Title
System Design and Management with Flexible Structures and Mechanisms

Author
Xu, Ye

Publication Date
2012

Peer reviewed|Thesis/dissertation

System Design and Management with Flexible Structures and Mechanisms

by

Ye Xu

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 Robert C. Leachman
Associate Professor Xuanming Su
Assistant Professor Ying-Ju Chen

Fall 2012

System Design and Management with Flexible Structures and Mechanisms

Abstract

System Design and Management with Flexible Structures and Mechanisms

by

Ye Xu

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun Shen, Chair

Flexible system design has received increasingly more attention in the last a few decades. Flexibility can increase systems' ability to adjust against fast-changing environment, and thereby improves efficiency and reliability, and avoids huge cost from rare but severe disruptions, or loss due to congestions caused by system uncertainties. In this dissertation, we focus on the design and management of flexible systems. In particular, we study three types of flexibility: process flexibility, network flexibility, and payment flexibility. We present quantitative formulations for these problems, and develop different methodologies to solve them. We further conduct numerical studies to generate insights as guidelines for the design of flexible systems in practice.

Flexible supply chains have been widely used by companies to deal with uncertainties. It is well known that chaining structure is very efficient in balanced supply chains. However, it is not clear whether it will work well when supply chains are unbalanced. We study the flexibility design problem of a general supply chain with unbalanced and nonhomogeneous structure. Both demand uncertainty and disruptions are considered in our model. We derive exact solutions for several special cases of the uncapacitated problem where the number of links is fixed, propose an efficient algorithm for solving the general uncapacitated problem, and use simulations to derive some managerial insights for the capacitated problem.

A similar idea is applied to network design. Air transportation networks suffer a lot from disruptions caused by severe weather, natural disasters, power outage, etc. We propose a flexible hub-and-spoke structure in which airports are allowed to have up to N hubs, and formulate the problem as a mixed-integer program that minimizes fixed cost, flexibility cost, and expected transportation cost and penalty cost. Benders decomposition algorithm is applied to solve this problem. Numerical studies show that the performance of the network can be improved substantially with flexible hub assignment, and a flexible structure with $N = 2$ can achieve most of the benefit of those with greater N. We also demonstrate the impact of the correlation between airport disruptions and address the importance of considering it in

stochastic air transportation models.

Trade credit, as a form of flexible payment, is a major tool used by small businesses to obtain external finance. It benefits the buyer and the supplier in multiple ways, and brings risk to them at the same time. We investigate the impact of trade credit on growing small businesses and their suppliers. By looking at a one-supplier-one-retailer supply chain, we study the expansion and inventory policies of the retailer when trade credit is extended or not. It is shown that the retailer grows faster and orders more with trade credit. It is also shown by numerical study that the effect of trade credit depends on demand correlation. When demand is positively correlated, trade credit makes the retailer more likely to go bankrupt, and thereby lowers the supplier's long-term profit and may even cause the failure of the supplier.

To my parents
Fenglou Xu
Lihua Luo
who taught me to work hard and think positive.

Contents

List of Figures

List of Tables

Acknowledgments

I owe my deepest gratitude to my advisor, Zuo-Jun Max Shen, for his guidance, encouragement, and understanding. I am also grateful to Professor Robert Leachman, Professor Xuanming Su, and Professor Ying-Ju Chen for their valuable comments. Last but not least, I thank my family, friends and colleagues for their help and support.

Chapter 1

Introduction

On March 11, 2011, the Tohoku earthquake took place in northeastern Japan. It not only caused damages, costs, and loss of lives within the country, but also had great and long-lasting impact on the global semiconductor industry. Taiwan's chipmakers, as major suppliers of the manufacturers of electronic products around the world, lost their supply of raw materials and key components from Japanese wholesale electronics suppliers. Although several wafer producers in Taiwan and South Korea were running in full capacity to meet demand, only 30% to 50% of the shortage created by the earthquake was fulfilled ([78]). Taiwan's semiconductor manufacturers suffered a lot from shortage in supply. As a result, the price of electronic components and products raised a lot. To a large extent, all these losses are caused by the lack of flexibility in the semiconductor supply chain. Taiwan's semiconductor manufacturers rely heavily on the supply from Japan. For example, 70% of their imported 12-inch silicon wafers were from Japan ([20]). In addition, the Japanese suppliers located most of their plants in the northeastern part of Japan, where happened to be close to the center of the earthquake. Having plants close to each other does have some advantages such as the economies of scale. However, it makes the system vulnerable to disruptions at the same time. The example of the Tohoku earthquake one more time reveals that flexibility is indispensable in supply chains.

Flexible supply chain management has long been studied in the literature, and is commonly used in many industries. Actually, the application of flexibility design is not restricted to supply chain management. It can also be applied in transportation system, financial system, product design, military planning, etc. – every system in general. The flexibility of a system is its ability to adjust itself when environment changes. It measures how much and how fast it is able to take actions in response to the change. The world is changing in a much faster pace, and there is much more uncertainty compared to the past. Hence, no one is able to predict what will happen, and being flexible is even more crucial than ever. Flexibility can help to hedge against uncertainty and to better utilize limited resources, and eventually save cost and improve reliability. In addition, flexibility needs to be incorporated into the strategic level decisions to enable systems to make flexible operations under different circumstances.

Therefore, the design of flexibility systems is an important issue that deserves a lot attention.

In this dissertation, we study three types of flexibility: (1) process flexibility, (2) network flexibility, and (3) payment flexibility. Chapter 2 discusses the flexibility design of manufacturing and service systems. We look at a two-layer supply chain which consists of nonhomogeneous suppliers and retailers. Two types of uncertainty exist: the randomness in demand faced by retailers, and the disruptions that may happen at suppliers and at the links between suppliers and retailers. A dedicated structure is usually used in traditional supply chains, in which each supplier only supplies one retailer, and each retailer only orders from one supplier. This structure incurs huge lost sales when disruptions take place. To mitigate the effects of disruptions, we build additional links so that each retailer can obtain products from multiple suppliers. A maintenance cost is incurred for each link at the same time. Our model seeks to find the optimal link configuration that minimizes the expected total cost.

In Chapter 3, we study the flexibility design of air transportation networks. The traditional hub-and-spoke structure has a lot of advantages and is widely used in transportation and communication networks. However, it is vulnerable to disruptions. As a result, the air transportation network in the United States suffers a lot from disruptions caused by severe weather, natural disasters, power outage, etc. We present a scenario-based flexible hub location model that deals with correlated airport disruptions. In this model, each spoke airport is allowed to select up to N hubs and to decide how much flow to transport via each hub in each scenario. This structure is referred to as a N-flexible hub-and-spoke structure. It incorporates flexibility in both strategic level and operational level decisions, and thus greatly reduces the loss of traveling demand caused by disruptions.

At last, Chapter 4 investigates the impact of trade credit on the growth of small businesses and their suppliers. Trade credit, as a form of flexible payment, serves as a major tool for small businesses to obtain external financial resource. Because the growth of small businesses is often constrained by financial shortage, trade credit accelerates their growths by providing extra cash without any interest. Suppliers of trade credit also benefit from it because it creates new business and promote sales. On the other hand, trade credit brings risk to both the buyer and the supplier. We build a one-supplier-one-retailer supply chain model and study the effect of trade credit on the supplier and the retailer. It is shown that the retailer expands more aggressively and the supplier sells more with trade credit. However, when demand is positively correlated, trade credit makes the retailer more likely to go bankrupt and thereby hurts the profitability of the supplier.

Chapter 2

Flexibility Design of Nonhomogeneous Supply Chains with Disruptions

2.1 Introduction

"Everyone has to become more flexible," said Richard Morris, vice president of BMW Manufacturing Co. ([33]). To cope with rapidly-changing demand, manufacturers have to be able to shift production of different products among different plants. Since 2007, because of the steady increase in gas prices, there has been a greater demand in the U.S. for cars that are more fuel-efficient and affordable. Due to the change of the market, Honda decided to build more 4-cylinder Accords and reduce the production of large vehicles such as trucks. To do this, they move the production of V-6 Accords from a plant in Ohio to a plant in Alabama, and use the Ohio plant to produce more 4-cylinder Accords. The plant in Alabama was a truck plant that had never been used to produce cars like Accord. It took Honda a few months to change over, but the duration was much shorter compared to that of its competitors. The reason why Honda could adjust faster is that their vehicles share some basic design structures. Flexibility can also be within the same factory. Since 2006, Honda has been using the same production line to produce Civic compacts and CR-V crossover. The setup time is only 5 minutes. In most recent years, Honda has become one of the most flexible automakers in North America. Its market share in the U.S. is steadily increasing. Honda's example shows that flexibility has become one of the keys to improve the competitiveness of manufacturing and service companies.

One of the most important questions in designing a flexible supply chain is how much flexibility is enough. Networks with full flexibility usually have the best performance in terms of inventory and service level. However, this performance is obtained at the cost of building a huge number of links. E.g. a supply chain with N suppliers and M demands needs MN links to be fully flexible. [37] introduce the concept of *chaining* to achieve substantial benefits from limited flexibility. They show that the chaining structure can achieve perfor-

mance almost as good as that of the full flexibility structure by only doubling the number of links in a dedicated structure. This result is of great significance because they find a highly connected structure that smartly balances between system performance and cost.

The next question is where to build the links. The balanced nature of chaining implies that it works best in balanced[1] and homogenous[2] supply chains. It may not work well when supply chains are unbalanced and nonhomogeneous, which are more common in practice. Suppose D_1 and D_2 are random demands in a supply chain with means equal to 100 and 1, respectively. It is unwise to equalize the capacities allocated to the two demands because D_1 is far more critical than D_2. Furthermore, in most unbalanced supply chains, it is even impossible to have a balanced link structure other than full flexibility.

Supply disruptions are considered in our model. There has been much research on dealing with demand uncertainty using flexible networks, such as, [37], [1] and [45], while only a few consider supply uncertainty. Supply disruptions have big impact on a supply chain's performance, which has been addressed in much literature such as [65], [71], and [76]. Moreover, it is shown that the optimal strategies for supply chains under disruptions are often the opposite of those under demand uncertainty ([69], [40]). Hence, it is important to consider the impacts of disruptions on supply chain flexibility design.

The rest of this chapter is structured as follows. Section 2.2 briefly reviews some related literature. Section 2.3 describes the general model. Section 2.4 studies a series of uncapacitated problems and provide algorithms to solve them optimally. Section 2.5 discusses some nice properties of capacitated supply chains. Computational studies are performed in section 2.6. A summary is presented in section 2.7. All proofs can be found in the Appendix.

2.2 Literature Review

[37] study the process flexibility in an N-by-N manufacturing system and model it as a bipartite graph. They are the first to introduce the chaining structure, in which all nodes are chained together to form a circle. With full flexibility as a benchmark, they analyze different levels of flexibility in the system and conclude that (1) partial flexibility could achieve most of the benefits of full flexibility, (2) a single long chain is more desirable than several short chains. For supply chains that already exist, principles for adding additional flexibility are proposed: add links that further balance total demand faced by each plant, further balance the capacity allocated to each product, and chain as many nodes as possible. They also investigate the interaction between flexibility and production capacity. It is demonstrated that adding flexibility could be fully substituted by increasing capacity. Moreover, there is

[1] A two-layer supply chain is balanced if it has equal numbers of supply nodes and demand nodes.

[2] A supply chain is homogeneous if it has identical supply nodes, identical demand nodes, and identical links.

no benefit to add flexibility if each supplier's capacity is either no greater than the minimum demand or no less than the maximum total demand.

[1] justify the results of [37]. They generalize the concept of chaining to a "k-chain", which is a bipartite symmetric network in which each node has degree k and the graph is connected in a circular manner. They show that the expected throughput is increasing and concave in k. This result consolidates the advantage of chaining because 2-chain has the most marginal value. They at last prove that it is better to balance the degrees of plant nodes (product nodes) if production capacities (demands) are equal, which is consistent with the principles proposed by [37].

Other research following Jordan and Graves' work provides more perceptions about the chaining concept. [36] evaluate the degree of flexibility of a cross-training CONWIP ([34]) system using the proposed concept *structural flexibility matrix*. [22] adopt the concept of graph expander to show that there exist sparse structures which can perform as good as fully flexible systems, and [21] show analytically that the chaining structure can achieve about 90% of the benefit of a fully flexible system. [31] extend chaining to a multistage supply chain setting. They identify the floating and stage-spanning bottlenecks as factors that cause additional inefficiency when products require multiple process activities, and show that chaining structure is effective even in large-size supply chains. [24] propose guidelines on flexibility design in unbalanced symmetric supply chains based on the Chaining principles.

Another stream of research studies the flexibility structure in nonhomogeneous networks. [45] use a two-stage integer stochastic programming model to formulate the flexibility structure problem. The paper examines a nonhomogeneous system with imperfect resources. They use "marginal cross production costs" to model the efficiency loss due to resource sharing. An efficient heuristic algorithm using Lagrangian relaxation is proposed. Computational results show that their approach generates better solutions than other chaining heuristics when the system is nonhomogeneous, and/or the cost structure is nonhomogeneous. However, the flexibility design problem is treated as a pure mathematical programming problem once the model is set up, which is not very helpful for understanding the underlying nature of the problem.

[5] consider the question of investing in flexibility in a firm that produces N types of products. Unlike all the previous work in which resources (or suppliers and their capacities) are already given, their model assumes that the firm can freely invest in all kinds of resources. A resource has level k if it is able to produce k types of products, and is further characterized by the product set it is able to produce. This problem is formulated as a linear two-stage stochastic program. They prove the decreasing return of flexibility, and show that there are at most two, adjacent levels of resources in the optimal configuration for supply chains with symmetric demand distributions. [4] study a similar model in symmetric queueing systems and show that the chaining structure (with level-2 resources only) performs well.

While all the above literature aims at using flexibility to mitigate demand uncertainty, [40] look at the effects of disruptions in a balanced homogeneous system with cost for extra flexibility. It is assumed that the capacity of each supply node equals to the mean demand at each demand node. It is demonstrated that the common belief that longer chains are more robust may fail under the existence of disruptions. Based on chain structures, they optimize over the size of each subnetwork which has a chain structure. They show that one should have shorter chains if link failures dominate and longer chains if node failures dominate. The paper also studies the networks facing multiple failures through decomposition, and shows that focusing on shorter chains are always more preferable when the probability of disruptions at nodes and/or links is high.

Other literature on flexible systems with disruptions includes: [2] study the problem of maximizing the capacity of a queueing system with fully flexible servers and develop an algorithm to obtain timed, generalized round-robin policies that approach the maximal capacity arbitrarily closely. [63] show that a "W" structure can achieve most of the flexibility of a fully flexible structure in a parallel queueing system with unreliable and nonhomogeneous servers. [62] model a supply chain with 2 products and unreliable suppliers, and reveal that it is beneficial to contracting with a backup supplier, especially when the risk of primary suppliers are well estimated or when the backup premium is high.

Our work contributes to the literature by studying the flexibility design problem of general unbalanced and nonhomogeneous supply chains with both demand uncertainty and disruptions. We develop an efficient algorithm to solve for the optimal flexible structure when supply capacity is abundant, and show by numerical study that our algorithm may work well for supply chains with limited capacity as well. We also show the diminishing returns of flexibility and capacity, and study the interaction between them.

2.3 General Model and Assumptions

We consider a single-period problem in a supply chain with M retailers and N external suppliers as displayed in Figure 2.1. All retailers are owned by a firm and sell the same product. They face random demands $D = (D_1, \ldots, D_M)$ having joint distribution function F and mean $\mu = (\mu_1, \ldots, \mu_M)$. Unmet demand is lost at the end and incurs a unit penalty cost p. Retailers acquire product from the external suppliers. However, suppliers are subject to disruptions. When a supplier is disrupted, it is not able to ship any product. Supplier i fails with probability q_i, for $i = 1, \ldots, N$. We assume that suppliers are independent of each other, so the availability of one supplier does not affect that of the others. We also assume that each supplier has a capacity limit which may result from limited inventory space or production rate.

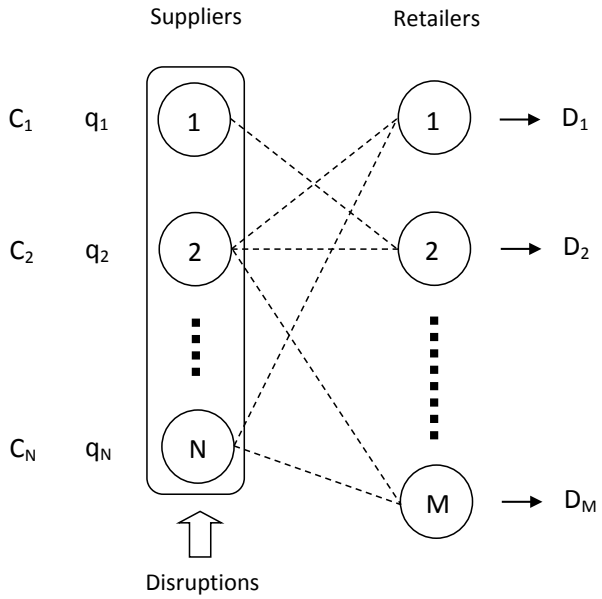


Figure 2.1: An N-by-M Supply Chain

A retailer can obtain product from a supplier only if there is a "link" between them. A link can represent the ability of a plant to produce a product or a contract between two firms. The dashed lines in Figure 2.1 represent potential links to add. We assume that each retailer can only be supplied by a subset of suppliers which is referred to as the *supplier set*. Let $S_j = \left\{ i^j_{(1)}, \ldots, i^j_{(u_j)} \right\}$ be the supplier set of retailer j, where u_j is the cardinality of S_j. There is a centralized decision maker who decides which links to build at the beginning. Each link costs the firm c to maintain it. The maintenance cost can be the cost of membership, communication, or transportation. Afterwards, demands are realized and the status of suppliers are observed. The decision maker then decides how much each retailer should order from each supplier. We assume that there is no production or transportation lead time, so orders arrive immediately. Demands are then fulfilled and penalty and maintenance costs are charged. Our objective is to find the optimal link configuration that minimizes the expected total cost. Now we summarize the notation used in our model.

Stage 1 decision variables:

$$Y_{ij} = \begin{cases} 1 & \text{if build link } (i,j) \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, \ldots, M, \quad i \in S_j.$$

Stage 2 decision variables:

f_{ij} = amount of product from supplier i to retailer j, $\quad j = 1, \ldots, M, \quad i \in S_j$.

Parameters:

c: maintenance cost for each link;
p: penalty cost for each unit of lost sale;
$D = (D_1, \ldots, D_M)$: demands at retailers;
$\mu = (\mu_1, \ldots, \mu_M)$: mean demands at retailers;
$d = (d_1, \ldots, d_M)$: realized demands at retailers;
$q = (q_1, \ldots, q_N)$: failure probabilities of suppliers;
$C = (C_1, \ldots, C_N)$: capacities of suppliers;

State Variables:

$R = (R_1, \ldots, R_N)$: suppliers' states. $R_i = 1$ if supplier i is not disrupted, and 0 otherwise;
$r = (r_1, \ldots, r_N)$: realized states of suppliers;

The problem is formulated as a two-stage mixed integer stochastic programming problem:

- Stage 1:

$$\begin{aligned} \min \quad & c\sum_{j=1}^{M}\sum_{i\in S_j} Y_{ij} + p\mathbb{E}_{D,R}[Q(D,R,Y)] \\ s.t. \quad & Y_{ij} \in \{0,1\} \qquad j = 1,\ldots,M, \quad i \in S_j \end{aligned}$$

- Stage 2:

$$\begin{aligned} Q(d,r,Y) = \quad \min \quad & \sum_j d_j - \sum_{j=1}^{M}\sum_{i\in S_j} f_{ij} \\ s.t. \quad & \sum_{i\in S_j} f_{ij} \le d_j \qquad j = 1,\ldots,M & (2.1) \\ & \sum_{j:i\in S_j} f_{ij} \le C_i r_i \qquad i = 1,\ldots,N & (2.2) \\ & 0 \le f_{ij} \le d_j \cdot Y_{ij} \qquad j = 1,\ldots,M, \quad i \in S_j & (2.3) \end{aligned}$$

Stage 1 problem minimizes the expected total cost (which consists of the link maintenance cost and the expected penalty cost) over all possible link configurations when only the demand distribution and suppliers' failure probabilities are known. Function $Q(D, R, Y)$ denotes the minimum lost sale given demand D, suppliers' states R, and link structure Y. Its value in each scenario is obtained by solving the stage 2 problem.

In stage 2, demands and suppliers' states are realized. Orders are placed for each retailer on suppliers in order to minimize lost sale. f_{ij} denotes the size of the order from supplier i to retailer j. (2.1) requires that the amount of product shipped to retailer j is less than or equal to the demand at retailer j. (2.2) guarantees that the product is shipped from supplier i only if supplier i is not disrupted, and the total amount should not exceed supplier i's capacity. (2.3) makes sure that an order can be placed on supplier i for retailer j only if link (i, j) is built. Since the expected total demand is constant, minimizing lost sale is equivalent to maximizing sales, and thereby stage 2 problem can be solved by solving a max-flow problem.

