAC) protocols in communication networks; see, e.g., Gibbens et al. [1995] and Lewis et al. [1998] and the references therein. An example is the RSVP (Resource Reservation Protocol) used for the ATM (Asynchronous Transfer Mode) network. In this protocol, different groups of consumers (packets such as email, ftp, voice data, etc.) are given choices over a number of flags (classes); each class is associated with a price, a priority class, and the probability of being dropped (that is analogous to the probabilistic admission control). See, e.g., Chang and Petr [2001] and Zhang et al. [1993] for the detailed descriptions of the protocol. (These protocols are proposed primarily from the system efficiency standpoint. On the contrary, in our model, the joint admission control/priority pricing is adopted to maximizing the server's revenue.)

Our model falls in the category of mechanism design problems with *multi-dimensional* private information (willingness to pay and willingness to wait) and screening tools (admission control and priority classes). Multi-dimensional mechanism design problems have long been recognized to be notoriously complicated and sometimes analytically intractable. The main challenges arise from the lack of complete ordering among the multi-dimensional types. Unlike the classical uni-dimensional framework, there is simply no unified way to ex ante identify redundant/ binding incentive compatibility constraints, thereby breaking down the systematic approach that has been prominently adopted in the literature; see the recent survey by Rochet and Stole [2005]. Moreover, the unique capacity constraint that arises from our queueing framework brings in new challenges and results in novel insights that would not occur in other contexts.

Our model is related to the vast literature on pricing, scheduling, and admission control in queueing systems. Classical papers in this field typically treat this problem in a centralized manner (i.e., a central planner is able to control all the behavior of the server, customers, etc.); see, e.g., Coffman and Mitrani [1980], Shanthikumar and Yao [1992], and Stidham [2002] for an excellent survey. In contrast, we incorporate the strategic customer behavior and asymmetric information. The strategic customer behavior has also been incorporated in the design of queueing systems at least dating back to Naor [1969]; see the monograph by Hassin and Haviv [2002] for a review of this literature. Mendelson [1985] and Mendelson and Whang [1990] are among the first to study socially optimal and incentive compatible priority pricing strategies in queueing systems. As aforementioned, Afeche [2006], Katta and Sethuraman [2005], and Yahalom et al. [2005] focus on incentive compatible priority pricing policies that maximizes the server's revenue. In line with this research stream, we introduce the freedom of choosing the (probabilistic) admission control that allows the server to extract more revenue from the customers effectively. Furthermore, by in-

corporating the possibility of probabilistic admission control, this paper expands the multi-dimensional nature of the classical incentive compatible revenue management to its full force.

Since we adopt the mechanism design approach to study this joint pricing, scheduling, and admission control problem, our work is also related to the principal-agent problems in which the principal intends to design an appropriate mechanism (contract) for the agents with private information to self-select. This framework has been extensively applied to various contexts in the operations research field, mostly within the single-dimensional framework, including capacity allocation (Cachon and Lariviere), supplier-retailer contracting (Corbett and de Groote [2000], Ha [2001]), product specification and production planning (Iyer et al. [2005]), inventory risk mitigation through promised lead time (Lutze and Ozer [2008]), pricing information goods (Wu and Chen [2008]), and long-term contract design (Zhang and Zenios [2008]). In contrast with the aforementioned papers, the *multi-dimensional* nature of our queueing framework inevitably creates new challenges. Wilson [1993] and Armstrong [1996] are the first to solve the multi-dimensional problems in closed form under specific assumptions on the model characteristics. Armstrong and Rochet [1999] provide a unified algorithm to solve discrete (specifically, four-type) multi-dimensional problems. This four-type framework is later adopted by Armstrong [2000] and Asker and Cantillon [2008] to study the forward and reverse auctions. Our four-type framework is also motivated by this stream of research. Nevertheless, the unique resource constraint that arises from the queueing framework results in a number of novel insights/ phenomena that would not occur in other contexts.

The remainder of this chapter is organized as follows. In Section 5.2, we describe the model setup. In Section 5.3, we present socially optimal contracts under symmetric information, as a benchmark for our study of information asymmetry. We discuss revenue maximizing contracts under asymmetric information in Section 5.4, with structural properties of the optimal solutions in Section 5.4.1, full characterization of the exact optimal mechanism for some special cases in Section 5.4.2, and novel features of the revenue maximizing contracts in Section 5.4.3. In Section 5.5, we demonstrate the revenue gains from the admission control policy using two numerical examples. We summarize our findings and give future research directions in Section 5.6. All proofs and the detailed derivations for the special cases can be found in the appendices.

5.2 Formulation

We consider a stylized model in which a capacity-constrained server modelled as an $M/M/1$ queueing system intends to serve several segments of customers. Customers request the same amount of task but are heterogeneous in two attributes: their willingness to pay, and their willingness to wait. Specifically, we assume that the service time of each customer follows an exponential distribution with a common

rate μ. Nevertheless, their valuations, denoted by v, and their delay sensitivities, denoted by c, are different across different groups. The value attribute $v > 0$ characterizes the customer's willingness to pay (in the absence of delay) for one unit of service, while the delay sensitivity $c > 0$ specifies the penalty per unit of time while the customer is kept in the system (service time included).

We assume that both attributes can take only two values to simplify our analysis. Specifically, we assume that $v \in \{v_L,\ v_H\}$ and $c \in \{c_L,\ c_H\}$, where $\Delta_v = v_H - v_L > 0$ and $\Delta_c = c_H - c_L > 0$. Given these values, there are four combinations $\{(v_L, c_H), (v_H, c_H), (v_L, c_L), (v_H, c_L)\}$, which are denoted by LH, HH, LL, and HL, respectively in the sequel. A customer with valuation v_H is willing to pay more for the service than the one with v_L; likewise, a customer endowed with a delay sensitivity c_H incurs a higher penalty (compared to the case with c_L). Each group of customers arrive at the system following a Poisson process, and we use λ_{ij} to denote the aggregate arrival rate of group ij of customers, where $ij \in \{LH, HH, LL, HL\} \equiv T$. Notably, from the server's viewpoint, type-HL customers are the most favorable customers since they are willing to pay more for the service and do not mind waiting so much. On the contrary, the server can extract the least amount of profit from the type-LH customers due to their low willingness to pay and high delay sensitivity. In compliance with the literature on incentive compatible priority pricing, we assume that the arrival process, the value and cost distributions and the service procedure are common knowledge. However, a customer's valuation and delay sensitivity are privately observed by this customer but unknown to the server. Thus, this private preference profile also represents a customer's *type*.

The server's problem is to design an appropriate mechanism to maximize his long-run expected payoff. In the absence of the information about the customers' preference, the server faces an adverse selection problem. As suggested by the agency literature (Laffont and Martimort [2002]), a common approach is to offer the customers a *menu* of contracts and let her self-select. Furthermore, the revelation principle allows us to restrict our attention to the direct mechanism in which the server simply requests the customers to report their types and then choose the contracts on behalves of the customers. Thus, we assume that the server offers a menu of contracts $\{q_{ij}, w_{ij}, p_{ij}\}$, where $q_{ij} \in [0, 1]$ is the admission rate, w_{ij} is the expected delay, and p_{ij} denotes the associated price charged by the server. Upon arrival, each customer decides which service class to purchase, and is charged and scheduled as prescribed by the contract. Given a contract $\{q, w, p\}$, a type-ij customer receives an expected (net) utility: $q(v_i - c_j w - p)$. We assume that a customer receives a null (zero) expected utility upon walking away without loss of generality. Notably, based on the above descriptions, the server is allowed to adopt a *stochastic/ probabilistic* admission control policy for a specific type (this occurs when $q_{ij} \neq \{0, 1\}$). This is in strict contrast with the extant literature on priority pricing, see, e.g., Afeche [2006], Katta and Sethuraman [2005], and Yahalom et al. [2005].

We restrict our attention to static scheduling policy and allow preemption. Furthermore, we adopt the achievable-region approach introduced by Coffman and Mi-

trani [1980] and Shanthikumar and Yao [1992]. Under our queueing framework, the associated expected delay is confined with the following resource constraints:

$$\sum_{ij\in T} \lambda_{ij} q_{ij} < \mu, \text{ and } \sum_{ij\in S} \frac{\lambda_{ij} q_{ij} w_{ij}}{\mu} \geq \frac{\sum_{ij\in S} \lambda_{ij} q_{ij}/\mu}{\mu - \sum_{ij\in S} \lambda_{ij} q_{ij}}, \ \forall S \subseteq T. \qquad \text{(RE)}$$

The condition $\sum_{ij\in T} \lambda_{ij} q_{ij} < \mu$ guarantees that the system size does not explode (since the effective aggregate arrival rate $\sum_{ij\in T} \lambda_{ij} q_{ij}$ is less than the service rate). In the second inequality of (RE), the left-hand side is the expected steady-state virtual load (defined as the remaining processing time) in set S as we recall by Little's law that $\lambda_{ij} w_{ij}$ is the steady-state queue length of group ij; the right-hand side corresponds to the average sojourn time (the waiting time plus the service time) when the customers in set S are given the *absolute* priority over all other customers outside this set (i.e., $T \setminus S$). Thus, from the viewpoint of the customers in set S, it is as if those customers outside this set are never in the system (and thus no further congestion is incurred due to the presence of customers in $T \setminus S$). This derivation follows from the classical queueing theory (see, e.g., Coffman and Mitrani [1980], Shanthikumar and Yao [1992] and also Katta and Sethuraman [2005]). For ease of notation, we denote $RE(S)$ as the resource constraint associated with the set S.

It is worth mentioning that this achievable region can be regarded as a sort of resource constraints for this queueing system. Moreover, each extreme point of this achievable region corresponds to a specific absolute priority rule. In our four-group setting, an extreme point is determined by four binding constraints, each of which is associated a specific set. Further, these sets must be *nested* in order to avoid any conflict between the queueing discipline. For example, if the sets associated with an extreme point is $\{HH\}$, $\{HH, LH\}$, $\{HH, LH, HL\}$, and $\{HH, LH, HL, LL\}$, the corresponding absolute priority rule is HH, LH, HL, LL, in descending order. This peculiar property implies that the achievable region is a base of a *polymatroid* (Shanthikumar and Yao [1992]). Another interesting observation is that since any interior point can be represented as a convex combination of a finite number of extreme points and feasible directions (which correspond to the "*strategic idleness*" in the terminology of Afeche [2006]). This convex combination also gives rise to a detailed implementation through a (*randomized*) priority rule, i.e., a certain group of customers are given priority only probabilistically. See Shanthikumar and Yao [1992] for more discussions and algorithms that implement the priority rules.

By the revelation principle, we restrict our attention to direct revelation mechanisms. Let

$$u(i'j'|ij) = q_{i'j'}(v_i - c_j w_{i'j'} - p_{i'j'})$$

denote the expected utility of a type-ij customer who pretends to be type-$i'j'$. For ease of notation, define

$$W_{ij} = q_{ij} w_{ij}, \text{ and } P_{ij} = q_{ij} p_{ij},$$

the customer's expected utility can be rewritten as $u(i'j'|ij) = v_i q_{i'j'} - c_j W_{i'j'} - P_{i'j'}$. In order to induce customers to participate, the following *individual rational* (IR)

constraint has to hold:

$$v_i q_{ij} - c_j W_{ij} - P_{ij} \geq 0, \ \forall ij \in T, \tag{IR}$$

where the right-hand side corresponds to the customers' reservation utility (which is normalized to zero). Furthermore, the menu of contracts has to induce the customers to willingly reveal their types, thereby giving rise to the following *incentive compatible* (IC) constraint:

$$v_i q_{ij} - c_j W_{ij} - P_{ij} \geq v_i q_{i'j'} - c_j W_{i'j'} - P_{i'j'}, \ \forall ij, i'j' \in T, \tag{IC}$$

where the left-hand side, as aforementioned, is the expected utility of a type-ij customer under truth-telling, and the left-hand side corresponds to the case of misrepresentation. Note that even if a customer misreports her type, the actual valuation as well as the delay sensitivity remain genuine (v_i and c_j, respectively). We use $IC(ij - i^{'}j^{'})$ to denote the incentive compatibility constraint that guarantees that a type-ij customer does not want to pretend to be type-$i^{'}j^{'}$.

Having discussed the customers' incentive problems, we now turn to the server's side. The server's goal is to find an appropriate menu of contracts that maximize her expected revenue:

$$\begin{aligned} \max_{\{q_{ij}, w_{ij}, p_{ij}\}} \quad & \sum_{ij \in T} \lambda_{ij} P_{ij}, \\ \text{s.t.} \quad & \text{(IC), (IR), and (RE).} \end{aligned}$$

For our convenience, we can replace the decision variables $\{q_{ij}, w_{ij}, p_{ij}\}$ by $\{q_{ij}, W_{ij}, P_{ij}\}$ following the definitions of W_{ij} and P_{ij}. Moreover, we introduce the "*information rent*":

$$R_{ij} \equiv v_i q_{ij} - c_j W_{ij} - P_{ij}$$

for each $ij \in T$. From the definition of R_{ij}, we have $P_{ij} = v_i q_{ij} - c_j W_{ij} - R_{ij}$. After these substitutions, the server's problem is reformulated as below:

$$\begin{aligned} \max_{\{q_{ij}, W_{ij}, R_{ij}\}} \quad & \sum_{ij \in T} \lambda_{ij} (v_i q_{ij} - c_j W_{ij} - R_{ij}) \\ \text{s.t.} \quad & R_{ij} - R_{i'j'} \geq (v_i - v_{i'}) q_{i'j'} - (c_j - c_{j'}) W_{i'j'}, \ \forall ij, i'j' \in T, && (5.1) \\ & R_{ij} \geq 0, \ \forall ij \in T, && (5.2) \\ & \sum_{ij \in T} \lambda_{ij} q_{ij} < \mu, && (5.3) \\ & \sum_{ij \in S} \lambda_{ij} W_{ij} \geq \frac{\sum_{ij \in S} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in S} \lambda_{ij} q_{ij}}, \ \forall S \subseteq T, && (5.4) \\ & W_{ij} \geq 0, \ 0 \leq q_{ij} \leq 1, \ \forall ij \in T, \end{aligned}$$

where (5.1) follows from (IC), (5.2) follows from (IR), and (5.3) and (5.4) are simply a restatement of (RE).

In the sequel, we derive the optimal mechanism (from the server's perspective) with four groups of customers. As a benchmark, we first study optimal (socially optimal) contracts under symmetric information in Section 5.3. Revenue maximizing contracts under asymmetric information are discussed in Section 5.4.

5.3 Optimal Contracts under Symmetric Information

To demonstrate the impact of information asymmetry, we first derive the optimal menu of contracts when the server has perfect knowledge of the customers' type information. We refer to this benchmark case as the scenario with symmetric information.

As the server knows the customers' types, the incentive compatibility condition (5.1) is no longer required, and since the server can extract the entire social surplus, the problem reduces to social maximization. The optimal contract design problem in this scenario can be formulated as follows:

$$\max_{\{q_{ij},W_{ij}\}} \quad \sum_{ij\in T} \lambda_{ij}(v_i q_{ij} - c_j W_{ij}) \tag{5.5}$$

$$\text{s.t.} \quad \sum_{ij\in T} \lambda_{ij} q_{ij} < \mu, \tag{5.6}$$

$$\sum_{ij\in S} \lambda_{ij} W_{ij} \geq \frac{\sum_{ij\in S} \lambda_{ij} q_{ij}}{\mu - \sum_{ij\in S} \lambda_{ij} q_{ij}}, \ \forall S \subseteq T, \tag{5.7}$$

$$W_{ij} \geq 0, \ 0 \leq q_{ij} \leq 1, \ \forall ij \in T. \tag{5.8}$$

Let $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$ be an optimal solution to (5.5)-(5.8). The following propositions characterize the optimal admission control and priority ranking policies under symmetric information.

5.3.1 Admission Preference

As a profit maximizer, the server has preferences of admitting certain types of customers over the others. We say that the server has *strong preference* of type ij over type $i'j'$ if she does not admit any type $i'j'$ customers unless she fully admits all type ij customers; i.e. $q_{i'j'} = 0$ if $q_{ij} < 1$. The following proposition characterizes the component-wise strong preference of a socially optimal admission policy: among customers with the same delay sensitivity, the server has strong preference of types with higher valuation; analogously, among customers with same valuation, the server has strong preference of types with lower delay sensitivity.

Proposition 6. *A socially optimal menu of contracts* $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$ *satisfies the following*

properties:

$$\hat{q}_{LL} = 0, \ \textit{if } \hat{q}_{HL} < 1;$$
$$\hat{q}_{LH} = 0, \ \textit{if } \hat{q}_{HH} < 1;$$
$$\hat{q}_{HH} = 0, \ \textit{if } \hat{q}_{HL} < 1;$$
$$\hat{q}_{LH} = 0, \ \textit{if } \hat{q}_{LL} < 1.$$

For comparison, we define the *weak preference* of customer types to be the ordering of their admission probabilities; i.e. the server has week preference of type ij customers over type $i'j'$ customers if and only if $q_{ij} > q_{i'j'}$. It is easy to see that the strong preference always implies the weak preference. Later on, we will show that component-wise strong preference no longer holds in revenue maximizing contracts under information asymmetry; however, weak preference is preserved.

5.3.2 Priority Scheduling

We say that type ij customers has *absolute priority* over type $i'j'$ customers if type ij always has the preemptive advantage of service over type $i'j'$. In terms of the resource constraints, there exists $S \subseteq T$ such that $ij \in S$, $i'j' \notin S$, and $RE(S)$ is binding. We say that type ij has randomized priority over type $i'j'$ if type ij customers have shorter lead times than type $i'j'$, but does not have absolute priority over type $i'j'$. Finally, we say that type ij and type $i'j'$ have equal priority if their average lead times are the same, i.e., $w_{ij} = w_{i'j'}$.

The following proposition states that in a socially optimal contract, customers with higher delay sensitivity have absolute priority over the others; however, there's no need to differentiate customers with same delay sensitivity but different valuations. On the contrary, while the first assertion still holds in a revenue maximizing contract under information asymmetry, it maybe optimal to assign absolute or randomized priorities among customers with same delay sensitivity but different valuation.

Proposition 7. *There exists a socially optimal menu of contracts* $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$ *that satisfies the following properties:*

$$\frac{\hat{W}_{LH}}{\hat{q}_{LH}} = \frac{\hat{W}_{HH}}{\hat{q}_{HH}} < \frac{\hat{W}_{LL}}{\hat{q}_{LL}} = \frac{\hat{W}_{HL}}{\hat{q}_{HL}};$$
$$\lambda_{LH}\hat{W}_{LH} + \lambda_{HH}\hat{W}_{HH} = \frac{\lambda_{LH}\hat{q}_{LH} + \lambda_{HH}\hat{q}_{HH}}{\mu - \lambda_{LH}\hat{q}_{LH} - \lambda_{HH}\hat{q}_{HH}}.$$

5.3.3 Work Conservation

We say that a scheduling policy follows the *work conservation* rule if it never idles the server. In terms of the resource constraints, $RE(T)$ is always binding in any work conservation policy. Since we allow preemption, a socially optimal contract

that minimizes delay costs should always satisfy the work conservation condition. However, in the presence of information asymmetry, it may be optimal to insert unforced idleness to delay the service for customers with lower delay sensitivity, in order to induce customers with higher delay sensitivity to report their true types.

Proposition 8. *A socially optimal menu of contracts* $(\hat{\mathbf{q}}, \hat{\mathbf{W}})$ *satisfies*

$$\sum_{ij\in T} \lambda_{ij}\hat{W}_{ij} = \frac{\sum_{ij\in T} \lambda_{ij}\hat{q}_{ij}}{\mu - \sum_{ij\in\{LH,HH\}} \lambda_{ij}\hat{q}_{ij}}.$$

The proof of Propositions 6-8 is straightforward and thus is omitted for conciseness. In the next Section, we present solutions for the optimal contract design problem under asymmetric information, focusing on the novel features introduced by information asymmetry.

5.4 Optimal Contracts under Information Asymmetry

In this section, we first overview general properties of the server's optimal contracts under information asymmetry in Section 5.4.1. Then we discuss detailed solutions for different cases of parameter values in Section 5.4.2. In Section 5.4.3, we compare the optimal contracts to that under symmetric information.

5.4.1 General Properties of the Optimal Contracts

To characterize structural properties of the revenue maximizing contracts, we use $q^* \in \mathcal{R}_+^{|T|}$ to denote the optimal allocation rule and $W^* \in \mathcal{R}_+^{|T|}$ to denote the corresponding optimal schedule. We first focus on the admission control policies and summarize our results in the next two propositions. Proposition 9 shows that the server has strong preference of customers with higher evaluation among all that with lower delay sensitivity, and customers with lower delay sensitivity among all that with higher evaluation (*strong preference at top*).

Proposition 9. *A revenue maximizing menu of contracts* $(\mathbf{q}^*, \mathbf{W}^*, \mathbf{R}^*)$ *has the following properties:*

$$q^*_{LL} = 0, \ \textit{if } q^*_{HL} < 1;$$
$$q^*_{HH} = 0, \ \textit{if } q^*_{HL} < 1.$$

However, as demonstrated by the special cases in Section 5.4.2, strong preference no longer holds among customers with lower evaluations or higher delay sensitivities, as the server may partially admit both the HH and LH types or the LL and LH types. Nonetheless, Proposition 10 shows the monotonicity on the optimal admission probability among these types (*week preference at bottom*).

Proposition 10. *The following assertions hold for any revenue maximizing menu of contracts* $(\mathbf{q}^*, \mathbf{W}^*, \mathbf{R}^*)$*:*

$$q^*_{LH} \leq q^*_{HH};$$
$$q^*_{LH} \leq q^*_{LL}.$$

Next, we turn to the priority scheduling policy. Analogously to the case under symmetric information, Proposition 11 shows that an optimal menu of contracts should always grant customers with higher delay sensitivity the absolute service priority.

Proposition 11. *A revenue maximizing solution* $(\mathbf{q}^*, \mathbf{W}^*, \mathbf{R}^*)$ *satisfies*

$$\lambda_{HH} W^*_{HH} + \lambda_{LH} W^*_{LH} = \frac{\lambda_{HH} q^*_{HH} + \lambda_{LH} q^*_{LH}}{\mu - \lambda_{HH} q^*_{HH} - \lambda_{LH} q^*_{LH}}.$$

As opposed to the symmetric information case, the server may have to use more than two priority classes in order to differentiate customers with the same delay sensitivity but different valuations. Proposition 12 shows that the HH type should have absolute priority over the LH type if neither of them are fully admitted.

Proposition 12. *The resource constraint* $RE(\{HH\})$ *is binding if* $q^*_{HH} < 1$*; i.e.,*

$$\lambda_{HH} W^*_{HH} = \frac{\lambda_{HH} q^*_{HH}}{\mu - \lambda_{HH} q^*_{HH}}.$$

Depending on the system configurations, it may be optimal to use randomized or reversed priority ranking between the HH and the LH types. It is also possible that the optimal contract schedules the HL and LL types at different priorities, along with strategic idleness to ensure incentive compatibility. These interesting phenomena are demonstrated by the special cases discussed in Section 5.4.2, and are discussed in detail in Section 5.4.3.

5.4.2 Some special cases

In this section, we present some special cases for which intriguing phenomena arise in terms of the optimal admission control and priority rules. Following Proposition 9 and 10, the server has strong preference of the HL type over any other customer types. In other words, the server will not admit customers of any other types if he does not fully admit the HL type. However, if the server fully admits the HL type, in general he could partially admit all three inferior types of customers, since the server has no clear-cut preference between the customers of two (intermediate) types HH and LL, although type-HL (type-LH) customers are always the most (least) favorable from the seller's perspective. For simplicity, we focus on the extreme cases in which the server partially admits at most two customer types.

When $\frac{\Delta_c}{\Delta_v}$, the ratio of the difference of delay sensitivity and the valuation difference, exceeds certain threshold level, upon fully admitting the most favorable type-HL customers, the server admits the type-LL customers until they are exhausted before admitting the type-HH and LH customers. When $\frac{\Delta_c}{\Delta_v}$ is below certain critical level, the preference is reversed: the server now intends to exhaust the type-HH customers before admitting any type-LL and LH customers. It would be difficult to determine the ranking between these two types in the intermediate cases; thus, in the sequel, we restrict ourselves to these two extreme cases.

Our results are summarized in Table 1, which are categorized by the the number of admitted groups (types) in the optimal solution. For each special case, we specify the admission policy, the number of priority classes required, and the detailed scheduling policy in each case. In the sequel we classify them by the number of admitted groups and elaborate on these cases. Detailed derivations of the optimal solutions for the special cases are provided in the online appendix.

Table 5.1. Summary of optimal contracts under asymmetric information

Number of admitted classes	$\frac{\Delta_c}{\Delta_v} \geq \mu(1+\frac{\lambda_{HL}}{\lambda_{LL}})$	$\frac{\Delta_c}{\Delta_v} \leq \frac{(\mu-\lambda_{HH})^2}{\mu(1+\lambda_{HL}/\lambda_{LL})}$
1	**Case 1a** $q_{HL}=1$, $q_{LL}=q_{HH}=q_{LH}=0$ Single queue.	**Case 1b** $q_{HL}=1$, $q_{HH}=q_{LL}=q_{LH}=0$ Single queue.
2	**Case 2a** $q_{HL}=1$, $0<q_{LL}\leq 1$, $q_{HH}=q_{LH}=0$ Single queue, $w_{HL}=w_{LL}$.	**Case 2b** $q_{HL}=1$, $0<q_{HH}\leq 1$, $q_{LL}=q_{LH}=0$ Two queues, $w_{HH}<w_{HL}$. Absolute priority rule.
3	**Case 3a** $q_{HL}=q_{LL}=1$, $0<q_{HH}\leq 1$, $q_{LH}=0$ Two queues, $w_{HH}<w_{HL}=w_{LL}$. Absolute priority rule.	**Case 3b** $q_{HL}=q_{HH}=1$, $0<q_{LL}\leq 1$, $q_{LH}=0$ Three queues, $w_{HH}<w_{HL}<w_{LL}$. Randomized priorities between LL and HL. Strategic idleness may be optimal.
4	$q_{HL}=<q_{LL}=1$, $0<q_{LH}\leq q_{HH}\leq 1$ **Case 4a-1** If $q_{HH}<1$, three queues. $w_{HH}<w_{LH}<w_{HL}=w_{LL}$, Absolute priority rule. **Case 4a-2** If $q_{HH}=1$ and $q_{LH}<1$, three queues. $w_{HH}<w_{LH}<w_{HL}=w_{LL}$. Randomized priorities between HH and LH. **Case 4a-3** If $q_{HH}=q_{LH}=1$, two queues. $w_{HH}=w_{LH}<w_{HL}=w_{LL}$. Absolute priority rule.	$q_{HL}=q_{HH}=1$, $0<q_{LH}\leq q_{LL}\leq 1$ **Case 4b-1** If $q_{LL}<1$, four queues. $w_{LH}<w_{HH}<w_{HL}<w_{LL}$ Randomized priorities between LL and HL. Strategic idleness may be optimal. **Case 4b-2** If $q_{LL}=1$ and $q_{LH}<1$, three queues. $w_{LH}<w_{HH}<w_{HL}=w_{LL}$. Randomized priorities between LH and HH. Strategic idleness may be optimal. **Case 4b-3** If $q_{LL}=q_{LH}=1$, two queues. $w_{LH}=w_{HH}<w_{HL}=w_{LL}$. Absolute priority rule.

One group of customers

Let us start with the simplest case in which only one group of customers are admitted (Case 1a and 1b). Since the server can extract the most revenue from the type-HL customers (who values the service highly and is less averse to the delay), this is the only group of customers that are admitted (as shown in Proposition 9). In this case, the "priority rule" degenerates since all admitted customers are identical.

Two groups of customers

A slightly more interesting case is when the server admits two groups of customers. By the above arguments, in addition to the type-HL customers, when the valuation difference is relatively small, the other admitted type is the type-LL (Case 2a); whereas the type-HH customers are admitted if the difference of delay sensitivity is relatively small (Case 2b).

In Case 2a, the two admitted types – HL and LL – have the same delay sensitivity. Thus, it makes no sense for the server to provide different priority rules; the only relevant parameter to differentiate between these two types is the admission probability. On the contrary, the server admits two types with different delay sensitivities (HL and HH) in Case 2b. In such a scenario, offering two priority classes allows the server to differentiate between them, since the type-HH customers are more averse to the delay and therefore are willing to pay more for a higher priority. We therefore observe that the server offers two priority classes in this case.

Three groups of customers

When the server admits three groups of customers, the admission and scheduling rules again critically depend on $\frac{\Delta_c}{\Delta_v}$; nevertheless, the set of admitted groups (HL, HH, and LL) is the same in the two extreme cases. When the valuation difference is relatively small (Case 3a), the server fully admits type LL and probabilistically admits type HH customers; the preference is reversed and the LL type is partially admitted if the difference of delay sensitivity is relatively small (Case 3b).

Following from Proposition 11, type-HH customers are given absolute priority over other types in both cases. However, the priority ranking between the HL and the LL types differs in the two extreme cases. In Case 3a, there is no need to offer different priority classes for the HL and LL types, since doing so will affect neither the system delay cost nor the customers' incentives. However, in Case 3b, the server intends to assign the type-HL customers a higher priority than the type-LL customers, even if these customers are homogeneous in terms of their delay sensitivity. In such a scenario, the purpose of this delay differentiation is to prevent the type-HH customers from misrepresenting themselves as either the LL or the HL type. The randomized priority rule provides the server with the desired flexibility to align the incentives of the customer's and minimize the information rent.

We further find that occasionally the server may insert unforced idleness to the queues with lower priority; in other words, *strategic idleness* may emerge as an optimal solution. This is because all work-conserving priority rules, although achieving system efficiency, result in severe incentive compatibility issues. Consequently, in order to differentiate the customers, the server must reduce the prices for the high priority queue significantly. Alternatively, the server may be better off by distorting the queueing discipline rather than adjusting the prices dramatically. This observation echoes the seminal work by Afeche [2006].

Four groups of customers

Finally, let us consider the case in which all four groups of customers are admitted, when the difference in delay sensitivity is relatively more significant (Case 4a-1 to 4a-3). We find that given the (soft) resource constraint, it may be in the server's best interest to probabilistically admit both the HH and LH types, even if ex ante the type-LH customers are perceived as the worst group. This probabilistic admission rule allows the server to maintain the customers' incentive compatibility in the least costly way. Since the difference in delay sensitivity is more significant than the difference in valuation, the server's main goal is to prevent customers with lower delay sensitivity from mimicking types with higher delay sensitivity. If the server solely admits type-HH customers besides the HL and the LL types, the contract intended for type-HH may look too appealing for the HL and LL types. In order to avoid this situation, the server may then be willing to allocate some capacity to serve the worst type (i.e., type-LH) customers.

If the server fully admits all customers, we find that only two priority classes are needed to differentiate between customers with high and low delay sensitivities. However, if at least one group of customers are admitted probabilistically, the server intends to offer higher priority to the HH type over the LH type, and he may randomize over the priority rule in order to achieve the best incentive provision.

When the difference in valuation is relatively more significant (Cases 4b-1 to 4b-3), the server partially admits both the LL and LH type, to leverage information rent paid to customers with higher valuation. Similarly to Case 3b, the type-HH customers may have the incentive to misreport as the LL and the HL type, and the server has to further differentiate between the LL and HL types with absolute or randomized priorities, in addition to the priority scheduling between the LH and HH types. Like the previously discussed cases, strategic idleness may also occur in an optimal solution, if the benefit outweighs the cost.

5.4.3 Novel Features of the Optimal Contracts

The aforementioned special cases reveal some interesting phenomena that arise from information asymmetry. For example, 1) the system may partially admit two customer groups, even though ex ante the server prefers one type to the other; 2)

customer groups with the same delay sensitivities (but different valuations) may be awarded priority over one another; 3) instead of fully exhausting the system resource, it may be optimal to idle the server intentionally. Although we relegate the detailed derivations to the online appendix, here we provide intuitive explanations for these novel features.

Mixed Admission Policy

We call an admission strategy a *mixed* policy, if it partially admits more than one customer types. Analogously, an *exclusive admission policy* is the one that partially admits at most one customer type. In Section 5.4.2, the mixed admission policy is applied in both Case 4a-1 and Case 4b-1.

Let us take Case 4a-1 and Case 3a as examples. In both cases, the type-HL and type-LL customers are fully admitted; however, the first case uses a mixed admission policy that partially admits both the type-HH and type-LH customers, while the admission policy for the second case is exclusive, with HH being the only partially admitted customer type. The main trade-off is as follows. In Case 4a-1, Δ_v, the difference in valuation, is extremely small as compare to Δ_c, the difference in delay sensitivity. The inclusion of type-LH customers helps reduce the admission rate of the type-HH customers while maintaining the same throughput level, which in turn helps to reduce the incentive of the type-HL and type-LL customers to misreport as the HH type. On the other hand, this mixed admission strategy also results in a decrease in valuation, as well as an increase in the information rent to the type-HH customer, both of which are determined by Δ_v. Since the reduction in information rent of type-HL and type-LL customer is proportional to Δ_c and is thus more significant than the decrease in valuation, the mixed admission policy can extract higher revenue for the server. Whereas in Case 3a, Δ_v is not sufficiently small as compared to Δ_c, and the cost of using the mixed admission policy outweighs the benefit. Therefore the exclusive admission policy is optimal in this case. The reasoning for the mixed admission policy in Case 4b-1 and the exclusive admission policy in Case 3b is similar.

To demonstrate the major trade-off, we illustrate the binding IC constraints of Case 4a-1 and Case 4b-1 in Figures 5.1 - 5.2. Here and in all the following diagrams, an edge from type-ij to type-$i'j'$ represents the IC constraints that type-$i'j'$ does not not be tempted to choose the contract intended for type-ij. The value along an edge is the right-hand side of the corresponding IC constraint; in other words, it is a lower bound on the difference in the information rents received by the origin and destination customer types. The solid lines represent unique binding IC constraints, while the dotted lines imply multiple binding IC constraints.

Randomized Priority Scheduling

In a system under information symmetry, it is not necessary to differentiate customers with the same delay sensitivity but different valuations. Under information

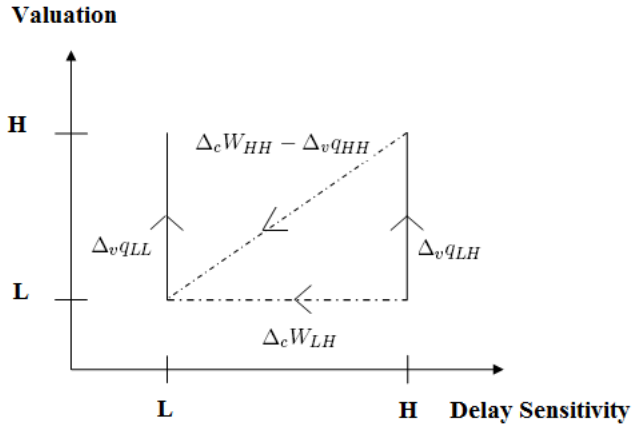


Figure 5.1. Binding IC constraints in Case 4a-1.

asymmetry, however, it may be optimal to give different priority rankings to these customers. For example, in Case 4a-2, the type-HH customers have higher service priority over the type-LH customers. Moreover, the priority ranking between the HH and the LH is randomized to achieve the best information provision. Figure 5.3 shows the binding IC constraints in this case. In order to minimize the information rent of the type-HL customers, the optimal menu of contracts needs to satisfy the following equation:

$$\Delta_c W_{LH} + \Delta_v q_{LL} = \Delta_v q_{LH} + \Delta_c W_{HH},$$

which can only be achieved through randomized priority ranking between the type-LH and type-HH customers.

Strategic Idleness

If the server has complete knowledge of the customers' valuations and delay sensitivities, the optimal strategy is to exhaust the system resources, since any idleness will result in excessive costs. However, when the valuation and delay sensitivities are the customers' private information, it may be optimal to strategically delay the service for customers with lower delay sensitivities in order to provide appropriate incentives for customers with higher delay sensitivities. We call this non-exhaustion of system capacity the strategic idleness. This phenomena may arise in Case 3b, Case 4b-1 and Case 4b-2.

Let us take Case 3b as an example, in which the binding IC constraints are illustrated in Figure 5.4. If we increase W_{LL} by a small amount $\epsilon > 0$, R_{HH}, the information rent received by type-HH customers, will decrease by $\Delta_c \epsilon$, so will R_{HL} since it is equal to $R_{HH} + \Delta_c W_{HH}$. However, the delay cost will increase by $c_L \epsilon$. The benefit and cost trade-off depends on the relative significance of Δ_c as compared to

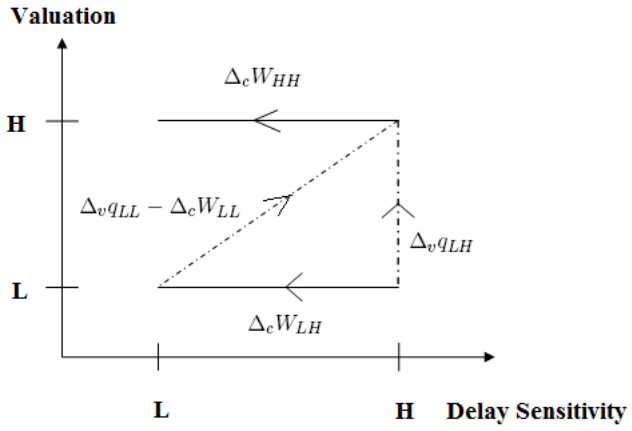


Figure 5.2. Binding IC constraints in Case 4b-1.

c_L, as well as the arrival rates λ_{LL}, λ_{HH} and λ_{HL}. In situations where the benefit outweighs the cost, strategic idleness may emerge as an optimal solution.

5.5 Revenue Gains from the Probabilistic Admission Policy

One of the key differences of our model and that of Afeche [2006] is the probabilistic admission control that allows the server to obtain more revenue from the customers. The revenue gains from a probabilistic admission policy follow from two sources. First, it helps the server to better utilize the limited resource; second, it provides the server the desired flexibility to leverage the information rents that must be paid to the customers. A natural question, that arises, is how much incremental revenue the probabilistic admission control can raise. To this end, we construct two examples to demonstrate the revenue gains from the probabilistic admission control policy, using the 0-1 (deterministic) admission policy (as in Afeche [2006]) as a benchmark. Example 2 demonstrates how the probabilistic admission policy gives the server the flexibility to accept the optimal level of work load, in order to balance the trade-off between the revenue gains from serving the customers and the delay costs due to congestion. Example 3 shows the revenue gains from leveraging the information rent by partially admitting two of the less favorable types (LH and LL).

Example 2. *In this example, we use the following key parameter values:* $\mu = 100$, $\lambda_{HL} = 30$, $\lambda_{LL} = 30$, $\lambda_{HH} = 10$, $\lambda_{LH} = 10$, $v_L = 5$, $v_H = 6$, $c_L = 100$, *and* $c_H = 300$. *Because* $v_H > c_L \frac{\mu}{(\mu-\lambda_{HH})^2}$, *the system should fully admit the type-*HL *(the most favorable) customers; i.e.,* $q_{HL} = 1$. *Also, because* $\frac{\Delta_c}{\Delta_v} > \mu(1 + \frac{\lambda_{HL}}{\lambda_{LL}})$, *the server*

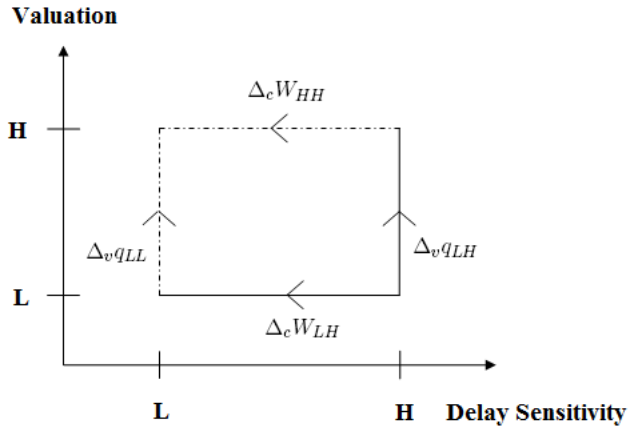


Figure 5.3. Binding IC constraints in Case 4a-2.

always prefers type-LL customers to type-HH customers. Since $v_L - c_L \frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL})^2} - \Delta_v \frac{\lambda_{HL}}{\lambda_{LL}} < 0$, $q_{HH} = q_{LH} = 0$.

Deterministic admission policy. If the server uses the deterministic admission policy, then following the above argument we obtain that either $q_{LL} = 0$ or $q_{LL} = 1$. If $q_{LL} = 0$, it follows that $w_{HL} = \frac{1}{\mu-\lambda_{HL}} = \frac{1}{70}$, and $R_{HL} = 0$. The server's revenue is:

$$\pi_s = v_H \lambda_{HL} - c_L \lambda_{HL} w_{HL} = 137.$$

On the other hand, if $q_{LL} = 1$, $w_{HL} = w_{LL} = \frac{1}{\mu-\lambda_{HL}-\lambda_{LL}} = \frac{1}{40}$, $R_{LL} = 0$, and $R_{HL} = \Delta_v q_{LL} = 1$. The server's corresponding revenue is:

$$\pi_s = v_H \lambda_{HL} + v_L \lambda_{LL} - c_L(\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) - \lambda_{HL} R_{HL} = 150.$$

Thus, under the deterministic admission policy, the server's optimal revenue is 150.

Probabilistic admission policy. If instead the server adopts the probabilistic admission control policy, the optimal admission probability q^*_{LL} should be the solution to the following equation:

$$v_L - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} q_{LL})^2} - \Delta_v \frac{\lambda_{HL}}{\lambda_{LL}} = 0. \tag{5.9}$$

Solving (5.9) yields $q^*_{LL} = \frac{2}{3}$. It then follows that $w^*_{HL} = w^*_{LL} = \frac{1}{\mu-\lambda_{HL}-\lambda_{LL} q^*_{LL}} = \frac{1}{50}$, $R^*_{HL} = \Delta_v q^*_{LL} = \frac{2}{3}$, and the server's optimal revenue is:

$$\pi^*_s = v_H \lambda_{HL} + v_L \lambda_{LL} q^*_{LL} - c_L(\lambda_{HL} w^*_{HL} + \lambda_{LL} w^*_{LL}) - \lambda_{HL} R^*_{HL} = 160,$$

which is clearly higher than the highest revenue (150) achieved by the deterministic admission policy.

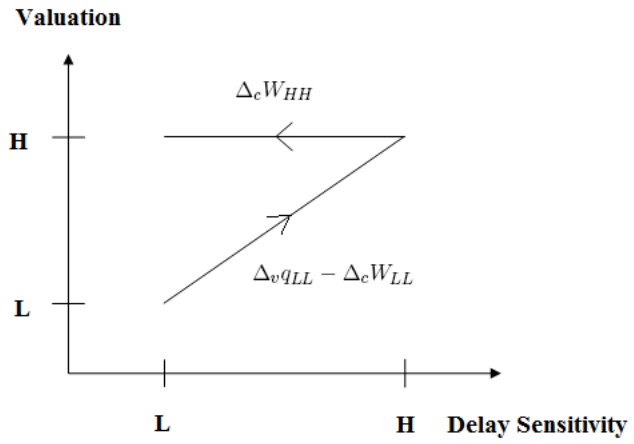


Figure 5.4. Binding IC constraints in Case 3b.

Example 3. *The following parameters are adopted in this example:* $\mu = 100$, $\lambda_{HL} = 30$, $\lambda_{LL} = 20$, $\lambda_{HH} = 10$, $\lambda_{LH} = 20$; $v_L = 20$, $v_H = 21$, $c_L = 100$, *and* $c_H = 400$. *In this case, since* $\frac{\Delta_c}{\Delta_v} > \mu(1 + \frac{\lambda_{HL}}{\lambda_{LL}})$, *the server always prefers type-*LL *customers to type-*HH *customers. Also, since* $v_L - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_v \frac{\lambda_{HL}}{\lambda_{LL}} > 0$, $q_{HL} = q_{LL} = 1$.

Deterministic admission policy. If the server uses the deterministic admission policy, she has three options: 1) admitting neither the type-HH nor the type-LH customers ($q_{HH} = q_{LH} = 0$); 2) admitting only the type-HH customers ($q_{HH} = 1$, $q_{LH} = 0$); and 3) admitting both type-HH and type-LH customers ($q_{HH} = q_{LH} = 1$). Her revenue associated with the three policies are calculated below.

When $q_{HH} = q_{LH} = 0$, it follows that $w_{HL} = w_{LL} = \frac{1}{\mu - \lambda_{HL} - \lambda_{LL}} = \frac{1}{50}$, $R_{LL} = 0$, and $R_{HL} = \Delta_v q_{LL} = 1$. The server's corresponding revenue is:

$$\pi_s = v_H \lambda_{HL} + v_L \lambda_{LL} - c_L(\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) - \lambda_{HL} R_{HL} = 900.$$

In the second case where $q_{HH} = 1$, $q_{LH} = 0$, we obtain that $w_{HH} = \frac{1}{\mu - \lambda_{HH}} = \frac{1}{90}$, and $w_{HL} = w_{LL} = \frac{1}{\lambda_{HL} + \lambda_{LL}} (\frac{\lambda_{HH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH}}{\mu - \lambda_{HH}}) = \frac{1}{36}$. The information rents in this case are $R_{HH} = 0$, $R_{LL} = \Delta_c w_{HH} q_{HH} - \Delta_v q_{HH} = \frac{7}{3}$, and $R_{HL} = R_{LL} + \Delta_v q_{LL} = \frac{10}{3}$. Accordingly, the server's revenue is:

$$\pi_s = v_H(\lambda_{HL} + \lambda_{HH}) + v_L \lambda_{LL} - c_L(\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) - c_H \lambda_{HH} w_{HH} - \lambda_{LL} R_{LL} - \lambda_{HL} R_{HL} = 910.$$

Finally, when $q_{HH} = q_{LH} = 1$, $w_{HH} = w_{LH} = \frac{1}{\mu - \lambda_{HH} - \lambda_{LH}} = \frac{1}{70}$, and

$$w_{HL} = w_{LL} = \frac{1}{\lambda_{HL} + \lambda_{LL}} (\frac{\lambda_{HH} + \lambda_{LH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{LH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH} + \lambda_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}}) = \frac{1}{14}.$$

The information rents in this case are $R_{LH} = 0$, $R_{HH} = \Delta_v q_{LH} = 1$, $R_{LL} = \Delta_c w_{LH} q_{LH} = \frac{30}{7}$, and $R_{HL} = R_{LL} + \Delta_v q_{LL} = \frac{37}{7}$. The server's revenue is:

$$\begin{aligned}\pi_s &= v_H(\lambda_{HL} + \lambda_{HH}) + v_L(\lambda_{LL} + \lambda_{LH}) - c_L(\lambda_{HL} w_{HL} + \lambda_{LL} w_{LL}) \\ &- c_H(\lambda_{HH} w_{HH} + \lambda_{LH} w_{LH}) - \lambda_{HH} R_{HH} - \lambda_{LL} R_{LL} - \lambda_{HL} R_{HL} = 857.\end{aligned}$$

Collectively, the server's optimal revenue under the deterministic admission policy is 910.

Probabilistic admission policy. Now we investigate the case with the probabilistic admission policy. In such a scenario, the optimal contract should satisfy the following equations:

$$\begin{aligned}&w_{HH} = \frac{1}{\mu - \lambda_{HH} q_{HH}}, \\ &w_{LH} = \frac{1}{\lambda_{LH} q_{LH}} \left(\frac{\lambda_{HH} q_{HH} + \lambda_{LH} q_{LH}}{\mu - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH}} - \frac{\lambda_{HH} q_{HH}}{\mu - \lambda_{HH} q_{HH}}\right), \\ &\Delta_c w_{LH} q_{LH} = \Delta_v q_{LH} + \Delta_c w_{HH} q_{HH} - \Delta_v q_{HH}, \\ &v_L - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH})^2} \\ &-\Delta_c \frac{\mu}{(\mu - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH})^2} \left(1 + \frac{\lambda_{HL} + \lambda_{LL}}{\lambda_{LH}}\right) - \Delta_v \frac{\lambda_{HH} + \lambda_{HL}}{\lambda_{LH}} = 0.\end{aligned}$$

Solving the system of equations above yields

$$q^*_{LH} = 0.3535,\ q^*_{HH} = 0.4194,\ w^*_{LH} = 0.0128, w^*_{HH} = 0.0104.$$

It follows that

$$\begin{aligned}w^*_{HL} = w^*_{LL} &= \frac{1}{\lambda_{HL} + \lambda_{LL}} \left(\frac{\lambda_{HH} q^*_{HH} + \lambda_{LH} q^*_{LH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} q^*_{HH} - \lambda_{LH} q^*_{LH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH} q^*_{HH} + \lambda_{LH} q^*_{LH}}{\mu - \lambda_{HH} q^*_{HH} - \lambda_{LH} q^*_{LH}}\right) \\ &= 0.029,\end{aligned}$$

$R^*_{LH} = 0$, $R^*_{HH} = \Delta_v q^*_{LH} = 0.3535$, $R^*_{LL} = \Delta_c w^*_{LH} q^*_{LH} = 1.3545$, and

$$R^*_{HL} = R^*_{LL} + \Delta_v q^*_{LL} = 2.3545$$

. The server's optimal revenue is:

$$\begin{aligned}\pi^*_s &= v_H(\lambda_{HL} + \lambda_{HH} q^*_{HH}) + v_L(\lambda_{LL} + \lambda_{LH} q^*_{LH}) - c_L(\lambda_{HL} w^*_{HL} + \lambda_{LL} w^*_{LL}) \\ &- c_H(\lambda_{HH} w^*_{HH} + \lambda_{LH} w^*_{LH}) - \lambda_{HH} R^*_{HH} - \lambda_{LL} R^*_{LL} - \lambda_{HL} R^*_{HL} = 962,\end{aligned}$$

which is clearly higher than that achieved by the best deterministic admission policy (910).

5.6 Conclusions

In this paper, we characterize the optimal joint pricing, scheduling, and admission control policy when the server faces customers with heterogeneous valuations and delay sensitivities. We show that the server always exhausts the most favorable type of customers (that have the highest valuation and are least sensitive to the delay) before admitting any other type of customers. Moreover, we find that even if the customers have identical valuations, in determining the admission policy, the server may still prefer one to another based on their delay sensitivity. Except the most favorable type of customers, the server's preference over the customers is endogenous. In particular, we find that the server may probabilistically admit more than one type.

We also characterize the optimal mechanisms in a number of special cases to gain further insights. Specifically, we find that the server may intend to offer different admission probabilities for the customers with common valuations, and may pool some groups of customers into one priority queue. Finally, occasionally a randomized priority may be adopted to prevent different types of customers to misrepresent themselves. Regarding the priority rules, we find that the server always optimally gives the customers with high delay sensitivity the absolute priority over those with low delay sensitivity. Moreover, to distinguish among different groups of customers, it may be the server's best interest to insert some strategic idleness and use randomized priority rules.

As we intend to provide a simple framework to illustrate the above managerial implications, our stylized model certainly has its own limitations and may be extended in various dimensions. In this paper, we focus exclusively on the case when the server makes a one-time decision on managing his business. In reality, there might be situations in which this can be done dynamically. For example, if the server is able to adjust dynamically the admission control based on the current queueing status, he may be able to strategically select the appropriate customers to serve. Also, if the scheduling policy can be adjusted over time, if the server attempts to serve a specific type of customers, he may be able to (temporarily) give the highest priority to those customers, thereby reducing their disutility that arises from the congestion. Extending our framework to a dynamic setting is a crucial step as well as a challenging task.

Another possible direction is to consider nonlinear delay cost. In contrast with our current setting in which only the expected delay is active, higher moments of the delay may also affect the customers' utility in such a nonlinear environment (Yahalom et al. [2005]). In this scenario, while designing the priority rule, the server may also intend to minimize, for example, the variance of the delay a customer may encounter. In particular, static priority rule may be suboptimal (Yahalom et al. [2005]) and one inevitably needs to search among those dynamic scheduling policy (e.g., the generalized $c\mu$ rule). Including these effects may broaden the applicability of our framework, and it definitely deserves further investigation.

Chapter 6

Robust Auction Mechanism Design

6.1 Introduction

In the area of mechanism design, an emerging trend of research attempts to address issues of the very restrictive common knowledge assumption. For example, Bergemann and Morris [2005] argue that the notion of robustness should be examined in the richer universal type space. Thus, they proceed to study the equivalence/difference between Bayesian and Ex post implementations and find that the equivalence holds in some settings that are termed "separable," but the discrepancy may arise beyond these environments; see also Bergemann and Morris [2008, 2009] along the same vein. Chung and Ely [2007] argue that a mechanism should be robust against the principal's (auctioneer's) own belief about the agents' (bidders') types. They show that, for every detail-free mechanism, there must exist a belief for which it is outperformed by the optimal dominant strategy mechanism; thus, this provides a normative justification for dominant strategy mechanisms. Bose et al. [2006] study the optimal auction design when the auctioneer exhibits ambiguity aversion. Most recently, Lopomo et al. [2007] introduce Knightian uncertainty to the players' beliefs, and identify the conditions on the uncertainty set under which the Ex post formulation is equivalent to the robust formulation in the single-dimensional mechanism design problems.

Despite the insightful elaborations in the aforementioned literature on the robust implmentability of auction mechanisms, surprisingly little attention has been paid to the revenue loss associated with a (more) robust formulation. In this paper, we adopt the Knightian uncertainty model introduced by Lopomo et al. [2007] and revisit the optimal auction design problem. We first provide a unified framework that allows us to formally define the bidders' belief uncertainty. This leads to a continuum of formulations/games, each of which corresponds to different levels of belief uncertainties. In this general framework, the bidders' valuations are allowed to be interdependent, and the auctioneer may sell more than one object, thereby giving rise to the multi-dimensional mechanism design problem.

We apply the network approach developed by Malakhov and Vohra [2005] to reformulate our optimal auction design problems, and use this to characterize the necessary and sufficient conditions (for a fixed allocation) to achieve the same revenue in a robust formulation as in the Bayesian formulation. This allows us to spot the possible distinction between different formulations and investigate the impact of bidders' belief uncertainty on the auctioneer. We apply our framework to study the two classical problems, namely the single-object auction introduced by Myerson [1981] and the multi-object auction introduced by Armstrong [2000]. Our results show that there is no discrepancy between any pair of formulations in the single-object auction even if the bidders' types are discrete. This is admittedly intuitive, but it serves as a benchmark that validates our result. Furthermore, we show that in a multiple-object auction, the auctioneer's expected revenue may strictly decrease as the bidders' beliefs become more uncertain. Following this, we provide a concrete example for which the ex post formulation gives rise to a strictly lower expected revenue in the multi-dimensional setting.

Our paper belongs to a long-standing literature on auction theory, including the seminal paper by Vickrey [1961], the paper on interdependent values by Milgrom and Weber [1982], the mechanism design approaches of Myerson [1981] and Maskin and Riley [1989], the survey by Klemperer [1999], and the recent book by Krishna [2002]. Recent advances are on the design of multi-object auctions with multi-dimensional valuations (Armstrong [2000]), on the competing auctions (Moldovanu et al. [2008]), and on the characterization of bidding behavior when there are interactions among bidders (Figueroa and Skreta [2009] and Jehiel et al. [1999]). Unlike all the aforementioned papers, we introduce the robust formulation in which the bidders may exhibit belief uncertainty. Our paper is also related to a broader class of papers on mechanism design. Stemming from Mirrlees [1971], this framework with information asymmetry has been applied extensively in a variety of areas, including product line design (Moorthy [1984] and Mussa and Rosen [1978]), taxation policies (Salanie [2003]), managerial compensation schemes (Holmstrom and Milgrom [1991]), and government regulation (Laffont and Tirole [1993]). Please see Laffont and Martimort [2002] for detailed discussions. Our approach to model the belief uncertainty may find its applications in these domains as well.

As aforementioned, we attempt to relax the common knowledge assumption and evaluate the robustness of mechanism design. Thus, our paper also belongs to a rising stream on robust or "detail-free" mechanism design, including Bergemann and Morris [2005, 2008, 2009], Bose et al. [2006], Chung and Ely [2007], and Lopomo et al. [2007]. Unlike Bergemann and Morris [2005], we focus on the conventional "first-order" belief space. While Bose et al. [2006] and Chung and Ely [2007] allow the auctioneer to exhibit belief uncertainty, we assume that the belief uncertainty arises solely from the bidders' side. A closely related recent paper by Heydenreich et al. [2009] applies the graph theory to mechanism design problems with a continuous type space and a discrete outcome space; their main objective is to identify the conditions under which all Ex post incentive compatible payment schemes can only vary by a constant. Our results complement theirs as we focus on the characterization of conditions for the

existence of Ex post incentive compatible payment scheme that is revenue equivalent to the Bayesian formulation; it can be verified that their conditions imply ours, but not vice versa. Moreover, in the terminology of network approach, their condition is equivalent to that all paths being equal between any pair of nodes, whereas we only require there be overlapping between the sets of longest paths; once again, the former condition implies the latter. Finally, as a relatively minor difference, we focus on a discrete type space and unrestricted outcome space rather than the environment studied by Heydenreich et al. [2009].

The rest of the paper is organized as follows. In Section 6.2, we describe the basic setting. In Section 6.3, we formulate the robust mechanism design problems, of which the Bayesian and the Ex post formulations can be taken as special cases. In Section 6.4, we introduce the network approach and the necessary and sufficient conditions for a fixed allocation mechanism to achieve the same revenue in a robust formulation as in the nominal Bayesian formulation. We then consider robust formulations of Bayesian optimal auction mechanisms in Section 6.5, in both the single-object and multiple-object settings. Finally, we conclude in Section 6.6.

6.2 The model

Let us first introduce the general model in which an auctioneer faces a number of bidders that possess privately known valuations. All players, including the auctioneer and the bidders, are risk neutral. Let F be the set of feasible allocations of the resources amongst the bidders and the auctioneer, and $T = \{t_1, t_2, ..., t_m\}$ is a finite set of a bidder's types (possibly multi-dimensional). That is, each bidder privately observes the realization of this signal $t_i \in T$. A collection of types one for each (of n bidders) will be called a "profile" t, and a profile involving only $n-1$ bidders will be denoted t^{n-1}. Let T^n denote the set of all possible profiles.

Given an allocation $a \in F$, if a bidder has type t_i while other bidders have type profile t^{n-1}, she assigns monetary valuation $v(a|t_i, t^{n-1})$ to the allocation a. In other words, $v(a|t_i, t^{n-1})$ is the gross "utility" a bidder receives from the allocation. Here, we do not limit our formulation to the private value model, i.e., the valuation of each bidder, $v(a|t_i, t^{n-1})$, can depend on the realized types of the bidder herself and that of the other bidders. For the purpose of this paper, we assume that bidders are ex ante symmetric. In other words, we assume that bidders' types are independent draws from a common distribution that is commonly known. Specifically, we let $f_i > 0$ denote the probability that a bidder has type t_i. The probability of a profile $t^{n-1} \in T^{n-1}$ being realized is $\pi(t^{n-1}) > 0$.