This two-stage problem is computational intractable. We apply the integer L-Shape method ([7]), but it is not efficient for large-size supply chains. An alternative approach is proposed by [45]. They develop an efficient heuristic algorithm using Lagrangian Relaxation method which can deal with large scale supply chains with continuous demand distributions. Although their model does not consider supply disruptions, we believe that their algorithm still works for our model as the formulation is similar. However, there is no lower bound provided for their heuristic solution, and it is shown that their solution is outperformed by chaining in some cases. More importantly, besides looking for high-quality solutions, we also seek for more insights on the structure of the problem and to develop better solution methods based on its properties.

The main difficulty in solving the general model lies in the estimation of the expected lost sale – we have to solve a max-flow problem for each scenario which has no close form solution. Since we are more interested in the impacts of heterogeneity and disruptions on supply chain design, we relax the constraints on suppliers' capacities for a while. When suppliers are uncapacitated, retailers do not interact with each other so that the expected lost sale at each retailer can be estimated individually. The uncapacitated version of our problem is examined in the next section.

2.4 Uncapacitated Supply Chains

Since the general model is difficult to solve, we relax the capacity constraints by assuming that suppliers have unlimited capacities. Under this assumption, the lost sale at retailer j in a specific scenario (d, r) is equal to d_j if all its suppliers fail, and 0 otherwise. So the lost sale (or sales) at each retailer only depends on its demand and the states of its suppliers. It is not affected by any other retailer or supplier. Given configuration Y, the expected lost sale at retailer j is

$$\mathbb{E}_{D_j,R}\left[D_j \cdot I\Big\{\sum_{i\in S_j} R_i Y_{ij} = 0\Big\}\right] = \mathbb{E}[D_j] \cdot \mathbb{P}\left(\sum_{i\in S_j} R_i Y_{ij} = 0\right) = \mu_j \cdot \prod_{i\in S_j} q_i^{Y_{ij}}$$

where $I\{\cdot\}$ is the indicator function. It turns out that the expected lost sale at retailer j is equal to its mean demand times the probability that all its suppliers are disrupted. Note that we make no assumption on demand distributions. What only matter are the mean demands. Our problem can then be formulated as a nonlinear $0-1$ integer program.

$$\begin{aligned} (PG) \quad \min \quad & c\sum_{j=1}^{M}\sum_{i\in S_j} Y_{ij} + p\left(\sum_{j=1}^{M}\mu_j \cdot \prod_{i\in S_j} q_i^{Y_{ij}}\right) \\ s.t. \quad & Y_{ij}\in\{0,1\} \qquad j=1,\ldots,M, \quad i\in S_j \end{aligned}$$

We note that the maintenance cost is proportional to the summation of Y_{ij}'s. This enables us to construct subproblems of (PG) by fixing the total number of links. In each subproblem, we remove the first term of the objective function and impose a constraint on the total number of links instead. (This constraint might be necessary in some case, e.g. a company may have a budget limit on link maintenance.) We refer to T as the total number of links, so T is an integer between 0 and $\sum_{j=1}^{M} u_j$. Given $T=t$, a subproblem decides where to locate the t links in order to minimize the expected lost sale, or equivalently, to maximize the expected amount of fulfilled demand. The formulation of the subproblem with $T=t$ is given below.

$$\begin{aligned} (PS_t) \quad S(t) \quad = \quad \max \quad & \sum_{j=1}^{M}\mu_j\left(1-\prod_{i\in S_j} q_i^{Y_{ij}}\right) \\ s.t. \quad & \sum_{j=1}^{M}\sum_{i\in S_j} Y_{ij} = t \\ & Y_{ij}\in\{0,1\} \qquad j=1,\ldots,M, \quad i\in S_j \end{aligned}$$

where $S(t)$ denotes the maximum expected sales if the total number of links is equal to t. (PG) can then be reformulated as

$$\begin{aligned} (PG) \quad \min \quad & c\cdot T - pS(T) \\ s.t. \quad & 0\le T\le \sum_{j=1}^{M} u_j, integer \end{aligned}$$

Once we find an efficient approach to solve (PS_t), we can solve (PG) efficiently.

In the rest of this section, we derive solution methods for subproblems, and develop an efficient algorithm to solve the master problem. At last, we discuss the flexibility design problem of an uncapacitated nonhomogeneous supply chain with both supply disruptions and link disruptions.

Solving the Subproblems

To make the analysis easier, we rank the suppliers in each supplier set in increasing order of failure probability. So $q_{i^j_{(1)}} \leq q_{i^j_{(2)}} \leq \ldots \leq q_{i^j_{(u_j)}}$, for any j. Because the suppliers for each retailer are ranked from most reliable to least reliable, it is not reasonable for a retailer to choose a supplier with higher index instead of one with lower index in its supplier set. Then it is natural to have the following lemma.

Lemma 1. *There exists an optimal solution to (PS_t) in which if a retailer is supplied by k suppliers, then it must be supplied by the first k suppliers in its supplier set.*

Lemma 1 implies that we only need to decide the number of suppliers assigned to each retailer. Let K_j be the number of suppliers assigned to retailer j. Then a solution $K = (K_1, \ldots, K_M)$ represents the link structure in which retailer j is supplied by supplier $i^j_{(1)}$ through supplier $i^j_{(K_j)}$, for $j = 1, \ldots, M$. (PS_t) can then be reformulated as

$$
\begin{aligned}
(PS_t) \quad S(t) \quad = \quad \max \quad & \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j} q_{i^j_{(k)}}\right) \\
s.t. \quad & \sum_{j=1}^{M} K_j = t \\
& 0 \leq K_j \leq u_j, integer \qquad j = 1, \ldots, M
\end{aligned}
$$

We first develop analytical approaches to solve (PS_t) for two special cases of uncapacitated nonhomogeneous supply chains, and then extend them to the general case.

Special Case 1: Homogeneous Mean Demands and Unrestricted Supplier Sets

We begin with a simple case in which all demands have identical mean $\bar{\mu}$ as illustrated in Figure 2.2. Recall that the expected lost sale at a retailer is equal to its mean demand times the probability that all its suppliers are disrupted. So we do not have to distinguish retailers from each other in this special case. In addition, we assume that $S_j = \{1, \ldots, N\}$ for any j, and $q_1 \leq q_2 \leq \ldots \leq q_N$. Then (PS_t) has a simplified formulation as below.

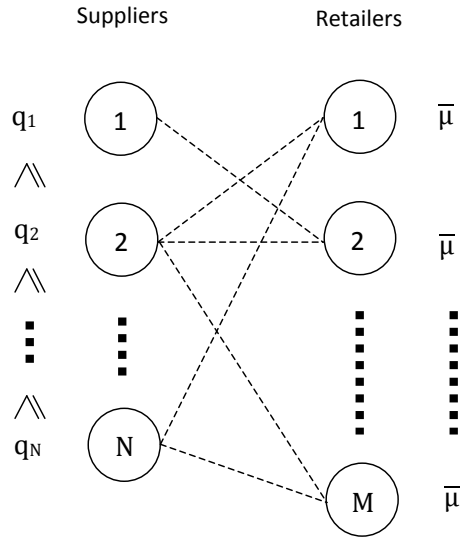


Figure 2.2: An N-by-M Supply Chain with Homogeneous Mean Demands, Nonhomogeneous Supply Reliabilities, and Unrestricted Supplier Sets.

$$
\begin{aligned}
(PS_t^I) \qquad \max \quad & \sum_{k=1}^{N} \left(1 - \prod_{i=1}^{k} q_i\right) L_k \\
s.t. \quad & \sum_{k=1}^{N} k \cdot L_k = t \qquad (1) \\
& \sum_{k=1}^{N} L_k \leq M \qquad (2) \\
& L_k \geq 0, integer \qquad k = 1, \ldots, N
\end{aligned}
$$

where L_k denotes the number of retailers that are supported by k suppliers, where $k = 1, \ldots, N$. (PS_t^I) maximizes the expected sales while satisfying two constraints: (1) the total number of links is equal to t, and (2) the total number of retailers that are covered is at most M.

If $t \leq M$, any non-negative L that satisfies (1) always satisfies (2), so (2) is redundant and can be removed. Now (PS_t^I) is reduced to an unbounded knapsack problem, in which there are N types of goods with unit value $(1 - \prod_{i=1}^{k} q_i)$ and unit cost k, $k = 1, \ldots, N$, and we need to decide how many of each goods to take to maximize the total value with a budget limit is t. This problem is NP-complete in general, but our problem can be solved in polynomial time because of its special property. We relax the integer constraints of the

knapsack problem, then we only need to find the type of goods with the highest 'bang per buck' and make its quantity as large as possible. The following claim shows that the first type of goods has the highest 'bang per buck'.

Claim 2. $1 - q_1 \geq \frac{1-\prod_{i=1}^{k} q_i}{k}$, *for any* $k \geq 2$.

Hence, the optimal solution to the LP relaxation of the knapsack problem is $L_1^* = t$, and $L_k^* = 0$, for $k = 2, \ldots, N$. Note that L^* is integral and thereby feasible and optimal for the knapsack problem as well. Therefore, the optimal link structure when $t \leq M$ is to have supplier 1 supply t retailers.

The above solution, although only solves a special case of (PS_t^I), demonstrates a strategy of link allocation: try to make full use of more reliable suppliers before considering those less reliable. Without loss of generality, we can write any t between 0 and MN as $t = aM + b$, where $a = 0, 1, \ldots, N$, $b = 0, 1, \ldots, M-1$. Then using the above strategy, we should have each of the first a suppliers supply all M retailers, supplier $a+1$ supply b of the retailers, and the other suppliers supply no retailers. Theorem 3 shows that this link structure is actually optimal for (PS_t^I).

Theorem 3. *Let* $a = \max\{k \in \mathbb{Z} : kM \leq t\}$, $b = t - aM$, *then there exists an optimal solution to* (PS_t^I), *denoted by* L^*, *s.t.* $L_a^* = M - b$, $L_{a+1}^* = b$, *and* $L_k^* = 0$ *for any* $k \neq a, a+1$.

Special Case 2: Homogeneous Supply Reliabilities

In this section, we study another special case of (PS_t) in which all suppliers have the same failure probability $\bar{q}$ while retailers can have different average demands and restricted supplier sets. The formulation of this problem is

$$
\begin{aligned}
(PS_t^{II}) \qquad \max \quad & \sum_{j=1}^{M} \mu_j \left(1 - \bar{q}^{K_j}\right) \\
s.t. \quad & \sum_{j=1}^{M} K_j = t \\
& 0 \leq K_j \leq u_j, integer, \qquad j = 1, \ldots, M
\end{aligned}
$$

Problem (PS_t^{II}) has a concave objective function and affine constraints. If the integer constraints are relaxed, the problem becomes a convex optimization problem with strong duality property. However, we can show that the solution to the relaxation problem is not necessarily integral. To find the optimal solution to (PS_t^{II}), or at least to narrow down the feasible region, we further investigate the properties of an optimal solution.

Lemma 4. *There exists an optimal solution to* (PS_t^{II}), *denoted by* K^*, *s.t. for any* $j \neq l$, $K_l^* \leq K_j^*$, *if* $K_j^* < u_j$ *and* $\mu_l \leq \mu_j$.

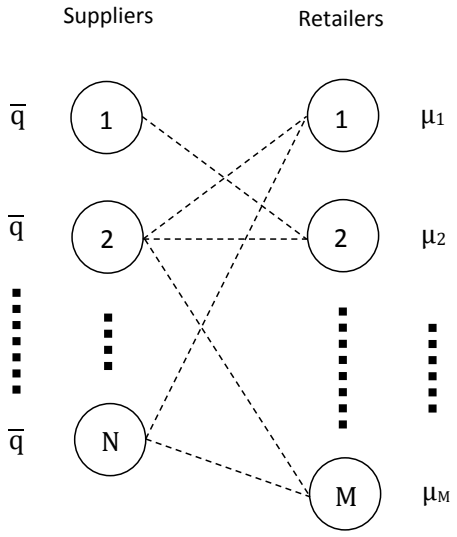


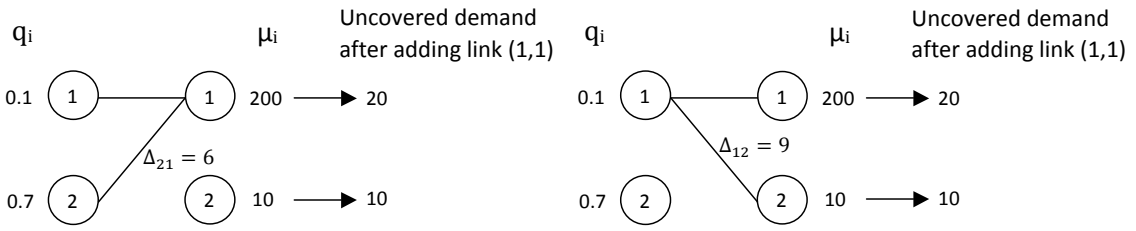
Figure 2.3: An N-by-M Supply Chain with Nonhomogeneous Mean Demands, Homogeneous Supply Reliabilities, and Restricted Supplier Sets.

In other words, a retailer with a greater mean demand should not have less suppliers than one with a smaller mean demand, unless it has been connected to all the suppliers in its supplier set. This is consistent with the intuition that important retailers deserve more reliable supply. From Lemma 4, we immediately arrive at the following theorem.

Theorem 5. *For* (PS_t^{II}) *with* $t > 0$*, there exists an optimal solution* K^* *s.t.* $K^*_{\hat{j}} \geq 1$ *where* $\hat{j} = \text{argmax}_{\{j:u_j>0\}}\{\mu_j\}$.

In light of Theorem 5, we develop an algorithm to solve (PS_t^{II}). The general idea of this algorithm is to add one link at a time until all t links are assigned. In each iteration, we find the retailer with the largest expected unmet demand, and connect it to a supplier that has not been assigned to it if there is any. Here is the notation used in the algorithm:

- ν: number of links that have been built, also the iteration number. $\nu \in \{0, 1, \ldots, t\}$.
- $K = (K_1, K_2, \ldots, K_M)$: number of suppliers that have been connected to each retailer.
- u_j^ν: upper bound on the number of suppliers that can be assigned to retailer j after iteration $\nu - 1$.
- μ_j^ν: expected amount of uncovered demand at retailer j in iteration ν.
- t^ν: number of links to be built after iteration $\nu - 1$.



(a) Algorithm 1 solution (b) Optimal solution

Figure 2.4: Solutions to a 2-by-2 Problem

Next, we propose an algorithm to solve problem (PS_t^{II}).

Algorithm 1:

Step 0:

$\mu_j^0 \leftarrow \mu_j$, $u_j^0 \leftarrow N$, for $j = 1, \ldots, M$

$t^0 \leftarrow t$, $\nu \leftarrow 0$, $K \leftarrow (0, 0, \ldots, 0)$

Step 1:

If $t^\nu = 0$, stop. K is the optimal solution.

Otherwise, go to **Step 2**.

Step 2:

$\hat{j} \leftarrow \text{argmax}_{\{j:u_j^\nu > 0\}}\{\mu_j^\nu\}$

$\nu \leftarrow \nu + 1$, $K \leftarrow K + e_{\hat{j}}$

$$u_j^\nu \leftarrow \begin{cases} u_j^{\nu-1} - 1 & \text{if } j = \hat{j} \\ u_j^{\nu-1} & \text{otherwise} \end{cases}, \mu_j^\nu \leftarrow \begin{cases} \bar{q}\mu_j^{\nu-1} & \text{if } j = \hat{j} \\ \mu_j^{\nu-1} & \text{otherwise} \end{cases}$$

$t^\nu \leftarrow t^{\nu-1} - 1$

Repeat from **Step 1**.

where e_j is the jth unit vector. The complexity of the algorithm is $O(t)$. Since (PS_t^{II}) is a special case of (PS_t), and Algorithm 2 in next section reduces to Algorithm 1 when applied to (PS_t^{II}), the optimality of Algorithm 1 is justified by that of Algorithm 2.

General Case: Nonhomogeneous Supply Reliabilities, Nonhomogeneous Mean Demands, and Restricted Supplier Sets

After examining the two special cases, we are ready to move on to (PS_t). Now both suppliers and retailers are nonhomogeneous. As expected, Algorithm 1 is not guaranteed to find the optimal solution to (PS_t). We illustrate this with a small example. Figure 2.4 describes a 2-by-2 uncapacitated supply chain in which $\mu = (200, 10)$, $q = (0.1, 0.7)$, $t = 2$, and $S_1 = S_2 = \{1, 2\}$. Lemma 1 suggests that link $(1, 1)$ must be built in an optimal solution, and the other link is either $(2, 1)$ or $(1, 2)$. According to Algorithm 1, we should build link $(2, 1)$ because the expected uncovered demand at retailer 1 (which is equal to $200 \times 0.1 = 20$) is greater than that at retailer 2 (which is equal to 10). However, the optimal solution is to build link $(1, 2)$ instead because the expected sales increases by $\Delta_{12} = 10 \times (1 - 0.1) = 9$ if adding link $(1, 2)$, and $\Delta_{21} = 20 \times (1 - 0.7) = 6$ if adding link $(2, 1)$. This example reveals that, to find an optimal solution to (PS_t), it is not enough to compare the expected remaining demands at retailers. The marginal benefit of adding a link is a more reasonable measure, and it is affected by both the expected remaining demand and the candidate supplier's reliability. Then we have the following theorem analogous to Theorem 5.

Theorem 6. *For (PS_t) with $t > 0$, there exists an optimal solution K^* s.t. $K^*_{\hat{j}} \geq 1$ where*

$$\hat{j} = \underset{\{j:u_j>0\}}{\operatorname{argmax}} \left\{ \mu_j \left(1 - q_{i^j_{(1)}} \right) \right\}.$$

Theorem 6 suggests a new criteria for the selection of the additional link in each iteration, and we use it to develop an algorithm to solve (PS_t). We keep the notation of Algorithm 1. Recall that vector K denotes the numbers of suppliers that have been assigned to retailers. In the following algorithm, it also specifies which suppliers have been assigned to retailers because retailers always choose their suppliers in increasing order of failure probability. The algorithm is depicted below.

Algorithm 2:

Step 0:

$\mu^0_j \leftarrow d_j$, $u^0_j \leftarrow u_j$, for any j
$t^0 \leftarrow t$, $\nu \leftarrow 0$, $K \leftarrow (0, 0, \ldots, 0)$

Step 1:

If $t^\nu = 0$, stop. K is the optimal solution.

Otherwise, go to **Step 2**.

Step 2:

$\hat{j} \leftarrow \text{argmax}_{\{j:u_j^\nu>0\}} \left\{ \mu_j^\nu \left(1 - q_{i^j_{(K_j+1)}} \right) \right\}$

$\nu \leftarrow \nu + 1$, $K \leftarrow K + e_{\hat{j}}$

$$u_j^\nu \leftarrow \begin{cases} u_j^{\nu-1} - 1 & \text{if } j = \hat{j} \\ u_j^{\nu-1} & \text{otherwise} \end{cases}, \mu_j^\nu \leftarrow \begin{cases} q_{i^j_{(K_j)}} \mu_j^{\nu-1} & \text{if } j = \hat{j} \\ \mu_j^{\nu-1} & \text{otherwise} \end{cases}$$

$t^\nu \leftarrow t^{\nu-1} - 1$

Repeat from **Step 1**.

The complexity of this algorithm is also $O(t)$. Algorithm 2 is essentially a myopic algorithm because we add the link with the most marginal benefit in each iteration. Nevertheless, in the following theorem, we show that Algorithm 2 is actually global optimal.

Theorem 7. *Algorithm* 2 *gives an optimal solution to problem* (PS_t).

(PS_t^{II}) is a special case of (PS_t), and Algorithm 1 is a special version of Algorithm 2 when applied to (PS_t^{II}). Therefore, the optimality of Algorithm 1 is guaranteed by that of Algorithm 2.

Back to the Master Problem

Recall that $S(T)$ is the maximum expected sales when the number of links equal to T, and the formulation of the master problem is

$$\begin{aligned} (PG) \quad \min \quad & c \cdot T - pS(T) \\ s.t. \quad & 0 \le T \le \sum_{j=1}^{M} u_j, integer \end{aligned}$$

$S(T)$ can be estimated efficiently by Algorithm 2, thus we can apply Algorithm 2 for every T between 0 and $\sum_{j=1}^M u_j$, and identify the T value, called T^*, that minimizes the objective function. The optimal link structure with $T = T^*$ is the optimal solution to (PG). In this way, (PG) can be solved in $O\left(\left(\sum_{j=1}^M u_j\right)^2 + \sum_{j=1}^M u_j\right)$. Next, we propose a more efficient algorithm to solve (PG) based on the following two propositions.

Proposition 8. *Solutions given by Algorithm* 2 *are nested. That is, a solution with* $T = t$ *is fully contained in a solution with* $T = t + 1$, *for* $t = 1, \ldots, \sum_{j=1}^M u_j - 1$.

Proposition 8 shows that, in order to find the optimal solution to (PS_t), we only need to start from an optimal solution to (PS_{t-1}). Put another way, we only need to apply Algorithm 2 once with $T = \sum_{j=1}^M u_j$ to find the optimal solutions and $S(T)$ values for all values of T.