By the revelation principle, we can restrict our attention to the direct revelation mechanisms (this holds regardless of the common prior assumption). In such a mechanism, each bidder is asked to announce her own type. The auctioneer, as a function of the announcements, decides what element (allocation) of F to pick and what payments each bidder has to make. As aforementioned, our primary goal is to provide

a unified framework that incorporates various types of belief systems depending on how confident the bidders are regarding their estimation of the game. To this end, in the following we first review the classical Bayesian and Ex post formulations of the optimal auction design problem. We then introduce the general framework, of which the two formulations can be regarded as two polar cases.

6.3 Formulations and uncertainty set

In this section, we provide the formulations for the classical solution concepts and introduce our solution concepts.

6.3.1 Bayesian formulation

We now formulate the auctioneer's optimization problem. To this end, we need to introduce the solution concept for the games played by the bidders. In the standard literature on mechanism design, the solution concept is *Bayesian Nash equilibrium*. Despite its popularity, this solution concept requires a relatively strong assumption on the common prior beliefs. Specifically, it requires that each bidder possesses the correct belief about other bidders' types, each bidder knows that each bidder possesses the correct belief about other bidders' types, and so on. Putting it in this particular problem, it translates to the following Bayesian incentive compatibility (BIC) constraint:

$$\begin{aligned} &\sum_{t^{n-1}\in T^{n-1}} v(a_i[t_i,t^{n-1}]|t_i,t^{n-1})\pi(t^{n-1}) - \sum_{t^{n-1}\in T^{n-1}} P(t_i,t^{n-1})\pi(t^{n-1}) \\ \geq &\sum_{t^{n-1}\in T^{n-1}} v(a_j[t_j,t^{n-1}]|t_i,t^{n-1})\pi(t^{n-1}) - \sum_{t^{n-1}\in T^{n-1}} P(t_j,t^{n-1})\pi(t^{n-1}),\ \forall t_i,t_j\in T, \end{aligned}$$

where, on the right-hand side,

$$\sum_{t^{n-1}\in T^{n-1}} v(a_j[t_j,t^{n-1}]|t_i,t^{n-1})\pi(t^{n-1})$$

is the expected utility the type-t_i bidder receives if she pretends to be type-t_j, and $P(t_j,t^{n-1})$ is the payment associated with this misreporting if other bidders' report profile is t^{n-1}. The left-hand side is a special case in which the bidder reports truthfully. This inequality guarantees that the bidder is willing to disclose her type. Moreover, we have to include the Bayesian individual rationality (BIR) condition to ensure that each bidder receives at least a null payoff:

$$\sum_{t^{n-1}\in T^{n-1}} v(a_i[t_i,t^{n-1}]|t_i,t^{n-1})\pi(t^{n-1}) - \sum_{t^{n-1}\in T^{n-1}} P(t_i,t^{n-1})\pi(t^{n-1}) \geq 0,\ \forall t_i\in T.$$

For convenience of our analysis, we define $R(t_i, t^{n-1})$ to be the "*information rent*" to a bidder with type t_i while other bidders have type t^{n-1}:

$$R(t_i, t^{n-1}) \equiv v(a_i[t_i, t^{n-1}]|t_i, t^{n-1}) - P(t_i, t^{n-1}),\ \forall t_i \in T.$$

Using this new notation, we can reformulate the (BIC) and (BIR) constraints as follows:

$$\sum_{t^{n-1}\in T^{n-1}} R(t_i, t^{n-1})\pi(t^{n-1}) - \sum_{t^{n-1}\in T^{n-1}} R(t_j, t^{n-1})\pi(t^{n-1}) \tag{BIC}$$

$$\geq \sum_{t^{n-1}\in T^{n-1}} v(a_j[t_j, t^{n-1}]|t_i, t^{n-1})\pi(t^{n-1}) - \sum_{t^{n-1}\in T^{n-1}} v(a_j[t_j, t^{n-1}]|t_j, t^{n-1})\pi(t^{n-1}),\ \forall t_i, t_j \in T,$$

$$\sum_{t^{n-1}\in T^{n-1}} R(t_i, t^{n-1})\pi(t^{n-1}) \geq 0,\ \forall t_i \in T. \tag{BIR}$$

The auctioneer's problem is to maximize his expected revenue subject to the constraints (BIC) and (BIR). Therefore, assuming that the allocation $a \in F$ has been fixed, the auctioneer's optimization problem under the Bayesian formulation is the following:

$$\begin{aligned} S^b(a) = \max_{P(t_i, t^{n-1})} \quad & \sum_{t_i\in T} f_i \sum_{t^{n-1}\in T^{n-1}} \left\{v(a_i[t_i, t^{n-1}]|t_i, t^{n-1}) - R(t_i, t^{n-1})\right\}\pi(t^{n-1}) \\ \text{s.t.} \quad & \text{(BIC) and (BIR)}, \end{aligned}$$

where the objective function is simply the expected revenue the auctioneer gets from the bidders. Note that from this formulation, $f_i\pi(t^{n-1})$ is the probability that a specific bidder's type is t_i, and other bidders' type profile is $\pi(t^{n-1})$. Furthermore, we can write down the auctioneer's expected payoff from a specific bidder's viewpoint precisely because all these bidders are ex ante symmetric. It is worth mentioning that if the allocation is fixed, the problem degenerates to a simple linear program.

Note that in this Bayesian formulation, we have assume that the bidder knows perfectly well the distributions of realized type profile t^{n-1} of other bidders. This is exactly where the common prior (common knowledge) assumption is used in this particular context. We shall call this case the "*nominal model*."

As a remark, with this representation, we force the auctioneer to specify the payment scheme (or equivalently information rent) for *every realization* of the report profile. Thus, there are much more decision variables (compared to that in Malakhov and Vohra [2005]). However, in our formulation, $\sum_{t^{n-1}\in T^{n-1}} R(t_i, t^{n-1})\pi(t^{n-1})$ can be conveniently redefined as R_i, which is what really matters to the bidders as well as the auctioneer under the Bayesian formulation. This handy change of variables is adopted by Malakhov and Vohra [2005] and most of the papers using the mechanism design approach, and it gives rise to the classical "reduced form" of the auctioneer's optimization problem.

6.3.2 Ex post formulation

The second extreme case is when the bidders completely have no idea of the realizations of other bidders' types. In such a scenario, their equilibrium bidding strategies have to be the best responses, regardless of what other bidders' strategies are. Equivalently, we can write down the following Ex post incentive compatibility (EPIC) constraint to represent this solution concept:

$$R(t_i, t^{n-1}) - R(t_j, t^{n-1}) \geq v(a_j[t_j, t^{n-1}]|t_i, t^{n-1}) - v(a_j[t_j, t^{n-1}]|t_j, t^{n-1}), \quad \text{(EPIC)}$$
$$\forall t_i, t_j \in T,\ t^{n-1} \in T^{n-1}.$$

Furthermore, to ensure that each bidder receives at least a null payoff, we need to impose the Ex post individual rationality condition (EPIR):

$$R(t_i, t^{n-1}) \geq 0,\ \forall t_i \in T,\ t^{n-1} \in T^{n-1}. \quad \text{(EPIR)}$$

Note that both (EPIC) and (EPIR) are much stronger constraints than (BIC) and (BIR) in the Bayesian formulation because they require the inequalities to hold for every instance rather than in expectation. Clearly, (EPIC) implies (BIC) and (EPIR) implies (BIR) as we aggregate these Ex post constraints weighted by the probabilities in the nominal model.

Given these incentive constraints, the auctioneer's optimization problem in this case (Ex post formulation) is:

$$\begin{aligned} S^e(a) = \max_{R(t_i, t^{n-1})} \quad & \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} [v(a_i[t_i, t^{n-1}]|t_i, t^{n-1}) - R(t_i, t^{n-1})]\pi(t^{n-1}) \\ \text{s.t.} \quad & \text{(EPIC) and (EPIR).} \end{aligned}$$

Apparently, this optimization should yield a weakly lower expected profit ($S^e(a)$) for the auctioneer since the constraints are tighter than those in the Bayesian formulation.

The Ex post incentive compatibility is a commonly accepted response by economists to Wilson doctrine (see, e.g., Bergemann and Morris [2005], Bikhchandani et al. [2006], and Chung and Ely [2007]). The primary reason may be that the mechanisms derived from this ex post formulation is "belief-free" as the bidders' (agents') beliefs regarding others and the game do not factor into the formulation. However, under this solution concept, an implicit assumption is that each bidder (agent) has *completely no* information of other bidders' types (or equivalently, other bidders' strategies based on their realized type profile).

If instead, a bidder more or less has an estimation regarding roughly the possible type realizations of others and how others perceive the game, she would not have completely abandoned her one estimation while determining her best response. In such a scenario, it seems appropriate to explicitly model the confidence and uncertainty of a bidder and formally incorporate this into the formulation of optimal auction design.

Incidentally, in Wilson [1987], the statement is that the players may not completely know the correct prior distributions of other players' types. It is certainly legitimate to push this argument to the extreme and derive the Ex post formulation (as is typically done in the economics literature for the past two decades). Nevertheless, it might be also useful to construct a flexible formulation that allows for the intermediate cases in which the bidders do not possess the completely correct beliefs but yet still retain some reasonable expectation/estimations. This is precisely our primary objective in this paper, as we describe next.

6.3.3 Uncertainty set and robust formulation

In the sequel, we propose a continuum of solution concepts that lie in between the two extreme cases – Bayesian and Ex post formulations; these solution concepts allow us to model various situations that account for the bidders' confidence or uncertainty about the beliefs. From the aforementioned two formulations, we find that only the incentive constraints differ in different solution concepts ((BIC), (BIR), (EPIC), and (EPIR)), and the probability distributions $q(t^{n-1})$, $\pi(t^{n-1})$ are actually the coefficients of linear programs. This motivates us to propose the following formulation for bidders' belief systems. Specifically, we shall take $\pi(t^{n-1})$ as the nominal model and define the following "*uncertainty set*":

$$U^{\varepsilon} = \{\mathbf{q} \in [0,1]^{n-1} : \ |q(t^{n-1}) - \pi(t^{n-1})| \leq \varepsilon, \sum_{t^{n-1} \in T^{n-1}} q(t^{n-1}) = 1\}.$$

Note that we require $\sum_{t^{n-1} \in T^{n-1}} q(t^{n-1}) = 1$ since a bidder's belief regarding other bidders' types have to be consistent even if it does not coincide with the correct prior. This uncertainty set nicely provides a ground for us to represent the confidence, conservatism, and uncertainty the bidders are endowed with. The larger the value of ε, the more conservative the bidders are, (or equivalently, the less confident they are regarding their beliefs on other bidders' types). An interpretation for the existence of uncertainty set is that bidders are endowed with incomplete preferences (see Lopomo et al. [2007]).

Based on the uncertainty set U^{ε}, we define the robust incentive compatibility (RIC) constraints as:

$$\sum_{t^{n-1} \in T^{n-1}} R(t_i, t^{n-1}) q(t^{n-1}) - \sum_{t^{n-1} \in T^{n-1}} R(t_j, t^{n-1}) q(t^{n-1}) \tag{RIC}$$

$$\geq \sum_{t^{n-1} \in T^{n-1}} [v(a_j[t_j, t^{n-1}]|t_i, t^{n-1}) - v(a_j[t_j, t^{n-1}]|t_j, t^{n-1})] q(t^{n-1}), \ \forall t_i, t_j \in T, \ \mathbf{q} \in U^{\varepsilon}$$

Similarly, the robust individual rationality (RIR) constraints are defined as:

$$\sum_{t^{n-1} \in T^{n-1}} R(t_i, t^{n-1}) q(t^{n-1}) \geq 0, \ \forall t_i \in T, \ \mathbf{q} \in U^{\varepsilon}. \tag{RIR}$$

The auctioneer's *robust mechanism design problem* is defined as follows:

$$S^{\varepsilon}(a) = \max_{R(t_i,t^{n-1})} \quad \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} [v(a_i[t_i, t^{n-1}]|t_i) - R(t_i, t^{n-1})]\pi(t^{n-1})$$
$$\text{s.t.} \quad \text{(RIC) and (RIR)}.$$

If $\varepsilon = 0$, the uncertainty set is a singleton that contains only the correct prior: $U^0(\pi) = \{\pi\}$. This corresponds to our nominal model and the corresponding solution concept is Bayesian Nash equilibrium, i.e.,

$$S^0(a) = S^b(a), \ \forall a \in F.$$

On the other hand, if $\varepsilon = 1$, any prior is contained in the uncertainty set: $U^1(\pi) = \{q(t^{n-1}) \in [0,1]^{n-1} : \sum_{t^{n-1} \in T^{n-1}} q(t^{n-1}) = 1\}$ because $|q(t^{n-1}) - \pi(t^{n-1})| \leq 1$ is redundant. The auctioneer's robust optimization problem becomes equivalent to the Ex post formulation:

$$S^1(a) = S^e(a), \ \forall a \in F.$$

Therefore, the existing two solution concepts are actually two extreme cases of this general class of solution concepts. In general, we should be able to identify a continuum of optimization problems, each of which corresponds to a different solution concept that captures the conservativeness of bidders' beliefs.

¿From the above formulations, the objective functions are identical, but the incentive constraints ((RIC) and (RIR)) are more restrictive as we increase the value of ε. This implies that the feasible region of this class of optimization problems becomes larger when we increase the measure of belief uncertainty, ε. This observation immediately leads to the following result:

Lemma 1. *In the auction game among bidders with belief uncertainty, 1) The set of equilibria is nested for all ε, and is enlarging as ε becomes larger; 2)The auctioneer's expected revenue is decreasing in ε.*

Lemma 1 shows that the auctioneer is averse to the uncertainty the bidders possess regarding their beliefs, and it has a clear economic intuition. When bidders are more uncertain regarding which types of bidders they are bidding with (as ε becomes larger), their strategies are more conservative. Thus, the set of mechanisms that the auctioneer can select amongst is smaller; consequently, he collects a (weakly) lower expected revenue.

It is worth mentioning that the idea of using robust optimization to formulate a solution concept was first proposed in Aghassi and Bertsimas [2006]. However, in their paper, the uncertainty is on the players' payoff profiles, in games with perfect or private information. In our setting, this corresponds to the case in which the form of $v(\cdot|t_i, t^{n-1})$ is unknown, but *the beliefs are correct*. Since the players in Aghassi and Bertsimas [2006] are uncertain about their own payoffs (even though they have observed their types), the players' goal is to find a strategy that provides

the optimal worst case performance guarantee; on the other hand, the uncertainty in our formulation generates a larger set of constraints that the auctioneer has to encounter while designing the optimal auction. Another key difference of our paper and Aghassi and Bertsimas [2006] is that Aghassi and Bertsimas [2006] mainly focus on proving the existence of "robust optimization equilibrium" in finite games, while we focus on the revenue difference under different solution concepts.

Having described the general robust formulation of the auction design problem, we next proceed to characterize the difference between the intermediate case and the two extreme cases, namely the Bayesian and Ex post formulations. Following this, we then provide some examples to illustrate the discrepancies explicitly.

6.4 Robust formulations of fixed allocations

In this section, we first introduce the network approach developed by Malakhov and Vohra [2005] to reformulate our optimal auction design problems. We then apply it to characterize the necessary and sufficient conditions (for a fixed allocation) to achieve the same revenue in a robust formulation as in the nominal Bayesian formulation.

6.4.1 The network approach

First, we introduce the network approach to solve the robust auction design problem. Our analysis closely follows the elegant framework by Malakhov and Vohra [2005], except that we work with the information rent $\mathbf{R}$, instead of the payment $\mathbf{P}$. Our goal is to recast the optimal auction design as a network design problem and utilize the established techniques in the network flow literature, as we elaborate in the sequel.

For fixed allocation $a \in F$, let us define a complete directed graph $G(a, \mathbf{q}) = (N, A, w(a, \mathbf{q}))$, for each $\mathbf{q} \in U^{\varepsilon}$. Here, $N = \{n_0,\ n_1,\ \cdots, n_{|T|}\}$ is the set of nodes, with $n_1,\ \cdots,\ n_{|T|}$ each corresponding to a type in T, and n_0 being the "pseudo" node. The set $A = \{(i, j) :\ 0 \leq i, j \leq |T|\}$ contains the edges of the graph, each corresponding to an incentive compatible constraint. To transform our optimal auction design problem to a network design problem, we can denote the weights on the edges as

$$w_{ji}(a, \mathbf{q}) = \sum_{t^{n-1} \in T^{n-1}} [v(a_j[t_j, t^{n-1}]|t_i) - v(a_j[t_j, t^{n-1}]|t_j)]q(t^{n-1}),\ \forall 1 \leq i, j \leq |T|,$$

and

$$w_{0i}(a, \mathbf{q}) = w_{i0}(a, \mathbf{q}) = 0,\ \forall 1 \leq i \leq |T|.$$

Let $d_i(a, \mathbf{q})$ be the length of the longest path from n_0 to n_i in $G(a, \mathbf{q})$, if one exists. We say that an allocation $a \in F$ is *robust incentive compatible with respect to uncertainty set* U^{ε}, if we can find a pricing/rent mechanism that satisfies (RIC). Lemma

2 reveals the connection between the robust auction design problem and the *longest path problem* in G.

Lemma 2. *If $a \in F$ is robust incentive compatible with respect to U^ε, the following results hold:*

1. *There is no positive cost cycle and a longest path exits between each pair of nodes in $G(a, \mathbf{q})$, for all $\mathbf{q} \in U^\varepsilon$;*
2. *For all $t_i \in T$ and $\mathbf{q} \in U^\varepsilon$, any feasible pricing/rent scheme $\mathbf{R}$ satisfies*

$$\sum_{t^{n-1} \in T^{n-1}} R(t_i, t^{n-1}) q(t^{n-1}) \geq d_i(a, \mathbf{q}).$$

Proof. Let us first prove the first claim. For an arbitrary cycle $C = \{(i_1, i_2), \cdots, (i_{k-1}, i_k), (i_k, i_1)\}$ in $G(a, \mathbf{q})$, it follows that

$$\begin{aligned}
\sum_{(i,j) \in C} w_{ij}(a, \mathbf{q}) &= \sum_{t^{n-1} \in T^{n-1}} [v(a_{i_1}[t_{i_1}, t^{n-1}]|t_{i_2}) - v(a_{i_1}[t_{i_1}, t^{n-1}]|t_{i_1})] q(t^{n-1}) + \cdots \\
&\quad + \sum_{t^{n-1} \in T^{n-1}} [v(a_{i_{k-1}}[t_{i_{k-1}}, t^{n-1}]|t_{i_k}) - v(a_{i_{k-1}}[t_{i_{k-1}}, t^{n-1}]|t_{i_{k-1}})] q(t^{n-1}) \\
&\quad + \sum_{t^{n-1} \in T^{n-1}} [v(a_{i_{k-1}}[v(a_{i_k}[t_{i_k}, t^{n-1}]|t_{i_1}) - v(a_{i_k}[t_{i_k}, t^{n-1}]|t_{i_k})] q(t^{n-1}) \\
&\leq [R(t_{i_2}, t^{n-1}) - R(t_{i_1}, t^{n-1})] + \cdots + [R(t_{i_k}, t^{n-1}) - R(t_{i_{k-1}}, t^{n-1})] \\
&\quad + [R(t_{i_1}, t^{n-1}) - R(t_{i_k}, t^{n-1})] \\
&= 0,
\end{aligned}$$

where the first inequality is implied by (RIC). This suggests that there should be no positive cost cycle in $G(a, t^{n-1})$ and asserts the first claim.

Let us now switch to the second claim. The proof is by induction. Specifically, from the first claim we know that there exist a longest path between each pair of nodes in $G(a, \mathbf{q})$. For each $t_i \in T$, let $P_i(a, \mathbf{q})$ be a longest path from n_0 to n_i in $G(a, \mathbf{q})$, and let

$$A(a, \mathbf{q}) \equiv \bigcup_{t_i \in T} P_i(a, \mathbf{q}).$$

¿From the first claim, $A(a, \mathbf{q})$ is acyclic, thus there exists a topological ordering of the nodes $0, j_1, \cdots, j_{|T|}$, such that for all edges $(j_u, j_v) \in A(a, \mathbf{q})$, $u < v$.

Since $n_0 \to n_{j_1}$ is the only path to n_{j_1} in $A(a, \mathbf{q})$, it follows that

$$\sum_{t^{n-1} \in T^{n-1}} R(t_{j_1}, t^{n-1}) q(t^{n-1}) \geq 0 = w_{0j_1}(a, \mathbf{q}) = d_{j_1}(a, \mathbf{q}),$$

where the first inequality is implied by (RIR). For $u > 1$, let j_v ($v < u$) be the predecessor of j_u in $A(a, \mathbf{q})$. It follows that

$$\begin{aligned}
& \sum_{t^{n-1} \in T^{n-1}} R(t_{j_u}, t^{n-1}) q(t^{n-1}) \\
\geq & \sum_{t^{n-1} \in T^{n-1}} R(t_{j_v}, t^{n-1}) q(t^{n-1}) + \sum_{t^{n-1} \in T^{n-1}} [v(a_{j_v}[t_{j_v}, t^{n-1}]|t_{j_u}) - v(a_{j_v}[t_{j_v}, t^{n-1}]|t_{j_v})] q(t^{n-1}) \\
\geq & \; d_{j_v}(a, \mathbf{q}) + w_{j_v j_u}(a, \mathbf{q}) \\
= & \; d_{j_u}(a, \mathbf{q}),
\end{aligned}$$

where the first inequality follows from (RIC), and the second one is implied by the inductive assumption. □

With the help of Lemma 2, we are now ready to state our main results.

6.4.2 Main results

Let $a \in F$ be a robust incentive compatible allocation with respect to U^ε. We say that the robust formulation *achieves the same expected revenue* as the Bayesian formulation under allocation a, if there exists a robust incentive compatible payment/rent scheme $\mathbf{R}$ that generates the Bayesian optimal revenue, i.e. $S^\varepsilon(a) = S^b(a)$. Note that this condition is weaker than the characterization of *revenue equivalence* in Heydenreich et al. [2009], which requires all Ex post incentive compatible payment schemes to generate the Bayesian optimal revenue.

For each $\mathbf{q} \in U^\varepsilon$, we define $\mathfrak{L}_i(a, \mathbf{q})$ to be the set of longest paths from the pseudo node n_0 to n_i in graph $G(a, \mathbf{q})$. Theorem 2 characterizes the necessary and sufficient conditions for a robust formulation to achieve the same revenue as a Bayesian formulation (the nominal case).

Theorem 2. *For a fixed allocation $a \in F$ that is robust incentive compatible with respect to U^ε, $S^\varepsilon(a) = S^b(a)$ if and only if*

$$\bigcap_{\mathbf{q} \in U^\varepsilon} \mathfrak{L}_i(a, \mathbf{q}) \neq \phi, \; \forall t_i \in T. \tag{6.1}$$

Proof. If (6.1) holds, there must exist a path $P_i(a) \in \mathfrak{L}_i(a, \mathbf{q})$, for each $t_i \in T$ and $\mathbf{q} \in U^\varepsilon$. Let $d_i(a)$ represent the length of $P_i(a)$. We claim that the optimal solution to the robust formulation is

$$R_a^\varepsilon(t_i, t^{n-1}) = d_i(a), \; \forall t^{n-1} \in T^{n-1},$$

and the proof goes as follows.

First, the above solution satisfies (RIC) because for all $\mathbf{q} \in U^\varepsilon$,

$$\begin{aligned}
& \sum_{t^{n-1} \in T^{n-1}} R(t_i, t^{n-1}) q(t^{n-1}) - \sum_{t^{n-1} \in T^{n-1}} R(t_j, t^{n-1}) q(t^{n-1}) = d_i(a) - d_j(a) \geq w_{ji}(a, \mathbf{q}) \\
= & \sum_{t^{n-1} \in T^{n-1}} [v(a_j[t_j, t^{n-1}]|t_i, t^{n-1}) - v(a_j[t_j, t^{n-1}]|t_j, t^{n-1})] q(t^{n-1}),
\end{aligned}$$

where the inequality follows from the properties of longest paths in a network.

Let R_a^B be the optimal solution to the Bayesian formulation. It follows that

$$\begin{aligned} S^{\varepsilon}(a) &= \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} \pi(t^{n-1}) v(a_i[t_i, t^{n-1}] | t_i, t^{n-1}) - \sum_{t_i \in T} f_i d_i(a) \\ &\geq \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} \pi(t^{n-1}) v(a_i[t_i, t^{n-1}] | t_i, t^{n-1}) - \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} \pi(t^{n-1}) R_a^B(t_i, t^{n-1}) \\ &= S^b(a), \end{aligned}$$

where the inequality follows from Lemma 2. However, $S^{\varepsilon}(a) \leq S^b(a)$ from Lemma 1, which indicates that $S^{\varepsilon}(a) = S^b(a)$.

Conversely, if (6.1) does not hold, there exists $\mathbf{q}^1 \in U^{\varepsilon}$ such that

$$\mathfrak{L}_i(a, \mathbf{q}^1) \bigcap \mathfrak{L}_i(a, \pi) = \phi.$$

Let $\mathbf{q}^2 \in U^{\varepsilon}$ be a probability vector such that $\pi = \alpha \mathbf{q}^1 + (1 - \alpha)\mathbf{q}^2$ for some $0 < \alpha < 1$ (since π is an interior point of U^{ε}, such a vector always exists). Also let $P_i(a, \mathbf{q})$ be a longest path from n_0 to n_i in $G(a, \mathbf{q})$, i.e. $P_i(\mathbf{q}) \in \mathfrak{L}_i(a, \mathbf{q})$, for each $1 \leq i \leq |T|$ and $q \in U^{\varepsilon}$.

Since the social surplus is the same in the robust and the Bayesian formulations for fixed allocation, it suffices just to compare the expected information rent. Let $\mathbf{R}_a^B$ and $\mathbf{R}_a^{\varepsilon}$ be the optimal rent in the Bayesian and the robust formulations, respectively.

Following an argument similar to the first part of the proof, $\mathbf{R}_a^B$ satisfies

$$\sum_{t^{n-1} \in T^{n-1}} R_a^B(t_i, t^{n-1}) \pi(t^{n-1}) = d_i(a, \pi), \ \forall t_i \in T.$$

Therefore, the expected rent in the Bayesian formulation is:

$$E[\mathbf{R}_a^B] = \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} R_a^B(t_i, t^{n-1}) \pi(t^{n-1}) = \sum_{t_i \in T} f_i d_i(a, \pi),$$

and the expect rent in the robust formulation is:

$$\begin{aligned} E[\mathbf{R}_a^{\varepsilon}] &= \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} R_a^{\varepsilon}(t_i, t^{n-1}) \pi(t^{n-1}) \\ &= \sum_{t_i \in T} f_i \sum_{t^{n-1} \in T^{n-1}} R_a^{\varepsilon}(t_i, t^{n-1}) [(\alpha) q^1(t^{n-1}) + (1 - \alpha) q^2(t^{n-1})] \\ &\geq \sum_{t_i \in T} f_i [(\alpha) d_i(a, \mathbf{q}^1) + (1 - \alpha) d_i(a, \mathbf{q}^2)], \end{aligned}$$

where the inequality follows from Lemma 2.

Since $d_i(a, \mathbf{q}^2)$ is the length of the longest path to n_i in $G(a, \mathbf{q}^2)$, it is greater than the length of $P_i(a, \pi)$, which is a longest path in $G(a, \pi)$, but not necessarily the longest in $G(a, \mathbf{q}^2)$, i.e.,

$$d_i(a, \mathbf{q}^2) \geq \sum_{(u,v) \in P_i(\pi)} w_{uv}(a, \mathbf{q}^2).$$

Moreover, as $P_i(a, \pi)$ is not a longest path in $G(a, \mathbf{q}^1)$, it follows that

$$d_i(a, \mathbf{q}^1) > \sum_{(u,v) \in P_i(\pi)} w_{uv}(a, \mathbf{q}^1).$$

Putting everything together, we have

$$\begin{aligned} E[\mathbf{R}_a^{\varepsilon}] &> \sum_{t_i n \in T} f_i \sum_{(u,v) \in P_i(\pi)} [(\alpha) w_{uv}(a, \mathbf{q}^1) + (1-\alpha) w_{uv}(a, \mathbf{q}^2)] \\ &= \sum_{t_i n \in T} f_i \sum_{(u,v) \in P_i(\pi)} w_{uv}(a, \pi) \\ &= \sum_{t_i n \in T} f_i d_i(a, \pi) \\ &= E[\mathbf{R}_a^{B}]. \end{aligned}$$

This implies that $S^{\varepsilon}(a) < S^{b}(a)$ and completes our proof. □

Because the Ex post formulation can be taken as a special case of the robust formulation when $\varepsilon = 1$, we can derive the conditions under which it achieves the same expected revenue as in the nominal case, as a corollary to Theorem 2. To simplify the notation in this case, we define $G(a, t^{n-1}) \equiv G(a, \mathbf{q}^{t^{n-1}})$ and $\mathfrak{L}_i(a, t^{n-1}) \equiv \mathfrak{L}_i(a, \mathbf{q}^{t^{n-1}})$, where $\mathbf{q}^{t^{n-1}}(t^{n-1}) = 1$ and $\mathbf{q}^{t^{n-1}}(s^{n-1}) = 0$ for all $s^{n-1} \in T^{n-1} \setminus \{t^{n-1}\}$.