Proposition 9. *The marginal increment in $S(T)$ is decreasing in T.*

The marginal maintenance cost is constant, and Proposition 9 shows that the marginal increment in expected revenue is decreasing in T. Hence, we can solve (PG) optimally by implementing Algorithm 2 with $T = \sum_{j=1}^{M} u_j$, and having it terminate once the marginal increment in expected revenue is less than or equal to the marginal maintenance cost. The algorithm is summarized as follow.

Algorithm 3:

Step 0:

$\mu_j^0 \leftarrow d_j$, $u_j^0 \leftarrow u_j$, for any j

$\nu \leftarrow 0$, $K \leftarrow (0, 0, \ldots, 0)$

Step 1:

If $\nu = \sum_{j=1}^{M} u_j$, stop. K is the optimal solution.

Otherwise, go to **Step 2**.

Step 2:

$\hat{j} \leftarrow \text{argmax}_{\{j:u_j^\nu > 0\}} \left\{\mu_j^\nu \left(1 - q_{K_j+1}\right)\right\}$

$\Delta S \leftarrow \mu_{\hat{j}}^\nu \left(1 - q_{K_{\hat{j}+1}}\right)$

If $\Delta S \leq c/p$, stop. K is the optimal solution.

Otherwise, go to **Step 3**.

Step 3:

$\nu \leftarrow \nu + 1$, $K \leftarrow K + e_{\hat{j}}$

$$u_j^\nu \leftarrow \begin{cases} u_j^{\nu-1} - 1 & \text{if } j = \hat{j} \\ u_j^{\nu-1} & \text{otherwise} \end{cases}, \mu_j^\nu \leftarrow \begin{cases} q_{K_j} \mu_j^{\nu-1} & \text{if } j = \hat{j} \\ \mu_j^{\nu-1} & \text{otherwise} \end{cases}$$

Repeat from **Step 1**.

The complexity of the algorithm is $O\left(\sum_{j=1}^{M} u_j\right)$ so it is very efficient.

Link Disruptions

It is assumed for all the previous models that links are perfectly reliable. However, besides suppliers, links may also fail. A lot of literature (such as [40]) have shown that link disruptions may affect the performance of supply chains a lot and in a different way from supply

disruptions. Therefore, we model the flexibility design problem of an uncapacitated nonhomogeneous supply chain with both supply disruptions and link disruptions in this section. We show that the new model is equivalent to a model with supply disruptions only and can be solved by our algorithm.

To discriminate between supply disruptions and links disruptions, we replace q_i and R_i with q_i^s and R_i^s, respectively, and define the following new notation:
q_{ij}^l: failure probability of the link between supplier i and retailer j;
R_{ij}^l: state of the link between supplier i and retailer j;
ρ_{ij} : correlation coefficient between R_i^s and R_{ij}^l.

With non-zero ρ_{ij}'s, we allow the states of a supplier to be correlated with the states of the links connected to them. This is usually true because sometimes the disruption at a supplier and that at a connected link are caused by the same event. Without loss of generality, assume that supplier sets are unrestricted. As before, because supply capacities are unlimited, the expected lost sale at each retailer can be calculated individually. At any retailer j, its demand D_j cannot be fulfilled if, for every supplier connected to it, either the supplier fails itself or the link connecting them fails. Then the expected lost sale at retailer j is

$$\mathbb{E}_{D_j,R^s,R^l}\left[D_j \cdot I\Big\{\sum_{i=1}^{N} R_i^s R_{ij}^l Y_{ij} = 0\Big\}\right] = \mathbb{E}[D_j]\cdot\mathbb{P}\left(\sum_{i=1}^{N} R_i^s R_{ij}^l Y_{ij} = 0\right) = \mu_j \cdot \prod_{i=1}^{N} P(R_i^s R_{ij}^l = 0)^{Y_{ij}}.$$

Define $\widetilde{q}_{ij} = P(R_i^s R_{ij}^l = 0)$, and by calculation, we have

$$\widetilde{q}_{ij} = q_i^s + q_{ij}^l - q_i^s q_{ij}^l - \rho_{ij}\left[q_i^s(1-q_i^s)q_{ij}^l(1-q_{ij}^l)\right]^{\frac{1}{2}}.$$

Then the expected lost sale at retailer j is equal to $\mu_j \prod_{i=1}^{N} \widetilde{q}_{ij}^{Y_{ij}}$, and the formulation of the problem is

$$\begin{aligned} (PL) \qquad \min \quad & c\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij} + p\left(\sum_{j=1}^{M}\mu_j \prod_{i=1}^{N} \widetilde{q}_{ij}^{Y_{ij}}\right) \\ s.t. \quad & Y_{ij} \in \{0,1\} \qquad i = 1,\ldots,N, j = 1,\ldots,M \end{aligned}$$

This formulation looks very similar to (PG), but now the failure probabilities are defined for (i,j) pairs in stead of suppliers.

Let $S_j = \{(i,j) : i = 1,\dots,N\}$ for any j, and $\widetilde{Y}_{(i,j)j} = Y_{ij}$ for any i, j. Then by substitution, (PL) can be expressed as

$$\begin{aligned} \min \quad & c\sum_{j=1}^{M}\sum_{k\in S_j}\widetilde{Y}_{kj} + p\left(\sum_{j=1}^{M}\mu_j\prod_{k\in S_j}\widetilde{q}_k^{\widetilde{Y}_{kj}}\right) \\ s.t. \quad & \widetilde{Y}_{kj}\in\{0,1\} \qquad j=1,\dots,M, k\in S_j \end{aligned}$$

This is exactly the same formulation as (PG) for an MN-by-M supply chain with mean demands μ, failure probabilities $\widetilde{q}$, and supplier sets S_j's. Therefore, an uncapacitated model with both supply and link disruptions can always be transformed to an equivalent one with supply disruptions only. Moreover, it can be solved by Algorithm 3 in $O(MN)$. Because there is a one-to-one correspondence between Y and $\widetilde{Y}$, the optimal solution to (PL) can be easily restored from the optimal $\widetilde{Y}$.

2.5 When Capacity is Limited

So far we have developed algorithms to solve the network configuration problem with uncapacitated suppliers. In this section, we discuss some properties of capacitated supply chains.

We have shown in Proposition 9 that the marginal benefit of adding flexibility is decreasing when suppliers are uncapacitated. A natural question to ask is: does this property hold when suppliers have limited capacities? The answer seems to be YES according to the numerical experiments in [37] and [38]. Besides, [1] provide analytical justification for the concavity of the throughput in the degree of each node under the assumptions of homogeneous nodes and symmetric network configuration. Nevertheless, it is very difficult to give a general proof because the optimal configurations are not necessarily nested in general. Rather than construct a supply chain in one step, companies usually improve the existing supply chain when it is not flexible enough. Most of the time, the improvement is basically adding more links because it is expensive to rebuild the network from scratch, even though it is not optimal to do so. Based on Corollary 1 in [1], we have a similar proposition as Proposition 9 for capacitated supply chains.

Proposition 10. *In a capacitated nonhomogeneous supply chain, if links are added sequentially following a myopic policy, then the increment in expected sales is non-increasing.*

Another way of improving the performance of a supply chain is to expand supply capacity. We are also interested in how the expected sales change as supply capacity increases.

Proposition 11. *In a capacitated nonhomogeneous supply chain, the expected sales is increasing and concave in supply capacity.*

These two propositions reveal the diminishing return from investment in flexibility or supply capacity. They shed some light on the question how much flexibility (supply capacity) is necessary in a supply chain. The interaction between the two is further investigated in the computational study.

2.6 Numerical Study

In this section, we carry out a series of numerical experiments to demonstrate a potential heuristic algorithm for capacitated supply chains, and to answer some important questions. The first study discusses about the selection of T for capacitated supply chains, the second study tests the performance of Algorithm 2 in capacitated supply chains, and the last study studies the interaction between flexibility and supply capacity.

Unless otherwise noted, we set the number of suppliers to be 3 and the number of retailers 4. This 3-by-4 structure has several advantages: (a) it has an unbalanced structure, (b) it has more retailers than suppliers, which is true for most supply chains, (c) it can have three levels of flexibility: no flexibility ($T \leq 4$), mild flexibility ($5 \leq T \leq 8$), and high flexibility ($T \geq 9$), and (d) its small size allows efficient simulation. In all these studies, each supplier's state is independently generated according to a Bernoulli distribution with success probability $1 - q_i$, $i = 1, 2, 3$. Demands at retailers are sampled from a 4-dimensional truncated normal distribution with mean μ, standard deviation $\sigma = \mu/4$, and the demand distribution is truncated at $\mu \pm 0.5\mu$. We consider both independent demands and positively correlated demands. In the positively correlated demand case, the correlation coefficient between each pair of demands is set at 0.5.

A basic block of all the experiments is the estimation of the expected sales per period. If suppliers' capacities are unlimited, the expected sales can be calculated analytically. However, simulation tools are needed for capacitated supply chains. To obtain a sharp estimation, the simulation horizon is set at $10,000$ periods. At the beginning of each period, we generate a set of demand observations and the suppliers' states. Given the link structure and the realized random variables, the maximum throughput of the supply chain is obtained by a max-flow algorithm. By taking average of the throughputs over the simulation horizon, we obtain an unbiased estimation of the expected sales.

Selection of the Total Number of Links

We have derived an efficient way to find the optimal number of links in uncapacitated supply chains, but we have not shown how to identify this value for capacitated supply chains. This section performs a numerical study on this issue.

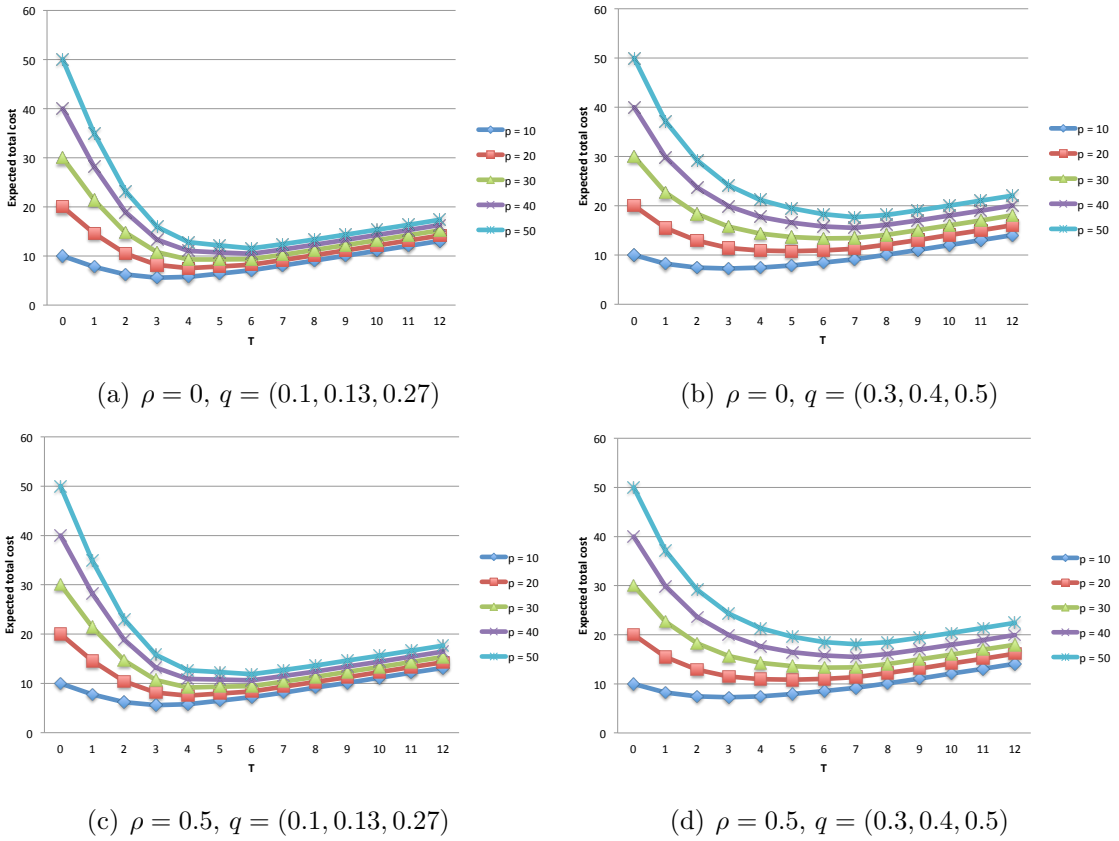
As always argued, flexibility is beneficial but could be costly. Suppose that each link incurs a cost c per period, and each percent of unmet demand incurs a penalty cost p per period. Note that here we interpret p as the penalty cost per percentage lost sale instead of per unit lost sale, so that the effect of demand magnitude is eliminated. If T is small, supply and demand uncertainties may cause large amount of lost sale and thereby a high penalty cost. On the other hand, the link maintenance cost may be high if T is large. So either too much or too little flexibility is not preferable. The tradeoff between penalty cost and link maintenance cost indicates that T should be set at a proper value in between.

To better understand the effect of T, we simulate the expected total cost of unbalanced nonhomogeneous supply chains with $\mu = (40, 30, 20, 10)$ and various T, q, C, and p. The expected total cost of a supply chain is obtained by enumerating all possible link configurations with the given T value and selecting the minimum average total cost. Demands are independent, or have a 0.5 correlation between each other. For simplicity, we set C such that every supplier has the same capacity and the expected total capacity of the supply chain is equal to the expected total demand.

We plot the expected total cost as functions of T in Figure 2.5. Demands are independent in Figure 2.5(a) and 2.5(b) and are correlated with $\rho = 0.5$ in Figure 2.5(c) and 2.5(d). Because we only care about the optimal value of T, denoted by T^*, and we can always rescale the expected total cost by $1/c$ without changing T^*, we fix $c = 1$ to reduce the dimension of the factors affecting T^*. All curves show that the cost function is convex in T. This is because that the expected total cost is the sum of a linear function (maintenance cost) and a convex function (penalty cost of lost sale) of T. The convexity of total cost makes it flat around T^*, hence it only increases slightly if T deviates from T^*. It is also notable that the curves with $\rho = 0.5$ are almost identical to those with $\rho = 0$, which means that the expect total cost does not increase substantially if demands are positively correlated.

Figure 2.5 also reveals that T^* depends on q and p. Consider the extreme cases: if q is the zero vector, all suppliers are perfectly reliable, then not much flexibility is needed to achieve a desirable level of expected sales; on the other hand, if q is $(1, 1, 1)$, no supplier works at any time, then no link should be built. Hence, T^* might be concave in q. Similarly, T^* might be increasing with respect to p by intuition, because the harder lost sale is penalized, the more flexibility is desirable.

To justify the conjectures above, we design an experiment to investigate T^*'s dependency on q and p. We first consider a 3-by-4 nonhomogeneous supply chain with $\mu = (40, 30, 20, 10)$. To make life easier, we let q_i's change proportionally, and C changes accordingly to keep the expected total supply capacity equal to the expected total demand. Figure 2.6 shows T^* v.s. q under different p values with independent or correlated demands. It turns out that T^* is not concave in q as we expected. However, T^* is *quasi-concave* in q. Comparing the curves in the same figure, we find that a curve with a larger p is always above one with a smaller



(a) $\rho = 0$, $q = (0.1, 0.13, 0.27)$

(b) $\rho = 0$, $q = (0.3, 0.4, 0.5)$

(c) $\rho = 0.5$, $q = (0.1, 0.13, 0.27)$

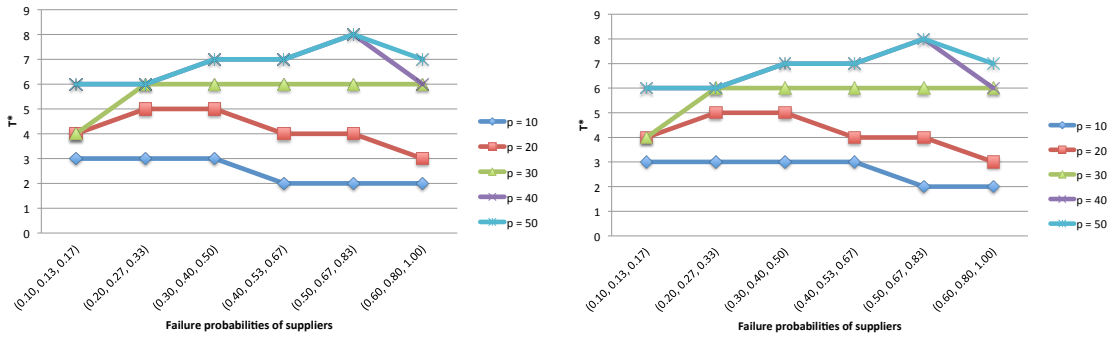
(d) $\rho = 0.5$, $q = (0.3, 0.4, 0.5)$

Figure 2.5: Expected Total Cost v.s. Number of Links

p, which is consistent with our intuition that T^* is increasing in p. Again, T^* has the same value no matter demands are independent or not, except for the instance when $p = 10$ and $q = (0.4, 0.53, 0.67)$. Therefore, positive correlation between demands has little impact on T^*.

Algorithm Test

It has been shown in the last study that a capacitated supply chain's expected total cost is robust against the variability of T in a neighborhood of T^*, and there are simple relationships between T^*, and failure probabilities and the penalty cost. Thus, it is possible to find a T that generates close-to-optimal expected total cost using some heuristics. Suppose T^* is given, we only need an algorithm to determine where to locate these links. In this study, extensive simulations are carried out to test how Algorithm 2 works with limited supply capacity. Our results show that Algorithm 2 provides reasonably good solutions for capacitated supply chains under weak conditions. We demonstrate its performance with the following numerical examples.



(a) $\rho = 0$ (b) $\rho = 0.5$

Figure 2.6: Optimal Number of Links v.s. Failure Probabilities in Unbalanced Nonhomogeneous Supply Chains with $\mu = (40, 30, 20, 10)$

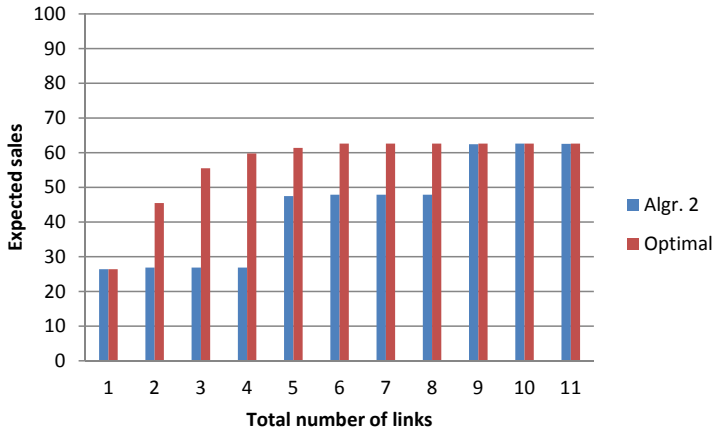
Consider a 3-by-4 system with $\mu = (40, 30, 20, 10)$, $q = (0.1, 0.3, 0.5)$, and $\rho = 0$. Let suppliers' capacities be (30, 30, 30), (50, 50, 50) and (80, 80, 80). These three capacity portfolios represent the cases that the expected total capacity is (1) less than, (2) roughly equal to, and (3) greater than the expected total demand, respectively. T increases from 1 to 11 with increment 1 ($T = 0$ or 12 is not studied because the problem has unique feasible solution in these cases). For each instance, the optimal link structure is identified by enumerating all possible structures and choose the one that achieves the maximum average sales in simulation. As a byproduct of simulation, the expected sales when using the link structure given by Algorithm 2 are also estimated.

Figure 2.7 compares the expected sales using optimal link configurations and those using link configurations given by Algorithm 2. In both figures, Algorithm 2 configurations are dominated by the optimal ones. The difference between the two sets of performance is substantial when $C = (30, 30, 30)$, and is almost negligible when $C = (80, 80, 80)$. This result is not surprising – since the difference is caused by limited capacities, it must go to zero as capacities become sufficient. Therefore, the link configurations given by Algorithm 2 may perform almost as good as the optimal ones when supply capacity is large enough.

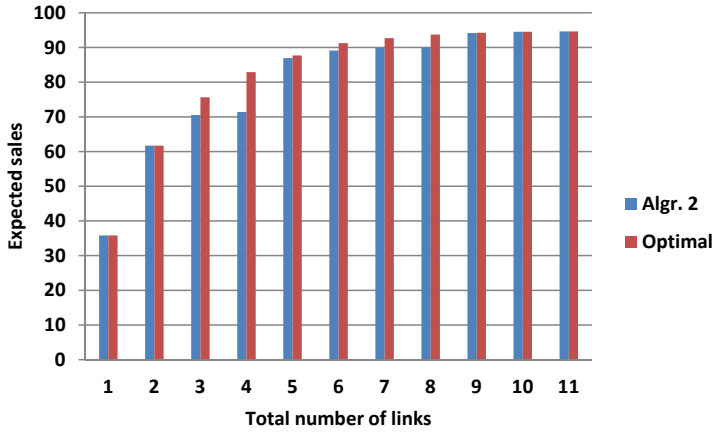
To gain more insights, we calculate the relative error in expected sales, denoted by er, which is given by

$$er = \frac{\text{optimal average sales} - \text{average sales using Algr.2 solutions}}{\text{optimal average sales}} \times 100\%.$$

Figure 2.8 shows er with different T and C. As shown in Figure 2.7(a) and Figure 2.7(b), er decreases as C increases – it can be over 55% when $C = (30, 30, 30)$ while it is within 41% and 14% when $C = (50, 50, 50)$ and $(80, 80, 80)$, respectively. Although er is obviously not



(a) $C = (30, 30, 30)$



(b) $C = (80, 80, 80)$

Figure 2.7: Performance Comparison of Algorithm 2 Solutions and Optimal Solutions with Different Values of T and C

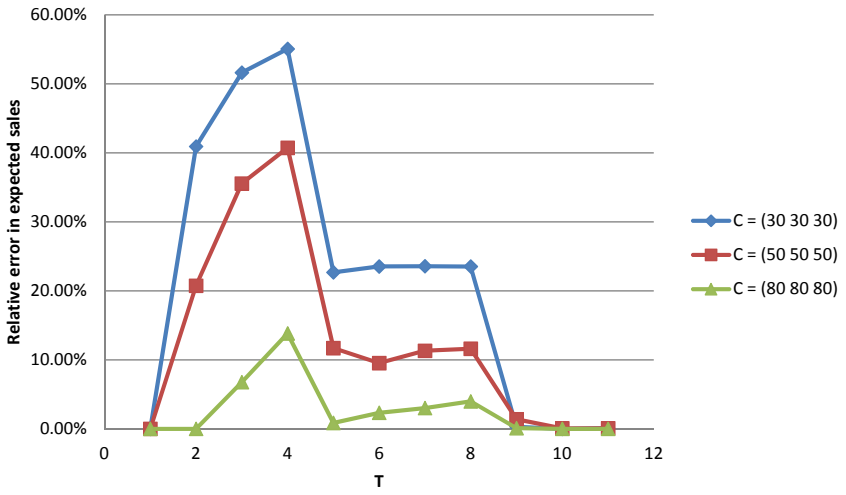


Figure 2.8: Relative Error in Expected Sales when Applying Algorithm 2 to Capacitated Supply Chains

monotone with respect to T, the general trend is that er is small when T is large, except for $T = 1$. When T is restricted to be greater than 4, the number of retailers, the performance of Algorithm 2 configurations can be greatly improved. For instance, er is within 12% instead of 41% when $C = (50, 50, 50)$ and $T > 4$. It is reasonable to believe that Algorithm 2 provides reasonably good solutions for capacitated supply chains when $T^* > M$, and when the expected total capacity of the supply chain is close to the expected total demand.

Supply chain	μ	q	ρ
A1	$(40, 30, 20, 10)$	$(0.1, 0.2, 0.3)$	0
A2	$(40, 30, 20, 10)$	$(0.1, 0.2, 0.3)$	0.5
B1	$(40, 30, 20, 10)$	$(0.1, 0.3, 0.5)$	0
B2	$(40, 30, 20, 10)$	$(0.1, 0.3, 0.5)$	0.5
C1	$(40, 30, 20, 10)$	$(0.7, 0.8, 0.9)$	0
C2	$(40, 30, 20, 10)$	$(0.7, 0.8, 0.9)$	0.5
D1	$(1000, 100, 10, 1)$	$(0.1, 0.2, 0.3)$	0
D2	$(1000, 100, 10, 1)$	$(0.1, 0.2, 0.3)$	0.5

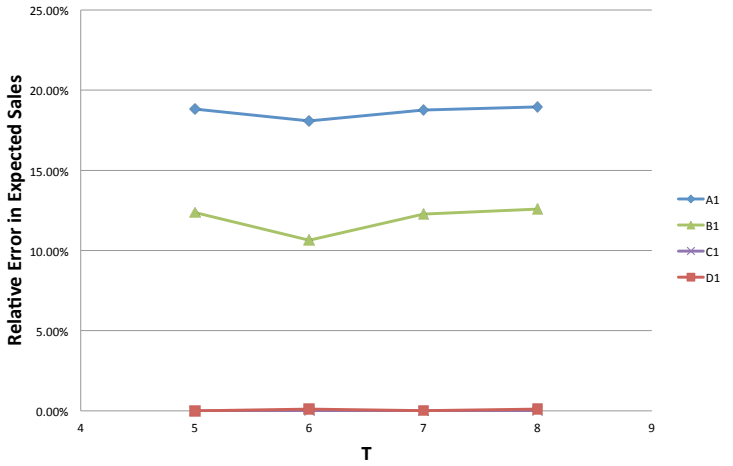
Table 2.1: Typical Supply Chains and Their Specifications

To better evaluate Algorithm 2's performance when supply capacity is limited, we test it in several more supply chains while keeping the expected total capacity equal to the expected total demand and increasing T from 5 to 8 with increment 1. Those T values greater than 8 are not tested because the relative error is negligible in these cases according to Figure

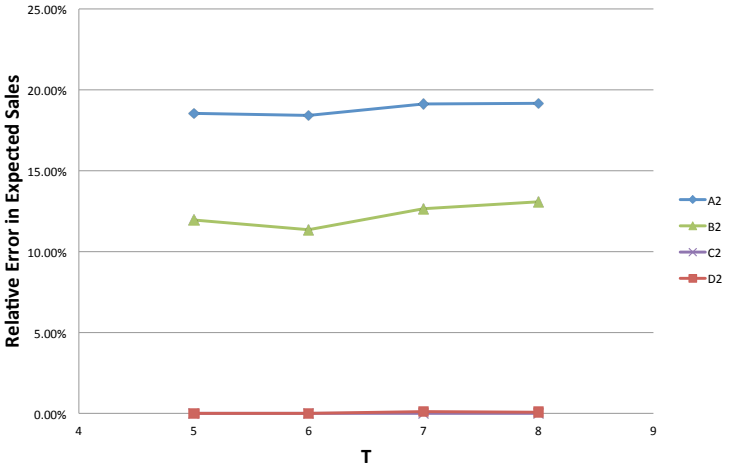
2.8. Table 2.1 summarizes the specifications of the supply chains considered here. They are separated into two groups: group *1* (*A1*, *B1*, *C1*, *D1*) have independent demands, while group *2* (*A2*, *B2*, *C2*, *D2*) have positive correlated demands with correlation coefficient 0.5. Within each group, we consider the four cases: (A) similar demands, and small and similar failure probabilities, (B) similar demands, and small and very different failure probabilities, (C) similar demands, and large and similar failure probabilities, and (D) very different demands, and small and similar failure probabilities. The result is described in fig:errors. We notice that Algorithm 2 performs almost the same no matter demand is independent or positively correlated. We also find that the relative error is relatively high ($\approx 18\%$) in *A1*, *A2*, *B1* and *B2*, and is almost zero in *C1*, *D1*, *C2*, and *D2*. This is because that Algorithm 2 tends to assign too many retailers to the most reliable suppliers when demands are close to each other, and failure probabilities are small and/or very different. When this happens, the demand allocated to the most reliable suppliers outweighs their capacity a lot, leading to more lost sale. Therefore, Algorithm 2 works better for capacitated supply chains when demands are significantly different from each other, and suppliers are frequently disrupted.

Another observation from Figure 2.8 is that the value of *er* drops dramatically at $T = 5$ and $T = 9$. This is because of the introduce of new suppliers. In the link configuration given by Algorithm 2, only supplier 1 is used when $T \leq 4$, supplier 2 is also used when $5 \leq T \leq 8$, and all three suppliers are engaged when $T \geq 9$. So the total capacity of the system raises by 78% when T increases from 4 to 5, and raises by 31% when T increases from 8 to 9. These sudden increases in capacity cause the sudden improvements in the performance of Algorithm 2 configurations. This reminds us about the importance of matching demand and capacity. More precisely, we should (1) *try to match the expected capacity at each supplier and the expected total demand it faces*, and (2) *try to match the expected demand at each retailer and the expected capacity of its suppliers*. These principles are general versions of those proposed by [37]. This idea can be incorporated into Algorithm 2 to develop an heuristic approach to solve capacitated problems. In each iteration, we find the supplier with the largest expected remaining capacity, and connect it to the retailer with the largest expected unmet demand. This algorithm is greedy and not necessarily optimal. We will study its performance in the future.

These studies on selection of T^* together with the sensitivity analysis of expected total cost over T lead to an idea of solving capacitated nonhomogeneous problems heuristically. If we can narrow down the range of possible values of T^* based on q, p, and C, and apply Algorithm 2 with a T properly chosen from that range, then the solution's performance could be very close to optimal unless C is very limited.



(a) $\rho = 0$



(b) $\rho = 0.5$

Figure 2.9: Relative Error in Expected Sales when Applying Algorithm 2 to Capacitated Supply Chains

Interaction between Flexibility and Supply Capacity

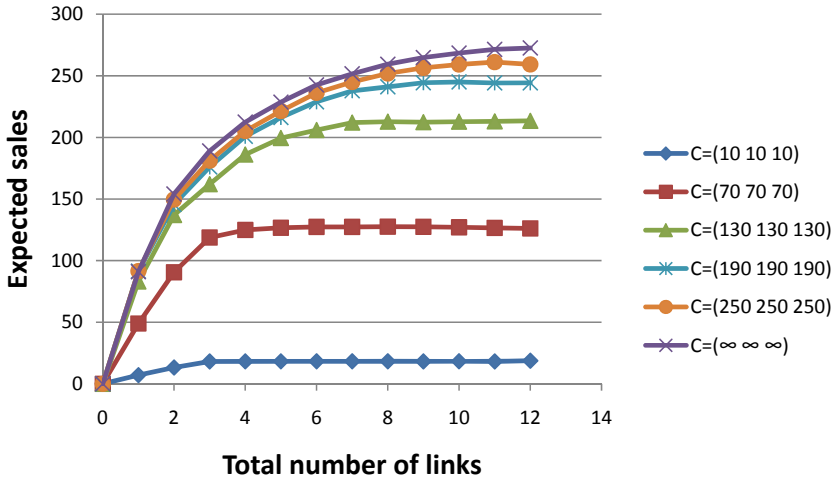
It is demonstrated in [37] that the value of flexibility may depend on supply capacity. They conclude that flexibility is most beneficial when total capacity is equal to total demand, but has little or even no value when capacity is very small or very large. We undertake a similar study on our model, and find some interesting results.

We look at a 3-by-4 supply chain with $\mu = (130, 90, 50, 20)$ and $q = (0.3, 0.4, 0.5)$, and increase T from 0 to 12, C from $(10, 10, 10)$ with step size $(60, 60, 60)$. From Figure 2.10(a), we see that the expected sales are increasing and concave in T regardless of the value of C ([1] prove this result for homogeneous supply chains). This is a strong evidence that Proposition 9 holds in more general settings. We also note that adding additional links into a system that already has 3 or more links cause no improvement when $C = (10, 10, 10)$. This comes from the same reason as [37] argue, "If each plant's capacity is less than the minimum possible demand for each product, each plant is fully utilized under any possible demand". Nevertheless, we observe the opposite result when C is sufficiently large. In Figure 2.10(a), the curve corresponding to $C = (\infty, \infty, \infty)$ represents the case that the capacity of each supplier is no less than the maximum total demand (435). We see that the expected sales is strictly increasing in T in this case, which suggesting that adding flexibility is beneficial even when capacity is arbitrarily large. This phenomenon reveals that *adding flexibility could NOT be fully substituted by increasing supply capacities.*

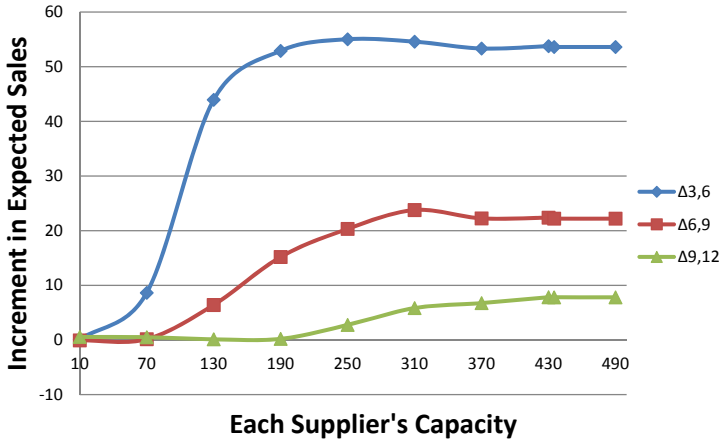
To see how the benefits of flexibility changes with respect to capacity more clearly, we plot the increment in expected sales when adding flexibility in Figure 2.10(b) as a function of each supplier's capacity. The curves $\Delta_{3,6}$, $\Delta_{6,9}$ and $\Delta_{9,12}$ represent the increases in expected sales when T increases from 3 to 6, 6 to 9 and 9 to 12, respectively. When each supplier's capacity is less than 70, extra flexibility after adding 6 links has no value, when each supplier's capacity is between 70 and 190, the expected sales of the supply chain can be increased by adding 3 more links, and when each supplier's capacity is greater than 190, the expected sales can be further raised by making the supply chain fully flexible. Therefore, rather than weakening the effect of flexibility, adding capacity actually magnifies the benefits of flexibility.

2.7 Summary

In this chapter, we look at a flexibility design problem in a capacitated nonhomogeneous supply chain with disruptions. Two types of cost are considered: penalty cost for lost sale and the cost of link maintenance. Our objective is to find the network configuration that minimizes the expected total cost. This problem is formulated as a two-stage mixed integer stochastic program which is hard to solve for exact optimal solutions. Since the primary interest of our work is to investigate the properties of optimal network configurations when



(a)



(b)

Figure 2.10: Expected Sales v.s. T under Different Values of C

supply chains have unbalanced nonhomogeneous structures and are exposed to disruptions, we relax the capacity constraints for a while to make the model solvable while keeping those key factors. The maintenance cost of links is proportional to the total number of links. This enables us to decompose the problem by fixing the number of links at T, and try to maximize the expected sales. A myopic algorithm (Algorithm 2) is proposed to solve the subproblem and is proved to be global optimal as well. The complexity of the algorithm is $O(T)$. Based on Algorithm 2, an efficient algorithm (Algorithm 3) is proposed for solving the uncapacitated model. Next, we show that a problem with both supply disruptions and link disruptions can always be transformed to an equivalent one with supply disruptions only, and thus can be solved by Algorithm 3. For capacitated problems, we prove that the marginal benefit of adding flexibility or increasing supply capacity is decreasing.

We gain some major insights from numerical experiments. First of all, we find that the expected total cost of the supply chain is convex in T, so the system performance is robust in a neighborhood of T^*. Moreover, T^* appears to be quasi-concave in q, and increasing in p. These provide some guidelines on identifying T^*. Secondly, we show that Algorithm 2 may give good solutions for capacitated supply chains, if the expected total capacity is roughly equal to or greater than the expected total demand and if T is greater than the number of retailers. Lastly, In the investigation of the interaction between capacity and flexibility, we observe decreasing marginal benefits of flexibility in a capacitated supply chain with supply disruptions. Additional flexibility is valueless when supply capacity is very small. Nevertheless, when suppliers are unreliable, adding flexibility is encouraged when supply capacity is large.

We identify several possible directions for future research. Firstly, the ideas obtained from numerical studies can be incorporated into Algorithm 2 to extend it to a heuristic algorithm for capacitated problems. Secondly, our model can be enriched by generalizing the cost structure. Instead of assuming each link incurs the same maintenance cost, we can allow the costs to be nonhomogeneous and/or degree-dependent. Lastly, besides the expected value of sales, its variance is also an important measure of supply chain performance. We can bring variance in to our model by either imposing a hard constraint or adopting the CVaR model.

Chapter 3

Flexible Hub Location Model for Air Transportation Networks

3.1 Introduction and Literature Review

From late October to early November, 2012, Hurricane Sandy forced the closure of major airports all over the northeast coast of the United States, and flights around the world from or to those airports were canceled and huge losses had incurred to airlines. It has been noticed that in addition to paying for meals and accommodations for the travelers who got stuck in airports, airlines have also reported significant losses towards canceling flights that use northeast coast airports as intermediate stops. Ideally, their losses could have been mitigated if they were able to reroute some of their flights through operating airports such as Logan International Airport at Boston. This story gives rise to the following research questions: how should we design an air transportation network that possesses necessary flexibility to hedge against the risk of disruptions at airports, what is the associated benefit and costs, and how to operate and plan the routes in a flexible network. In the remainder of this chapter, we focus on the design of a flexible air transportation network, and answer these research questions. However, we should point out that the same design methodology applies not only for the air transportation problem, but also for the general hub-and-spoke systems, which is also known as hub location systems.

Hub location models have important applications in industry companies such as airlines, logistic companies, and telecommunication firms. A common feature of these applications is that there exists demand of either the movement of physical goods or the usage of links from origins to destinations. To satisfy the demand, it is necessary to build connections for each of the origin-destination (OD) pairs. The OD pairs can be either connected directly, or connected via a series of links. Then, routing decisions need to be made, based on the network topology. Usually, there are costs associated with building links, and there are also transportation costs as functions of demand as well as route lengths for the flows between

each of the OD pairs. The objective can be various, and the majority of existing research sets minimizing costs as the objective.

[16] publish a review paper to celebrate the 25th anniversary of two seminal transportation hub location models ([53] and [52]). The paper gives an insightful review of the origins of hub-location problems, and comments on the status quo of related research. It can be observed from [16] that the four most unique features of hub location models are the following: (1) demand is characterized by OD pairs, (2) there is benefit of routing via hubs, (3) a route for any of the OD pairs can pass at most two hubs, and (4) there is no direct connection between spokes. The first feature distinguishes hub location models from facility location models, in which demand is characterized by the node of interests. The second feature provides incentives for using hubs. In fact, the reason for employing a hub-and-spoke system rather than point-to-point systems has three folds: hub-and-spoke systems exert economy of scale by consolidating flows at hubs and save operating costs; hub-and-spoke systems have fewer links than the point-to-point systems; hub-and-spoke systems also help to increase the reliability of the network ([41]).

The last two features of hub location models are made by [15]. The third one actually comes from the assumptions of fully connectivity between hubs and the triangle inequality. Specifically, suppose that a route passes three hubs, then by the triangle inequality, the costs associated with transportation can be decreased by using the link that connects the first and the last hubs in the original route. The last feature is justified based on the removal of flows that are large enough to be routed directly. There exists work that treats the flow more sophisticated that adding links between spokes is necessary, see for example [3], however in this paper, we follow the more general setting and only focus on demand flows that are beneficial to be transported via hubs.

In a nutshell, given their benefits and unique features, hub location models aim to solve for the optimal hub locations, assignment of spokes to hubs, and the routes for flows between each of the OD pairs. Here, it is worth mentioning the categorization of hub location models based on assignment assumptions. If each spoke is allowed to be connected to only one hub, then the corresponding model belongs to the category of *Single Allocation* models. For single allocation models, all routes starting from the same origin share the same first hub visited. On the other hand, if there is no limit on the number of hubs each spoke can be connected to, then the corresponding model is a *Multiple Allocation* one. For multiple allocation models, routes staring from the same origin can have different first hub visited, depending on the destinations. In this chapter, we adopt a variant of multiple allocation models.

Hub location models can be viewed as a hybrid of facility location models and network flow models. Selecting hub locations and assignment is analogous to facility location problems, while choosing the optimal routings is a network flow problem. However, decisions are made jointly, which makes hub location models more difficult to solve compared to the other

two problems. Nonetheless, the connections, especially that with facility location models, inspire new research of hub location problems. For example, there are four fundamental facility location problems: p-center, p-median, set covering, and uncapacitated facility location problems, similarly, [15] formulate four fundamental hub location models: p-hub center, p-hub median, hub covering, and uncapacitated hub location models. The fundamental hub location models are more complicated than their facility location counterparts. For instance, [51] presents an integer formulation for the single allocation p-hub median problem, using assignment variables $Z_{ij} = 1$ if spoke i is allocate to hub j. Although the dimension of decision variables is similar to that of the p-median problem, the formulation is quadratic. Meanwhile, [46] presents the multiple allocation model for p-hub median problem, in which decision variables take form of X_{ijkm}, representing the fraction of flow through path $(i - k - m - j)$ for the OD pair (i, j), and thus the dimension of decision variables is much greater. Other fundamental hub location problems can be formulated in similar ways, see for example [15], [28], and [61].

Hub location problems have received much attention in the past quarter century. Different model formulations and various solution approaches have been studied. Here we briefly summarize the work that is more related to this chapter. The first quantitative model for hub location problems appears in 1986 ([53], [52], and [51]). [51] presents the first quantitative analysis on the p-hub median location problem using Civil Aeronautics Board (CAB) data set, which has then become the standard test-bed for hub location models. Later on, motivated by the desire to solve larger instances, [74] and [75] formulate single allocation and multiple allocation p-hub median models based on the idea of multi-commodity flow, and manage to reduce the size of the formulations to solve problems with up to 200 nodes. Since then, one important thread of research extension focuses on solution approaches for the fundamental problems and their variations, see for example [47], [13], [27], [67], and [57]. [28] and [61] provide comprehensive surveys of modeling techniques and solution methods. Nonetheless, as noted by [16], hub location problems usually solve for strategic level decisions, and thus the fast CPU time is less important. Another thread of research focuses on extending hub location models by incorporating more realistic assumptions. For instance, the discount in operating costs between hubs is used to be treated as independent of flow quantities. [23], [54], and [12] formulate new models in which unit transportation costs are flow-dependent. Similar extensions include considering capacity limits ([73]), adding service level constraints ([60], [14]), extending to dynamic hub location problems ([25], [35]), and modeling competitive hub location problems ([79]).