Corollary 1. *For an Ex post incentive compatible allocation $a \in F$, $S^{e}(a) = S^{b}(a)$, if and only if*

$$\bigcap_{t^{n-1} \in T^{n-1}} \mathfrak{L}_i(a, t^{n-1}) \neq \phi, \ \forall t_i \in T. \tag{6.2}$$

The next corollary provides sufficient conditions for a robust formulation of $a \in F$ to achieve the same expected revenue for the auctioneer as in the Ex post formulation. Here, we introduce new notation: define Θ^{ε} as the set of extreme points of U^{ε}, and $\tau(\mathbf{q}) = \{t^{n-1} \in T^{n-1} : \ q(t^{n-1}) > 0\}$.

Corollary 2. *For an Ex post incentive compatible allocation $a \in F$, $S^{\varepsilon}(a) = S^{e}(a)$ if*

$$\bigcap_{t^{n-1} \in \tau(\mathbf{q})} \mathfrak{L}_i(a, t^{n-1}) \neq \phi, \ \forall t_i \in T, \ \mathbf{q} \in \Theta^{\varepsilon}. \tag{6.3}$$

The proofs of Corollaries 1 and 2 follow directly from that of Theorem 2 and thus are omitted from our discussion. It should be noted that condition (6.2) in Corollary 1 is weaker than the characterization of revenue equivalence in Theorem 1 of Heydenreich et al. [2009]. Specifically, in Heydenreich et al. [2009], the necessary and sufficient condition (under our setting) for an Ex post incentive compatible allocation $a \in F$ to satisfy revenue equivalence is

$$d_{ij}(a, t^{n-1}) = -d_{ji}(a, t^{n-1}),\ \forall t_i, t_j \in T,\ t^{n-1} \in T^{n-1}, \tag{6.4}$$

where $d_{ij}(a, t^{n-1})$ is the length of the longest path from n_i to n_j in $G(a, t^{n-1})$. It can be verified that (6.4) is equivalent to the condition that all paths have equal length between any pair of nodes in $G(a, t^{n-1})$ for all $t^{n-1} \in T^{n-1}$. Clearly, (6.4) indicates (6.2), but not vice versa. Thus, our results can be regarded as complementary to those in Heydenreich et al. [2009].

In the next section, we analyze Bayesian optimal allocations in single and multiple object auctions; we then focus on the revenue difference between various robust formulations.

6.5 Robust formulations given Bayesian optimal allocations

In this section, we demonstrate whether and when the different formulations give rise to different expected revenues.

6.5.1 Single object auction- Myerson's case

We first investigate the setting of Myerson [1981], in which the auctioneer intends to sell a single object to bidders with private valuations. The valuation follows a continuous distribution that satisfies the monotone hazard rate condition. Malakhov and Vohra [2005] consider a similar problem under the discrete setting that allows for more transparent network representations. In this section, we adopt this discrete setting and their assumptions:

$$v(\rho_i|t_i) = t_i\rho_i,\ \text{and}\ \frac{1-F_i}{f_i} \geq \frac{1-F_j}{f_j},\ \text{if}\ t_i \geq t_j,$$

where ρ_i is the expected quantity allocation to a type t_i bidder, and $F_i = \sum_{j:\ t_j \leq t_i} f_j$ is the cumulative distribution of a bidder's type. Malakhov and Vohra [2005] show that the optimal allocation for the Baysian formulation is a standard auction with reservation price $x^* = \min\{t_i \in T :\ t_i - \frac{1-F_i}{f_i} \geq 0\}$, which coincides with results in the continuous setting in Myerson [1981]. Here, we show that this optimal allocation has a robust formulation that achieves the same expected revenue, irrespective of the associated uncertainty set.

Theorem 3. *The Bayesian optimal allocation for Myerson's single-object auction, a^B, achieves the same optimal expected revenue in any robust formulation, i.e.,*

$$S^{\varepsilon}(a^B) = S^{b}(a^B),\ \forall 0 < 1 < \varepsilon.$$

Proof. Without loss of generality, we assume that the types are ordered such that $t_1 < t_2 < \cdots < t_m$**.** We first show that for each i $(1 \leq i \leq |T|)$, a longest path from n_0 to n_i in $G(a^B, \mathbf{q})$ is $n_0 \to n_1 \to \cdots \to n_i$. That is, the longest path is the same for all $\mathbf{q} \in U^{\varepsilon}$, or equivalently,

$$d_i(a^B, \mathbf{q}) = \sum_{k=1}^{i} w_{k-1,k}(a^B, \mathbf{q}),\ \forall \mathbf{q} \in U^{\varepsilon}.$$

To show that $d_i(a^B, \mathbf{q})$ is indeed the length of a longest path, let $t_i > t_j$ be two arbitrary types. It follows that

$$\begin{aligned}
d_i(a^B, \mathbf{q}) - d_j(a^B, \mathbf{q}) &= \sum_{k=1}^{i} w_{k-1,k}(a^B, \mathbf{q}) - \sum_{k=1}^{j} w_{k-1,k}(a^B, \mathbf{q}) \\
&= \sum_{k=j+1}^{i} w_{k-1,k}(a^B, \mathbf{q}) \\
&= \sum_{k=j+1}^{i} \sum_{t^{n-1} \in T^{n-1}} [v(a^B_{k-1}[t_{k-1}, t^{n-1}]|t_k) - v(a^B_{k-1}[t_{k-1}, t^{n-1}]|t_{k-1})] q(t^{n-1}) \\
&\geq \sum_{k=j+1}^{i} \sum_{t^{n-1} \in T^{n-1}} [v(a^B_{j}[t_{j}, t^{n-1}]|t_k) - v(a^B_{j}[t_{j}, t^{n-1}]|t_{k-1})] q(t^{n-1}) \\
&= \sum_{t^{n-1} \in T^{n-1}} v(a^B_{j}[t_{j}, t^{n-1}]|t_i) - v(a^B_{j}[t_{j}, t^{n-1}]|t_j) q(t^{n-1}) \\
&= w_{ij}(a^B, \mathbf{q}),
\end{aligned}$$

where the inequality follows from the monotonicity of a^B in a standard auction.

Following a similar argument, we can show that $d_i(a^B, \mathbf{q}) - d_j(a^B, \mathbf{q}) \geq w_{ij}(a^B, \mathbf{q})$ if $t_i < t_j$. This implies that $d_i(a^B, \mathbf{q})$ is indeed the length of a longest path in $G(a^b B, \mathbf{q})$. Clearly, $\bigcap_{\mathbf{q} \in U^{\varepsilon}} \mathfrak{L}_i(a^B, \mathbf{q}) \neq \phi$, for all $t_i \in T$. Following Theorem 2, $S^b(a^B) = S^{\varepsilon}(a^B)$. □

It is well known that in the single-object auction design problem, when the bidders' valuations are private-valued and follow a continuous distribution, the optimal auction can be implemented by a second-price auction with an appropriately chosen reservation price (Myerson [1981]). Since the second-price auction can sustain truth-telling as a dominant strategy equilibrium, it is not surprising that all the intermediate case under any robust formulation should yield the same expected revenue for the auctioneer in the continuous setting. Note that the proof of Theorem 3 also

applies to arbitrary feasible allocation, as we do not explicitly use any property of the optimal allocation. Thus, as anticipated, it can be slightly generalized to establish the equivalence between the Bayesian solution and any other robust formulation. On a related note, Theorem 3 also follows as a corollary of Theorem 1 in Heydenreich et al. [2009], which implies that any Bayesian incentive compatible mechanism can be implemented in dominant strategy, in a single object auction with discrete value. However, as we show in Section 6.5.2, this result no longer holds in a multiple object auction.

6.5.2 Multiple object auction- Armstrong's case

We now consider the multi-object auction design problem introduced by Armstrong [2000]. In this setting, the auctioneer intends to sell two objects, denoted as A and B. A bidder's type t is described by a pair (ν^A, ν^B), where ν^ℓ is the bidder's valuation for object ℓ. If an allocation $a = (p^A, p^B)$ awards object A (B) to this bidder with probability q^A (q^B), her gross utility is

$$v(a|t) = p^A \nu^A + p^B \nu^A.$$

In Armstrong [2000], it is assumed that $\nu^\ell \in \{\nu_L^\ell,\ \nu_H^\ell\}$, where $\Delta^\ell = \nu_H^\ell - \nu_L^\ell > 0$. Thus there are four types of bidder corresponding to the four realizations $\{(\nu_L^A,\ \nu_L^B), (\nu_L^A,\ \nu_H^B), (\nu_H^A,\ \nu_L^B), (\nu_H^A,\ \nu_H^B)\}$, and these types are denoted by LL, LH, HL, and HH. The probability that a bidder has type ij is f_{ij}, where $f_{LL} + f_{LH} + f_{HL} + f_{HH} = 1$. Let f_L^ℓ and $f_H^\ell = 1 - f_L^\ell$ denote the marginal probability of having a high or low valuation for object ℓ, respectively; in other words, $f_L^A = f_{LL} + f_{LH}$, $f_H^A = f_{HL} + f_{HH}$, and likewise for object B. This is arguably the simplest possible setting that fully demonstrates the complicated nature of the multi-object auction design problem. Furthermore, this four-type framework has been adopted as a fixture in the multi-dimensional mechanism design problem, see, e.g., Armstrong and Rochet [1999] and Asker and Cantillon [2008].

Results from Armstrong [2000]

Before we demonstrate how the various solution concepts apply to this multi-object auction design problem, let us first revisit the results in Armstrong [2000]. Following the convention in the auction theory, Armstrong [2000] focuses exclusively on the Bayesian formulation of this problem. He finds that, the optimal auction crucially depends on the "correlation" between the bidders' valuations for the two objects, defined as follows:

$$\lambda^A = \frac{f_{HH} f_L^A}{f_{LH} f_H^A},\ \lambda^B = \frac{f_{HH} f_L^B}{f_{HL} f_H^B}.$$

Based on the above model characteristics, Armstrong [2000] identifies the structural properties of the optimal auctions. If $1/\lambda^A + 1/\lambda^B \leq 1$, i.e., there is strong

positive correlation, the optimal allocation takes the format of an *Independent Auction*, in which ρ_{ij}^{ℓ}, the expected quantity of object ℓ that is allocated to a type-ij bidder is prescribed as:

$$\begin{aligned}\rho_{HH}^{A} &= \rho_{HL}^{A} = \frac{1-(f_L^A)^n}{nf_H^A}, \quad \text{and } \rho_{LL}^{A} = \rho_{LH}^{A} = \frac{(f_L^A)^{n-1}}{n};\\ \rho_{HH}^{B} &= \rho_{LH}^{B} = \frac{1-(f_L^B)^n}{nf_H^B}, \quad \text{and } \rho_{LL}^{B} = \rho_{HL}^{B} = \frac{(f_L^B)^{n-1}}{n}.\end{aligned}$$

When $1/\lambda^A + 1/\lambda^B \geq 2$, there is negative correlation. Armstrong [2000] characterizes some other technical conditions (for example, the symmetric case and the case in which there are sufficiently many bidders) and shows that the optimal allocation takes the format of a *Bundling Auction*. Under this format, the allocation of an object to a bidder with high valuation for the object is the same as in the Independent Auction, and the allocation to low-valuation bidders is given below:

$$\begin{aligned}\rho_{LH}^{A} &= \frac{(f_L^A)^n - f_{LL}^n}{nf_{LH}}, \text{ and } \rho_{LL}^{A} = \frac{f_{LL}^{n-1}}{n};\\ \rho_{HL}^{B} &= \frac{(f_L^B)^n - f_{LL}^n}{nf_{HL}}; \text{ and } \rho_{LL}^{B} = \frac{f_{LL}^{n-1}}{n}.\end{aligned}$$

In all other cases, the optimal allocation is a *Mixed Auction*, which is a convex combination of the Independent Auction and the Bundling Auction.

Difference between the formulations

We now introduce other solution concepts for this problem. Our first result is that the auctioneer's expected revenue may be strictly less if the Ex Post formulation is used instead of the Baysian formulation. To this end, let us introduce a new notation, in which each profile $t^{n-1} \in T^{n-1}$ is specified with a triplet (k_1, k_2, k_3), where k_1, k_2, and k_3 respectively are the number of bidders with types LH, HL and HH, and $0 \leq k_1 + k_2 + k_3 \leq n-1$.

Theorem 4. *Unless the Bayesian optimal allocation is the Independent Auction for both objects, the auctioneer's expected revenue under the Ex post formulation is strictly lower than that under the Bayesian formulation, i.e., $S^e(a^B) > S^b(a^B)$.*

Proof. Suppose that a^B is an Bayesian optimal allocation (specified by the expected quantities ρ). We can then construct a detailed allocation $\mathbf{p}$ that satisfies

$$\rho_{ij}^{\ell} = \sum_{t^{n-1} \in T^{n-1}} p_{ij}^{\ell}(t^{n-1})\pi(t^{n-1}), \ \forall\ (i,j) \in T, \ \ell \in \{A, B\}.$$

It is straightforward that an optimal allocation to the high-valuation bidders should satisfy

$$p_{HH}^{A}(k_1,k_2,k_3) = p_{HL}^{A}(k_1,k_2,k_3) = \frac{1}{k_2+k_3+1}, \text{ and}$$

$$p^B_{HH}(k_1,k_2,k_3) = p^B_{LH}(k_1,k_2,k_3) = \frac{1}{k_1+k_3+1}.$$

To characterize the allocation to the low-valuation bidders, we restrict our attention to the case where $k_3 = 0$ and $k_1k_2 = 0$, since otherwise the solution is trivial. Now there are three possible scenarios, depending on the structural properties of the Bayesian optimal allocation a^B. In the following we discuss them separately.

Case 1. *Independent Auction for both objects.*

It is clear that if the Bayesian optimal allocation is an independent auction, the auctioneer can obtain the same expected revenue under the Ex post formulation by utilizing "the same" allocation. Specifically, we can construct a detailed allocation that is equivalent to the independent auction as follows

$$p^A_{LL}(k,0,0) = p^A_{LH}(k,0,0) = \frac{1}{n}, \text{ and } p^B_{LL}(0,k,0) = p^B_{HL}(0,k,0) = \frac{1}{n}.$$

It is easy to verify that the detailed allocation $\mathbf{p}$ satisfy condition (6.2) in Corollary 1; therefore, it achieves the same expected revenue under the Ex post formulation.

Case 2. *Bundling Auction for both objects.*
In this case, the unique equivalent detailed allocation is given by

$$\begin{aligned}
p^A_{LL}(0,0,0) &= \frac{1}{n},\ p^A_{LL}(k,0,0) = 0,\ \forall\ 1 \le k \le n-1,\\
p^A_{LH}(k,0,0) &= \frac{1}{k+1},\ \forall\ 1 \le k \le n-1,\\
p^B_{LL}(0,0,0) &= \frac{1}{n},\ p^B_{LL}(0,k,0) = 0,\ \forall\ 1 \le k \le n-1,\\
p^B_{HL}(0,k,0) &= \frac{1}{k+1},\ \forall\ 1 \le k \le n-1.
\end{aligned}$$

It is straightforward that for $k \ge 1$ and $t^{n-1} = (k,0,0)$, $p^A_{LH}(t^{n-1}) > p^A_{LL}(t^{n-1})$. Thus, $LL \to LH \to HH$ is the unique longest path from the LL type to the HH type in graph $G(a^b, t^{n-1})$. Similarly, for $k \ge 1$ and $t^{n-1} = (0,k,0)$, $p^B_{HL}(t^{n-1}) > p^B_{LL}(t^{n-1})$. Therefore, $LL \to HL \to HH$ is the unique longest path. Clearly, condition (6.2) in Corollary 1 is violated, which indicate that $S^e(a^B) < S^b(a^B)$.

Case 3. *Bundling Auction for one object and Mixed Auction for the other.*
According to Armstrong [2000], the expected allocation in this case satisfies

$$\Delta^A(\rho^A_{LH} - \rho^A_{LL}) = \Delta^B(\rho^B_{HL} - \rho^B_{LL}). \tag{6.5}$$

Suppose that there exists an equivalent detailed allocation $\mathbf{p}$ which achieves the same expected revenue in the Ex post formulation. We first consider the case in which the optimal allocation for object A is Bundling Auction.

For $t^{n-1} = (k,0,0)$, $1 \le k \le n-1$, the detailed allocation of the two objects

$$p^A_{LL}(k,0,0) = 0,\ p^A_{LH}(k,0,0) = \frac{1}{k+1},\ p^B_{LL}(k,0,0) = 0,\ p^B_{HL}(k,0,0) = 0,$$

which implies that

$$\Delta^B[p^B_{HL}(0,k,0) - p^B_{LL}(0,k,0)] = 0 < \Delta^A[p^A_{LH}(0,k,0) - p^A_{LL}(0,k,0)]. \tag{6.6}$$

This detailed allocation also indicates that $LL - LH - HH$ is the unique longest path in $G(a^b, t^{n-1})$ if $t^{n-1} = (k,0,0)$. From Corollary 1, $LL - LH - HH$ must also be a longest path in $G(a^b, t^{n-1})$ for all $t^{n-1} \in T^{n-1}$.

For $t^{n-1} = (0,k,0)$, $1 \leq k \leq n-1$, this implies

$$\Delta^A p^A_{LL}(0,k,0) + \Delta^B p^B_{HL}(0,k,0) \geq \Delta^B p^B_{LL}(0,k,0) + \Delta^A p^A_{LH}(0,k,0),$$

where the left-hand side is the length of path $LL - LH - HH$ and the right-hand side is the length of $LL - HL - HH$.

Since $p^A_{LL}(0,k,0) = p^A_{LH}(0,k,0) = 0$, it follows that

$$\Delta^B[p^B_{HL}(0,k,0) - p^B_{LL}(0,k,0)] \leq 0 = \Delta^A[p^A_{LH}(0,k,0) - p^A_{LL}(0,k,0)]. \tag{6.7}$$

Moreover, for $t^{n-1} = (0,0,0)$, since $LL - LH - HH$ is no shorter than $LL - HL - HH$,

$$\Delta^A p^A_{LL}(0,0,0) + \Delta^B p^B_{HL}(0,0,0) \geq \Delta^B p^B_{LL}(0,0,0) + \Delta^A p^A_{LH}(0,0,0),$$

or equivalently,

$$\Delta^B[p^B_{HL}(0,0,0) - p^B_{LL}(0,0,0)] \leq \Delta^A[p^A_{LH}(0,0,0) - p^A_{LL}(0,0,0)]. \tag{6.8}$$

Finally, for $t^{n-1} = (k_1,k_2,k_3)$ where $k_1k_2 > 0$ or $k_3 > 0$, the detailed allocation is

$$p^A_{LL}(k_1,k_2,k_3) = p^A_{LH}(k_1,k_2,k_3) = 0,\ p^B_{LL}(k_1,k_2,k_3) = p^B_{HL}(k_1,k_2,k_3) = 0.$$

It follows that

$$\Delta^B[p^B_{HL}(k_1,k_2,k_3) - p^B_{LL}(k_1,k_2,k_3)] = 0 = \Delta^A[p^A_{LH}(k_1,k_2,k_3) - p^A_{LL}(k_1,k_2,k_3)]. \tag{6.9}$$

Adding (6.6)-(6.9) leads to

$$\Delta^B(\rho^B_{HL} - \rho^B_{LL}) < \Delta^A(\rho^A_{LH} - \rho^A_{LL}),$$

which leads to a contradiction to (6.5).

Following a similar argument, one can show that there does not exist a detailed allocation that achieves the same revenue in the Ex post formulation, in the case where the optimal allocation is Bundling Auction for object B. □

Theorem 4 indicates the possibility of discrepancies between different formulations in terms of the auctioneer's expected revenue. A natural question, that follows, is whether there is indeed a gap between any two formulations. To this end, we present an example for which different formulations give rise to distinct maximum expected revenues, thereby asserting this possibility.

Example 4. *In this example, there are two bidders, i.e., $n = 2$. The differences in bidders' valuations are $\Delta^A = \Delta^B = \Delta$. The probabilities of bidder types are*

$$f_{LL} = \frac{1}{6},\ f_{LH} = \frac{1}{4},\ f_{HL} = \frac{5}{12},\ \textit{and}\ f_{HH} = \frac{1}{6},$$

and the common prior π is given by

$$\pi(LL) = \frac{1}{6},\ \pi(LH) = \frac{1}{6},\ \pi(HL) = \frac{1}{2},\ \textit{and}\ \pi(HH) = \frac{1}{6}.$$

Since $1/\lambda^A + 1/\lambda^B > 2$, the Bayesian optimal allocation is Bundling Auction for both objects according to Armstrong [2000]. Following Theorem 4, the Ex post formulation achieves a strictly lower expected revenue than the Bayesian formulation. In the following steps, we demonstrate how the auctioneer's revenue changes according to the level of uncertainty ε. Since both objects are always sold, the social surpluses under the two formulations are the same. It suffices to only consider the difference in information rents.

According to Armstrong [2000], the Bayesian optimal (expected) allocation in this case is:

$$\begin{aligned}
\rho^A_{LL} &= \frac{f_{LL}^{n-1}}{n} = \frac{1}{12}, \rho^A_{LH} = \frac{(f_L^A)^n - f_{LL}^n}{n f_{LH}} = \frac{7}{24};\\
\rho^B_{LL} &= \frac{f_{LL}^{n-1}}{n} = \frac{1}{12}, \rho^B_{HL} = \frac{(f_L^B)^n - f_{LL}^n}{n f_{HL}} = \frac{3}{8}.
\end{aligned}$$

Thus, R^b_{ij}, the minimum expected information rent received by a type-ij bidder, is:

$$\begin{aligned}
R^b_{LL} &= 0,\ R^b_{LH} = \Delta^B \rho^B_{LL} = \frac{1}{12}\Delta;\\
R^b_{HL} &= \Delta^A \rho^A_{LL} = \frac{1}{12}\Delta,\ R^b_{HH} = \Delta^B \rho^B_{LL} + \Delta^A \rho^A_{LH} = \frac{3}{8}\Delta.
\end{aligned}$$

Ex ante, the expected information rent for any bidder in the Bayesian formulation is therefore

$$E[R^b] = \sum_{(i,j)\in T} f_{ij} R_{ij} = \frac{17}{144}\Delta.$$

For the Ex post formulation, consider the following equivalent detailed allocation:

$$p^A_{LL}(0,0,0) = \frac{1}{2},\ p^A_{LH}(0,0,0) = 1,\ p^B_{LL}(0,0,0) = \frac{1}{2},\ p^B_{HL}(0,0,0) = 1;$$

$$p^A_{LL}(1,0,0) = 0,\ p^A_{LH}(0,0,0) = \frac{1}{2},\ p^B_{LL}(1,0,0) = 0,\ p^B_{HL}(1,0,0) = 0;$$

$$p^A_{LL}(0,1,0) = 0,\ p^A_{LH}(0,1,0) = 0,\ p^B_{LL}(0,1,0) = 0,\ p^B_{HL}(0,1,0) = \frac{1}{2};$$

$$p^A_{LL}(0,0,1)=0,\ p^A_{LH}(0,0,1)=0,\ p^B_{LL}(0,0,1)=0,\ p^B_{HL}(0,0,1)=0.$$

In this case, $R^e_{ij}(t^{n-1})$, the minimum information rent in the Ex post formulation, is given by:

$$R^e_{LL}(0,0,0)=0,\ R^e_{LH}(0,0,0)=\frac{1}{2}\Delta,\ R^e_{HL}(0,0,0)=\frac{1}{2}\Delta,\ R^e_{HH}(0,0,0)=\frac{3}{2}\Delta;$$

$$R^e_{LL}(1,0,0)=0,\ R^e_{LH}(1,0,0)=0,\ R^e_{HL}(1,0,0)=0,\ R^e_{HH}(1,0,0)=\frac{1}{2}\Delta;$$

$$R^e_{LL}(0,1,0)=0,\ R^e_{LH}(0,1,0)=0,\ R^e_{HL}(0,1,0)=0,\ R^e_{HH}(0,1,0)=\frac{1}{2}\Delta;$$

$$R^e_{LL}(0,0,1)=0,\ R^e_{LH}(0,0,1)=0,\ R^e_{HL}(0,0,1)=0,\ R^e_{HH}(0,0,1)=0.$$

The expected information rent received by a bidder in the Ex post formulation is:

$$E[R^e]=\sum_{(i,j)\in T} f_{ij}\sum_{t^{n-1}\in T^{n-1}} R^e_{ij}(t^{n-1})\pi(t^{n-1})=\frac{11}{72}\Delta,$$

which is clearly higher than that of the Bayesian formulation ($E[R^b]$). This indicates that the auctioneer's expected revenue is strictly lower in the Ex post formulation, i.e. $S^e(a)>S^b(a)$.

Next, let us consider a more general case where $0<\varepsilon<1$. It can be verified that condition (6.1) in Theorem 2 holds if and only if $\varepsilon\le\frac{1}{12}$, which indicates $S^e(a)<S^\varepsilon(a)=S^b(a)$. Moreover, condition (6.3) in Corollary 2 holds if $\varepsilon\ge\frac{3}{4}$, which implies that $S^e(a)=S^\varepsilon(a)<S^b(a)$. It should be noted that for $\frac{1}{12}<\varepsilon<\frac{3}{4}$, the robust formulation might be different from both the Bayesian and the Ex post formulations. For example, if $\varepsilon=\frac{1}{6}$, the extreme points of U^ε are:

$$\mathbf{q}^1=(1/3,\ 1/12,\ 1/4,\ 1/3),\ \mathbf{q}^2=(0,\ 5/12,\ 7/12,\ 0),\ \mathbf{q}^3=(0,\ 1/12,\ 7/12,\ 1/3),$$

$$\mathbf{q}^4=(1/3,\ 1/12,\ 7/12,\ 0),\ \mathbf{q}^5=(0,\ 5/12,\ 1/4,\ 1/3),\ \mathbf{q}^6=(1/3,\ 5/12,\ 1/4,\ 0).$$

Furthermore, it can be verified that a payment/rent scheme is robust incentive compatible if and only if (RIC) holds for $\mathbf{q}^k$, $k=1,\cdots,6$, since (RIC) is linear and all probability vectors in U^ε can be written as a convex combination of the extreme points. Solving the corresponding linear program yields the following robust optimal rent $\mathbf{R}^\varepsilon$:

$$R^\varepsilon_{LL}(0,0,0)=0,\ R^\varepsilon_{LH}(0,0,0)=\frac{1}{2}\Delta,\ R^\varepsilon_{HL}(0,0,0)=\frac{1}{2}\Delta,\ R^\varepsilon_{HH}(0,0,0)=\frac{3}{2}\Delta;$$

$$R^\varepsilon_{LL}(1,0,0)=0,\ R^\varepsilon_{LH}(1,0,0)=0,\ R^\varepsilon_{HL}(1,0,0)=0,\ R^\varepsilon_{HH}(1,0,0)=\frac{7}{32}\Delta;$$

$$R^\varepsilon_{LL}(0,1,0)=0,\ R^\varepsilon_{LH}(0,1,0)=0,\ R^\varepsilon_{HL}(0,1,0)=0,\ R^\varepsilon_{HH}(0,1,0)=\frac{15}{32}\Delta;$$

$$R^\varepsilon_{LL}(0,0,1)=0,\ R^\varepsilon_{LH}(0,0,1)=0,\ R^\varepsilon_{HL}(0,0,1)=0,\ R^\varepsilon_{HH}(0,0,1)=0.$$

Accordingly, the expected information rent received by a bidder in this robust formulation is:

$$E[R^{\varepsilon}] = \sum_{(i,j)\in T} f_{ij} \sum_{t^{n-1}\in T^{n-1}} R_{ij}^{\varepsilon}(t^{n-1})\pi(t^{n-1}) = \frac{5}{36}\Delta,$$

which indicates that $S^e(a) < S^{\varepsilon}(a) < S^b(a)$. Thus, we have constructed an example for which all the three formulations (Bayesian, Ex post, and general robust formulations) lead to different expected revenues for the auctioneer. It is worth mentioning that in Lopomo et al. [2007], they identify the conditions on the uncertainty set under which the Ex post formulation is equivalent to the robust formulation in *single-dimensional* mechanism design problems. Specifically, they demonstrate that as long as the uncertainty set has full dimensionality, then the incentive compatibility constraints in fact expand to cover every possible scenario of belief systems; thus, the set of feasible solutions is the same under the two formulations. Our result complements theirs by explicitly finding an example for which the results differ in the multiple-dimensional setting.

6.6 Conclusions

In this chapter, we revisit the optimal auction design problem and propose a robust formulation that incorporates the bidders' belief uncertainty. For a fixed allocation mechanism, we identify the necessary and sufficient conditions under which the robust formulation achieves the same expected revenue as the Bayesian formulation. We apply this result to show that the optimal allocation of a single-object auction achieves the same expected revenue in any robust formulation, even though the bidders' valuations are discrete rather than continuous. We also find that, there may be a discrepancy between different formulations in a multiple-object auction, and we provide a concrete example for which the ex post formulation gives rise to a strictly lower expected revenue in the multi-dimensional setting. Our results imply that the auctioneer may have to sacrifice the expected revenue for a more robust formulation.

Bibliography

P. Afeche. Incentive compatible revenue management in queueing systems: Optimal strategic idleness and other delay tactics. Working paper, Northwestern University, 2006.