Another important extension is modeling the reliability of hubs. Almost all hub location research assumes that hubs will work as planned. However, as in our motivating example, severe weather, natural disasters, terrorist attacks, and labor strikes can all cause temporary unavailability of hubs. Hub-and-spoke systems are more vulnerable to disruptions than point-to-point systems. In a point-to-point system, one failed node affects only flows from or to that node, and one failed link affects only the flow between one OD pair. But in a

hub-and-spoke system, one failed hub affects flows from or to all its spokes. In other words, optimal hub locations and assignment obtained by assuming hubs are perfectly reliable may contain hubs that have high failure rates, and thus the corresponding system may have high "failure costs".

The current practice in the air transportation system of the United States is to assign a nominal route to each flight and to reroute it if the intermediate hubs are disrupted. Rerouting decisions are made primarily according the National Playbook which contains alternative routes for common scenarios. These routes are mostly obtained from historical experience and are not necessarily optimal. In addition, they are lack of flexibility and may cause congestions at hubs. Therefore, we need a joint optimization model that considers the outcome in every scenario when making strategic level decisions. The main idea of our model is to designe a flexible hub-and-spoke network to hedge against the risk of hub disruptions. However, flexibility comes at the costs of additional link expenses. Therefore, we seek for the optimal trade-off between flexibility and link costs.

In the facility location community, there have been growing interests in facility location models under disruptions. Specifically, both hubs and links are possible to fail. The first reliable facility model was introduced by [70]. They study the stochastic p-median problem and assume that all nodes have equal failure probabilities. The assumption of uniform failure probabilities is relaxed in [66], and several heuristic solution algorithms are provided. Other work considering facility disruptions includes [58], [59], and [50].

Unlike the fast growing situation of reliable facility location research, there is only a few papers study hub location problems under hub disruptions. The main difficulty rises when considering backup hubs, as for each of the OD pairs in a p-hub system, if one of the hubs on the designated route fails, there are $(p-1)^2$ alternate routes to choose from, and choosing the hub locations is even harder after taking into consideration of backup hubs and alternative routes. [39] is the first to study reliable hub location problems. In [39], both single allocation and multiple allocation models are formulated, and the objective is to maximize the expected network flow, in the absence of backup hubs and alternative routes. [81] also formulate both single and multiple allocation p-hub median location problems, and the objective is to minimize the expected total operating costs. The paper makes several assumptions to enable a Lagrangian relaxation solution approach: (1) it is assumed that no more than one hub will be disrupted at any time, (2) the model does not consider loss of flow due to the failures of spokes. However, these assumptions fail if the failure probabilities are large or positively correlated. We relax these assumptions in our model.

This chapter contributes to the literature of hub location problems in the following aspects: (1) to the best knowledge of the authors, this chapter is the first that studies the flexible capacitated hub location problem, which deals with correlated airport disruptions, (2) the model proposed in this chapter adopts a variation of multiple allocation to allow

more flexible routing decisions, (3) we reveal that a hub-and-spoke structure in which each spoke are assigned with 2 hubs can perform almost as good as a fully flexible one in which each spoke are assigned with all hubs, and (4) principles on deciding hub locations and assignment are proposed to serve as guidelines on the design of networks with disruptions.

The remainder of the chapter is organized as follows. Section 3.2 presents our model formulation and describes the solution method. In Section 3.3, we carry out a series of numerical studies to evaluate the benefit of flexibility, to investigate the effects of capacity and disruption correlation, and to get insights on the design of networks with disruptions. We conclude in Section 3.4 and point out several future directions.

3.2 Model

We look at the design problem of an air transportation system with airports $i \in I$. The hub-and-spoke framework is adopted. We follow the following assumptions in the traditional hub location models. The transportation cost matrix is symmetric. That is, the transportation cost from airport i to airport j is the same as that from j to i. Based on this assumption, we define the traveling demand between i and j as the summation of the traveling demand in both directions. The traveling demand between each pair of airports is deterministic and constant. Each airport has a finite capacity and incurs a fixed cost if it is selected as a hub. The subgraph of hubs is complete. Transportation costs in hub-to-hub routes are discounted by a factor $\beta \in (0, 1)$.

Unlike in traditional hub location problems, airports are not reliable in our model. Disruptions may happen at airports with respective probabilities, and the state of one airport may be dependent on those of the others. When an airport is disrupted, no passenger can pass, leave from, or arrive at it. Penalty cost is incurred whenever traveling demand is lost. To cope with hub disruptions, we allow each spoke airport to have at most N hubs, and to choose which of these hubs to use in each scenario. We refer to this network structure as an *N-flexible hub-and-spoke* structure. To measure the increment in operational cost at hubs due to added flexibility, we incorporate a variable cost into the cost of each hub, which is referred to as the *flexibility* cost. Define the *degree of flexibility* of a hub as the number of spokes connected to it. We assume that the flexibility cost of a hub is proportional to its degree of flexibility.

The goal of our model is to determine the hub locations and assignment in order to minimize the expected total cost of the air transportation network, which consists of the fixed cost, the flexibility cost, the transportation cost, and the penalty cost. We choose to use a scenario-based model in order to characterize the correlation between airport disruptions[1].

[1]In the rest of this chapter, we use 'correlation between airports' for 'correlation between airport disrup-

The total number of possible scenarios can be huge in practice, but when the failure probabilities of airports are small, we can limit the number of scenarios by only including those with substantial probabilities. This will be further demonstrated in numerical studies.

Our model differs from a similar work, [81], in several ways. First of all, our model is more flexible in both network structure and operations. Instead of restricting the number of alternative hubs of each airport (or OD pair) by 2, we allow each airport to have up to N hubs, for $N = 1, 2, \ldots$. In this way, our model is also able to represent a wide range of network structures, from the traditional hub-and-spoke structure with single allocations ($N = 1$) to a fully flexible hub-and-spoke structure ($N = I$). In the operational level, [81] assumes that airports need to use hubs in the pre-determined primary-backup order. In our model, airports choose which hubs to use and decide how much flow to transport via each hub in each scenario. Thus, each OD pair is able to adjust its routes according to the availability of hubs. Secondly, our model captures the correlation between airports, which is fairly common in practice. It is shown in numerical studies that airport correlation may have great impact on the network structure. Thirdly, our model considers the capacity limits at airports. Lots of literature on flexibility has proved that capacity affects the value of flexibility, and flexibility in turn changes the capacity utilization rate. Lastly, we use fixed cost to control the total number of hubs instead of fixing it at a constant P. Actually, as we will show, flexibility may reduce the number of hubs needed.

Formulation

The notation used in our model are summarized below.
Parameters:

- I: set of airports.
- S: set of scenarios.
- f_i = fixed cost of a hub at i, for $i \in I$.
- w_i = unit flexibility cost of a hub at i, for $i \in I$.
- c_i = capacity of airport i, for $i \in I$.
- d_{ij} = unit traveling cost of route $i - j$, for $i, j \in I$, $i < j$.
- h_{ij} = traveling demand[2] between airport i and j, for $i, j \in I$, $i < j$.
- p_{ij} = unit penalty cost for the loss of traveling demand between i and j, for $i, j \in I$, $i < j$.

tions'.

[2]Denoted as demand in the rest of this chapter.

- q^s = probability of scenario s, for $s \in S$.
- a_i^s: indicator of the availability of airport i in scenario s, 1 if available, and 0 otherwise, for $i \in I$, $s \in S$.
- β = discount factor of hub-to-hub transportation cost.

Decision variables:

- $X_i = \begin{cases} 1 & \text{if } i \text{ is a hub} \\ 0 & \text{otherwise} \end{cases}$ for $i \in I$.
- $Y_{ji} = \begin{cases} 1 & \text{if hub } i \text{ is assigned to } j \\ 0 & \text{otherwise} \end{cases}$ for $i, j \in I$.
- F_{ijkl}^s = fraction of h_{ij} transported via route $i - k - l - j$ in scenario s, for $i < j, \quad i, j, k, l \in I, \quad s \in S$.

Note that d_{ij}, h_{ij}, p_{ij}, and F^s_{ijkl} are defined only for $i < j$ to avoid double counting. Then the flexible hub location problem is formulated as the following MIP.

$$
\begin{aligned}
(P) \quad \min \quad & \sum_i f_i X_i + \sum_i \sum_j w_i Y_{ji} + \sum_s q^s \left(\sum_i \sum_{j>i} \sum_k \sum_l h_{ij}(d_{ik} + \beta d_{kl} + d_{kj}) F^s_{ijkl} \right) \\
& + \sum_s q^s \left(\sum_i \sum_{j>i} h_{ij} p_{ij} \left(1 - \sum_k \sum_l F^s_{ijkl} \right) \right) && \text{(3.1a)} \\
s.t. \quad & Y_{ji} \le X_i \qquad i \ne j, \quad i,j \in I && \text{(3.1b)} \\
& Y_{ii} = X_i \qquad i \in I && \text{(3.1c)} \\
& \sum_i Y_{ji} + (N-1) X_j \le N \qquad j \in I && \text{(3.1d)} \\
& \sum_k \sum_l F^s_{ijkl} \le 1 \qquad i < j, \quad i,j \in I, \quad s \in S && \text{(3.1e)} \\
& \sum_{j<i} \sum_k \sum_l h_{ji} F^s_{jikl} + \sum_{j>i} \sum_k \sum_l h_{ij} F^s_{ijkl} \\
& + \sum_{k \ne i} \sum_{\substack{l \ne i \\ l > k}} h_{kl} \left(\sum_{j \ne i} (F^s_{klij} + F^s_{klji}) + F^s_{klii} \right) \le c_i \qquad i \in I, \quad s \in S && \text{(3.1f)} \\
& \sum_l F^s_{ijkl} \le a^s_i a^s_j a^s_k Y_{ik} \qquad i < j, \quad i,j,k \in I, \quad s \in S && \text{(3.1g)} \\
& \sum_k F^s_{ijkl} \le a^s_i a^s_j a^s_l Y_{jl} \qquad i < j, \quad i,j,l \in I, \quad s \in S && \text{(3.1h)} \\
& X_i \in \{0,1\} \qquad i \in I && \text{(3.1i)} \\
& Y_{ji} \in \{0,1\} \qquad i,j \in I && \text{(3.1j)} \\
& F^s_{ijkl} \ge 0 \qquad i < j, \quad i,j,k,l \in I, \quad s \in S && \text{(3.1k)}
\end{aligned}
$$

(3.1a) minimizes the expected total cost consisting of the fixed cost, the flexibility cost, the transportation cost, and the penalty cost. (3.1b) ensures that an airport can be assigned to other airports only if it is a hub. (3.1c) forces that a hub must be assigned to itself. (3.1d) is the flexibility constraint. It restricts the total number of hubs assigned to an airport by N if it is a spoke, and by 1 if it is a hub. (3.1e), (3.1f), (3.1g) and (3.1h) guarantee the feasibility of flows in each scenario. (3.1e) ensures F^s_{ijkl}'s are fractions. (3.1f) enforces that the total flow going through an airport cannot exceed its capacity. (3.1g) and (3.1h) ensures that there is a positive flow in a route only if the route is built and no airport in it is disrupted.

Define $u_{ijkl} = h_{ij}(d_{ik} + \beta d_{kl} + d_{lj} - p_{ij})$, then (3.1a) can be rewritten as

$$
\min \quad \sum_i f_i X_i + \sum_i \sum_j w_i Y_{ji} + \sum_s q^s \left(\sum_i \sum_{j>i} \sum_k \sum_l u_{ijkl} F^s_{ijkl} + \sum_i \sum_{j>i} h_{ij} p_{ij} \right) \quad \text{(3.2)}
$$

Then we have the following theorem on the global sensitivity of the problem.

Theorem 12. *The minimum expected total cost is concave in f, w, and q.*

Our numerical studies show that the minimum expected total cost is convexly decreasing in N, but is not necessarily convex in c.

The formulation (P) has large numbers of decision variables ($O(SI^4)$) and constraints ($O(SI^3)$), so it is not easy to be solved using commercial solvers. We apply a Benders decomposition algorithm to solve it. The general idea of Benders decomposition is to partition the original problem into a restricted master problem and a series of subproblems. By solving the subproblems, we generate Benders cuts that guarantee either the feasibility of subproblems or the optimality of the master problem, and iteratively add them into the restricted master problem. The algorithm terminates when no more cuts can be found. The details of the Benders decomposition algorithm is described in next section.

Benders Decomposition Algorithm and Pareto-Optimal Cuts

In our problem, the restricted master problem decides hub locations and assignment, and subproblems solve the transportation problems in all scenarios. When hub locations X and assignment Y are fixed at x and y, respectively, the remaining problem can be decomposed by scenario. The subproblem of scenario s is as follow.

$$(SP^s) \quad z^s(x,y) = \min \quad \sum_i \sum_{j>i} \sum_k \sum_l u_{ijkl} F^s_{ijkl} + \sum_i \sum_{j>i} h_{ij} p_{ij} \tag{3.3a}$$

$$s.t. \quad \sum_k \sum_l F^s_{ijkl} \le 1 \qquad i<j, \quad i,j \in I \tag{3.3b}$$

$$\sum_{j<i} \sum_k \sum_l h_{ji} F^s_{jikl} + \sum_{j>i} \sum_k \sum_l h_{ij} F^s_{ijkl} + \sum_{k\neq i} \sum_{\substack{l\neq i \\ l>k}} h_{kl} \left(\sum_{j\neq i} (F^s_{klij} + F^s_{klji}) + F^s_{klii} \right) \le c_i \quad i \in I \tag{3.3c}$$

$$\sum_l F^s_{ijkl} \le a^s_i a^s_j a^s_k y_{ik} \qquad i<j, \quad i,j,k \in I \tag{3.3d}$$

$$\sum_k F^s_{ijkl} \le a^s_i a^s_j a^s_l y_{jl} \qquad i<j, \quad i,j,l \in I \tag{3.3e}$$

$$F^s_{ijkl} \ge 0 \qquad i<j, \quad i,j,k,l \in I \tag{3.3f}$$

(SP^s) decides how much flow to transport via each route subject to the availability of airports and hub capacities. $z^s(x,y)$ is the minimum transportation cost and penalty cost in scenario s given hub locations x and assignment y. Let λ, τ, μ, π be the dual variables associated

with constraints (3.3b), (3.3c), (3.3d), and (3.3e), respectively. The dual problem of (SP^s) is

$$
\begin{aligned}
(DSP^s) \quad z^s(x,y) = \max \quad & \sum_i \sum_{j>i} \lambda_{ij}^s + \sum_i c_i \tau_i^s + \sum_i \sum_{j>i} \sum_k a_i^s a_j^s a_k^s y_{ik} \mu_{ijk}^s \\
& + \sum_i \sum_{j>i} \sum_k a_i^s a_j^s a_k^s y_{jk} \pi_{ijk}^s + \sum_i \sum_{j>i} h_{ij} p_{ij} \\
s.t. \quad & \lambda_{ij}^s + h_{ij}\tau_j^s + h_{ij}\tau_i^s + \mu_{ijk}^s + \pi_{ijl}^s \le u_{ijkl} \\
& \qquad i,j,k,l \in I \text{ such that } i<j, \text{ and condition (a)} \quad (3.4a) \\
& \lambda_{ij}^s + h_{ij}\tau_j^s + h_{ij}\tau_i^s + \mu_{ijk}^s + \pi_{ijl}^s + h_{ij}\tau_k^s \le u_{ijkl} \\
& \qquad i,j,k,l \in I \text{ such that } i<j, \text{ and condition (b)} \quad (3.4b) \\
& \lambda_{ij}^s + h_{ij}\tau_j^s + h_{ij}\tau_i^s + \mu_{ijk}^s + \pi_{ijl}^s + h_{ij}\tau_l^s \le u_{ijkl} \\
& \qquad i,j,k,l \in I \text{ such that } i<j, \text{ and condition (c)} \quad (3.4c) \\
& \lambda_{ij}^s + h_{ij}\tau_j^s + h_{ij}\tau_i^s + \mu_{ijk}^s + \pi_{ijl}^s + h_{ij}\tau_k^s + h_{ij}\tau_l^s \le u_{ijkl} \\
& \qquad i,j,k,l \in I \text{ such that } i<j, \text{ and condition (d)} \quad (3.4d) \\
& \lambda_{ij}^s + h_{ij}\tau_j^s + h_{ij}\tau_i^s + \mu_{ijk}^s + \pi_{ijk}^s + h_{ij}\tau_k^s \le u_{ijkl} \\
& \qquad i,j,k,l \in I \text{ such that } i<j, \text{ and condition (e)} \quad (3.4e) \\
& \lambda, \tau, \mu, \pi \le 0 \quad (3.4f)
\end{aligned}
$$

where condition (a) is $i = k$ or $j = k$, and $i = l$ or $j = l$, condition (b) is $i \neq k$, $j \neq k$, and $i = l$ or $j = l$, condition (c) is $i \neq l$, $j \neq l$, and $i = k$ or $j = k$, condision (d) is $i \neq k$, $j \neq k$, $i \neq l$, $j \neq l$, and $k \neq l$, and condition (e) is $i \neq k$ and $j \neq k$.

Define η^s as the decision variable representing the sum of transportation cost and penalty cost in scenario $s \in S$. Then the restricted master problem is

$$
\begin{aligned}
(RMP) \quad \min \quad & \sum_i f_i X_i + \sum_i \sum_j w_i Y_{ji} + \sum_s q^s \eta^s \quad (3.5a) \\
s.t. \quad & (3.1b), (3.1c), (3.1d), (3.1i), (3.1j) \quad (3.5b)
\end{aligned}
$$

The connection between η^s and $z^s(X,Y)$ is temporarily broken, so feasibility cuts and optimality cuts need to be added in order to force them equal to each other. Note that (SP^s) is always feasible and (DSP^s) is always bounded, so only optimality cuts will be added. Let $(\bar{\lambda}^{(t)}, \bar{\tau}^{(t)}, \bar{\mu}^{(t)}, \bar{\pi}^{(t)})$, $t = 1, \ldots, T$, be the extreme points of the feasible region of (DSP^s), then the optimality cuts generated by solving (DSP^s) are

$$
\eta^s \ge \sum_i \sum_{j>i} \sum_k a_i^s a_j^s a_k^s \bar{\mu}_{ijkl}^{s(t)} Y_{ik} + \sum_i \sum_{j>i} \sum_k a_i^s a_j^s a_k^s \bar{\pi}_{ijkl}^{s(t)} Y_{jk} + \sum_i \sum_{j>i} \bar{\lambda}_{ij}^{s(t)} + \sum_i c_i \bar{\tau}^{s(t)} \quad (3.6)
$$

for $t = 1, \ldots, T$. (3.6) are the so called natural Benders cuts. T can be very large, so only the active ones of (3.6) are added into (RMP).

In an optimal solution to the primal subproblem (SP^s), each OD pair may use only one to two routes, so most of F^s_{ijkl}'s are 0. Therefore, the primal subproblem (SP^s) is usually degenerate, and then the dual subproblem (DSP^s) usually has multiple optimal solutions. However, only one of these solutions is returned by the solver in the algrithm and it might be dominated by other optimal solutions. Adding weak cuts may result in large number of iterations and eventually slows down the algorithm.

To strengthen the Benders cuts, in addition to the natural cuts, we also add Pareto-optimal cuts as proposed by [44]. A Pareto-optimal cut is identified by solving an auxiliary problem of (DSP^s) as below.

$$
\begin{aligned}
(AP^s) \quad \max \quad & \sum_i \sum_{j>i} \lambda^s_{ij} + \sum_i c_i \tau^s_i + \sum_i \sum_{j>i} \sum_k a^s_i a^s_j a^s_k y^0_{ik} \mu^s_{ijk} + \sum_i \sum_{j>i} \sum_k a^s_i a^s_j a^s_k y^0_{jk} \pi^s_{ijk} \\
s.t. \quad & \sum_i \sum_{j>i} \lambda^s_{ij} + \sum_i c_i \tau^s_i + \sum_i \sum_{j>i} \sum_k a^s_i a^s_j a^s_k y_{ik} \mu^s_{ijk} \\
& + \sum_i \sum_{j>i} \sum_k a^s_i a^s_j a^s_k y_{jk} \pi^s_{ijk} \geq z^s(x,y) \qquad (3.7) \\
& (3.4a), (3.4b), (3.4c), (3.4d), (3.4e), (3.4f)
\end{aligned}
$$

(3.7) ensures that every feasible solution to (AP^s) is an optimal solution to (DSP^s). We choose y^0 such that

$$
y^0_{ji} = \begin{cases} \frac{N}{2I} & i \neq j \\ \frac{1}{2} & \text{otherwise} \end{cases}
$$

for $i, j \in I$, and define x^0 such that $x^0_i = 1/2$, for $i \in I$. Then it is easy to show that (x^0, y^0) forms a relative interior point of the feasible region of (X, Y). It has been proved by [44] that the cut generated by (AP^s) is Pareto-optimal, so it is not dominated by any other cut generated by (DSP^s). By adding the Pareto-optimal cuts together with the natural cuts, the total number of iterations is greatly reduced and the overall solution time is also improved.

3.3 Numerical Studies

In this section, we present some numerical results to demonstrate the value of flexibility in air transportation networks and to develop more insights on the design of flexible hub-and-spoke structures. The data of the first 15 cities in the CAB data ([51]) is used. We take the flow in CAB as demand and the distance as transportation cost. The fixed cost of each airport is randomly generated from the interval $[4 \times 10^7, 5 \times 10^7]$, and its flexibility cost is set at 1/10 of its fixed cost. The capacity of each airport is set at twice of the total demand from and to it, i.e. $c_i = 2 \sum_j h_{ij}$, for $i \in I$. The unit penalty cost of each OD pair is is randomly generated from the interval $[4 \times 10^3, 6 \times 10^3]$. The discount factor of hub-to-hub

route is 0.9. Later, some of these parameters are adjusted in a controlled experiment to examine their effects on the network topology.

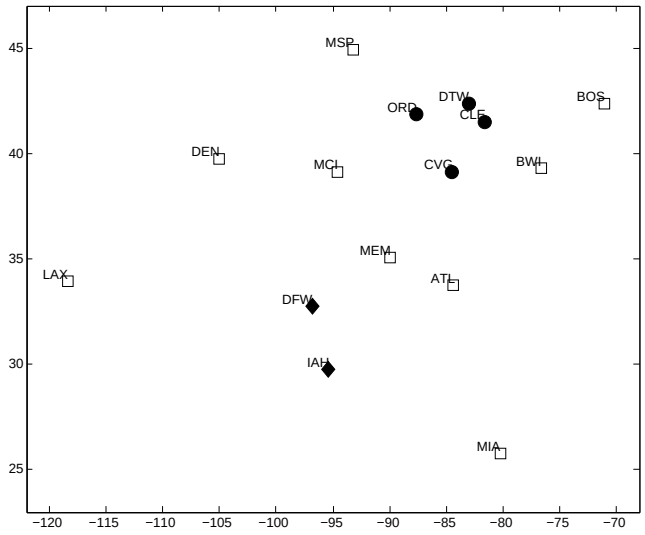


Figure 3.1: Clustering of Airports

Because of the lack of data on the correlation between airports, and because airports close to each other are usually positively correlated, we partition airports into clusters according to distance, and assume that the airports within the same cluster are perfectly positive correlated. Among the first 15 cities in the CAB data, Chicago (ORD), Cincinnati (CVG), Cleveland (CLE), and Detroit (DTW) are identified as one cluster, Dallas (DFW) and Houston (IAH) as another cluster, and each of the other nine airports forms a cluster by itself, as depicted in Figure 3.1, in which singleton clusters are denoted by hollow square makers and other airports in the same cluster share the same type of solid marker. Clusters are assumed to be independent of each other. The failure probabilities of the clusters are independently generated from a uniform distribution over $[0, 0.1]$. Because the failure probabilities are small, only the scenarios with no more than one cluster disrupted are considered. Note that compared with [81], our assumption allows more than one airport to be disrupted.

Value of Flexibility in Air Transportation Networks

When There Is No Flexibility Cost

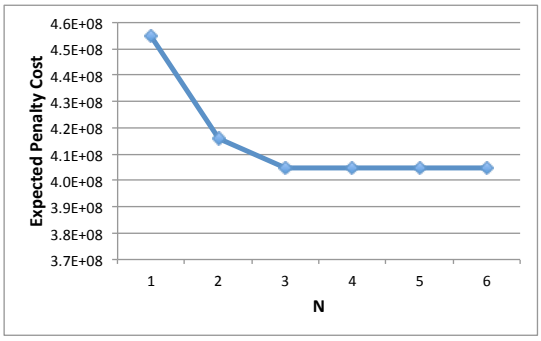
Our model proposes to use flexibility in hub assignment to hedge against the risk of airport disruptions. Thus, it is necessary to examine whether flexibility can improve the performance of air transportation networks with unreliable airports. Considering the cost of flexibility,

it is also interesting to investigate how much flexibility is needed. It has been proved for process flexibility design that, the marginal benefit of flexibility is decreasing, and a 2-chain is able to perform almost as well as a fully flexible structure ([1]). In this study, we check whether the above properties hold for hub-and-spoke structures.

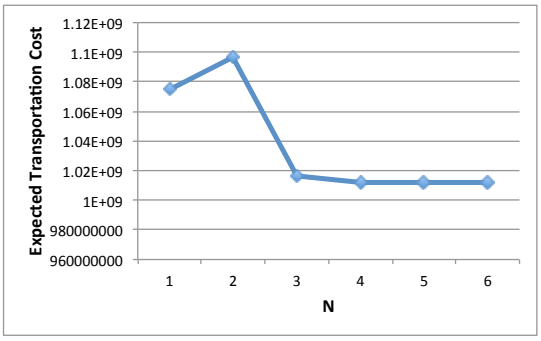
We first set the cost of flexibility at zero in order to reveal the pure benefit of flexibility. Figure 3.2 shows the changes in expected costs as N increases from 1 to 6. Note that the network has a traditional hub-spoke structure with single allocation when $N = 1$, and has a flexible hub-and-spoke structure with all the other N values. Figure 3.2(a) demonstrates the decreasing in penalty cost as a result of increased reliability when there is more flexibility in the network. The figure also suggests that the improvement in reliability is diminishing as N increases. Figure 3.2(b) shows that there may be higher transportation cost when changing from a inflexible network to a flexible network, but adding more flexibility will eventually return lower transportation cost. We also find that the expected total cost is convexly decreasing in N, as shown in Figure 3.2(c). It means that flexibility can reduce expected total cost, and the marginal value of flexibility diminishes as N increases. Moreover, the expected total cost is constant when $N \geq 5$, so adding flexibility has no value when N is sufficiently large. On the contrary, when N increases from 1 to 2, the expected total cost drops dramatically. In out simulation, the drop in expected total cost from $N = 1$ to $N = 2$ is 77.56% of the total drop from $N = 1$ to $N = 6$. Therefore, by having a 2-flexible hub-and-spoke structure, we get almost all the benefit of a fully flexible structure.

Recall that the flexibility constraint requires that each spoke airport uses *at most* N hubs, so some may use N hubs and some may use less. Obviously, since there is no flexibility cost, it does not hurt for spoke airports to use exactly N hubs whenever the total number of hubs is at least N. However, our simulation results provide evidence for that the flexibility constraint is not always binding. In Figure 3.3, we plot the total number of hubs (TH), the average number of hubs assigned to a spoke airport(AH), and the maximum number of hubs assigned to a spoke airport (MH) with different values of N. First of all, AH and MH is restricted by TH, and TH is determined by all parameters, so increasing N does not have any effect on the network structure once N is large enough. Secondly, when $N = 4$, 5, or 6, even if some spoke airports use exactly N hubs, others use less, so AH is less than MH. Therefore, the flexibility constraint is not binding in these cases. There are several possible reasons for not using flexibility when it incurs no cost. One possible reason is that the hub is too far away such that the transportation cost if using it outweighs the penalty cost. Another possible reason is that, several hubs belongs to the same cluster and some of them have enough capacity to serve a spoke airport, then not all these hubs will be used by the spoke airport. Other reasons may include insufficient hub capacity.

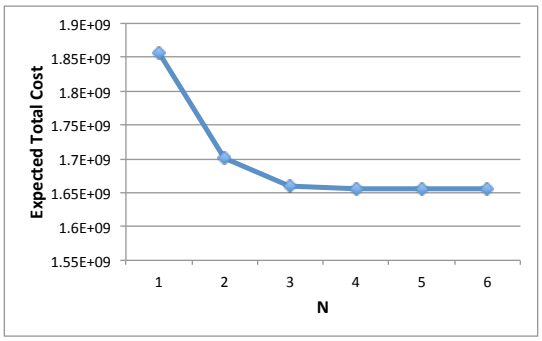
We also observe that much fewer hubs are needed in flexible hub-and-spoke structures (When $N > 1$). As noted before, $N = 1$ is equivalent to single allocation. Under disruptions, it is natural for a single allocation network to build more hubs, and each hub serves a small



(a) Expected Penalty cost v.s. N with zero flexibility cost



(b) Expected Transportation cost v.s. N with zero flexibility cost



(c) Expected total cost v.s. N with zero flexibility cost

Figure 3.2: Expected Costs v.s. N with Zero Flexibility Cost

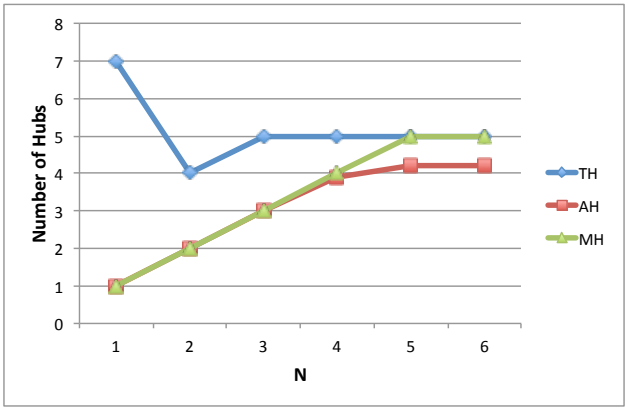


Figure 3.3: Total Number of Hubs, Average Hubs Assigned to Each Spoke, and Maximum Hubs Assigned to One Spoke v.s. N, with Zero Flexibility Cost

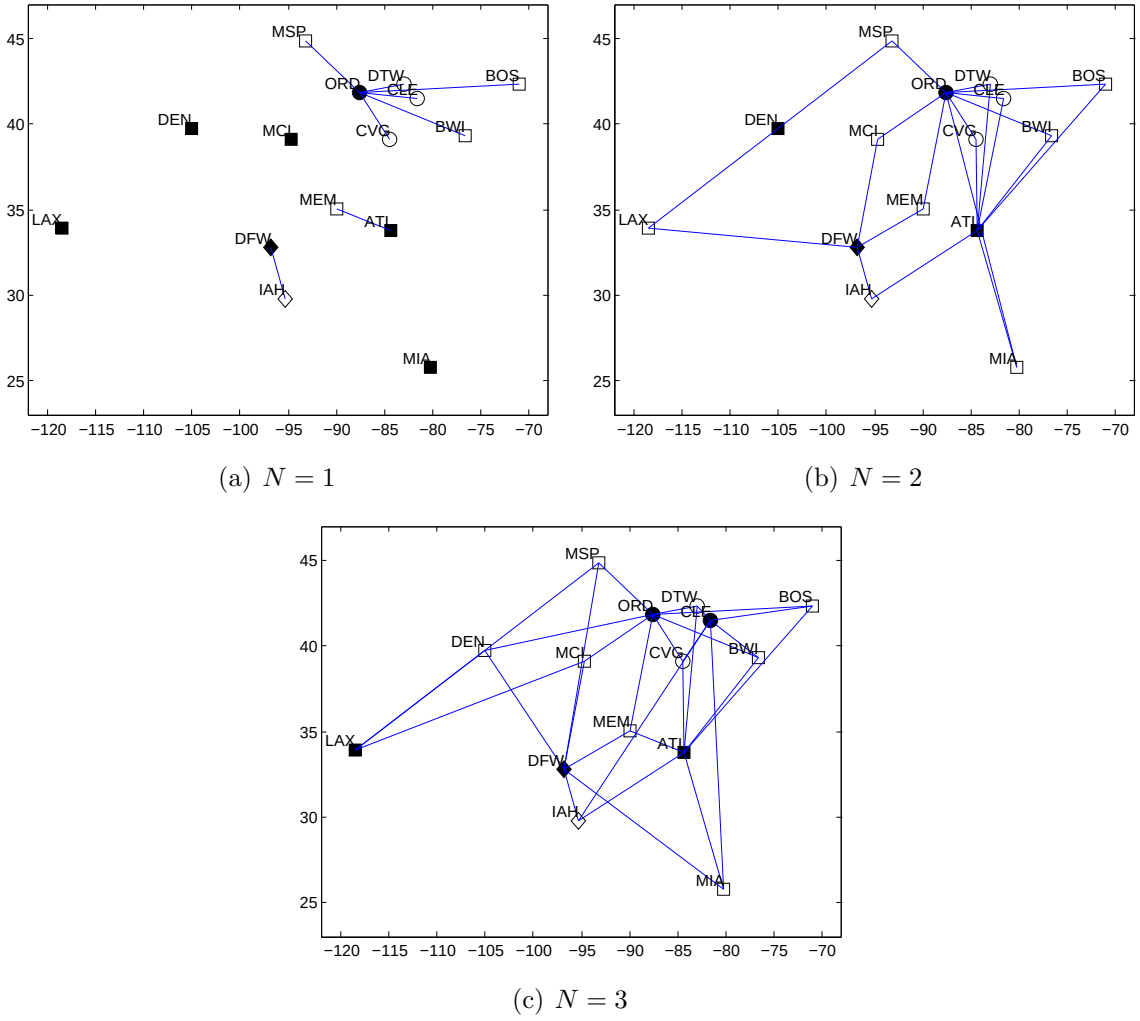
number of spokes in order to spread the risk, as if lots of spokes are assigned to one hub, then the failure of that hub will incur penalty costs for all passengers associated with that

hub and the spokes connecting to it. When N is greater than 1, the risk of hub failures is mitigated by the extra flexibility of the spokes. Moreover, because each spoke is connected to two or more hubs, decreasing the number of hubs will not cause sharp increase in the transportation cost – similar to multiple allocation models, flows from the same origin can use different hubs for different destinations. It is worthwhile to point out here that one major difference between our model and traditional multiple allocation models is that in our model, the hubs used for each OD pair in different scenarios can be different. Then, building fewer hubs saves fixed costs. It can be noticed that as N increases from 2 to 3, the total number of hubs increases from 4 to 5. This is mainly because the decrement in expected penalty cost and transportation cost outweigh the fixed cost of an additional hub, and the available capacity when using four hubs limits the ability to serve more passengers under certain scenarios.

At last, AH and MH are increasing in N. Actually, this result can be verified by contradiction. For example, suppose that as N increases from 4 to 5, the MH of the optimal topology decreases from 4 to 3, then, the network topology with MH equals to 3 should return lower total expected cost than any topology with MH equals to 4. However, note that the new network is a feasible configuration for the case of $N = 4$, and this contradicts with the assumption that the optimal topology when $N = 4$ has an MH equals to 4.

Figure 3.4 summarizes the network topologies when $N = 1$, $N = 2$, and $N = 3$. In these figures and the figure hereafter, we use solid markers to denote hub locations, and hollow markers for spoke locations. As shown in Figure 3.4(a), when $N = 1$, most of airports, except for those in the northeast, are hubs themselves to take the advantage of discounts in transportation costs between hubs, because they are far away from each other. Airports in the northeast choose ORD to be their common hub, instead of employing different hubs to improve reliability as discussed above, due to the following reasons. Firstly, since ORD, CVG, CLE, and DTW belong to the same cluster and have perfectly positive correlation, having more than one hub in these four cities will not return higher reliability. Secondly, having a common hub for northeastern airports expresses economy of scale, as at the fixed cost of one hub, the flows associated with northeastern airports are aggregated together to save transportation cost. Thirdly, CAB data suggests that the flows between ORD and northeastern airports are significant, so choosing ORD as the common hub results in substantial savings in transportation cost compared to using other airports in the same region, such as DTW. Last but not least, according to our assumption, ORD's reliability is above average and its capacity is sufficient to serve all northeastern airports; thus, it is not economical to connect a northeastern airport to a hub that is in a different cluster from ORD, such as connecting Baltimore (BWI) to Atlanta (ATL).

When flexibility exists, the network topology expresses a more balanced pattern. As shown in Figure 3.4(b) and Figure 3.4(c), hubs are evenly spread and each of them serves several spokes. The most important observation is that, each spoke node tends to choose



(a) $N = 1$

(b) $N = 2$

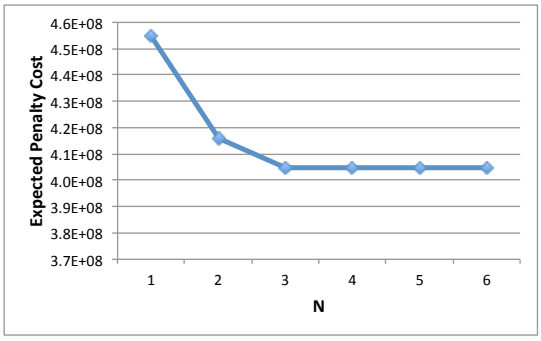
(c) $N = 3$

Figure 3.4: Network Topologies with Zero Flexibility Cost

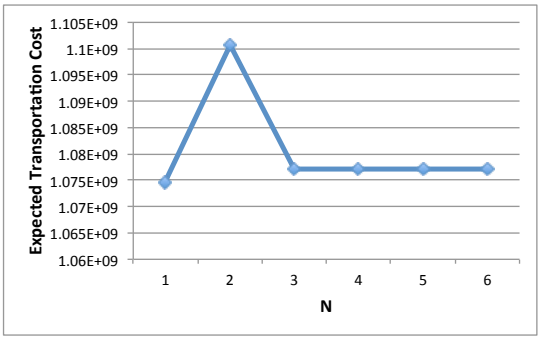
hubs from different clusters, for example when $N = 3$, Miami (MIA) chooses ATL, CLE, and DFW as its three hubs. Connecting with hubs in different clusters improve the reliability of each spoke, because the hubs belonging to different clusters are independent. Moreover, since each OD pair can choose different intermediate hubs, it helps shorten routes by having scattered hubs. At last, it is worth noting that increasing the flexibility level may not only change the number of hubs, but also the hub locations. Therefore when more hubs are needed, simply adding hubs to the old configuration does not guarantee optimality. For instance, Denver (DEN) is a hub when TH= 4 (when $N = 2$), and a spoke when TH= 5 (when $N = 3$).

When There Is Positive Flexibility Cost

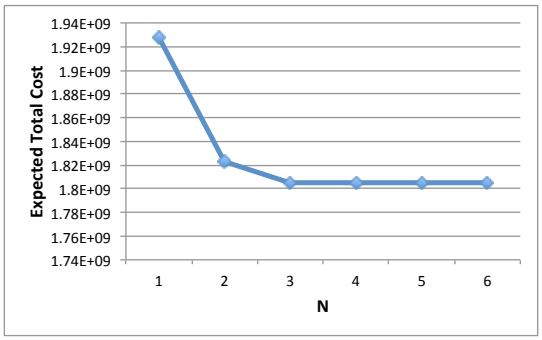
Next, since flexibility comes at the cost of establishing connections between spokes and hubs, we restore the positive flexibility cost and examine the changes in network topology. Figure 3.5 summarizes the expected costs with non-zero flexibility costs. Obviously, the expected total cost after taking into consideration of the flexibility cost should be higher than that in the case of zero flexibility cost, as shown in Figure 3.5(c). Similar to the case with zero flexibility cost, having a 2-flexible hub-and-spoke structure achieves about 83.3% of the benefit of a fully flexible structure.



(a) Expected Penalty cost v.s. N with flexibility cost



(b) Expected Transportation cost v.s. N with flexibility cost



(c) Expected total cost v.s. N with flexibility cost

Figure 3.5: Expected Costs v.s. N with Flexibility Cost

The first observation is that when is positive flexibility cost, the expected penalty and the expected total costs express similar trends as those with zero flexibility cost, as N increases. However, Figure 3.5(b) shows that the expected transportation cost increases after bringing in flexibility. This is because when $N \geq 2$, less hubs and less links are employed, as shown in Figure 3.6. Therefore, the savings in the expected total cost can be explained by the savings in penalty cost and fixed cost.