M. Aghassi and D. Bertsimas. Robust game theory. *Mathematical Programming*, 107 (1):231–273, 2006.

M. Armstrong. Optimal multi-object auctions. *Review of Economic Studies*, 67(3): 455–481, 2000.

M. Armstrong. Multiproduct nonlinear pricing. *Econometrica*, 64(1):51–75, 1996.

M. Armstrong and J. Rochet. Multi-dimensional screening: A user's guide. *European Economic Review*, 43:959–979, 1999.

J. Asker and E. Cantillon. Procurement when price and quality matter. Working paper, New York University, 2008.

M.O. Ball and F.L. Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, pages 18–36, 1993.

R. Batta, J.M. Dolan, and N.N. Krishnamurthy. The maximal expected covering location problem: Revisited. *Transportation Science*, 23(4):277, 1989.

D. Bergemann and S. Morris. The role of the common prior in robust implementation. *Journal of the European Economic Association*, 6(2-3):551–559, 04-05 2008.

D. Bergemann and S. Morris. Robust implementation in direct mechanisms. *Review of Economic Studies*, 76:1175–1204, 2009.

D. Bergemann and S. Morris. Robust mechanism design. *Econometrica*, 73(6):1771–1813, November 2005.

O. Berman, D. Krass, and M.B.C. Menezes. Facility reliability issues in network p-median problems: Strategic centralization and co-location effects. *Operations research*, 55(2):332, 2007.

S. Bikhchandani, S. Chatterji, R. Lavi, A. Mu'alem, N. Nisan, and A. Sen. Weak monotonicity characterizes deterministic dominant-strategy implementation. *Econometrica*, 74(4):1109–1132, 07 2006.

S. Bose, E. Ozdenoren, and A. Pape. Optimal auctions with ambiguity. *Theoretical Economics*, 1(4):411–438, December 2006.

G. Cachon and M. Lariviere. Capacity choice and allocation: strategic behavior and supply chain performance. *Management Science*, 45.

J.F. Campbell. Continuous and discrete demand hub location problems. *TRANSPORTATION RESEARCH PART B METHODOLOGICAL*, 27:473–473, 1993a.

J.F. Campbell. One-to-many distribution with transshipments: An analytic model. *Transportation Science*, 27(4):330, 1993b.

X. Chang and D. Petr. A survey of pricing for integrated service networks. *Computer Communications*, 24:1808–1818, 2001.

K. Chung and J.C. Ely. Foundations of dominant-strategy mechanisms. *Review of Economic Studies*, 74(2):447–476, 04 2007.

E. Coffman and I. Mitrani. A characterization of waiting time performance realizable by single-server queues. *Operations Research*, 28:810–821, 1980.

C. Corbett and X. de Groote. A supplier's optimal quantity discount policy under asymmetric information. *Management Science*, 46:444–450, 2000.

C.F. Daganzo. *Logistics systems analysis*. Springer, 2005.

C.F. Daganzo. The length of tours in zones of different shapes. *TRANSPORT. RES.*, 18(2):135–146, 1984a.

C.F. Daganzo. The distance traveled to visit N points with a maximum of C stops per vehicle: An analytic model and an application. *Transportation Science*, 18(4): 331, 1984b.

C.F. Daganzo and A.L. Erera. On planning and design of logistics systems for uncertain environments. *University of California, Berkeley*, 1999.

C.F. Daganzo and G.F. Newell. Configuration of physical distribution networks. *Networks*, 16(2), 1986.

A. Dasci and V. Verter. A continuous model for production–distribution system design. *European Journal of Operational Research*, 129(2):287–298, 2001.

M. Daskin. Network and discrete location: models, algorithms and applications. *Journal of the Operational Research Society*, 48(7):763–764, 1997.

M.S. Daskin. Application of an expected covering model to emergency medical service system design. *Decision Sciences*, 13(3):416–439, 1982.

M.S. Daskin, National Science Foundation (US, and T. Center. A maximum expected covering location model: formulation, properties and heuristic solution. *TRANSPORT. SCI.*, 17(1):48–70, 1983.

Z. Drezner. *Facility location: a survey of applications and methods*. Springer Verlag, 1995.

N. Figueroa and V. Skreta. The role of optimal threats in auction design. *Journal of Economic Theory*, 144(2):884–897, March 2009.

M.L. Fisher. The Lagrangian relaxation method for solving integer programming problems. *Management science*, pages 1861–1871, 2004.

A. Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information Theory*, 25(4):373–380, 1979.

R. Gibbens, F. Kelly, and P. Key. A decision-theoretic approach to call admission control in atmnetworks. *IEEE Journal of Selected Areas in Communications*, 13: 1101–1114, 1995.

B. Goldengorin, G. Sierksma, G.A. Tijssen, and M. Tso. The data-correcting algorithm for the minimization of supermodular functions. *Management Science*, pages 1539–1551, 1999.

A. Ha. Supplier-buyer contracting: Asymmetric cost information and cutoff level policy for buyer participation. *Naval Research Logistics*, 48:41–64, 2001.

R.W. Hall. Travel distance through transportation terminals on a rectangular grid. *Journal of the Operational Research Society*, pages 1067–1078, 1984.

R.W. Hall. Discrete models/continuous models. *Omega International Journal of Management Science*, 14:213–220, 1986.

R.W. Hall. Configuration of an overnight package air network. *Transportation Research*, 23(2):139–149, 1989.

R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA, 2002.

B. Heydenreich, R. Muller, M. Uetz, and R.V. Vohra. Characterization of revenue equivalence. *Econometrica*, 77(1):307–316, 01 2009.

B. Holmstrom and P. Milgrom. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organization*, 7(0):24–52, Special I 1991.

A. Iyer, L. Schwarz, and S. Zenios. A principal-agent model for product specification and production. *Management Science*, 51:106–119, 2005.

P. Jehiel, B. Moldovanu, and E. Stacchetti. Multidimensional mechanism design for auctions with externalities. *Journal of Economic Theory*, 85(2):258–293, 1999.

W.C. Johnson. *Mathematical and physical principles of engineering analysis*. McGraw-Hill book company, inc., 1944.

A.-K. Katta and J. Sethuraman. Pricing strategies and service differentiation in queues - a profit maximization perspective. Working paper, Columbia University, 2005.

P. Klemperer. Auction theory: A guide to the literature. *Journal of Economic Surveys*, 13:227–286, 1999.

V. Krishna. *Auction theory*. Academic Press, USA, 2002.

HW Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics*, 52(1), 2005.

J. Laffont and D. Martimort. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, USA, 2002.

J. Laffont and J. Tirole. *A Theory of Incentives in Regulation and Procurement*. The MIT Press, Cambridge, Massachussetts, 1993.

A. Langevin, P. Mbaraga, and J.F. Campbell. Continuous approximation models in freight distribution: An overview. *Transportation Research Part B*, 30(3):163–188, 1996.

J.T. Lewis, R. Russell, F. Toomey, B. McGurk, S. Crosby, and I. Leslie. Practical connection admission control for ATM networks based on on-line measurements. *Computer Communications*, 21(17):1585–1596, 1998.

M. Lim, M. Daskin, S. Chopra, and A. Bassamboo. Managing risks of facility disruptions. Technical report, Working paper, Northwestern University.

G. Lopomo, L. Rigotti, and C. Shannon. Uncertainty in mechanism design. *Unpublished working paper, Duke University*, 2007.

H. Lutze and O. Ozer. Promised lead time contracts under asymmetric information. Forthcoming in Operations Research, 2008.

A. Malakhov and R. Vohra. Single and multi-dimensional optimal auctions – a network approach. Working paper, Northwestern University, 2005.

E. Maskin and J. Riley. Optimal multi-unit auctions. In Frank Hahn, editor, *The Economics of Missing Markets, Information, and Games*. Oxford University Press, New York, 1989.

H. Mendelson. Pricing computer services: Queueing effects. *Communications of ACM*, 28:312–321, 1985.

H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research*, 38:870–883, 1990.

P. Milgrom and R. Weber. A theory of auctions and competitive bidding. *Econometrica*, 50:1089–1122, 1982.

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

B. Moldovanu, A. Sela, and X. Shi. Competing auctions with endogenous quantities. *Journal of Economic Theory*, 141(1):1–27, July 2008.

K. Moorthy. Market segmentation, self-selection, and product line design. *Marketing Science*, 3:288–307, 1984.

M. Mussa and S. Rosen. Monopoly and product quality. *Journal of Economic Theory*, 18:301–317, 1978.

R.B. Myerson. Optimal auction design. *Mathematics of operations research*, pages 58–73, 1981.

P. Naor. On the regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

GF Newell. Dispatching policies for a transportation route. *Transportation Science*, 5(1):91, 1971.

GF Newell. Scheduling, location, transportation, and continuum mechanics; some simple approximations to optimization problems. *SIAM Journal on Applied Mathematics*, pages 346–360, 1973.

A. Okabe, B. Boots, K. Sugihara, and S.N. Chiu. *Spatial tessellations: concepts and applications of Voronoi diagrams*. Wiley New York, 1992.

Y. Ouyang. Design of vehicle routing zones for large-scale distribution systems. *Transportation Research Part B*, 41(10):1079–1093, 2007.

Y. Ouyang and C.F. Daganzo. Discretization and validation of the continuum approximation scheme for terminal system design. *Transportation Science*, 40(1):89–98, 2006.

L. Qi and Z.-J.M. Shen. A supply chain design model with unreliable supply. *Naval Research Logistics*, 54(8), 2007.

L. Qi, Z.-J.M. Shen, and L.V. Snyder. The Effect of Supply Disruptions on Supply Chain Design Decisions. Technical report, Working paper, University of Missouri Rolla, 2008.

C. ReVelle and K. Hogan. The maximum availability location problem. *Transportation Science*, 23(3):192, 1989.

J. Rochet and L. Stole. The economics of multidimensional screening. In M. Dewatripont, L. Hansen, and S. Turnovsky, editors, *Advances in Economics and Econometrics: Theory and Applications - Eight World Congress*, Econometric Society Monographs. Cambridge University Press, New York, 2005.

B. Salanie. *The economics of taxation.* The MIT Press, Cambridge, Massachussetts, 2003.

J. Shanthikumar and D. Yao. Multiclass queuing systems: Polymatroidal structure and optimal scheduling control. *Operations Research*, 40:293–299, 1992.

Z.-J.M. Shen, L. Zhan, and J. Zhang. The Reliable Facility Location Problem: Formulations, Heuristics, and Approximation Algorithms. Technical report, Working paper, University of California, Berkeley, 2009.

H.D. Sherali and A. Alameddine. A new reformulation-linearization technique for bilinear programming problems. *Journal of Global Optimization*, 2(4):379–410, 1992.

L.V. Snyder. `http://www.lehigh.edu/lvs2`.

L.V. Snyder and M.S. Daskin. Reliability models for facility location: The expected failure cost case. *Transportation Science*, 39(3):400–416, 2005.

S. Stidham. Analysis, design and control of queuing systems. *Operations Research*, 50:197–216, 2002.

L.F. Toth. Sur la representation dune population infinie par un nombre fini defilements. *Acta Math. Acad. Sci. Hungar*, 25:76–81, 1959.

W. Vickrey. Counterspeculation, auctions and competitive sealed tenders. *Journal of Finance*, 16:8–37, 1961.

A. Weber and C.J. Friedrich. *Theory of the Location of Industries.* University of Chicago Press, 1929.

R. Wilson. *Nonlinear Pricing.* Oxford University Press, New York, 1993.

R. Wilson. Game-theoretic analyses of trading processes. In Advances in Economic Theory, ed. Truman Bewley. Cambridge: Cambridge University Press, 1987.

S.-Y. Wu and P.-Y. Chen. Versioning and piracy control for digital information goods. Forthcoming in Operations Research, 2008.

T. Yahalom, J. M. Harrison, and S. Kumar. Designing and pricing incentive compatible grades of service in queueing systems. Working paper 1032, Stanford University, 2005.

H. Zhang and S. Zenios. A dynamic principal-agent model with hidden information: Sequential optimality through truthful state revelation. Forthcoming in Operations Research, 2008.

L. Zhang, S. Deering, D. Estrin, and S. Shenker. Rsvp: A new resource reservation protocol. *IEEE Network*, 7:8–18, 1993.

Appendix A

Proof of Propositions in Chapter 2

A.1 Proof of Proposition 1

Let $\Omega = \{0,1\}^J$ be the set of failure scenarios. For each $\omega \in \Omega$, let p_ω be the probability that scenario ω will occur, also let $\delta_{j\omega}$ be the binary parameter indicating whether or not facility j is operational in scenario ω. The scenario based stochastic program (SSP) is formulated as follows

$$\text{(SSP)} \quad \text{Min} \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1}\sum_{j=0}^{J}\sum_{\omega\in\Omega} \lambda_i d_{ij} p_\omega Y_{ij\omega} \tag{A.1a}$$

$$\text{s.t.} \sum_{j=0}^{J} Y_{ij\omega} = 1 \quad \forall\ 0 \le i \le I-1,\ \omega \in \Omega \tag{A.1b}$$

$$\sum_{i=0}^{I-1} Y_{ij\omega} \le \delta_{j\omega} X_j \quad \forall\ 0 \le j \le J-1,\ \omega \in \Omega \tag{A.1c}$$

$$X_j,\ Y_{ij\omega} \in \{0,1\} \quad \forall\ 0 \le i \le I-1,\ 0 \le j \le J,\ \omega \in \Omega, \tag{A.1d}$$

where X_j is the binary variable indicating whether or not a facility is build at location j, and $Y_{ij\omega}$ is equal to one if and only if customer i is served by facility j in scenario ω ($Y_{iJ\omega} = 1$ indicates that the penalty cost is incurred in scenario ω).

To verify that the scenario based formulation (A.1a)-(A.1d) is equivalent to the compact (RUFL) formulation (2.1a)-(2.1g), we first show how to map an optimal solution of (RUFL) to a feasible solution of (SSP). Let $(\mathbf{X}, \mathbf{Y}, \mathbf{P})$ be an optimal solution of (RUFL), we construct a solution $(\mathbf{X}', \mathbf{Y}')$ for (SSP) by letting $\mathbf{X}' = \mathbf{X}$. For each $0 \le i \le I-1$ and $0 \le r \le J$, let $j(i,r) \in \{0 \le j \le J :\ Y_{ijr} = 1\}$, *i.e.* $j(i,r)$ is the unique facility that serves customer i at level r. The customer assignment in each scenario is determined as follows (by convention, we let $\delta_{J\omega} = 1$ for all $\omega \in \Omega$)

$$Y'_{ij\omega} \quad = \quad \begin{cases} 1 & \text{if } j = j(i,r) \text{ for some } 0 \le r \le J,\ \delta_{j\omega} = 1, \\ & \text{and } \delta_{j(i,\ell)\omega} = 0,\ \forall\ 0 \le \ell \le r-1 \\ 0 & \text{otherwise.} \end{cases}$$

By construction, $(\mathbf{X}', \mathbf{Y}')$ is feasible to (SSP). Next, we show that $(\mathbf{X}', \mathbf{Y}')$ achieves the same object value as $(\mathbf{X}, \mathbf{Y}, \mathbf{P})$. Let $\Phi(\mathbf{X}, \mathbf{Y}, \mathbf{P})$ and $\Psi(\mathbf{X}', \mathbf{Y}')$ be the objective function of (RUFL) and (SSP) respectively. Also, define $\Omega(i, r) = \{\omega \in \Omega : \ Y'_{ij(i,r)\omega} = 1\}$, *i.e.* $\Omega(i, r)$ is the set of scenarios in which customer i is served by facility $j(i, r)$. It follows that

$$\begin{aligned}
\Psi(\mathbf{X}', \mathbf{Y}') &= \sum_{j=0}^{J-1} f_j X'_j + \sum_{i=0}^{I-1} \sum_{j=0}^{J} \sum_{\omega \in \Omega} \lambda_i d_{ij} p_\omega Y'_{ij\omega} \\
&= \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{r=0}^{J} \lambda_i d_{i,j(i,r)} \sum_{\omega \in \Omega(i,r)} p_\omega \\
&= \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{r=0}^{J} \lambda_i d_{i,j(i,r)} (1 - q_{j(i,r)}) \prod_{\ell=0}^{r-1} q_{j(i,\ell)} \\
&= \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{r=0}^{J} \sum_{j=0}^{J} \lambda_i d_{ij} P_{ijr} Y_{ijr} \\
&= \Phi(\mathbf{X}, \mathbf{Y}, \mathbf{P}),
\end{aligned}$$

which implies that solving (SSP) yields a lower bound to (RUFL).

Conversely, given an optimal solution $(\mathbf{X}, \mathbf{Y})$ to (SSP), we construct a feasible solution $(\mathbf{X}', \mathbf{Y}', \mathbf{P}')$ to (RUFL), also by letting $\mathbf{X}' = \mathbf{X}$. Without loss of generality, we assume that $Y_{ij\omega} = 1$ if and only if $j = \min\{0 \le k \le J : \ \delta_{k\omega} X_k = 1, \ d_{ik} \le d_{ik'} \ \forall \ k' \ne k \text{ s.t. } \delta_{k'\omega} X_{k'} = 1\}$ (by convention, we assume $X_J = 1$), *i.e.* each customer is always served by her closest open facility, and if there are more than one facilities that are equally close, we break the tie by choosing the facility with the lowest index.

Let $N = \{0 \le j \le J - 1 : \ X_j = 1\}$ be the set of facilities that are constructed in the optimal solution to (SSP). For each customer i, let $j(i, 0), \ j(i, 1), \ \cdots, \ j(i, |N|)$ be an ordering of the facilities in $N \bigcup \{J\}$ such that for all $1 \le r \le |N|$, $d_{i,j(i,r-1)} \le d_{i,j(i,r)}$, and if $d_{i,j(i,r-1)} = d_{i,j(i,r)}$, then $j(i, r-1) < j(i, r)$. Also, define $\Omega(i, r) = \{\omega \in \Omega : \ \delta_{j(i,r)\omega} = 1, \text{ and } \delta_{j(i,\ell)\omega} = 0, \ \forall \ 0 \le \ell \le r - 1\}$. We set the values of $\mathbf{Y}'$ and $\mathbf{P}'$ as follows

$$\begin{aligned}
Y'_{ijr} &= \begin{cases} 1 & \text{if } j = j(i, r) \text{ and } d_{ij} \le d_{iJ} \\ 0 & \text{otherwise} \end{cases} \\
P'_{ijr} &= \begin{cases} (1 - q_{j(i,r)}) \prod_{\ell=0}^{r-1} q_{j(i,\ell)} & \text{if } j = j(i, r) \text{ and } d_{ij} \le d_{iJ} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

It is clear that by construction $(\mathbf{X}', \mathbf{Y}', \mathbf{P}')$ is feasible to (RUFL). The objective

value associated with solution is

$$\begin{aligned}
\Phi(\mathbf{X}', \mathbf{Y}', \mathbf{P}') &= \sum_{j=0}^{J-1} f_j X'_j + \sum_{i=0}^{I-1} \sum_{r=0}^{J-1} \sum_{j=0}^{J-1} \lambda_i d_{ij} P'_{ijr} Y'_{ijr} \\
&= \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{r=0}^{|N|} \lambda_i d_{i,j(i,r)} (1 - q_{j(i,r)}) \prod_{\ell=0}^{r-1} q_{j(i,\ell)} \\
&= \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{r=0}^{J} \lambda_i d_{i,j(i,r)} \sum_{\omega \in \Omega(i,r)} p_\omega \\
&= \sum_{j=0}^{J-1} f_j X_j + \sum_{i=0}^{I-1} \sum_{j=0}^{J} \sum_{\omega \in \Omega} \lambda_i d_{ij} p_\omega Y_{ij\omega} \\
&= \Psi(\mathbf{X}, \mathbf{Y}).
\end{aligned}$$

Therefore, the optimal solution to (SSP) is also a lower bound to (RUFL). This completes our proof.

A.2 Proof of Proposition 2

Suppose, for a contradiction, that $(\mathbf{X}, \mathbf{Y}, \mathbf{P})$ is optimal for (RUFL) where $Y_{ijr} = Y_{ik,r+1} = 1$ and $d_{ij} > d_{ik}$ for some $0 \le i \le I-1,\ 0 \le j \le J$, and $0 \le r \le R$. We will show that by "swapping" j and k the objective value will decrease. Obviously $j \le J-1$, otherwise j is the pseudo facility and customer i cannot be assigned to facility k as a backup. We consider two cases based on whether or not k is the pseudo facility.

If $k \le J-1$ we construct a different solution $(\mathbf{X}', \mathbf{Y}', \mathbf{P}')$ as follows:

$$\begin{aligned}
X' &= X; \\
Y'_{h\ell s} &= \begin{cases} 1 & \text{if } h = i, \ell = k,\ s = r \text{ or } h = i, \ell = j,\ s = r+1, \\ 0 & \text{if } h = i, \ell = j,\ s = r \text{ or } h = i, \ell = k,\ s = r+1, \\ Y_{h\ell s} & \text{otherwise;} \end{cases} \\
P'_{h\ell s} &= \begin{cases} \frac{1-q_k}{1-q_j} P_{jr} & \text{if } h = i, \ell = k,\ s = r, \\ \frac{q_k(1-q_j)}{1-q_k} P'_{kr} = q_k P_{jr} & \text{if } h = i, \ell = j,\ s = r+1, \\ 0 & \text{if } h = i, \ell = j,\ s = r \text{ or } h = i, \ell = k,\ s = r+1, \\ P_{h\ell s} & \text{otherwise.} \end{cases}
\end{aligned}$$

By construction, $(\mathbf{X}', \mathbf{Y}', \mathbf{P}')$ is a feasible solution. Let $\Phi(\mathbf{X}, \mathbf{Y}, \mathbf{P})$ be the objective

value associated with $(\mathbf{X}, \mathbf{Y}, \mathbf{P})$, it follows that:

$$\begin{aligned}
\Phi(\mathbf{X}', \mathbf{Y}', \mathbf{P}') - \Phi(\mathbf{X}, \mathbf{Y}, \mathbf{P}) &= \lambda_i(P'_{kr} d_{ik} + P'_{j,r+1} d_{ij} - P_{jr} d_{ij} - P_{k,r+1} d_{ik}) \\
&= \lambda_i[d_{ik}(P'_{kr} - P_{k,r+1}) - d_{ij}(P_{jr} - P'_{j,r+1})] \\
&= \lambda_i\{d_{ik}[\frac{1-q_k}{1-q_j} P_{jr} - \frac{q_j(1-q_k)}{1-q_j} P_{jr}] - d_{ij}(P_{jr} - q_k P_{jr})\} \\
&= \lambda_i(1-q_k)(d_{ik} - d_{ij}) P_{jr} < 0.
\end{aligned}$$

The case in which $k = J$ is similar, except that $Y'_{ij,r+1} = P'_{ij,r+1} = 0$, which reduces the cost even more. This implies a contradiction to that $(\mathbf{X}, \mathbf{Y}, \mathbf{P})$ is optimal.

A.3 Proof of Proposition 3

Let $S \subseteq \{0, \cdots, J-1\}$ be a subset of candidate locations, and $u, v \in \{0, \cdots, J-1\} \setminus S$, we show that

$$\Phi_i(S \cup \{u, v\}) - \Phi_i(S \cup \{u\}) \;\geq\; \Phi_i(S \cup \{v\}) - \Phi_i(S). \tag{A.2}$$

Assume that $S = \{j_1,\ j_2,\ \cdots,\ j_n\}$ where $d_{ij_1} \leq d_{ij_2} \leq \cdots \leq d_{ij_n}$, i.e., we sort elements in S in nondecreasing order of their distance to customer i. Let

$$\begin{aligned}
\bar{n} &= \inf\{1 \leq k \leq n :\ d_{ij_k} \leq \phi_i\} \\
s &= \inf\{1 \leq k \leq n :\ d_{ij_k} \leq d_{iu}\} \\
t &= \inf\{1 \leq k \leq n :\ d_{ij_k} \leq d_{iv}\}.
\end{aligned}$$

In addition, define

$$\begin{aligned}
P_k &= \begin{cases} \prod_{\ell=1}^{k} q_{j_\ell} & 1 \leq k \leq \bar{n} \\ 1 & k = 0, \end{cases} \\
C_k &= \begin{cases} P_{k-1}(1-q_{j_k}) d_{ij_k} & 1 \leq k \leq \bar{n} \\ P_{\bar{n}} \phi_i & k = \bar{n}+1. \end{cases}
\end{aligned}$$

Following a similar argument as in the proof of Proposition 2, we know that it is optimal to assign the facilities level by level in increasing order of distance, until the transportation cost exceeds the penalty cost, i.e.,

$$\begin{aligned}
\Phi_i(S) &= \lambda_i \sum_{k=1}^{\bar{n}} P_{k-1}(1-q_{j_k}) d_{ij_k} + P_{\bar{n}} \phi_i + \sum_{j \in S} \mu_{ij} \\
&= \lambda_i \sum_{k=1}^{\bar{n}+1} C_k + \sum_{j \in S} \mu_{ij}.
\end{aligned}$$

Without loss of generality, we assume that d_{iu} and d_{iv} are less than the penalty cost ϕ_i, i.e. $s \leq \bar{n}$ and $t \leq \bar{n}$. It follows that

$$\begin{aligned}
& \Phi_i(S \cup \{v\}) - \Phi_i(S) \\
= \; & \lambda_i[\sum_{k=1}^{t} C_k + P_t(1-q_v)d_{iv} + q_v \sum_{k=t+1}^{\bar{n}+1} C_k] + \sum_{j \in S \cup \{v\}} \mu_{ij} - \lambda_i \sum_{k=1}^{\bar{n}+1} C_k - \sum_{j \in S} \mu_{ij} \\
= \; & \lambda_i[P_t(1-q_v)d_{iv} - (1-q_v) \sum_{k=t+1}^{\bar{n}+1} C_k] + \mu_{iv} \\
= \; & \lambda_i(1-q_v)[P_t d_{iv} - \sum_{k=t+1}^{\bar{n}+1} C_k] + \mu_{iv}.
\end{aligned}$$

Note that the first item in the last equation is negative, because

$$\begin{aligned}
P_t d_{iv} - \sum_{k=t+1}^{\bar{n}+1} C_k &= P_t[d_{iv} - \sum_{k=t+1}^{\bar{n}} (\prod_{\ell=t+1}^{k-1} q_{j_\ell})(1-q_{j_k})d_{ij_k} - (\prod_{\ell=t+1}^{\bar{n}} q_{j_\ell})\phi_i] \\
&< P_t d_{iv}[1 - (\prod_{\ell=t+1}^{k-1} q_{j_\ell})(1-q_{j_k}) - \prod_{\ell=t+1}^{\bar{n}} q_{j_\ell}] = 0.
\end{aligned}$$

To show that (A.2) holds, we consider the following two cases.

Case 1: $d_{iu} \leq d_{iv}$. In this case, it follows that

$$\begin{aligned}
& \Phi_i(S \cup \{u,\ v\}) - \Phi_i(S \cup \{u\}) \\
= \; & \lambda_i[\sum_{k=1}^{s} C_k + P_s(1-q_u)d_{iu} + q_u \sum_{k=s+1}^{t} C_k + q_u P_t(1-q_t)d_{iv} + q_u q_v \sum_{k=t+1}^{\bar{n}+1} C_k \\
+ \; & \sum_{j \in S \cup \{u,\ v\}} \mu_{ij} - \lambda_i[\sum_{k=1}^{s} C_k + P_s(1-q_u)d_{iu} + q_u \sum_{k=s+1}^{\bar{n}+1} C_k] - \sum_{j \in S \cup \{u\}} \mu_{ij} \\
= \; & \lambda_i[q_u P_t(1-q_v)d_{iv} - q_u(1-q_v) \sum_{k=t+1}^{\bar{n}+1} C_k] + u_{iv} \\
= \; & \lambda_i q_u(1-q_v)(P_t d_{iv} - \sum_{k=t+1}^{\bar{n}+1} C_k) + \mu_{iv}.
\end{aligned}$$

Clearly (A.2) holds in this case, since $0 \leq q_u \leq 1$ and $P_t d_{iv} - \sum_{k=t+1}^{\bar{n}+1} C_k < 0$.

Case 2: $d_{iu} > d_{iv}$. In this case $t \leq s$, and the following assertion holds:

$$\begin{aligned}
& \Phi_i(S \cup \{u,\ v\}) - \Phi_i(S \cup \{u\}) \\
= \ & \lambda_i[\sum_{k=1}^{t} C_k + P_t(1-q_v)d_{iv} + q_v \sum_{k=t+1}^{s} C_k + q_v P_s(1-q_u)d_{iu} + q_v q_u \sum_{k=s+1}^{\bar{n}+1} C_k] \\
+ \ & \sum_{j \in S \cup \{u,\ v\}} \mu_{ij} - \lambda_i[\sum_{k=1}^{s} C_k + P_s(1-q_u)d_{iu} + q_u \sum_{k=s+1}^{\bar{n}+1} C_k] - \sum_{j \in S \cup \{u\}} \mu_{ij} \\
= \ & \lambda_i\{P_t(1-q_v)d_{iv} - (1-q_v)[\sum_{k=t+1}^{s} C_k + P_s(1-q_u)d_{iu} + q_u \sum_{k=s+1}^{\bar{n}+1} C_k]\} + \mu_{iv} \\
= \ & \lambda_i(1-q_v)\{P_t d_{iv} - [\sum_{k=t+1}^{s} C_k + P_s(1-q_u)d_{iu} + q_u \sum_{k=s+1}^{\bar{n}+1} C_k]\} + \mu_{iv}.
\end{aligned}$$

We claim that (A.2) holds in this case, because

$$\begin{aligned}
& [\sum_{k=t+1}^{s} C_k + P_s(1-q_u)d_{iu} + q_u \sum_{k=s+1}^{\bar{n}+1} C_k] - \sum_{k=t+1}^{\bar{n}+1} C_k \\
\leq \ & [\sum_{k=t+1}^{s} C_k + P_s(1-q_u)d_{iu} + q_u \sum_{k=s+1}^{\bar{n}+1} C_k] - \sum_{k=s+1}^{\bar{n}+1} C_k \\
= \ & (1-q_u)[P_s d_{iu} - \sum_{k=s+1}^{\bar{n}+1} C_k] < 0.
\end{aligned}$$

A.4 Proof of Proposition 4

First, we introduce an equivalent formulation of (RSP) by "spliting" the decision variables:

$$Y_{jr} = \begin{cases} 1 & \text{if the level } r \text{ facility has the same transportation distance as facility } j \\ 0 & \text{otherwise.} \end{cases}$$

$$Z_{jr} = \begin{cases} 1 & \text{if the level } r \text{ facility has the same failure probability as facility } j \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that RSP is equivalent to the following problem:

$$\text{Min} \sum_{j=0}^{J}\sum_{r=0}^{R}\lambda_i d_{ij} W_{jr} + \sum_{j=0}^{J-1}\sum_{r=0}^{R-1}\mu_{ij} Y_{jr} \tag{A.3a}$$

$$\text{s.t. } (2.4\text{b}) - (2.4\text{d}) \tag{A.3b}$$

$$\sum_{j=0}^{J-1} Z_{jr} + \sum_{s=0}^{r-1} Z_{Js} = 1 \quad \forall\ 0 \le r \le R \tag{A.3c}$$

$$\sum_{r=0}^{R-1} Z_{jr} \le 1 \quad \forall\ 0 \le j \le J-1 \tag{A.3d}$$

$$\sum_{r=0}^{R} Z_{Jr} = 1 \tag{A.3e}$$

$$P_{j0} = 1 - q_j \quad \forall\ 0 \le j \le J \tag{A.3f}$$

$$P_{jr} = (1-q_j)\sum_{k=0}^{J-1}\frac{q_k}{1-q_k} W_{i,k,r-1} \quad \forall\ 0 \le j \le J,\ 1 \le r \le R \tag{A.3g}$$

$$W_{jr} \le P_{jr} \quad \forall\ 0 \le j \le J,\ 0 \le r \le R \tag{A.3h}$$

$$W_{jr} \le Z_{jr} \quad \forall\ 0 \le j \le J,\ 0 \le r \le R \tag{A.3i}$$

$$W_{jr} \ge 0 \quad \forall\ 0 \le j \le J,\ 0 \le r \le R \tag{A.3j}$$

$$W_{jr} \ge P_{jr} + Z_{jr} - 1 \quad \forall\ 0 \le j \le J,\ 0 \le r \le R \tag{A.3k}$$

$$Y_{jr},\ Z_{jr} \in \{0,1\} \quad \forall\ 0 \le j \le J,\ 0 \le r \le R \tag{A.3l}$$

$$Y_{jr} = Z_{jr} \quad \forall\ 0 \le j \le J,\ 0 \le r \le R. \tag{A.3m}$$

If we remove the last constraint (A.3m), the customer is allowed to choose an arbitrary combination of transportation cost and failure probability. Next, we show that the (RRSP) formulation (2.7a)-(2.7e) is equivalent to formulation (A.3a) - (A.3l), based on the following lemma.

Lemma 3. *There exists an optimal solution* $(\mathbf{Y}^*, \mathbf{Z}^*, \mathbf{P}^*)$ *to formulation* (A.3a) - (A.3l)*, such that if* $Z^*_{jr} = 1$*,* $Z^*_{k,r+1} = 1$ *and* $r+1 \le R-1$*, then* $q_j \le q_k$*.*

Proof of Lemma 3. Suppose that $(\mathbf{Y}, \mathbf{Z}, \mathbf{P})$ is an optimal solution to formulation (A.3a) - (A.3l), such that $Z_{jr} = 1$, $Z_{k,r+1} = 1$, $j, k \le R-1$ and $q_j > q_k$. Let u and v be the facilities assigned to this customer at level r and $r+1$, *i.e.* $Y_{ur} = 1$ and $Y_{v,r+1} = 1$. We construct a new solution $(\mathbf{Y}', \mathbf{Z}', \mathbf{P}')$ as follows:

$$\mathbf{Y}' = \mathbf{Y};$$

$$Z'_{\ell s} = \begin{cases} 1 & \text{if } \ell = k,\ s = r \text{ or } h = i, \ell = j,\ s = r+1, \\ 0 & \text{if } \ell = j,\ s = r \text{ or } h = i, \ell = k,\ s = r+1, \\ Z_{\ell s} & \text{otherwise;} \end{cases}$$

$$P'_{\ell s} = \begin{cases} \frac{1-q_k}{1-q_j} P_{jr} & \text{if } \ell = k,\ s = r, \\ \frac{q_k(1-q_j)}{1-q_k} P'_{kr} = q_k P_{jr} & \text{if } \ell = j,\ s = r+1, \\ 0 & \text{if } \ell = j,\ s = r \text{ or } h = i, \ell = k,\ s = r+1, \\ P_{\ell s} & \text{otherwise.} \end{cases}$$

By construction, $(\mathbf{Y}', \mathbf{Z}', \mathbf{P}')$ is a feasible solution to formulation (A.3a) - (A.3l). Define $G(\mathbf{Y}, \mathbf{Z}, \mathbf{P})$ to be the objective value of formulation (A.3a) - (A.3l) associated with solution $(\mathbf{Y}, \mathbf{Z}, \mathbf{P})$. The following assertion holds:

$$\begin{aligned} G(\mathbf{Y}', \mathbf{Z}', \mathbf{P}') - G(\mathbf{Y}, \mathbf{Z}, \mathbf{P}) &= \lambda_i(P'_{kr} d_{iu} + P'_{j,r+1} d_{iv} - P_{jr} d_{iu} - P_{k,r+1} d_{iv}) \\ &= \lambda_i[d_{iu}(P'_{kr} - P_{jr}) + d_{iv}(P'_{j,r+1} - P_{k,r+1})] \\ &= \lambda_i\{d_{iu}[\frac{1-q_k}{1-q_j} P_{jr} - P_{jr}] - d_{iv}(q_k P_{jr} - \frac{q_j(1-q_k)}{1-q_j} P_{jr})\} \\ &= \frac{q_j - q_k}{1-q_j} \lambda_i P_{jr}(d_{iu} - d_{iv}). \end{aligned}$$

Following a similar argument as in the proof of Proposition 2, $d_{iu} \leq d_{iv}$, implying $G(\mathbf{Y}', \mathbf{Z}', \mathbf{P}') \leq G(\mathbf{Y}, \mathbf{Z}, \mathbf{P})$. Therefore, if an optimal solution does not satisfy the condition in Lemma 3, we can always construct an alternative optimal solution by swapping j and k. This completes the proof of Lemma 3.

Without loss of generality, we can fix $\mathbf{Z} = \mathbf{Z}^*$ and $\mathbf{P} = \mathbf{P}^*$ in formulation (A.3a) - (A.3l), which leads to the (RRSP) formulation (2.7a)-(2.7e). Since formulation (A.3a) - (A.3l) is a relaxation of (RSP), it follows that (RRSP) yields a lower bound for (RSP).

Appendix B

Proof of Proposition in Chapter 4

B.1 Proof of Proposition 5

The following lemma gives a necessary optimality condition for facility location design and customer allocation.

Lemma 4. *The optimal facility locations should satisfy the following conditions:*

1. *the initial service areas* $\mathbf{R}$ *(i.e., initial customer allocation before any failure) should form a Voronoi tessellation;*

2. *the location of each facility should be the centroid of all customer demands weighted by this facility's service probability to the customers.*

Proof of Lemma 4. The first condition is obvious from the fact that for any given facility location design, every customer always goes to the nearest available facility. The second necessary condition can be proven by examining the cost objective with respect to an infinitesimal perturbation of one generic facility location, x_j, while holding $\mathcal{R}_{jk}, \forall j, k$, fixed. Let $\mathcal{F}(x_j)$ denote the expected service cost of a facility located at x_j to serve all its potential customers. Consider an arbitrary location perturbation Δ_x and a scalar $\epsilon > 0$. Consider an arbitrary location perturbation Δ_x and a scalar $\epsilon > 0$:

$$\begin{aligned}\mathcal{F}(x_j + \epsilon\Delta_x) - \mathcal{F}(x_j) &= \sum_{k=0}^{R-1} \int_{x \in R_{jk}} (1-q)q^k\lambda \left\{\|x - x_j - \epsilon\Delta_x\| - \|x - x_j\|\right\} dx \\ &= \sum_{k=0}^{R-1} \int_{x \in R_{jk}} (1-q)q^k\lambda \left\{ \frac{\|x - x_j - \epsilon\Delta_x\|^2 - \|x - x_j\|^2}{\|x - x_j - \epsilon\Delta_x\| + \|x - x_j\|} \right\} dx.\end{aligned}$$

It is easy to show that the first-order condition $\lim_{\epsilon \to 0} \frac{1}{\epsilon}\{\mathcal{F}(x_j + \epsilon\Delta_x) - \mathcal{F}(x_j)\} = 0$

requires that the optimal facility location x_j satisfies

$$x_j = \frac{\sum_{k=0}^{R-1} \int_{x \in R_{jk}} (1-q) q^k \lambda x dx}{\sum_{k=0}^{R-1} \int_{x \in R_{jk}} (1-q) q^k \lambda dx} = \frac{\sum_{k=0}^{R-1} \int_{x \in R_{jk}} P(x, x_j | \mathbf{x}) x dx}{\sum_{k=0}^{R-1} \int_{x \in R_{jk}} P(x, x_j | \mathbf{x}) dx}.$$

Hence, the optimal facility location x_j is the centroid of all customer demands weighted by the corresponding service probability. This completes the proof of Lemma 4.

It is worth noting that the above proof does not require S to be homogeneous and infinite. Hence, Lemma 4 holds also for finite and heterogeneous S.

Since the plane is infinite and homogeneous, the facility locations and all service areas should be translationally and rotationally symmetric. The initial service area of every facility (which, as a Voronoi polygon, must be convex Okabe et al. [1992]) should have the facility location as its centroid. Hence, collectively they should form a centroidal Voronoi tessellation—which should then minimize the total customer initial access cost (before any failure) to the facilities. As pointed out by Gersho Gersho [1979], Fejes Toth Toth [1959] proved that this cost is minimized under the Euclidean metric when the shape of the initial service areas are exactly congruent (i.e., of same shape and size) and form a regular hexagonal tessellation of the space. Gersho further proved that even in a finite 2-d plane, regular hexagonal tessellations should cover most of the space if the number of facilities is sufficiently large Gersho [1979]. This result leads to Proposition 5.

Appendix C

Proof for Propositions in Chapter 5

C.1 Technical Lemmas

Before we prove the propositions, we first introduce the following technical lemmas that are necessary for establishing our results.

Lemma 5. *Let* (q^1, W^1, R) *and* (q^2, W^2, R) *be two solutions such that for some* $ij \in T$*,*

$$\begin{aligned} q^2_{ij} &\geq q^1_{ij}, \text{ if } ij \in \{HL, HH\}, \\ q^2_{ij} &\leq q^1_{ij}, \text{ if } ij \in \{LL, LH\}, \\ W^2_{ij} &\leq W^1_{ij}, \text{ if } ij \in \{LH, HH\}, \\ W^2_{ij} &\geq W^1_{ij}, \text{ if } ij \in \{LL, HL\}. \end{aligned}$$

If (q^1, W^1, R) *satisfies constraint* $IC(i'j' - ij)$*, then* (q^2, W^2, R) *also satisfies it.*

Proof. Consider the following constraint $IC(i'j' - ij)$:

$$R_{i'j'} - R_{ij} \geq (v_{i'} - v_i)q_{ij} - (c_{j'} - c_j)W_{ij}.$$

Since the left hand side of the constraint only depends on R, if we could show that

$$(v_{i'} - v_i)q^2_{ij} - (c_{j'} - c_j)W^2_{ij} \leq (v_{i'} - v_i)q^1_{ij} - (c_{j'} - c_j)W^1_{ij}, \ \forall ij, i'j' \in T, \qquad \text{(C.1)}$$

then (q^2, W^2, R) must satisfy this constraint if (q^1, W^1, R) does. To see this is the case, we first consider $ij = LH$. In this case, $v_{i'} - v_i \geq 0$ and $c_{j'} - c_j \leq 0$, since $v_i = v_L$ and $c_j = c_H$. Because $q^2_{ij} \leq q^1_{ij}$ and $W^2_{ij} \leq W^1_{ij}$, (C.1) must hold. Following similar procedures, one can show that (C.1) holds if ij is LL, HH or HL. □

Lemma 6. $W_{LH} \leq W_{LL}$*, and* $W_{HH} \leq W_{HL}$*.*

Proof. From the IC constraints $IC(LL - LH)$ and $IC(LH - LL)$, we have $R_{LL} - R_{LH} \geq \Delta_c W_{LH}$, and $R_{LH} - R_{LL} \geq -\Delta_c W_{LL}$. Adding the above two inequalities together, it follows that $0 \geq \Delta_c(W_{LH} - W_{LL})$, which implies that $W_{LH} \leq W_{LL}$. Similarly, one can verify that $IC(HL - HH)$ and $IC(HH - HL)$ implies $W_{HH} \leq W_{HL}$. □

C.2 Proof of Proposition 9

We divided the proof in three parts, each dedicated to show that if the optimal solution does not fully admit the type-HL customers, then it is optimal not to admit any the type-LH, LL or HH customers respectfully. The proof is by contradiction. Thus, in the following three parts we posit hypotheses by negation and then show that they result in contradictions.

Part 1: $q^*_{LH} > 0$ *and* $q^*_{HL} < 1$.

In this case, for some positive number ϵ, we construct a new solution (q', W', R') as follows:

$$\begin{aligned} q'_{ij} &= \begin{cases} q^*_{ij} + \frac{\epsilon}{\lambda_{ij}} & ij = HL \\ q^*_{ij} - \frac{\epsilon}{\lambda_{ij}} & ij = LH \\ q^*_{ij} & ij \in \{HH, LL\} \end{cases}, \\ W' &= W^*, \\ R' &= R^*. \end{aligned}$$

By Lemma 5, the new solution satisfies all IC constraints. We only need to check the resource constraint $RE(S)$ for all $S \subseteq T$. If $S = \{HL\}$, then $RE(S)$ is not binding at (q^*, W^*, R^*), because from Proposition 11 $RE(\{LH, HH\})$ is binding, leading to a conflict against the properties of a polymatroid. In this case, we can always find an ϵ small enough so that $RE(\{HL\})$ is satisfied at (q', W', R'). On the other hand, if $S \neq \{HL\}$, the right-hand side of $RE(S)$ never increases in the new solution, while the left-hand side remains unchanged. Therefore, the new solution satisfies all resource constraints.

In summary, the new solution is feasible, and the objective increases by $\Delta_v \epsilon$ as compared to (q^*, W^*, R^*). This leads to a contradiction to the optimality of (q^*, W^*, R^*).

Part 2: $q^*_{LL} > 0$ *and* $q^*_{HL} < 1$.

In this case, for some $\epsilon > 0$, we define a new mechanism (q', W', R') as follows:

$$\begin{aligned} q'_{ij} &= \begin{cases} q^*_{ij} + \frac{\epsilon}{\lambda_{ij}} & ij = HL \\ q^*_{ij} - \frac{\epsilon}{\lambda_{ij}} & ij = LL \\ q^*_{ij} & ij \in \{HH, LL\} \end{cases}, \\ W' &= W^*, \\ R' &= R^*. \end{aligned}$$

Following similar procedures as in Case 1, one can verify that (q', W', R') satisfies all IC and resource constraints for sufficiently small ϵ. Furthermore, the objective value under this new solution increases by $\Delta_v \epsilon$, thus leading to a contradiction to the fact that (q^*, W^*, R^*) is optimal.

Part 3: $q^*_{HH} > 0$ and $q^*_{HL} < 1$.

Suppose that $q^*_{HH} > 0$ and $q^*_{HL} < 1$. From parts 1 and 2, we know that $q^*_{LH} = q^*_{LL} = 0$. In this case, $RE(\{HH\})$ and $RE(\{HH, HL\})$ must be binding, otherwise the solution would be suboptimal. This implies $W^*_{HH} = \frac{q^*_{HH}}{\mu - \lambda_{HH} q^*_{HH}}$, and

$$W^*_{HL} = \frac{q^*_{HL} + \frac{\lambda_{HH}}{\lambda_{HL}} q^*_{HH}}{\mu - \lambda_{HL} q^*_{HL} - \lambda_{HH} q^*_{HH}} - \frac{q^*_{HH}}{\mu - q^*_{HH} \lambda_{HH}}.$$

To disprove the optimality of (q^*, W^*, R^*) , we define a new solution (q', W', R') as below:

$$\begin{aligned}
q'_{ij} &= \begin{cases} q^*_{ij} + \frac{\epsilon}{\lambda_{ij}} & ij = HL \\ q^*_{ij} - \frac{\epsilon}{\lambda_{ij}} & ij = HH \\ 0 & ij \in \{LH, LL\} \end{cases}, \\
W'_{ij} &= \begin{cases} \frac{q'_{HL} + \frac{\lambda_{HH}}{\lambda_{HL}} q'_{HH}}{\mu - \lambda_{HL} q'_{HL} - \lambda_{HH} q'_{HH}} - \frac{q'_{HH}}{\mu - \lambda_{HH} q'_{HH}} & ij = HL \\ \frac{q'_{HH}}{\mu - \lambda_{HH} q'_{HH}} & ij = HH \\ 0 & ij \in \{LH, LL\}. \end{cases}, \\
R' &= R^*.
\end{aligned}$$

By construction, $\lambda_{HH} W'_{HH} + \lambda_{HL} W'_{HL} = \lambda_{HH} W^*_{HH} + \lambda_{HL} W^*_{HL}$, and it can be verified that the objective function, $\sum_{ij \in T} \lambda_{ij}(v_i q_{ij} - c_j W_{ij} - R_{ij})$, increases by a positive amount of $\Delta_c(\lambda_{HH} W^*_{HH} - \lambda_{HH} W'_{HH})$ in the new solution. As all the resource constraints are clearly satisfied by the new solution, we just need to check the IC constraints. Following Lemma 5, $IC(ij - HL)$ is satisfied, since $q'_{HL} > q^*_{HL}$ and $W'_{HL} > W^*_{HL}$. Furthermore, $IC(LH - HH)$ and $IC(HL - HH)$ can not be binding at (q^*, W^*, R^*), since $R^*_{HH} = 0$, which means the left-hand sides of these constraints are nonnegative, while the right-hand sides are negative. Therefore the small changes in q_{HH} and W_{HH} will not affect these constraints.

The only constraint that might be violated by the new solution is $IC(LL - HH)$, and this only happens when $IC(LL - HH)$ is binding at (q^*, W^*, R^*). Suppose this is the case, define $f(q_{HH})$ to be the right-hand side of $IC(LL - HH)$:

$$f(q_{HH}) \equiv \Delta_c \frac{q_{HH}}{\mu - \lambda_{HH} q_{HH}} - \Delta_v q_{HH}.$$

The derivative of f is $f'(q_{HH}) = \frac{\mu}{(\mu - \lambda_{HH} q_{HH})^2}$, which is increasing in q_{HH}. This suggests that f is modular. Because $f(0) = 0$, $f_(q^*_{HH}) = R^*_{LL} - R^*_{HH} \geq 0$, and $0 \leq q'_{HH} < q^*_{HH}$, it must be true that $f(q'_{HH}) \leq f(q^*_{HH})$, which implies $IC(LL - HH)$ is not violated by the new solution. □

C.3 Proof of Proposition 10

First, recall the incentive compatibility constraints $IC(HH-LH)$ and $IC(LH-HH)$, $R^*_{HH}-R^*_{LH}\geq\Delta_v q^*_{LH}$ and $R^*_{LH}-R^*_{HH}\geq-\Delta_v q^*_{HH}$. Adding the above two inequalities together, it follows that $0\geq\Delta_v(q^*_{LH}-q^*_{HH})$, which implies that $q^*_{LH}\leq q^*_{HH}$. Similarly, one can verify that $IC(HL-LL)$ and $IC(LL-HL)$ implies $q^*_{LL}\leq q^*_{HL}$.

Second, the claim $q^*_{LL}\leq q^*_{HL}$ follows straightly from Proposition 9, since either $q^*_{HH}=0$ or $q^*_{HL}=1$.

Finally, we show $q^*_{LH}\leq q^*_{LL}$ by contradiction. Suppose that $q^*_{LH}>q^*_{LL}$, we construct a new solution (q',W',R') as below

$$\begin{aligned}
q'_{ij} &= \begin{cases} q^*_{LH}-\frac{\epsilon}{\lambda_{LH}} & ij=LH \\ q^*_{LL}+\frac{\epsilon}{\lambda_{LL}} & ij=LL \\ q^*_{ij} & ij\in\{HH,HL\}\end{cases}, \\
W'_{ij} &= \begin{cases} W^*_{LH}-\frac{\epsilon}{\lambda_{LH}}\frac{1}{\mu-\lambda_{LH}q^*_{LH}} & ij=LH \\ W^*_{LL}+\frac{\epsilon}{\lambda_{LH}}\frac{1}{\mu-\lambda_{LH}q^*_{LH}} & ij=LL \\ W^*_{ij} & ij\in\{HH,HL\}\end{cases}, \\
R' &= R^*.
\end{aligned}$$

The objective value increased by $\Delta_v\frac{\epsilon}{\mu-\lambda_{LH}q^*_{LH}}$ under the new solution. Next we show that it satisfy all IC and resource constraints. Among all resource constraints, we limit our attention to $RE(\{LH\})$ and $RE(\{LH,HH\})$, All resource constraints are either non-binding at (q^*,W^*,R^*), or not affected by the change. To see $RE(\{LH\})$ still holds under the new solution:

$$\begin{aligned}
&\lambda_{LH}W'_{LH}=\lambda_{LH}W^*_{LH}-\frac{\epsilon}{\mu-\lambda_{LH}q^*_{LH}} \\
\geq\ & \frac{\lambda_{LH}q^*_{LH}}{\mu-\lambda_{LH}q^*_{LH}}-\frac{\epsilon}{\mu-\lambda_{LH}q^*_{LH}} \\
=\ & \frac{\lambda_{LH}q'_{LH}}{\mu-\lambda_{LH}q^*_{LH}} \\
\geq\ & \frac{\lambda_{LH}q'_{LH}}{\mu-\lambda_{LH}q'_{LH}}.
\end{aligned}$$

In addition, it follows that

$$\begin{aligned}
&\lambda_{LH}W'_{LH} + \lambda_{HH}W'_{HH} = \lambda_{LH}W^*_{LH} + \lambda_{HH}W^*_{HH} - \frac{\epsilon}{\mu - \lambda_{LH}q^*_{LH}} \\
\geq\ & \frac{\lambda_{LH}q^*_{LH} + \lambda_{HH}q^*_{HH}}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} - \frac{\epsilon}{\mu - \lambda_{LH}q^*_{LH}} \\
\geq\ & \frac{\lambda_{LH}q^*_{LH} + \lambda_{HH}q^*_{HH}}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} - \frac{\epsilon}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} \\
=\ & \frac{\lambda_{LH}q'_{LH} + \lambda_{HH}q'_{HH}}{\mu - \lambda_{LH}q^*_{LH} - \lambda_{HH}q^*_{HH}} \\
\geq\ & \frac{\lambda_{LH}q'_{LH} + \lambda_{HH}q'_{HH}}{\mu - \lambda_{LH}q'_{LH} - \lambda_{HH}q'_{HH}},
\end{aligned}$$

which implies that $RE(\{LH, HH\})$ is satisfied by the new solution.

Following from Lemma 5, all IC constraints remain valid under the new solution except for $IC(HH - LL)$ and $IC(HL - LL)$. Our next step is to show that both of them are non-binding at (q^*, W^*, R^*).

By contradiction, if $IC(HH - LL)$ is binding at (q^*, W^*, R^*), it follows that

$$R^*_{HH} = R^*_{LL} + \Delta_v q^*_{LL} - \Delta_c W^*_{LL} < R^*_{LL} - \Delta_c W^*_{LL} + \Delta_v q^*_{LH} \leq R^*_{LH} + \Delta_v q^*_{LH},$$

where the first inequality follows from the assumption that $q^*_{LL} < q^*_{LH}$, and the second one follows from $IC(LH - LL)$. Clearly, this leads to a contradiction to $IC(HH - LH)$.

Additionally, if $IC(HL - LL)$ is binding, we claim that either $IC(LL - LH)$ or $IC(LL - HH)$ must also binding. First of all, at least one of $IC(LL - LH)$, $IC(LL - HH)$ and $IC(LL - HL)$ has to be binding, otherwise (q^*, W^*, R^*) is suboptimal. Secondly, $IC(LL - HL)$ cannot be binding when $IC(HL - LL)$ is binding, due to the fact that $q^*_{LL} < q^*_{LH} \leq q^*_{HL}$.

If both $IC(HL - LL)$ and $IC(LL - LH)$ are binding, it follows that

$$R^*_{HL} = R^*_{LL} + \Delta_v q^*_{LL} = R^*_{LH} + \Delta_c W^*_{LH} + \Delta_v q^*_{LL} < R^*_{LH} + \Delta c W^*_{LH} + \Delta_v q^*_{LH},$$

which leads to a contradiction to $IC(HL - LH)$.

On the other hand, if both $IC(HL - LL)$ and $IC(LL - HH)$ are binding, it follows that

$$R^*_{HL} = R^*_{LL} + \Delta_v q^*_{LL} = R^*_{HH} + \Delta_c W^*_{HH} - \Delta_v q^*_{HH} + \Delta_v q^*_{LL} < R^*_{HH} + \Delta c W^*_{HH},$$

where the inequality follows from the fact that $q^*_{LL} < q^*_{LH} \leq q^*_{HH}$. Clearly, this leads to a contradiction to $IC(HL - HH)$. □

C.4 Proof of Proposition 11

Suppose that $RE(\{LH, HH\})$ is not binding at (q^*, W^*, R^*), then at most one of $RE(\{LH\})$ and $RE(\{HH\})$ can be binding. Because from the properties of a polymatroid, if both $RE(\{LH\})$ and $RE(\{HH\})$ are binding, it must be that $q^*_{LH} = q^*_{HH} = 0$, which implies that $RE(\{LH, HH\})$ is binding. Let $i_1j_1 \in \{LH, HH\}$ be a type such that $RE(\{ij\})$ is not binding. We consider the following two cases, depending on whether there exists a set $S \subseteq T$ such that $RE(S)$ is binding at (q^*, W^*, R^*) and $LH \in S$ (Case 1) or not (Case 2).

Case 1:

Suppose we can find a set $S \subseteq T$ such that $RE(S)$ is binding at (q^*, W^*, R^*) and $LH \in S$. We let S_0 to be the minimal among such sets. It must be true that $S_0 \cap \{LL, HL\} \neq \Phi$, since otherwise $S_0 = \{LH, HH\}$, which means $RE(\{LH, HH\})$ is binding. Let $i_2j_2 \in S_0 \cap \{LL, HL\}$. We construct a new solution (q', W', R') as follows:

$$\begin{aligned} q' &= q^*, \\ W'_{ij} &= \begin{cases} W^*_{ij} - \frac{\epsilon}{\lambda_{ij}} & ij = i_1j_1 \\ W^*_{ij} + \frac{\epsilon}{\lambda_{ij}} & ij = i_2j_2 \\ W^*_{ij} & ij \in T \setminus \{i_1j_1, i_2j_2\}, \end{cases} \\ R' &= R^*, \end{aligned}$$

where $\epsilon > 0$ is a sufficiently small number.

¿From Lemma 5, the new solution satisfies all IC constraints, we only need to it also satisfies the resource constraint $RE(S)$ for all $S \subseteq T$. As a reminder, the resource constraint is formulated as $\sum_{ij \in S} \lambda_{ij} W_{ij} \geq \frac{\sum_{ij \in S} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in S} \lambda_{ij} q_{ij}}$.

Clearly, $RE(S)$ is satisfied if $S_0 \subseteq S$, since both sides of the constraint remain unchanged under the new solution, as compare to (q^*, W^*, R^*). If $S \subset S_0$, but $i_1j_1 \notin S$, $RE(S)$ is not violated by the new solution, since its left-hand side can only increase, while its right-hand side remains the same. Finally, if $S \subset S_0$ and $i_1j_1 \in S$, it must be true that $RE(S)$ is not binding at (q^*, W^*, R^*). In this case, we could always find an ϵ small enough so that the new solution does not violate $RE(S)$.

However, the server's revenue, $\sum_{ij \in T} \lambda_{ij}(v_i q_{ij} - c_j W_{ij} - R_{ij})$, increases by $\Delta_c \epsilon$ in the new solution, which leads to a contradiction to the optimality of (q^*, W^*, R^*).