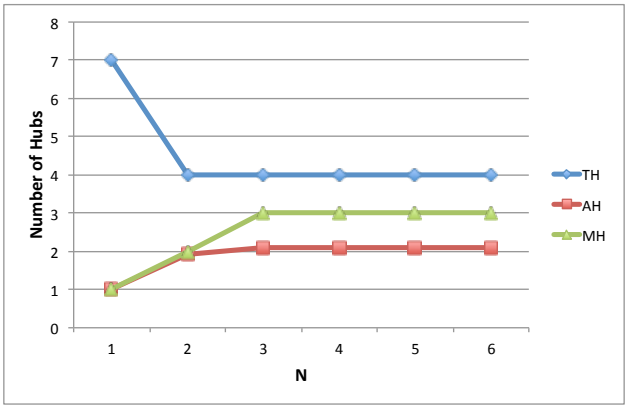
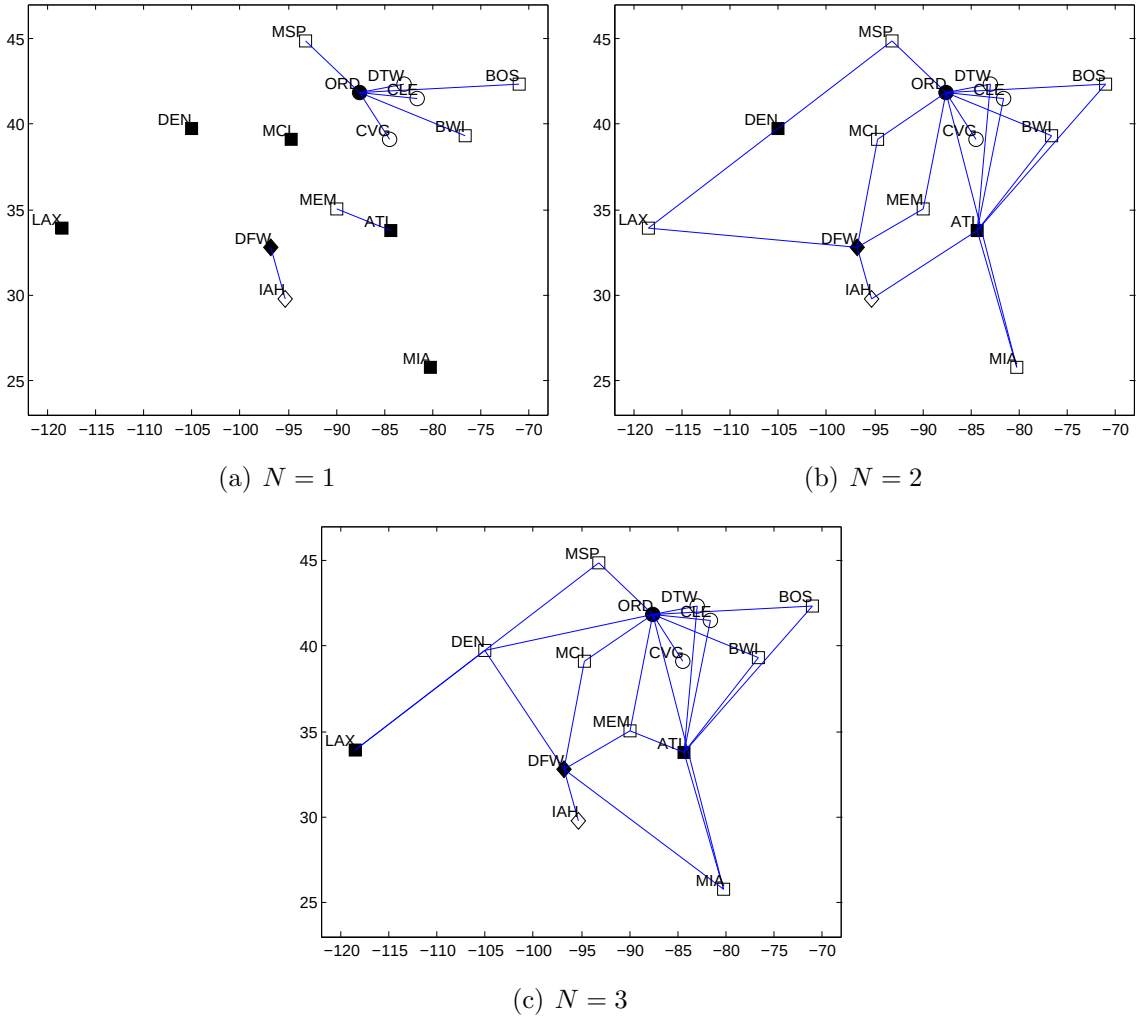


Figure 3.6: Total Number of Hubs, Average Hubs Assigned to Each Spoke, and Maximum Hubs Assigned to One Spoke v.s. N, with Flexibility Cost

Figure 3.7 depicts the network topology when there is positive flexibility cost. It is observed that the topology when $N = 1$ is exactly the same as the case with zero flexibility cost. However, when $N = 2$, CVG is only connected to one hub at ORD, as shown in Figure 3.7(b). When $N = 3$, CLE is no longer used as hub as in the case of zero flexibility, and most of spokes are connected to only two hubs. The main cause is the trade-off between decrement in transportation and penalty cost and the increment in flexibility cost. For instance when $N = 2$, the flexibility cost of connecting CVG to ATL is above the average flexibility cost, while the flows associated with CVG is much lower than the average flow, and the majority of CVG's flow goes to northern airports, thus the flexibility cost outweighs the potential savings in transportation and penalty cost.

Another important observation is based on Figure 3.7(c). When $N > 2$, a spoke node is rarely connected to a third hub node in order to be more reliable, because the two assigned hubs of that spoke usually belongs to different clusters and the probability for both of them fail is extremely small. Thus, the main reason for having a third hub is to save transportation cost, or to meet the capacity limits at its hubs. For instance, according to CAB data, Memphis (MEM) has high demand to north, west, and east, and thus it is connected to ATL, ORD, and DFW. IAH is another example. Although IAH and its hub DFW are in the same cluster, IAH is also connected to ATL mainly due to the capacity limit at DFW.

The comparison between Figure 3.4 and Figure 3.7 delivers another key finding: spokes tend to choose their first hub from airports in the same cluster, and when a spoke form a cluster by itself, it tends to choose a nearby hub as its first hub. For instance, CLE, CVG, and DTW choose ORD, and IAH chooses DFW. The main reason is that ORD is perfectly "reliable" to CLE, CVG, and DTW, because ORD fails only if all of CLE, CVG, and DTW fail. The other cause is the economies of scale. Combining nearby airports together (and selecting one of them to be a hub) returns saving in transportation cost through discounts



(a) $N = 1$

(b) $N = 2$

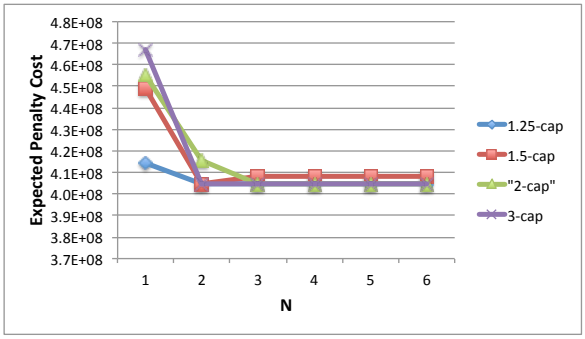
(c) $N = 3$

Figure 3.7: Network Topologies with Flexibility Cost

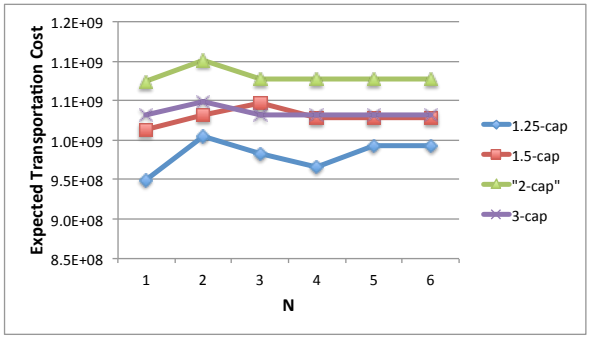
between hubs. However, there is always a tradeoff between the increment in fixed cost and the decrement in transportation cost, and this is why both Boston (BOS) and BWI are connect to ORD instead of having separate hubs.

Effect of Airport Capacities

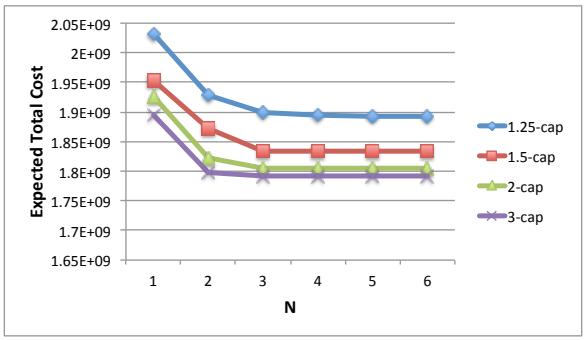
In previous simulation studies, the capacity of each airport i is set at twice of the total demand from and to it, i.e. $c_i = 2\sum_j h_{ij}$. For convenience, we denote "$\gamma - cap$" as the setting in which the capacity of each airport is set at $c_i = \gamma \sum_j h_{ij}$, for $\gamma > 0$. That is, previous simulation studies are conducted under $2 - cap$. In this section, we investigate the effect of γ.



(a) Expected Penalty cost v.s. N under Different $\gamma - cap$



(b) Expected Transportation cost v.s. N under Different $\gamma - cap$

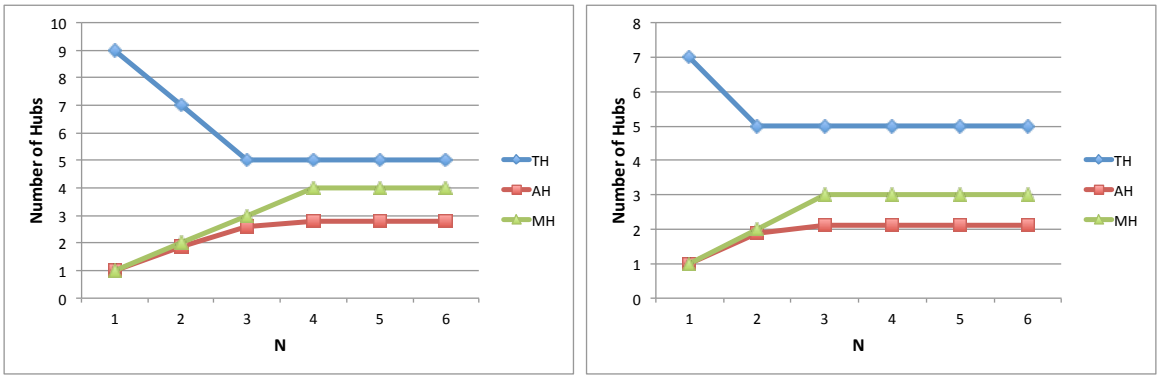


(c) Expected total cost v.s. N under Different $\gamma - cap$

Figure 3.8: Expected Costs under Different $\gamma - cap$ with Flexibility Cost

Figure 3.8(c) plots the expected total cost under different $\gamma - cap$, and naturally it shows that greater γ corresponds to lower expected total cost. It is also noticed from Figure 3.8(c) that under all $\gamma - cap$'s in our simulations, the expected total cost is decreasing in N. In particular, the decrement is diminishing as N increases, and a 2-flexible hub-and-spoke network is able to attain most of the cost saving benefit returned by a fully flexible network. Figure 3.8(a) suggests that flexibility reduces the expected penalty cost as well. Moreover, greater γ is more efficient in mitigating the risk of disruption. This is because when γ is greater, each available hub is able to handle more demand when other hubs are disrupted. However, Figure 3.8(b) shows that switching from no-flexibility to flexible networks does not necessarily lead to lower transportation cost. Actually, the simulation results indicate that when adding a little flexibility to a inflexible network, the transportation cost increases as a result of the reduction in the number of hubs. In addition, it is also showed by our simulation that for fixed N, the expected total cost is not necessarily convexly decreasing in γ.

Figure 3.9 further illustrates that when γ is small, more hubs is needed by each spoke.



(a) Total Number of Hubs, Average Hubs Assigned to Each Spoke, and Maximum Hubs assigned to One Spoke v.s. N, under $1.5 - cap$

(b) Total Number of Hubs, Average Hubs Assigned to Each Spoke, and Maximum Hubs assigned to One Spoke v.s. N, under $3 - cap$

Figure 3.9: Total Number of Hubs, Average Hubs Assigned to Each Spoke, and Maximum Hubs assigned to One Spoke under Different $\gamma - cap$

In addition, comparing Figure 3.9(a) with Figure 3.6 and Figure 3.9(b), we notice that when $N \geq 2$, AH always takes values close to 2, which implies that a 2-flexible hub-and-spoke structure is close to optimal, and adding more flexibility returns only marginal improvement. In other words, when capacities at hubs are large enough, a 2-flexiblehub-and-spoke structure can efficiently hedge against the risk of disruptions. Adding more flexibility only has marginal impact on transportation cost and fixed cost.

To get more insights, we plot in Figure 3.10 the network topologies with $N = 1$, $N = 2$, and $N = 3$ under $1.5 - cap$ and $3 - cap$. The figure suggests increased capacity also leads to savings in transportation cost. For instance, compared to Figure 3.7(b), the case of $N = 2$ under $3 - cap$ has one more hub at CLE, as shown in Figure 3.10(d). In the new topology, DTW is connected to CLE instead of ATL. Note that since DTW and CLE as well as ORD are in the same cluster, then once CLE and ORD fail, DTW fails as well, so the re-assignment of DTW is only due to the potential saving in transportation cost, which also explains why BOS and BWI are disconnected from ORD and re-assigned to CLE. In addition, compared to the case of $N = 2$ under $2 - cap$, MEM is no longer assigned to ORD, and IAH is only connected to DFW. The implication based on the above observations is that, there exists potential savings in flexibility cost and transportation cost from increasing hub capacities, because low hub capacities make it necessary to use routes through different intermediate hubs for the demand associated with the same OD pair.

Effect of Correlation between Airports

It is usually assumed that disruptions are independent in reliable facility location models or reliable hub location models. However, this assumption fails for air transportation networks. For example, snow storms often hit the whole Chicago-Minneapolis area, so a few airports in this area, including ORD, MSP, DTW, CLE, and IND, are often disrupted at the same time. In this study, we demonstrate how disruption correlation affects the topology of the network by a numerical example, and point out that overlooking correlation may result in inefficiency and higher cost.

In this study, we set the disruption probability of each airport at 0.2 and the unit penalty cost of each OD pair at 8000. Two cases of airport disruptions are studies: (1) the correlated disruption case, in which ORD, DTW, and ATL form a cluster, and each of the other airports forms a cluster by itself, and (2) the independent disruption case, in which airports are disrupted independently. Because disruption probabilities are small at each airport, we only consider the scenarios in which at most one cluster is disrupted for the correlated disruption case, and those in which at most one airport is disrupted for the independent disruption case. Scenario probabilities are normalized so that they sum up to one.

In Figure 3.11, we plot the optimal hub locations and assignment in these two cases when $N = 2$. It is notable that the network topology in the correlated disruption case is quite different from that in the independent disruption case. If airports are disrupted independently, ORD, DTW, and ATL are all selected as hubs, while in the correlated disruption case, only ORD among the three is selected, CLE takes the place of DTW, and MIA is also selected as a hub. The reason for the change in hub locations is because, it is not beneficial to have two perfectly correlated hubs, especially when they are close to each other and any of them has sufficient capacity. ORD and DTW make a typical example for this issue. First, they are always disrupted at the same time, so they are not able to serve as backup hubs of each other. Then the only possible advantages to have both of them as hubs are more sufficient hub capacity and lower transportation cost. However, ORD and DTW are close to each other, and the capacity of ORD is already large enough to serve all the nearby airports, so the benefit of having both of them as hubs is outweighed by the increased fixed cost.

Besides the change in hub locations, we also observe change in hub assignment. In the independent disruption case, each of the five spoke airports near the east coast, BOS, CLE, CVG, BWI, and MIA, connects to two of the three hubs, ORD, DTW, and ATL. In the correlated disruption case, because DTW is replaced by CLE, BOS, CVG and DTW connect to ORD and CLE, BWI switches from ATL to the closer hub CLE, and MIA becomes a hub. A natural question is, whether the change in hub assignment is just a result of the change in hub locations, or it is also caused by the correlation between airports. To answer this question, we conduct another experiment in which ORD, DTW, and ATL are perfectly correlated, and hub locations are fixed as in the independent disruption case. The optimal

	Loss of demand	Fixed Cost	Flex. Cost	Transp. Cost	Penalty Cost
Correlated	1.51×10^5	2.83×10^8	1.13×10^8	1.00×10^9	1.08×10^9
Independent	1.34×10^5	2.83×10^8	1.13×10^8	9.70×10^8	1.21×10^9
Relative Increase	12.26%	0.08%	0.17%	−2.99%	12.26%

Table 3.1: Cost of Overlooking Disruption Correlation

hub assignment is displayed in Figure 3.12. It is notable that the hub assignment is different from that in Figure 3.11(a) although the hub locations remain the same. In particular, none of the spoke airports uses more than one of the three correlated hubs. This is because that having two perfectly correlated hubs is not helpful with mitigating their disruptions.

At last, we investigate the cost of overlooking disruption correlation. We have shown that, in the correlated disruption case, the optimal network topology is as displayed in Figure 3.11(b) (denoted by the correlated topology). Now suppose that the planner treat the airports as independent ones, so he ends up with a network topology as shown in Figure 3.11(a) (denoted by the independent topology), which is optimal in the independent disruption case. Given that the hub locations and assignment have been fixed, we optimize over the flows in all scenarios and obtain various statistics. Some of the results are summarized in Table 3.1. The first column is the expected loss of demand caused by disruptions. Columns 2 to 5 are the fixed cost, the flexibility cost, the expected transportation cost, and the expected penalty cost, respectively. The statistics when using the correlated topology is listed in line 1, and those when using the independent topology in line 2. The relative increases in these statistics are listed in line 3. We note that the expected loss of demand raises by 12.26% from using the independent topology. Fixed cost and flexibility cost have little change, and the expected transportation cost slightly decreases because less demand is transported. The expected penalty cost increased by the same percentage as the expected loss of demand because the unit penalty cost is the same for every OD pair. The above results imply that, failing to consider correlations between airports when designing a network may result in more demand loss and higher penalty cost.

Summary of Key Findings and Insights

According to the results from the numerical studies, we propose the following principles on deciding hub locations and assignment.

1. A flexible network is better than a single allocation network, in terms of stronger reliability, less transportation cost, and fewer required hubs.

2. When there exists multiple clusters that are independent or weakly correlated with each other, a 2-flexible hub-and-spoke structure can efficiently hedge against the risk

of disruptions. Adding flexibility usually has diminishing benefits in expected total cost.

3. Avoid selecting airports that are positively correlated and close to each other as hubs at the same time unless the capacity of any of them is not large enough.

4. For any spoke airport, try to first assign it with the hubs that are strongly positively correlated with it.

5. If a spoke airport uses hubs that are all independent or weakly correlated with it, try not to choose those that are strongly correlated with each other.

6. When adding flexibility to any spoke airport which currently has only one hub, the choice of new hub should either be independent or has weak correlation with its first one.

7. If there are multiple hubs that satisfy principle 4-6, then choose the one that has greater capacity and is on the direction of the most significant demand associated with the spoke of interest.

8. Add a third hub to a spoke, only if it helps reduce the transportation cost, or the total capacity of current hubs cannot meet the demand associated with the spoke of interest.

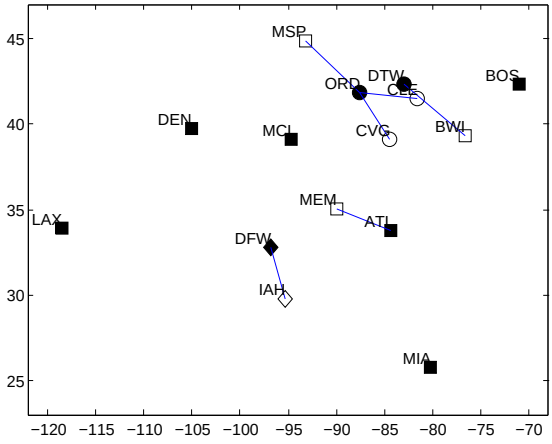
3.4 Summary

In this chapter, we present a flexible capacitated hub location model that deals with airport disruptions. Unlike the traditional hub location models, our model allows each spoke airport to have up to N hubs, all of which can be used in every scenario without any pre-determined order. In each scenario, after the states of airports are observed, each airport decides how much of its demand to be transported via each of its hubs. In this way, our model not only adopts a flexible network structure, but also uses a flexible mechanism to make routing decisions. To properly characterize airport disruptions, we assume that they can be correlated with each other, which is quite common in practice. Our problem is formulated as a mixed-integer program that minimizes the expected total cost. Because of the large number of decision variables and constraints, this problem is computational intractable. Benders decomposition is applied to solve it, and Pareto-optimal cuts are added to accelerate the algorithm.

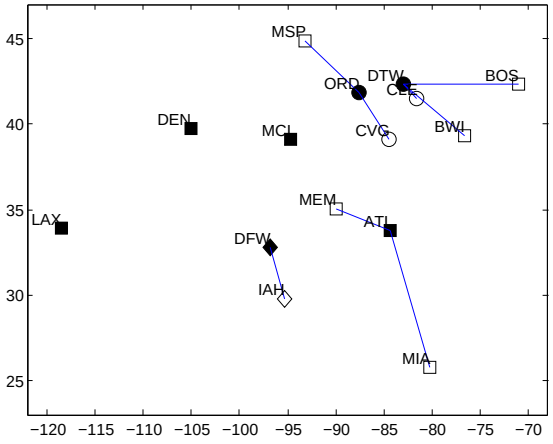
In numerical studies, we carry out a series of experiments on the first 15 cities in the CAB data, and obtain the following conclusions and insights. First of all, we reveal the diminishing return of flexibility in networks. As N increases, the minimum expected total cost is convexly decreasing. Analogous to 2-chain in process flexibility, a 2-flexible hub-and-spoke structure is able to achieve most of the benefit of a fully flexible structure. When failure

probabilities are small, having a 3-flexible structure is not very helpful in further improving the reliability of the network once we already have a 2-flexible structure. However, it may indeed lower the total cost by better matching demand with hub capacity and saving transportation cost. We also observe that, the more capacity the airports have, the less flexibility is needed. Of course, a 2-flexible structure is always beneficial as long as airports are not reliable. At last, we point out that the correlation between airports can greatly change the topology of the optimal network. Overlooking correlation may result in suboptimal hub locations and assignment and cause inefficiency and cost.