Case 2:

If we cannot find a set $S \subseteq T$ that contains the type-ij such that $RE(S)$ is binding at (q^*, W^*, R^*), for some $\epsilon > 0$, we construct a new solution (q', W', R') as below:

$$\begin{aligned} q' &= q^*, \\ W'_{ij} &= \begin{cases} W^*_{ij} - \epsilon & ij = i_1j_1 \\ W^*_{ij} & ij \in T \setminus \{i_1j_1\} \end{cases}, \\ R' &= R^*. \end{aligned}$$

This new solution clearly yields a higher revenue for the service provider. It also satisfies all the IC constraints, following Lemma 5. If we choose a small enough ϵ, the new solution also satisfy the resource constraint $RE(S)$ for all $S \subseteq T$, because either $RE(S)$ is not binding at (q^*, W^*, R^*), or S does not contain type-$i_1 j_1$. This again leads to a contradiction to the fact that (q^*, W^*, R^*) is optimal. □

C.5 Proof of Proposition 12

Without loss of generality, we assume that $q^*_{LH} > 0$, because if $q^*_{LH} = 0$, $RE(\{HH\})$ is equivalent to $RE(\{LH, HH\})$, which is binding by Proposition 11. Suppose $q^*_{HH} < 1$, but $RE(\{HH\})$ is non-binding. For some $\epsilon > 0$, we construct a new solution (q', W', R') as follows:

$$\begin{aligned} q'_{ij} &= \begin{cases} W^*_{LH} - \frac{\epsilon}{\lambda_{LH}} & ij = LH \\ W^*_{HH} + \frac{\epsilon}{\lambda_{LH}} & ij = HH \\ W^*_{ij} & ij \in \{LL, HL\} \end{cases}, \\ W' &= W^*, \\ R' &= R^*. \end{aligned}$$

By Lemma 5, all IC constraints are satisfied by the new solution. All resource constraints must also be satisfied, except for $RE(\{HH\})$. Since $RE(\{HH\})$ is non-binding at (q^*, W^*, R^*), we can always find an ϵ small enough so that this resource constraint is not violated by the new solution. However, the object value increase by $\Delta_v \epsilon$ under the new solution, leading to a contradiction to the optimality of (q^*, W^*, R^*). □

Appendix D

Detailed Solutions to Special Cases in Chapter 5

In this appendix, we construct detailed solutions to some special cases. All of proofs follow three identical steps. In *Step 1*, we construct an upper bound function $g(q)$ or $g(q, W)$ of the server's objective $Z(q, W, R)$ by replacing all the information rent R and the expected delay W of some customer types with their valid lower bounds. In *Step 2*, we then search for $\tilde{q}$ or $(\tilde{q}, \tilde{W})$, the optimal solution to $\max_q\{g(q) : 0 \leq q_{ij} \leq 1 \; \forall ij \; \in T\}$ or $\max_{q,W}\{g(q, W) : 0 \leq q_{ij} \leq 1 \; \forall ij \; \in T, \; (\text{RE})\}$. Based on the concavity of $g(q)$ or $g(q, W)$, it suffice to look for solutions that satisfy the first order conditions. In *Step 3*, we construct a feasible solution (q^*, W^*, R^*) where q^* equals $\tilde{q}$, and W^* and R^* are set to their lower bounds used in *Step 1*. Then we verify that $Z(q^*, W^*, R^*)$ is equal to $g(\tilde{q})$ or $g(q, \tilde{W})$, an upper bound on the server's maximum revenue, validating (q^*, W^*, R^*) as an optimal solution to the server's problem. In the sequel we present our results as propositions and prove them accordingly following the aforementioned procedures.

D.1 One Group of Customers

Proposition 13. *If* $v_H - \frac{c_L}{(\mu - \lambda_{HL})^2} \leq 0$*, the optimal mechanism is to admit only the type-*HL *customers.*

Proof. Step 1: Constructing an upper bound function.

The server's objective function satisfies

$$\begin{aligned}
&Z(q,W,R) \equiv \sum_{ij\in T} \lambda_{ij}(v_i q_{ij} - c_j W_{ij} - R_{ij}) \\
&= v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) \\
&-\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij\in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} \\
&\leq v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij\in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij\in T} \lambda_{ij} q_{ij}} \\
&\equiv g(q),
\end{aligned}$$

where the first inequality is due to the nonnegativity of q, W and R, and the second follows from $RE(T)$. Note that in this case the upper bound function only depends on q.

Step 2: Optimizing the upper bound function.

Define $\tilde{q}_{LH} = \tilde{q}_{LL} = \tilde{q}_{HH} = 0$, and $\tilde{q}_{HL}$ to be the solution of $v_H - \frac{c_L}{(\mu - \lambda_{HL} q_{HL})^2} = 0$. Because $v_H - \frac{c_L}{(\mu - \lambda_{HL})^2} \leq 0$, $\tilde{q}_{HL} \leq 1$. In addition,

$$\frac{\partial g}{\partial q_{ij}}|_{q=\tilde{q}} = \lambda_{ij}(v_H - \frac{c_L}{(\mu - \lambda_{HL}\tilde{q}_{HL})^2}) = 0, \ \forall ij \ \in T.$$

Since g is concave in q, $\tilde{q}$ is optimal to the optimization problem $\max_q\{g(q) : 0 \leq q_{ij} \leq 1 \ \forall ij \ \in T\}$.

Step 3: Constructing a feasible solution.

Given $\tilde{q}$,we construct a solution (q^*, W^*, R^*) as below

$$\begin{aligned}
q^*_{ij} &= \begin{cases} \frac{y^*}{\lambda_{HL}} & ij = HL \\ 0 & ij \in \{LH, LL, HH\} \end{cases}, \\
W^*_{ij} &= \begin{cases} \frac{q^*_{HL}}{\mu - \lambda_{HL} q^*_{HL}} & ij = HL \\ 0 & ij \in \{LH, LL, HH\} \end{cases}, \\
R^*_{ij} &= 0, \ \forall ij \in T.
\end{aligned}$$

Since $Z(q^*, W^*, R^*) = g(\tilde{q})$, it is clear an upper bound on any feasible solution. We only need to verify that (q^*, W^*, R^*) is feasible, which is obvious since all IC and resource constraints are satisfied.

D.2 Two Groups of Customers

Proposition 14. *The optimal mechanism is to fully admit the HL and partially admit the type-LL customers with equal priority, if the following conditions hold:*

- $\Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) \le \Delta_c \frac{1}{\mu}$,
- $v_H - c_L \frac{\mu}{(\mu-\lambda_{HL})^2} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) > 0$,
- $v_H - c_L \frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL})^2} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) \le 0$.

Proof. Step 1: Constructing an upper bound function.

$$
\begin{aligned}
& Z(q,W,R) \\
&= v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) \\
&-\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij\in T} \lambda_{ij} R_{ij} \\
&\le v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v \lambda_{LL} q_{LL} \\
&-\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \lambda_{HL} \Delta_v q_{LL} \\
&\le v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij\in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij\in T} \lambda_{ij} q_{ij}} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) \lambda_{LL} q_{LL} \\
&-\Delta_c \frac{\lambda_{LH} q_{LH} + \lambda_{HH} q_{HH}}{\mu - \lambda_{LH} q_{LH} - \lambda_{HH} q_{HH}} \\
&\equiv g(q).
\end{aligned}
$$

Here the first inequality is due to $IC(HL-LL)$ and the nonnegativity of q and R; the second follows from $RE(T)$ and $RE(\{LH, HH\})$.

Step 2: Optimizing the upper bound function.

Define $\tilde{q}_{HL} = 1$, $\tilde{q}_{LH} = \tilde{q}_{HH} = 0$, and $\tilde{q}_{LL}$ to be the solution to

$$
h(q_{LL}) \equiv v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} q_{LL})^2} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) = 0.
$$

We claim that $\tilde{q}$ is the optimal solution to $\max_q \{g(q) :\ 0 \le q_{ij} \le 1,\ \forall ij \in T\}$.

By assumption, we have

$$
\begin{aligned}
h(0) &= v_H - \frac{c_L}{(\mu - \lambda_{HL})^2} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) > 0, \\
h(1) &= v_H - \frac{c_L}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) \le 0.
\end{aligned}
$$

Because $h(q_{LL})$ is decreasing in q_{LL}, it must be true that $0 < \tilde{q}_{LL} \leq 1$.

We now claim that $\tilde{q}$ satisfies the KKT (Karush-Kuhn-Tucker) conditions of the optimization problem $\max_q \{g(q) : \ 0 \leq q_{ij} \leq 1, \ \forall ij \in T\}$. Since $g(q)$ is a concave function, $\tilde{q}$ must be an optimal solution to this problem and $g(\tilde{q})$ is an upper bound on $Z(q, W, R)$. To see that the KKT conditions are satisfied, we observe that

$$\begin{aligned}
& \frac{\partial g}{\partial q_{LL}}|_{q=\tilde{q}} = \lambda_{LL}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_v (1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
= \ & \lambda_{LL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_v (1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
= \ & 0,
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial g}{\partial q_{HL}}|_{q=\tilde{q}} = \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2}] \\
= \ & \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2}] \\
> \ & \lambda_{HL}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_v (1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
= \ & 0,
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} = \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij})^2} - \Delta_c \frac{\mu}{(\mu - \lambda_{LH} \tilde{q}_{LH} - \lambda_{HH} \tilde{q}_{HH})^2}] \\
= \ & \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_c \frac{1}{\mu}] \\
\leq \ & \lambda_{LH}[v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} \tilde{q}_{LL})^2} - \Delta_v (1 + \frac{\lambda_{HL}}{\lambda_{LL}})] \\
= \ & 0,
\end{aligned}$$

$$\frac{\partial g}{\partial q_{HH}}|_{q=\tilde{q}} = \frac{\lambda_{HH}}{\lambda_{LH}} \frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} \leq 0.$$

Step 3: Constructing a feasible solution.

Define (q^*, W^*, R^*) as follows:

$$\begin{aligned}
q^* &= \tilde{q}, \\
W^*_{ij} &= \begin{cases} \frac{q^*_{ij}}{\mu - \lambda_{HL} q^*_{HL} - \lambda_{LL} q^*_{LL}} & ij \in \{HL, LL\} \\ 0 & ij \in \{LH, HH\} \end{cases}, \\
R^*_{ij} &= \begin{cases} \Delta_v q^*_{LL} & ij = HL \\ 0 & ij \in \{LH, HH, LL\} \end{cases}.
\end{aligned}$$

By construction, (q^*, W^*, R^*) satisfies all IC and resource constraints. Furthermore, since $Z(q^*, W^*, R^*) = g(\tilde{q})$, (q^*, W^*, R^*) is an optimal solution to the server's problem.

Proposition 15. *The optimal mechanism is to fully admit the type-HL customers, partially admit the type-HH customers, and give the type-HH absolute priority, if the following conditions hold*

- $\Delta_v \geq \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{LL}})\frac{\mu}{(\mu-\lambda_{HH})^2}$,
- $v_H - c_L\frac{\mu}{(\mu-\lambda_{HL})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{HH}})\frac{1}{\mu} > 0$,
- $v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{HH})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH})^2} \leq 0$.

Proof. Step 1: Constructing an upper bound function.

$$\begin{aligned}
&Z(q,W,R)\\
&= v_H\sum_{ij\in T}\lambda_{ij}q_{ij} - c_L\sum_{ij\in T}\lambda_{ij}W_{ij} - \Delta_v(\lambda_{LH}q_{LH}+\lambda_{LL}q_{LL})\\
&-\Delta_c(\lambda_{LH}W_{LH}+\lambda_{HH}W_{HH}) - \sum_{ij\in T}\lambda_{ij}R_{ij}\\
&\leq v_H\sum_{ij\in T}\lambda_{ij}q_{ij} - c_L\sum_{ij\in T}\lambda_{ij}W_{ij} - \Delta_c\lambda_{HH}W_{HH}\\
&-\Delta_v(\lambda_{LH}q_{LH}+\lambda_{LL}q_{LL}) - \lambda_{HL}\Delta_c W_{HH}\\
&\leq v_H\sum_{ij\in T}\lambda_{ij}q_{ij} - c_L\frac{\sum_{ij\in T}\lambda_{ij}q_{ij}}{\mu-\sum_{ij\in T}\lambda_{ij}q_{ij}}\\
&-\Delta_c(1+\frac{\lambda_{HL}}{\lambda_{HH}})\frac{\lambda_{HH}q_{HH}}{\mu-\lambda_{HH}q_{HH}} - \Delta_v(\lambda_{LH}q_{LH}+\lambda_{LL}q_{LL})\\
&\equiv g(q).
\end{aligned}$$

Here the first inequality is due to $IC(HL-HH)$ and the nonnegativity of W and R, while the second one follows from $RE(T)$ and $RE(\{HH\})$.

Step 2: Optimizing the upper bound function.

Define $\tilde{q}_{HL}=1$, $\tilde{q}_{LH}=\tilde{q}_{LL}=0$, and $\tilde{q}_{HH}$ to be the solution to

$$h(q_{HH}) \equiv v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{HH}q_{HH})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{LL}})\frac{\mu}{(\mu-\lambda_{HH}q_{HH})^2} = 0.$$

We claim that $\tilde{q}$ is the optimal solution to $\max_q\{g(q):\ 0\leq q_{ij}\leq 1,\ \forall ij\in T\}$.

By assumption, we have

$$\begin{aligned}
h(0) &= v_H - c_L\frac{\mu}{(\mu-\lambda_{HL})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{HH}})\frac{1}{\mu} > 0,\\
h(1) &= v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{HH})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH})^2} \leq 0.
\end{aligned}$$

Because $h(q_{HH})$ is decreasing in q_{HH}, it must be true that $0<\tilde{q}_{HH}\leq 1$.

Moreover, it follows that

$$\begin{aligned}
&\frac{\partial g}{\partial q_{HH}}|_{q=\tilde{q}} = \lambda_{HH}[v_H - c_L\frac{\mu}{(\mu-\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{LL}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})}]\\
= \;& \lambda_{HH}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{HH}\tilde{q}_{HH})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{LL}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}]\\
= \;& 0,
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial g}{\partial q_{HL}}|_{q=\tilde{q}} = \lambda_{HL}[v_H - c_L\frac{\mu}{(\mu-\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij})^2}]\\
= \;& \lambda_{HL}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}\tilde{q}_{HH})^2}]\\
> \;& \lambda_{HL}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{HH}\tilde{q}_{HH})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{LL}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}]\\
= \;& 0,
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} = \lambda_{LH}[v_H - c_L\frac{\mu}{(\mu-\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij})^2} - \Delta_c\frac{\mu}{(\mu-\lambda_{LH}\tilde{q}_{LH}-\lambda_{HH}\tilde{q}_{HH})^2}]\\
= \;& \lambda_{LH}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{HH}\tilde{q}_{HH})^2} - \Delta_v]\\
\leq \;& \lambda_{LH}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}\tilde{q}_{LL})^2} - \Delta_c(1+\frac{\lambda_H L}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH})^2}]\\
\leq \;& \lambda_{LH}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{HH}\tilde{q}_{HH})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{LL}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}]\\
= \;& 0,
\end{aligned}$$

$$\frac{\partial g}{\partial q_{LL}}|_{q=\tilde{q}} = \frac{\lambda_{LL}}{\lambda_{LH}}\frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} \leq 0.$$

Clearly, $\tilde{q}$ satisfies the KKT conditions of the optimization problem $\max_q \{g(q) : 0 \leq q_{ij} \leq 1,\ \forall ij \in T\}$. Since $g(q)$ is a concave function, $\tilde{q}$ must be an optimal solution to this problem and $g(\tilde{q})$ is an upper bound on $Z(q, W, R)$.

Step 3: Constructing a feasible solution.

Define (q^*, W^*, R^*) as follows:

$$\begin{aligned}
q^* &= \tilde{q},\\
W^*_{ij} &= \begin{cases} \frac{q^*_{HH}}{\mu-\lambda_{HH}q^*_{HH}} & ij = HH \\ \frac{1}{\lambda_{HL}}(\frac{\lambda_{HH}q^*_{HH}+\lambda_{HL}q^*_{HL}}{\mu-\lambda_{HH}q^*_{HH}-\lambda_{HL}q^*_{HL}} - \frac{\lambda_{HH}q^*_{HH}}{\mu-\lambda_{HH}q^*_{HH}}) & ij = HL \\ 0 & ij \in \{LH, LL\} \end{cases},\\
R^*_{ij} &= \begin{cases} \Delta_c W^*_{HH} & ij = HL \\ 0 & ij \in \{LH, HH, LL\} \end{cases}.
\end{aligned}$$

By construction, (q^*, W^*, R^*) satisfies all IC and resource constraints. Furthermore, since $Z(q^*, W^*, R^*) = g(\tilde{q})$, (q^*, W^*, R^*) is an optimal solution to the server's problem.

D.3 Three Groups of Customers

Proposition 16. *The optimal mechanism is to fully admit the type-HL and type-LL customers, partially admit the type-HH customers, give the type-HH absolute priority, and treat the type-HL and type-LL customers equally, if the following conditions hold:*

- $\Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \leq \Delta_c \frac{1}{\mu}$,
- $\Delta_v(1 + \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} + \frac{\lambda_{HH}+\lambda_{LL}+\lambda_{HL}}{\lambda_{LH}}) \geq \Delta_c(1 + \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH})^2}$,
- $v_H - c_L \frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{HH}})\frac{1}{\mu} + \Delta_v \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} > 0$,
- $v_H - c_L \frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{HH})^2} - \Delta_c(1 + \frac{\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH})^2} + \Delta_v \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} \leq 0$.

Proof. Step 1: Constructing an upper bound function.

$$
\begin{aligned}
& Z(q, W, R) \\
& = v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) \\
& -\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij\in T} \lambda_{ij} R_{ij} \\
& \leq v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) \\
& -\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \lambda_{HH}\Delta_v q_{LH} - \lambda_{LL}(\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH}) \\
& -\lambda_{HL}(\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH} + \Delta_v q_{LL}) \\
& = v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})\lambda_{LL} q_{LL} \\
& -\Delta_c(1 + \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\lambda_{HH} W_{HH} + \Delta_v \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} \lambda_{HH} q_{HH} \\
& -\Delta_v(1 + \frac{\lambda_{HH}+\lambda_{LL}+\lambda_{HL}}{\lambda_{LH}})\lambda_{LH} q_{LH} \\
& \leq v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij\in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij\in T} \lambda_{ij} q_{ij}} - \Delta_v(1 + \frac{\lambda_{HL}}{\lambda_{LL}})\lambda_{LL} q_{LL} \\
& -\Delta_c(1 + \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{\lambda_{HH} q_{HH}}{\mu - \lambda_{HH} q_{HH}} + \Delta_v \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} \lambda_{HH} q_{HH} \\
& -\Delta_v(1 + \frac{\lambda_{HH}+\lambda_{LL}+\lambda_{HL}}{\lambda_{LH}})\lambda_{LH} q_{LH} \\
& \equiv g(q).
\end{aligned}
$$

Step 2: Optimizing the upper bound function. Define $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$, $\tilde{q}_{LH} = 0$ and $\tilde{q}_{HH}$ to be the solution to

$$h(q_{HH}) \equiv v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH} q_{HH})^2}$$
$$-\Delta_c (1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH} q_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} = 0.$$

By assumption,

$$\begin{aligned} h(0) &= v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL})^2} - \Delta_c (1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{1}{\mu} \\ &\quad + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} > 0, \\ h(1) &= v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{LL} - \lambda_{HH})^2} - \Delta_c (1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH})^2} \\ &\quad + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} \leq 0. \end{aligned}$$

Because $h(q_{HH})$ is decreasing in q_{HH}, it follows that $0 < \tilde{q}_{HH} \leq 1$. Next we verify that $\tilde{q}$ satisfies the KKT conditions of the optimization problem $\max_q \{g(q) : \ 0 \leq q_{ij} \leq 1, \ \forall ij \in T\}$. Because $g(q)$ is concave, these conditions are sufficient for optimality.

The partial derivatives at $\tilde{q}$ satisfy

$$\begin{aligned}
&\frac{\partial g}{\partial q_{HH}}|_{q=\tilde{q}} = \lambda_{HH}[v_H - c_L\frac{\mu}{(\mu-\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij})^2} \\
&-\Delta_c(1+\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}+\Delta_v\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}}] \\
= \;& \lambda_{HH}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{HH}\tilde{q}_{HH})^2} \\
&-\Delta_c(1+\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}+\Delta_v\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}}] \\
= \;& 0,
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial g}{\partial q_{HL}}|_{q=\tilde{q}} = \lambda_{HL}[v_H - c_L\frac{\mu}{(\mu-\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij})^2}] \\
= \;& \lambda_{HL}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{LL}\tilde{q}_{HH})^2}] \\
> \;& \lambda_{HL}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{HH}\tilde{q}_{HH})^2} \\
&-\Delta_c(1+\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}+\Delta_v\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}}] \\
= \;& 0,
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial g}{\partial q_{LL}}|_{q=\tilde{q}} = \lambda_{LL}[v_H - c_L\frac{\mu}{(\mu-\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij})^2} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}})] \\
= \;& \lambda_{LL}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{HH}\tilde{q}_{HH})^2} - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}})] \\
> \;& \lambda_{LL}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{HH}\tilde{q}_{HH})^2} \\
&-\Delta_c(1+\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}+\Delta_v\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}}] \\
= \;& 0,
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial g}{\partial q_{LH}}|_{q=\tilde{q}} = \lambda_{LH}[v_H - c_L\frac{\mu}{(\mu-\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij})^2} - \Delta_v(1+\frac{\lambda_{HH}+\lambda_{LL}\lambda_{HL}}{\lambda_{LH}})] \\
= \;& \lambda_{LH}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{HH}\tilde{q}_{HH})^2} - \Delta_v(1+\frac{\lambda_{HH}+\lambda_{LL}\lambda_{HL}}{\lambda_{LH}})] \\
\leq \;& \lambda_{LH}[v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL}-\lambda_{HH}\tilde{q}_{HH})^2} \\
&-\Delta_c(1+\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{\mu}{(\mu-\lambda_{HH}\tilde{q}_{HH})^2}+\Delta_v\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}}] \\
= \;& 0.
\end{aligned}$$

Step 3: Constructing a feasible solution

Given $\tilde{q}$, we define a solution (q^*, W^*, R^*) as follows:

$$
\begin{aligned}
q^* &= \tilde{q}, \\
W^*_{ij} &= \begin{cases} \frac{q^*_{HH}}{\mu - \lambda_{HH} q^*_{HH}} & ij = HH \\ \frac{1}{\lambda_{LL}+\lambda_{HL}} \left(\frac{\lambda_{HH} q^*_{HH} + \lambda_{LL} q^*_{LL} + \lambda_{HL} q^*_{HL}}{\mu - \lambda_{HH} q^*_{HH} - \lambda_{LL} q^*_{LL} - \lambda_{HL} q^*_{HL}} - \frac{\lambda_{HH} q^*_{HH}}{\mu - \lambda_{HH} q^*_{HH}} \right) & ij \in \{LL, HL\} \\ 0 & ij = LH, \end{cases} \\
R^*_{ij} &= \begin{cases} 0 & ij = LH \\ \Delta_v q^*_{LH} & ij = HH \\ \Delta_v q^*_{LH} + \Delta_c W^*_{HH} - \Delta_v q^*_{HH} & ij = HL \\ \Delta_v q^*_{LH} + \Delta_c W^*_{HH} - \Delta_v q^*_{HH} + \Delta_v q^*_{LL} & ij = HL. \end{cases}
\end{aligned}
$$

By construction, (q^*, W^*, R^*) satisfies all IC and resource constraints. Furthermore, since $Z(q^*, W^*, R^*) = g(\tilde{q})$, (q^*, W^*, R^*) is an optimal solution to the server's problem.

Proposition 17. *If all of the following conditions holds, the optimal admission policy is to fully admit the type-HL and type-LL customers and partially admit the type-HH customers. The optimal priority ranking is absolute, with type-HH at the highest, followed by type-HL, and type-LL at the lowest. Strategic idleness is always required.*

- $\Delta_v \geq \Delta_c (1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH})^2}$,
- $\Delta_v > \frac{\Delta_c}{\lambda_{LL}} \left(\frac{\lambda_{HH}+\lambda_{HL}+\lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH}+\lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} \right)$,
- $\Delta_v (1 + \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} + \frac{\lambda_{HH}+\lambda_{LL}+\lambda_{HL}}{\lambda_{LH}}) \leq \Delta_c \frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} \frac{1}{\mu}$,
- $c_L < \Delta_c \frac{\lambda_{HH}}{\lambda_{LL}+\lambda_{HL}}$,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH})^2} - \Delta_c (1 + \frac{\lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH})^2} > 0$.
- $v_L - c_L \frac{\Delta_c}{\Delta_v} > 0$,
- $v_H - \Delta_v (1 + \frac{\lambda_{HL}}{\lambda_{LL}}) - c_L \frac{\Delta_c}{\Delta_v} (1 + \frac{\lambda_{HL}}{\lambda_{LL}}) < 0$.

Proof. Step 1: Constructing an upper bound function.

$$
\begin{aligned}
& Z(q, W, R) \\
&= v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \sum_{ij \in T} \lambda_{ij} W_{ij} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) \\
&- \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij \in T} \lambda_{ij} R_{ij} \\
&\leq v_H \sum_{ij \in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \Delta_v (\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL}) \\
&- \Delta_c (\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \lambda_{LL} \Delta_c W_{LH} \\
&- \lambda_{HH} \max\{\Delta_v q_{LL} - \Delta_c W_{LL},\ \Delta_v q_{LL} - \Delta_c W_{HL},\ 0\} - \lambda_{HL} \max\{\Delta_v q_{LL},\ \Delta_c W_{HH}\} \\
&\equiv g(q, W).
\end{aligned}
$$

Step 2: Optimizing the upper bound function. Let $(\tilde{q}, \tilde{W})$ be an optimal solution to $\max_{q,W} \{g(q,W) : \ 0 \le q_{ij} \le 1, \ \forall ij \in T, \ (\text{RE})\}$. First of all, it can be verified that $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$, and $\tilde{q}_{LH} = 0$ if $\tilde{q}_{LL} < 1$. And since W_{HH} doesn't appear in $g(q,W)$, we arbitrarily set it to its lower bound, i.e. $\tilde{W}_{HH} = \frac{1}{\mu - \lambda_{HH}}$. The problem left is to determine $\tilde{q}_{LL}$, $\tilde{W}_{LL}$ and $\tilde{W}_{HL}$.

Because $c_L < \Delta_c \frac{\lambda_{HH}}{\lambda_{LL}+\lambda_{HL}}$, it must be true that $\max\{\Delta_v q_{LL} - \Delta_c W_{LL}, \ \Delta_v q_{LL} - \Delta_c W_{HL}, \ 0\} = 0$, otherwise we can increase W_{LL} and W_{HL} equivalently to increase $g(q,W)$. It follows that

$$\Delta_v \tilde{q}_{LL} - \Delta_c \tilde{W}_{LL} \le 0, \text{ and } \Delta_v \tilde{q}_{LL} - \Delta_c \tilde{W}_{HL} \le 0.$$

This implies that strategic idleness has to be applied to the optimal solution, because otherwise even if LL is given the lowest priority, the left hand side of the first equality will always be positive, under the condition that $\Delta_v > \frac{\Delta_c}{\lambda_{LL}}(\frac{\lambda_{HH}+\lambda_{HL}+\lambda_{LL}}{\mu-\lambda_{HH}-\lambda_{HL}-\lambda_{LL}} - \frac{\lambda_{HH}+\lambda_{HL}}{\mu-\lambda_{HH}-\lambda_{HL}})$. Therefore, we have

$$\begin{aligned}
\tilde{W}_{LL} &= \frac{\Delta_v}{\Delta_c}\tilde{q}_{LL}, \\
\tilde{W}_{HL} &= \max\{\frac{\Delta_v}{\Delta_c}\tilde{q}_{LL}, \ \frac{1}{\lambda_{HL}}(\frac{\lambda_{HH}+\lambda_{HL}}{\mu-\lambda_{HH}-\lambda_{HL}} - \frac{\lambda_{HH}}{\mu-\lambda_{HH}})\},
\end{aligned}$$

since $\tilde{W}_{HL}$ has to satisfy $RE(\{HH, HL\})$ at the same time.

After substituting W_{LL} and W_{HL} out, it can be verified that $g(q,W)$ has two breakpoint at $q^1_{LL} \equiv \frac{\Delta_c}{\Delta_v(\mu-\lambda_{HH})}$ and $q^2_{LL} \equiv \frac{\Delta_c}{\Delta_v \lambda_{HL}}(\frac{\lambda_{HH}+\lambda_{HL}}{\mu-\lambda_{HH}-\lambda_{HL}} - \frac{\lambda_{HH}}{\mu-\lambda_{HH}})$. Furthermore, $0 < q^1_{LL} < q^2_{LL}$.

The marginal benefit of increasing q_{LL} is

$$\begin{aligned}
&v_L - c_L\frac{\Delta_c}{\Delta_v}, \ 0 \le q_{LL} \le q^1_{LL}, \\
&v_H - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) - c_L\frac{\Delta_c}{\Delta_v}, \ q^1_{LL} < q_{LL} \le q^2_{LL}, \\
&v_H - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) - c_L\frac{\Delta_c}{\Delta_v}(1+\frac{\lambda_{HL}}{\lambda_{LL}}), \ q^2_{LL} < q_{LL} \le 1.
\end{aligned}$$

Because $v_L - c_L\frac{\Delta_c}{\Delta_v} > 0$ and $v_H - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) - c_L\frac{\Delta_c}{\Delta_v}(1+\frac{\lambda_{HL}}{\lambda_{LL}}) > 0$, the optimal value of q_{LL} is either q^1_{LL} or q^2_{LL}, depending on the sign of $v_H - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) - c_L\frac{\Delta_c}{\Delta_v}$. Specifically, if $v_H - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) - c_L\frac{\Delta_c}{\Delta_v} \le 0$, $(\tilde{q}, \tilde{W})$ is given below

$$\tilde{q}_{ij} = \begin{cases} 0 & ij = LH \\ 1 & ij \in \{HH, HL\} \\ \frac{\Delta_c}{\Delta_v(\mu-\lambda_{HH})} & ij = LL \end{cases},$$

$$\tilde{W}_{ij} = \begin{cases} 0 & ij = LH \\ \frac{1}{\mu-\lambda_{HH}} & ij = HH \\ \frac{1}{\lambda_{HL}}(\frac{\lambda_{HH}+\lambda_{HL}}{\mu-\lambda_{HH}-\lambda_{HL}} - \frac{\lambda_{HH}}{\mu-\lambda_{HH}}) & ij = HL \\ \frac{\Delta_v}{\Delta_c}\tilde{q}_{LL} & ij = LL \end{cases}.$$

If $v_H - \Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) - c_L\frac{\Delta_c}{\Delta_v} > 0$, $(\tilde{q}, \tilde{W})$ can be specified as below

$$\begin{aligned}
\tilde{q}_{ij} &= \begin{cases} 0 & ij = LH \\ 1 & ij \in \{HH, HL\} \\ \frac{\Delta_c}{\Delta_v \lambda_{HL}}\left(\frac{\lambda_{HH}+\lambda_{HL}}{\mu-\lambda_{HH}-\lambda_{HL}} - \frac{\lambda_{HH}}{\mu-\lambda_{HH}}\right) & ij = LL \end{cases} \\
\tilde{W}_{ij} &= \begin{cases} 0 & ij = LH \\ \frac{1}{\mu-\lambda_{HH}} & ij = HH \\ \frac{\Delta_v}{\Delta_c}\tilde{q}_{LL} & ij = \{HL, LL\} \end{cases}.
\end{aligned}$$

Step 3: Constructing a feasible solution.

Given $(\tilde{q}, \tilde{W})$, we define a solution (q^*, W^*, R^*) as follows:

$$\begin{aligned}
q^* &= \tilde{q}, \\
W^* &= \tilde{W}, \\
R^*_{ij} &= \begin{cases} 0 & ij \in \{LH, LL, HH\} \\ \max\{\Delta_v q^*_{LL},\ \Delta_c W^*_{HH}\} & ij = HL \end{cases}.
\end{aligned}$$

By construction, (q^*, W^*, R^*) satisfies all IC and resource constraints. Furthermore, since $Z(q^*, W^*, R^*) = g(\tilde{q}, \tilde{W})$, (q^*, W^*, R^*) is an optimal solution to the server's problem.

D.4 Four Groups of Customers

Proposition 18. *If the following conditions hold, the optimal admission control is to fully admit the type-HL and type-LL customers and partially admit the type-HH and type-LH customers. If $q^*_{HH} < 1$ and $q^*_{LH} < 1$, the optimal priority ranking is absolute, with type-HH at the highest, followed by type-LH at the second, and type-LL and type-HL equally at the lowest. If $q^*_{HH} = 1$ and $q^*_{LH} < 1$, the optimal mechanism uses randomized ranking between type-HH and type-LH. If $q^*_{HH} = 1$ and $q^*_{LH} = 1$, the optimal mechanism only uses two priority classes, with type-LH and type-HH equally at the higher, and type-LL and type-HL equally at the lower priority.*

- $\Delta_v(1+\frac{\lambda_{HL}}{\lambda_{LL}}) \le \Delta_c\frac{1}{\mu}$,
- $v_H - c_L\frac{\mu}{(\mu-\lambda_{HL}-\lambda_{LL})^2} - \Delta_c(1+\frac{\lambda_{HL}}{\lambda_{HH}})\frac{1}{\mu} > 0$,
- $\Delta_v(1+\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}} + \frac{\lambda_{HH}+\lambda_{LL}+\lambda_{HL}}{\lambda_{LH}}) < \Delta_c(1+\frac{\lambda_{LL}+\lambda_{HL}}{\lambda_{HH}})\frac{1}{\mu}$,

Proof. Step 1: Constructing an upper bound function.

$$
\begin{aligned}
&Z(q,W,R)\\
&= v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL})\\
&-\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij\in T} \lambda_{ij} R_{ij}\\
&\le v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij\in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij\in T} \lambda_{ij} q_{ij}} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL})\\
&-\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH})\\
&-\lambda_{HH}\Delta_v q_{LH} - \lambda_{LL} \max\{\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH},\ \Delta_c W_{LH}\}\\
&-\lambda_{HL} \max\{\Delta_v q_{LH} + \Delta_c W_{HH} - \Delta_v q_{HH} + \Delta_v q_{LL},\ \Delta_c W_{LH} + \Delta_v q_{LL}\}\\
&\equiv g(q,W).
\end{aligned}
$$

Step 2: Optimizing the upper bound function.

Let $(\tilde{q}, \tilde{W})$ be an optimal solution to $\max_{q,W} \{g(q,W) :\ 0 \le q_{ij} \le 1,\ \forall ij \in T,\ \text{(RE)}\}$. Clearly, $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$ in this case. The problem left is to find the optimal value of q_{HH}, q_{LH}, W_{HH} and W_{LH}. For fixed q, the optimal value of W depends on q_{HH}. If $\tilde{q}_{HH} < 1$, type-HH should be given absolute priority over LH, otherwise $(\tilde{q}, \tilde{W})$ would be suboptimal since we could re-balance the allocations between type-LH and type-HH to improve $g(q,W)$. This implies

$$\tilde{W}_{HH} = \frac{\tilde{q}_{HH}}{\mu - \lambda_{HH}\tilde{q}_{HH}}.$$

Further more, $RE(\{HH, LH\})$ should also be binding at optimality, which leads to

$$\tilde{W}_{LH} = \frac{1}{\lambda_{LH}}\left(\frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\lambda_{HH}\tilde{q}_{HH}}{\mu - \lambda_{HH}\tilde{q}_{HH}}\right)$$

In addition, the following condition has to be satisfied at optimality in order minimize the information rent:

$$\Delta_v \tilde{q}_{LH} + \Delta_c \tilde{W}_{HH} - \Delta_v \tilde{q}_{HH} = \Delta_c \tilde{W}_{LH}.$$

First, we assume $q_{HH} < 1$ and look for a solution that satisfies the first order conditions. Define $q^1_{LL} = q^1_{HL} = 1$, and (q^1_{HH}, q^1_{LH}) to be the solution of

$$
\begin{aligned}
&v_H - c_L \frac{\mu}{(\mu - \lambda_{LL} - \lambda_{HL} - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH})^2}\\
&-\Delta_c\left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}}\right)\frac{\mu}{(\mu - \lambda_{HH} q_{HH})^2} + \Delta_v \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{HH}} = 0,
\end{aligned}
$$

$$v_H - c_L \frac{\mu}{(\mu - \lambda_{LL} - \lambda_{HL} - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH})^2}$$
$$-\Delta_c \frac{\mu}{(\mu - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH})^2} - \Delta_v (1 + \frac{\lambda_{HH} + \lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) = 0,$$

$$\Delta_v q_{LH} + \Delta_c \frac{q_{HH}}{\mu - \lambda_{HH} q_{HH}} - \Delta_v q_{HH} = \Delta_c \frac{1}{\lambda_{LH}} (\frac{\lambda_{HH} q_{HH} + \lambda_{LH} q_{LH}}{\mu - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH}} - \frac{\lambda_{HH} q_{HH}}{\mu - \lambda_{HH} q_{HH}}),$$

where the left hands side of the first two equalities are the marginal benefits to increase q_{HH} and q_{LH} respectively.

If $q^1_{HH} \leq 1$, we claim that $(\tilde{q}, \tilde{W})$, specified as below, is optimal to $\max_{q,W} \{g(q, W) : \ 0 \leq q_{ij} \leq 1, \ \forall ij \in T, \ \text{(RE)}\}$:

$$\begin{aligned} \tilde{q} &= q^1, \\ \tilde{W}_{ij} &= \begin{cases} \frac{\tilde{q}_{HH}}{\mu - \tilde{q}_{HH}} & ij = HH \\ \frac{1}{\lambda_{LH}} (\frac{\lambda_{HH} \tilde{q}_{HH} + \lambda_{LH} \tilde{q}_{LH}}{\mu - \lambda_{HH} \tilde{q}_{HH} - \lambda_{LH} \tilde{q}_{LH}} - \frac{\lambda_{HH} \tilde{q}_{HH}}{\mu - \lambda_{HH} \tilde{q}_{HH}}) & ij = LH \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} (\frac{\sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}} - \frac{\lambda_{HH} \tilde{q}_{HH} + \lambda_{LH} \tilde{q}_{LH}}{\mu - \lambda_{HH} \tilde{q}_{HH} - \lambda_{LH} \tilde{q}_{LH}}) & ij = \{HL, LL\} \end{cases} . \end{aligned}$$

If $q^1_{HH} > 1$, it follows that $\tilde{q}_{HH} = 1$. In order to minimize the information rent and maximize the social surplus, we need to apply randomized priority rule to reduce the difference in W_{HH} and W_{LL} such that the following equalities holds:

$$\Delta_v \tilde{q}_{LH} + \Delta_c \tilde{W}_{HH} - \Delta_v \tilde{q}_{HH} = \Delta_c \tilde{W}_{LH}.$$

The resource constraint $RE(\{HH, LH\})$ should still be binding in this case. Therefore,

$$\lambda_{LH} \tilde{W}_{LH} + \lambda_{HH} \tilde{W}_{HH} = \frac{\lambda_{LH} \tilde{q}_{LH} + \lambda_{HH} \tilde{q}_{HH}}{\mu - \lambda_{LH} \tilde{q}_{LH} - \lambda_{HH} \tilde{q}_{HH}}.$$

Combining the above two equations yields

$$\tilde{W}_{LH} = (\frac{1}{\lambda_{LH} + \lambda_{HH}}) [\frac{\lambda_{HH} + \lambda_{LH} \tilde{q}_{LH}}{\mu - \lambda_{HH} - \lambda_{LH} \tilde{q}_{LH}} - \frac{\Delta_v}{\Delta_c} \lambda_{HH} (1 - \tilde{q}_{LH})].$$

To find the first order solution under the current setting, we define $q^2_{LL} = q^2_{HL} = q^2_{HH} = 1$ and q^2_{LH} to be the solution to the following equation

$$v_H - \Delta_v (1 + \frac{\lambda_{HH}}{\lambda_{LH}}) - \Delta_c (1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) (\frac{1}{\lambda_{LH} + \lambda_{HH}}) [\frac{\mu}{(\mu - \lambda_{HH} - \lambda_{LH} q_{LH})^2} + \frac{\Delta_v \lambda_{HH}}{\Delta_c}] = 0,$$

where the left hand side is the marginal benefit of increasing q_{LH} under this condition.

If $q^2_{LL} < 1$, $(\tilde{q}, \tilde{W})$ can be specified as below:

$$\begin{aligned} \tilde{q} &= q^2 \\ \tilde{W}_{ij} &= \begin{cases} (\frac{1}{\lambda_{LH} + \lambda_{HH}}) [\frac{\lambda_{HH} + \lambda_{LH} \tilde{q}_{LH}}{\mu - \lambda_{HH} - \lambda_{LH} \tilde{q}_{LH}} - \frac{\Delta_v}{\Delta_c} \lambda_{HH} (1 - \tilde{q}_{LH})] & ij = LH \\ \frac{1}{\lambda_{HH}} (\frac{\lambda_{HH} \tilde{HH} + \lambda_{LH} \tilde{LH}}{\mu - \lambda_{HH} \tilde{HH} - \lambda_{LH} \tilde{LH}} - \lambda_{LH} \tilde{W}_{LH}) & ij = HH \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} (\frac{\sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}} - \frac{\lambda_{HH} \tilde{q}_{HH} + \lambda_{LH} \tilde{q}_{LH}}{\mu - \lambda_{HH} \tilde{q}_{HH} - \lambda_{LH} \tilde{q}_{LH}}) & ij \in \{HL, LL\} \end{cases} . \end{aligned}$$

Finally, if $q_{LL}^2 \geq 1$, it is clear that $\tilde{q}_{LH} = \tilde{q}_{HH} = \tilde{q}_{LL} = \tilde{q}_{HL} = 1$, and $\tilde{W}$ can be specified as follows:

$$\tilde{W}_{ij} = \begin{cases} \frac{1}{\mu - \lambda_{LH} + \lambda_{HH}} & ij = \{LH, HH\} \\ \frac{1}{\lambda_{HL} + \lambda_{LL}} \left(\frac{\sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} \tilde{q}_{ij}} - \frac{\lambda_{HH} + \lambda_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}} \right) & ij = \{HL, LL\} \end{cases}.$$

Step 3: Constructing a feasible solution.

Given $(\tilde{q}, \tilde{W})$, we define a solution (q^*, W^*, R^*) as follows:

$$\begin{aligned} q^* &= \tilde{q}, \\ W^* &= \tilde{W}, \\ R_{ij}^* &= \begin{cases} 0 & ij = LH \\ \Delta_v q_{LH}^* & ij = HH \\ \max\{\Delta_v q_{LH}^* + \Delta_c W_{HH}^* - \Delta_v q_{HH}^*,\ \Delta_c W_{LH}^*\} & ij = LL \\ R_{LL}^* + \Delta_v q_{LL}^* & ij = HL \end{cases}. \end{aligned}$$

By construction, (q^*, W^*, R^*) satisfies all IC and resource constraints. Furthermore, since $Z(q^*, W^*, R^*) = g(\tilde{q}, \tilde{W})$, (q^*, W^*, R^*) is an optimal solution to the server's problem.

Proposition 19. *If the following conditions hold, the optimal admission control is to fully admit the type-HL and type-HH customers and partially admit the type-LL and type-LH customers. If $q_{LL}^* < 1$ and $q_{LH}^* < 1$, the optimal priority ranking is absolute, with type-LH at the highest, followed by type-HH at the second highest, type-HL at the third, and type-LL at the lowest. If $q_{HH}^* = 1$ and $q_{LH}^* < 1$, the optimal mechanism uses randomized ranking between LH and HH. If $q_{HH}^* = 1$ and $q_{LH}^* = 1$, the optimal mechanism only uses two priority classes, with type-LH and type-HH equally at the higher priority, and type-LL and type-HL equally at the lower priority.*

- $\Delta_v \geq \Delta_c (1 + \frac{\lambda_{HL}}{\lambda_{LL}}) \frac{\mu}{(\mu - \lambda_{HH})^2}$,
- $\Delta_v \geq \frac{\Delta_c}{\lambda_{LL}} \left(\frac{\lambda_{HH} + \lambda_{HL} + \lambda_{LL}}{\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL}} - \frac{\lambda_{HH} + \lambda_{HL}}{\mu - \lambda_{HH} - \lambda_{HL}} \right)$,
- $v_H - c_L \frac{\mu}{(\mu - \lambda_{HL} - \lambda_{HH})^2} - \Delta_c (1 + \frac{\lambda_{HL}}{\lambda_{HH}}) \frac{\mu}{(\mu - \lambda_{HH})^2} > 0$,
- $c_L \geq \Delta_c \frac{\lambda_{HH} + \lambda_{HL}}{\lambda_{LL}}$,
- $\Delta_v (\frac{\lambda_{HL}}{\lambda_{LL}} - \frac{\lambda_{HH}}{\lambda_{LH}}) > \Delta_c (1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}})$.

Proof. Step 1: Constructing an upper bound function.

$$
\begin{aligned}
&Z(q,W,R)\\
&= v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \sum_{ij\in T} \lambda_{ij} W_{ij} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL})\\
&-\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \sum_{ij\in T} \lambda_{ij} R_{ij}\\
&\le v_H \sum_{ij\in T} \lambda_{ij} q_{ij} - c_L \frac{\sum_{ij\in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij\in T} \lambda_{ij} q_{ij}} - \Delta_v(\lambda_{LH} q_{LH} + \lambda_{LL} q_{LL})\\
&-\Delta_c(\lambda_{LH} W_{LH} + \lambda_{HH} W_{HH}) - \lambda_{LL}\Delta_c W_{LH}\\
&-\lambda_{HH} \max\{\Delta_c W_{LH} + \Delta_v q_{LL} - \Delta_c W_{LL},\ \Delta_v q_{LH}\}\\
&-\lambda_{HL} \max\{\Delta_c W_{LH} + \Delta_v q_{LL} - \Delta_c W_{LL} + \Delta_c W_{HH},\ \Delta_c W_{LH} + \Delta_v q_{LL}\}\\
&\equiv g(q,W).
\end{aligned}
$$

Step 2: Optimizing the upper bound function.

Let $(\tilde{q}, \tilde{W})$ be an optimal solution to $\max_{q,W}\ \{g(q,W) :\ 0 \le q_{ij} \le 1,\ \forall ij \in T,\ \text{(RE)}\}$. It is obvious that $\tilde{q}_{HL} = \tilde{q}_{LL} = 1$. So the problem left is to find the optimal q_{LL}, q_{LH}, W_{LH}, W_{HH} and W_{LL}.

We note that strategic idleness should never be used since the cost overweighs the benefit. Therefore, for fixed q, the optimal value for W_{LL} is equal to its upper bound, which equals to

$$\tilde{W}_{LL} = \frac{1}{\lambda_{LL}}\left(\frac{\sum_{ij\in T} \lambda_{ij}\tilde{q}_{ij}}{\mu - \sum_{ij\in T} \lambda_{ij}\tilde{q}_{ij}} - \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH} + \lambda_{HL}\tilde{q}_{HL}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH} - \lambda_{HL}\tilde{q}_{HL}}\right).$$

The optimal value of W_{LH} and W_{HH} depends on q_{LL}. If $\tilde{q}_{LL} < 1$, it can be verified that LH should be given absolute priority over HH. This implies

$$
\begin{aligned}
\tilde{W}_{LH} &= \frac{\tilde{q}_{LH}}{\mu - \lambda_{LH}\tilde{q}_{LH}},\\
\tilde{W}_{HH} &= \frac{1}{\lambda_{HH}}\left(\frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{LH}\tilde{q}_{LH}}\right).
\end{aligned}
$$

If $q_{LL} = 1$, we may need to apply randomized priority ranking between LH and HH. In either case, $(\tilde{q}, \tilde{W})$ should satisfy the following equality in order to minimize the information rent:

$$\Delta_c \tilde{W}_{LH} + \Delta_v \tilde{q}_{LL} = \Delta_v \tilde{q}_{LH} + \Delta_c \min\{\tilde{W}_{LL},\ \tilde{W}_{HH}\}.$$

We note that $g(q,W)$ has two breakpoints. The first breakpoint (q^a, W^a) occurs when $W^a_{LL} \le W^a_{HH}$, while the second breakpoint (q^b, W^b) arises when $q^b_{LL} = 1$. Clearly $q^a_{LL} < q^b_{LL} = 1$, otherwise W^a_{LL} must be greater than W^a_{HH}. Our next step is to look

for first order solutions separately on the three regimes $0 \leq q_{LL} < q^a_{LL}$, $q^a_{LL} \leq q_{LL} < 1$, and $q_{LL} = 1$

For the first regime $0 \leq q_{LL} \leq q^a_{LL}$, let (q^1, W^1) be the first order solution. Clearly, $q^1_{HH} = q^1_{HL} = 1$, and (q^1_{LL}, q^1_{LH}) is the solution of

$$v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL} q_{LL} - \lambda_{LH} q_{LH})^2} - \Delta_v (1 + \frac{\lambda_{HL}}{\lambda_{LL}}) = 0,$$

$$\begin{aligned} &v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL} q_{LL} - \lambda_{LH} q_{LH})^2} - \Delta_v (1 + \frac{\lambda_{HH}}{\lambda_{LH}}) \\ &-\Delta_c (1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) \frac{\mu}{(\mu - \lambda_{LH} q_{LH})^2} = 0, \end{aligned}$$

$$\Delta_c \frac{q_{LH}}{\mu - \lambda_{LH} q_{LH}} + \Delta_v q_{LL} = \Delta_v q_{LL} + \Delta_c \frac{1}{\lambda_{LL}} \left(\frac{\sum_{ij \in T} \lambda_{ij} q_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q_{ij}} - \frac{\lambda_{HH} q_{HH} + \lambda_{LH} q_{LH} + \lambda_{HL} q_{HL}}{\mu - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH} - \lambda_{HL} q_{HL}} \right),$$

where the left hand sides of the first two equalities are the marginal benefits of increasing q_{LL} and q_{LH} respectively.

In this case, W^1 is determined by

$$W^1_{ij} = \begin{cases} \frac{q^1_{LH}}{\mu - q^1_{LH}} & ij = LH \\ \frac{1}{\lambda_{HH}} \left(\frac{\lambda_{HH} q^1_{HH} + \lambda_{LH} q^1_{LH}}{\mu - \lambda_{HH} q^1_{HH} - \lambda_{LH} q^1_{LH}} - \frac{\lambda_{HH} q^1_{LH}}{\mu - \lambda_{HH} q^1_{LH}} \right) & ij = HH \\ \frac{1}{\lambda_{HL}} \left(\frac{\lambda_{HL} q^1_{HL} + \lambda_{HH} q^1_{HH} + \lambda_{LH} q^1_{LH}}{\mu - \lambda_{HL} q^1_{HL} - \lambda_{HH} q^1_{HH} - \lambda_{LH} q^1_{LH}} - \frac{\lambda_{HH} q^1_{HH} + \lambda_{LH} q^1_{LH}}{\mu - \lambda_{HH} q^1_{HH} - \lambda_{LH} q^1_{LH}} \right) & ij = HL \\ \frac{1}{\lambda_{LL}} \left(\frac{\sum_{ij \in T} \lambda_{ij} q^1_{ij}}{\mu - \sum_{ij \in T} \lambda_{ij} q^1_{ij}} - \frac{\lambda_{HL} q^1_{HL} + \lambda_{HH} q^1_{HH} + \lambda_{LH} q^1_{LH}}{\mu - \lambda_{HL} q^1_{HL} - \lambda_{HH} q^1_{HH} - \lambda_{LH} q^1_{LH}} \right) & ij = LL \end{cases}.$$

If $W^1_{LL} \leq W^1_{HH}$, it follows that $q^1_{LL} \leq q^a_{LL}$. Because $g(q, W)$ is concave, $(\tilde{q}, \tilde{W}) = (q^1, W^1)$ is optimal. If $W^1_{LL} > W^1_{HH}$, we need to continuing searching in the second regime $q^a_{LL} < q_{LL} \leq q^b_{LL}$. Define $q^2_{HH} = q^2_{HL} = 1$, and (q^2_{LL}, q^2_{LH}) to be the solution of

$$v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL} q_{LL} - \lambda_{LH} q_{LH})^2} - \Delta_v (1 + \frac{\lambda_{HL}}{\lambda_{LL}}) = 0,$$

$$\begin{aligned} &v_H - c_L \frac{\mu}{(\mu - \lambda_{HH} - \lambda_{HL} - \lambda_{LL} q_{LL} - \lambda_{LH} q_{LH})^2} - \Delta_v (1 + \frac{\lambda_{HH}}{\lambda_{LH}}) \\ &-\Delta_c (1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}) \frac{\mu}{(\mu - \lambda_{LH} q_{LH})^2} = 0, \end{aligned}$$

$$\Delta_c \frac{q_{LH}}{\mu - \lambda_{LH} q_{LH}} + \Delta_v q_{LL} = \Delta_v q_{LL} + \Delta_c \frac{1}{\lambda_{HH}} \left(\frac{\lambda_{HH} q_{HH} + \lambda_{LH} q_{LH}}{\mu - \lambda_{HH} q_{HH} - \lambda_{LH} q_{LH}} - \frac{\lambda_{LH} q_{LH}}{\mu - \lambda_{LH} q_{LH}} \right).$$

Again, the left hand sides of the first two equalities are the marginal benefits of increasing q_{LL} and q_{LH} under the current assumptions.

If $q_{LL}^2 \leq 1$, the optimal solution $(\tilde{q}, \tilde{W})$ can be specified as below:

$$\begin{aligned}
\tilde{q} &= q^2, \\
\tilde{W}_{ij} &= \begin{cases}
\frac{q_{LH}^2}{\mu - q_{LH}^2} & ij = LH \\
\frac{1}{\lambda_{HH}}\left(\frac{\lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2} - \frac{\lambda_{HH}q_{LH}^2}{\mu - \lambda_{HH}q_{LH}^2}\right) & ij = HH \\
\frac{1}{\lambda_{HL}}\left(\frac{\lambda_{HL}q_{HL}^2 + \lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HL}q_{HL}^2 - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2} - \frac{\lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2}\right) & ij = HL \\
\frac{1}{\lambda_{LL}}\left(\frac{\sum_{ij\in T}\lambda_{ij}q_{ij}^2}{\mu - \sum_{ij\in T}\lambda_{ij}q_{ij}^2} - \frac{\lambda_{HL}q_{HL}^2 + \lambda_{HH}q_{HH}^2 + \lambda_{LH}q_{LH}^2}{\mu - \lambda_{HL}q_{HL}^2 - \lambda_{HH}q_{HH}^2 - \lambda_{LH}q_{LH}^2}\right) & ij = LL
\end{cases}.
\end{aligned}$$

However, if $q_{LL}^2 > 1$, we need to look for the optimal solution in the third regime $q_{LL} = 1$. Let (q^3, W^3) be the first order condition in this case. Apparently, $q_{LL}^3 = q_{HL}^3 = q_{HH}^3 = 1$. The problem left is to determine the optimal q_{LH}, W_{LH} and W_{HH}.

First, we fix q_{LH} and determine the optimal W_{LH} and W_{HH}. Recall that the optimal solution satisfies

$$\Delta_c W_{LH} + \Delta_v q_{LL} = \Delta_v q_{LH} + \Delta_c W_{HH}.$$

Additionally, $RE(\{HH, LH\})$ is binding at optimality, implying

$$\lambda_{LH}W_{LH} + \lambda_{HH}W_{HH} = \frac{\lambda_{LH}q_{LH} + \lambda_{HH}q_{HH}}{\mu - \lambda_{LH}q_{LH} - \lambda_{HH}q_{HH}}.$$

Combining the above two equations yields

$$W_{LH} = \left(\frac{1}{\lambda_{LH} + \lambda_{HH}}\right)\left[\frac{\lambda_{HH} + \lambda_{LH}q_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}q_{LH}} - \frac{\Delta_v}{\Delta_c}\lambda_{HH}(1 - q_{LH})\right].$$

Under this setting, q_{LH}^3 should be the solution of the following equation

$$v_H - \Delta_v\left(1 + \frac{\lambda_{HH}}{\lambda_{LH}}\right) - \Delta_c\left(1 + \frac{\lambda_{LL} + \lambda_{HL}}{\lambda_{LH}}\right)\left(\frac{1}{\lambda_{LH} + \lambda_{HH}}\right)\left[\frac{\mu}{(\mu - \lambda_{HH} - \lambda_{LH}q_{LH})^2} + \frac{\Delta_v\lambda_{HH}}{\Delta_c}\right] = 0,$$

where the left hand side is the marginal benefit of increasing q_{LH} in this case.

If $q_{LL}^3 < 1$, $(\tilde{q}, \tilde{W})$ can be specified as below:

$$\begin{aligned}
\tilde{q} &= q^3, \\
\tilde{W}_{ij} &= \begin{cases}
\left(\frac{1}{\lambda_{LH} + \lambda_{HH}}\right)\left[\frac{\lambda_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}\tilde{q}_{LH}} - \frac{\Delta_v}{\Delta_c}\lambda_{HH}(1 - \tilde{q}_{LH})\right] & ij = LH \\
\frac{1}{\lambda_{HH}}\left(\frac{\lambda_{HH}\tilde{HH} + \lambda_{LH}\tilde{LH}}{\mu - \lambda_{HH}\tilde{HH} - \lambda_{LH}\tilde{LH}} - \lambda_{LH}\tilde{W}_{LH}\right) & ij = HH \\
\frac{1}{\lambda_{HL} + \lambda_{LL}}\left(\frac{\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij}}{\mu - \sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij}} - \frac{\lambda_{HH}\tilde{q}_{HH} + \lambda_{LH}\tilde{q}_{LH}}{\mu - \lambda_{HH}\tilde{q}_{HH} - \lambda_{LH}\tilde{q}_{LH}}\right) & ij = \{HL, LL\}
\end{cases}.
\end{aligned}$$

Finally, if $q_{LL}^3 \geq 1$, it is obvious that $\tilde{q}_{LH} = \tilde{q}_{HH} = \tilde{q}_{LL} = \tilde{q}_{HL} = 1$, and $\tilde{W}$ satisfies

$$\tilde{W}_{ij} = \begin{cases}
\frac{1}{\mu - \lambda_{LH} + \lambda_{HH}} & ij = \{LH, HH\} \\
\frac{1}{\lambda_{HL} + \lambda_{LL}}\left(\frac{\sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij}}{\mu - \sum_{ij\in T}\lambda_{ij}\tilde{q}_{ij}} - \frac{\lambda_{HH} + \lambda_{LH}}{\mu - \lambda_{HH} - \lambda_{LH}}\right) & ij = \{HL, LL\}
\end{cases}.$$

Step 3: Constructing a feasible solution.

Given $(\tilde{q}, \tilde{W})$, we define a solution (q^*, W^*, R^*) as follows:

$$\begin{aligned} q^* &= \tilde{q}, \\ W^* &= \tilde{W}, \\ R^*_{ij} &= \begin{cases} 0 & ij = LH \\ \Delta_c W^*_{LH} & ij = LL \\ \Delta_v q^*_{LH} & ij = HH \\ \Delta_c W^*_{LH} + \Delta_v q^*_{LL} & ij = HL \end{cases} . \end{aligned}$$

By construction, (q^*, W^*, R^*) satisfies all IC and resource constraints. Furthermore, since $Z(q^*, W^*, R^*) = g(\tilde{q}, \tilde{W})$, (q^*, W^*, R^*) is an optimal solution to the server's problem.