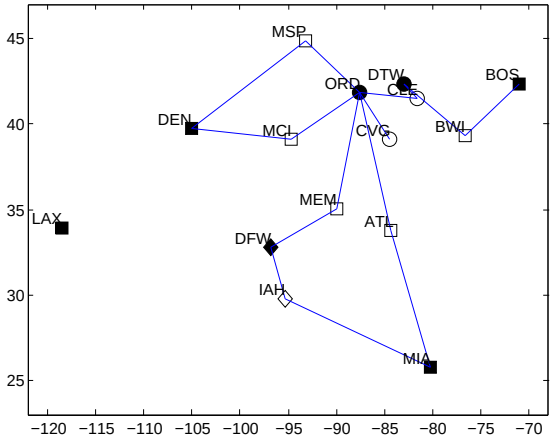
Our model also has several limitations. The approach we use is essentially a scenario-based two stage stochastic programming approach, and it suffers the drawbacks of the latter. For example, our model fails to work if the probabilities of scenarios are unknown (though it is unlikely that we do not have these probabilities). In addition, our model sets minimizing the expected total cost as objective, and it fails to capture the risk-aversion of decision makers. Therefore, a possible extension could be a robust hub location model that optimizes the worst case total cost when disruptions happen. At last, although the authors agree with [16] that the fast growing computing capabilities assure that computational time is no longer the major limitation, it is still interesting to see how our algorithm performs with a larger instance on cloud computing engines.



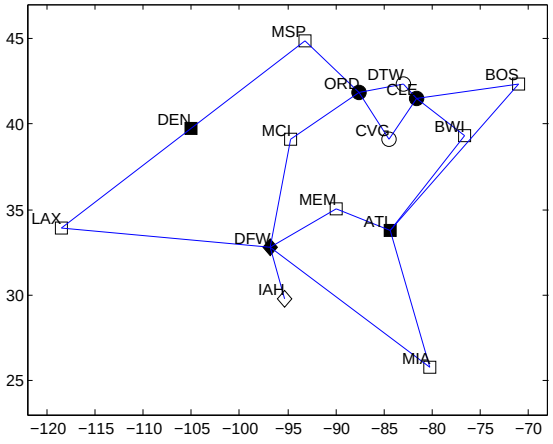
(a) $N = 1$, under $1.5 - cap$



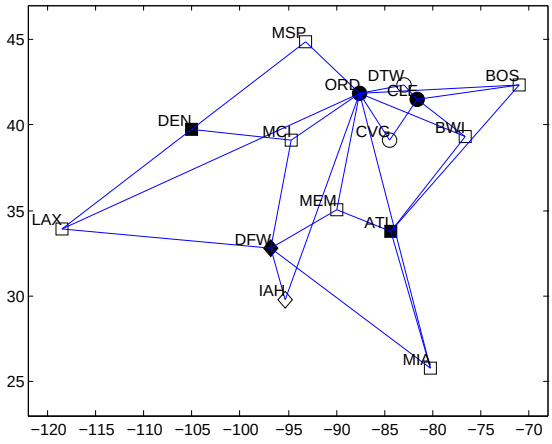
(b) $N = 1$, under $3 - cap$



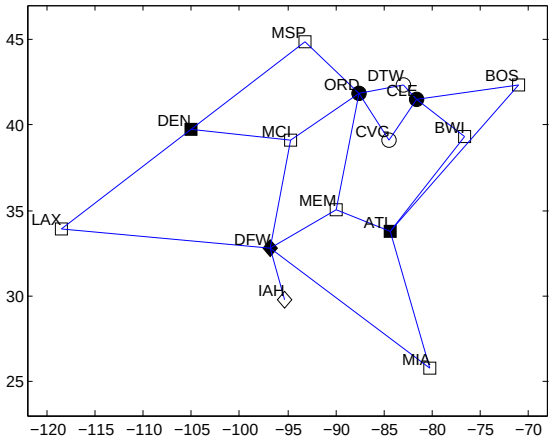
(c) $N = 2$, under $1.5 - cap$



(d) $N = 2$, under $3 - cap$

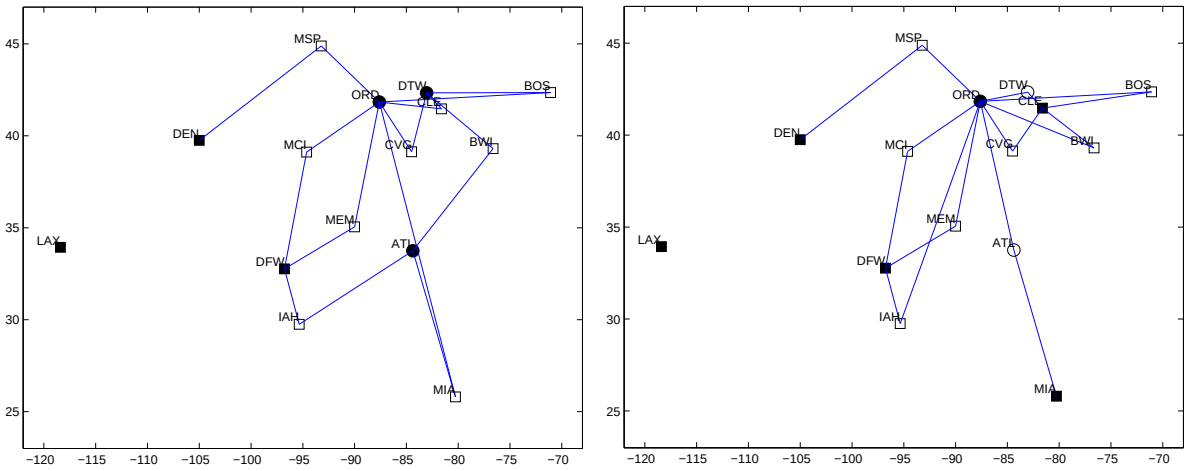


(e) $N = 3$, under $1.5 - cap$



(f) $N = 3$, under $3 - cap$

Figure 3.10: Network Topologies under Different $\gamma - cap$



(a) Independent Disruption Case (b) Correlated Disruption Case

Figure 3.11: Network Topologies in the Independent Disruptions Case and the Correlated Disruptions Case

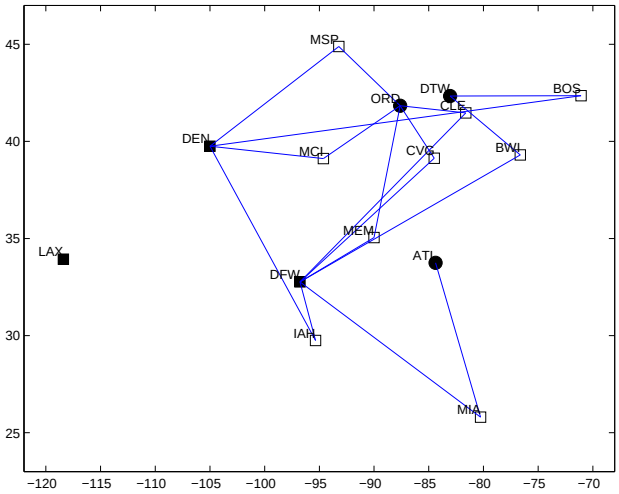


Figure 3.12: Network Topology in the Correlated Disruption Case with Hub Locations in the Independent Disruption Case

Chapter 4

Impact of Trade Credit on Retailers' Growth and Suppliers' Benefit

4.1 Introduction and Literature Review

Shortage in internal finance has been identified as the main obstacle for the growth of small firms ([17] and [11]). According to the World Business Environment Survey in 1999-2000, small (with 5 to 50 employees) and medium (with 51 to 500 employees) firms are significantly more financially constrained then large firms (with over 500 employees), and the growth of small firms are most negatively affected by financial constraints ([6]). It is usually difficult for small businesses to get loan from banks, so they use trade credit as a major tool to get external finance. A supplier extends trade credit to a buyer by allowing the buyer to order now and pay later. Trade credit provides an additional resource of funds to buyers by delaying payments. It takes two forms: net term and two-part term. A net k term requires that payment will be due k days after the delivery. An interest, which is usually high, will be charged if the buyer fails to pay on time. A two-part term not only specifies the due date of the payment, but also gives a discount percentage and a discount period. Hence, it provides the buyers with an option to pay early to get discount. We will focus on the net term in this chapter.

Trade credit is widely used, especially by small firms. According to [9], more than half of the small businesses in the United States in 1998 use trade credit. Trade credit is used more than all other financial services except checking account. In particular, the use of trade credit is most common among firms in manufacturing, construction, and wholesale and retail trade industries for which non-labor costs, such as the costs of equipment and inventory, are large relative to labor costs. Small firms not only use but also supply trade credit. Most small firms offer net terms trade credit, and about 20% offer two-part terms ([80]). Other evidence of the extensive use of trade credit can be found in [56] and [49].

The existence of trade credit has long been studied by economists. For buyers, the advantage of trade credit is obvious: they can borrow cash without paying any interest. This is especially desirable for those firms short in funds but unable to raise it from specialized financial institutions such as banks. In addition, trade credit provides buyers a period to inspect the product and make exchange before payment ([68] and [42]). Then why suppliers extend trade credit? The main reason is that suppliers can create new business and promote sales. They can also build stronger supplier-customer relationships. Besides, trade credit allows suppliers to price discriminate ([56]) though direct price discrimination is usually illegal. Other benefits include better control of buyers and lower transaction costs ([29] and [26]).

While trade credit brings all the benefits discussed above, it may increase the risk of the supplier as well. Delinquency cost arises when the buyer fails to pay by the end of the credit period. More importantly, all the past due become bad debt once the buyer goes bankrupt. [10] identifies the nonpayment of trade credit as a major cause of the failures of small businesses. Taking into account that buyers who seek trade credit are usually financially constrained and *lack of good credit record*, it is even more critical for the supplier to better evaluate the risk when determining whether to offer trade credit to a certain buyer.

This chapter aims to answer the following questions: How does trade credit facilitate the growth of small businesses? Does trade credit affect the chance that small businesses survive? And when should suppliers extend trade credit? We study a one-supplier-one-retailer supply chain that produces and sells a single product in a multi-period setting. The retailer starts with a small size and limited fund. He makes expansion decisions every a few periods based on his size and fund. The supplier may choose to extend or not to extend trade credit to the retailer. We analyze the growth of the retailer, and carry out numerical experiments to get insights on the benefits and risks of trade credit.

In general, there are two streams of research on trade credit: one uses empirical studies to identify the determinants of trade credit and build financial theories, the other focuses on inventory models with permissible delay of payment. Among the literature in the first stream, [55] find that small firms concentratedly borrow from a few financial institutions to build a stronger relationship and to increase their availability of financing, and they use less trade credit if having a longer relationship with financial institutions. This observation is later verified by [56]. They find that small firms use more trade credit if having limited access to credit from financial institutions. They also propose a price discrimination theory that trade credit is provided to firms with higher profit margin. [49] identify a firm's industry as an important determinant of trade credit terms, and find evidence supporting the product-quality-guarantee theory and the theory that trade credit provides suppliers with information on buyers' creditworthiness. We refer the readers to [30] for a thorough review of other related literature.

The second stream studies inventory problems when trade credit is extended. [64] pro-

pose an EOQ model when the supplier offers trade credit with price discount. All increasing deterministic demand models are discussed. [18] present an EOQ model with order-quantity-dependent trade credit, price dependent demand, and finite production rate that is proportional to demand rate. The total profit of the vendor-buyer integrated system is optimized over the retail price, the buyer's order quantity, and the vendor's production batch size. It is found that a longer trade credit term can increase the total profit. Other EOQ based trade credit models include [77].

The effects of trade credit in stochastic inventory models are also studied. [43] present an approximate method for calculating the cost function in a periodic review setting when the (s, S) policy is used. [32] study the inventory policy of a retailer confronted with random demand and general trade credit. The base-stock-policy is proved to be optimal and the optimal base stock level is derived. They also demonstrate with numerical examples how the supplier adjusts the discount period length or the discount rate when the other parameter is fixed in order to maximize his expected profit. [19] compare the order quantities in a newsvendor problem under different payment schemes including own-financing and supplier-financing, and experimental studies show that the retailer orders more under the own-financing payment scheme than under the supplier-financing payment scheme. While this phenomenon can not be explained by any utility model, it is consistent with the mental accounting in consumer behavior. Their results implies that trade credit may lower the order quantity of the retailer instead of increasing it. [72] demonstrate how payment schemes affect inventory policies in the EOQ model and the base stock model, and propose that payment scheme should match the supply chain strategy.

This chapter investigates the impact of trade credit on the growth of small businesses and their suppliers through a mathematical model, which has not been studied in the literature. In particular, our work is different from all the others in the following aspects. Firstly, we build a mathematical model of the growth of small businesses that jointly makes inventory and expansion decisions. Demand are assumed to be random and size-dependent. Both additive and multiplicative demand models are considered. Secondly, we study the effect of trade credit on the growth of small businesses and show that they tend to expand more aggressively and towards a higher target size when using trade credit. Thirdly, we quantitatively evaluate the risk of trade credit for both the supplier and the buyer. It is shown that the risk of trade credit is negligible under the additive demand model. However, under the multiplicative demand model, trade credit makes the buyer more likely to go bankrupt in his growing stage, and thereby negatively affects the supplier's profitability. At last, we reveal that trade credit has a higher risk when demand is positively correlated. We point out that it is important for suppliers to carefully investigate the market when making trade credit decisions.

The remainder of this chapter is organized as follow: section 4.2 describes the model and solves the retailer's problem analytically, section 4.3 uses simulation to evaluate the sup-

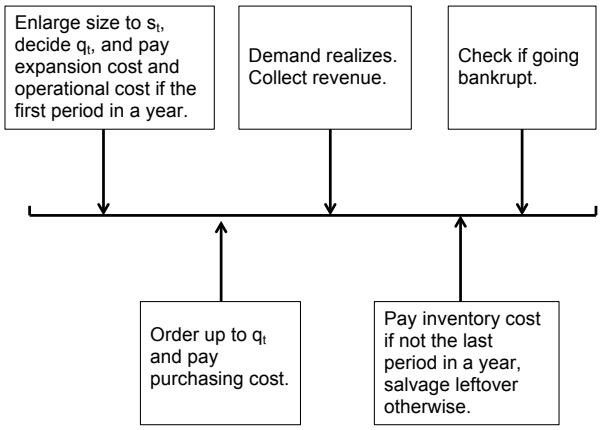
plier's payoff when trade credit is extended or not, and to provide insights on the value of trade credit, and section 4.4 summarizes our main results and points out several directions for future work.

4.2 Model and Assumptions

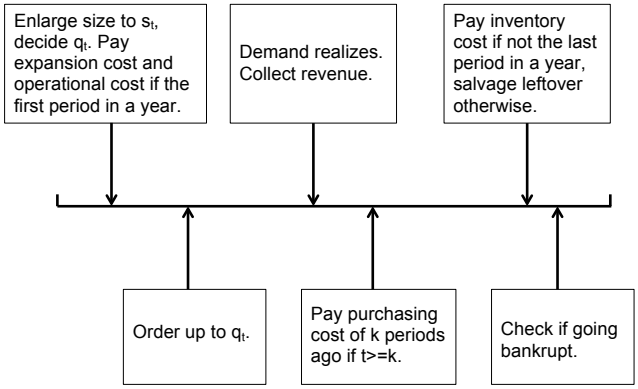
We consider a supply chain with one supplier and one retailer. The supplier produces a single product at unit cost c, and sells it to the retailer at wholesale price $w(> c)$. The retailer then sells the product to its customers at retail price $p(> w)$. Here we assume that the retail price p is fixed, because the retailer has little market power in his growing stage and hence is a price-taker. Demand is random and depends on the retailer's size. We adopt a linear mean demand function $d(s) = \alpha s$, where s is the size of the retailer, and $\alpha > 0$ is the average demand attracted by each unit of size. Both additive and multiplicative demand models are studied. In the additive model, demand $D(s) = d(s) + \epsilon$, with $\epsilon \sim F$ and $E[\epsilon] = 0$, while in the multiplicative model, $D(s) = d(s) \cdot \epsilon$, with $\epsilon \sim F$ and $E[\epsilon] = 1$. In both cases, the support of $D(s)$ is contained in $[0, +\infty)$. These two demand models represent two cases of demand correlation: negative correlation with $\rho = -0.5$, and positive correlation with $\rho = 1$.

The supplier can choose to charge the retailer on the delivery of orders, or to extend trade credit. We denote these two cases by scenario a and scenario b, respectively. For analytic simplicity, we only consider trade credit terms in the form of Net k, for $k = 1, 2, \ldots$. Then in scenario b, payment from the retailer can be delayed for k periods without penalty, and there is no discount for early payment. The supplier needs to decide whether to extend trade credit in order to maximize her expected profit.

The retailer starts at initial size $s_0 > 0$ with a very limited initial fund $f_0 > 0$. We model his growing process in a multi-period setting. He makes expansion decision every n periods which we call a *year*. Let the planning horizon be T years which is long enough for the retailer to reach his target size. We number the n periods between the tth and the $(t+1)$th expansion with $(t, 1), \ldots, (t, n)$. At the beginning of year t, the retailer may choose to enlarge his size at unit cost e. Let s_t denote the expanded size of year t, then an operational cost us_t^2 is charged. Note that the operational cost is convex increasing in his size, reflecting the loss of efficiency as he becomes bigger. This prevents the retailer from growing excessively large. Unmet demand at the end of each period is lost. Leftover is carried to the next period in the same year but not to the next year. This assumption works well for perishable products. The unit inventory cost is h, satisfying $h + w < p$, otherwise the retailer would rather to discard leftover and order in the next period. Without loss of generality, the salvage value of inventory at the end of each year is 0.



(a) Scenario a: No trade credit



(b) Scenario b: Trade credit on terms of Net k

Figure 4.1: Sequence of Events in Period t

Retailer's Problem

For the tractability of the problem, we assume that the retailer is myopic. That is, he maximizes his profit in the current year without planning for the future. Actually, this is not uncommon among small businesses. Then the retailer's problem can be solved for each individual year. Once the retailer's size is fixed, the remaining inventory problem is a dynamic linear inventory problem. It has been proved that a base-stock policy is optimal, but the optimal base-stock levels are difficult to calculate. Here we assume that the retailer follows a stationary base-stock policy with base-stock level q_t. Then the retailer needs to decide q_t at the beginning of period 1 in order to maximize his expected profit over the year.

Figure 4.1(a) and 4.1(b) describe the sequence of events happened at the retailer in one period in scenario a and b, respectively. If the current period is the first one in a year, the retailer first enlarges his size to s_t and decides the base-stock level q_t. At the same time,

expansion cost and operational cost is charged. He then observes his beginning inventory and order up to q_t. We assume that there is no production or transportation lead time, so the retailer receives the order instantly. Purchase cost of the current period is charged in scenario a, but it can be delayed in scenario b. Afterwards, demand is realized, and the retailer sells the product to its customers and collects revenue. Then the retailer pays for some previous delayed purchase cost if in scenario b. Inventory cost is charged for any leftover if this is not the last period of a year. Leftover is salvaged otherwise. At last, the retailer checks his financial status and either moves on to the next period or goes bankrupt.

In scenario b where trade credit on terms of net k is offered, the retailer always pays k periods later because there is no incentive for him to pay early. Note that payment is not delayed by exact k periods in our model. Rather, it is delayed until revenue has been collected k times. Hence, when $k = 1$, the retailer pays for the order after selling the product in the same period.

Let $\Pi_t(s_t, q_t)$ denote the retailer's profit in year t given expanded size s_t and base-stock level q_t, then

$$\begin{aligned}\Pi_t(s_t, q_t) &= p\sum_{i=1}^{n}\min\{D_{t,i}(s_t), q_t\} - w^{(k)}\left[q_t + \sum_{i=2}^{n}\left(q_t - [q_t - D_{t,i-1}(s_t)]^+\right)\right] \\ &\quad -h\sum_{i=1}^{n-1}[q_t - D_{t,i}(s_t)]^+ - us_t^2 - e(s_t - s_{t-1}).\end{aligned}$$

where $w^{(k)}$ is the k-period discounted wholesale price, for $k = 0, 1, 2, \ldots$. Note that $w^{(k)}$ only represents the *mentally* discounted wholesale price if payment can be delayed for k periods. The actually wholesale price is still w even if trade credit is extended. Thus, $w^{(0)} = w$, and we assume that $w^{(k+1)} < w^{(k)}$ for all k. In other words, the longer the payment is delayed, the less weight is has when the retailer makes the planning. Then the retailer's expected profit is

$$\begin{aligned}\pi_t(s_t, q_t) &= -\left[np + (n-1)(h - w^{(k)})\right]\mathbb{E}[D_{t,i}(s_t) - q_t]^+ - \left[w^{(k)} + (n-1)h\right]q_t \\ &\quad -us_t^2 + \left[n\alpha p + (n-1)\alpha(h - w^{(k)}) - e\right]s_t + es_{t-1}.\end{aligned}$$

Given f_{t-1}, the retailer's fund at the end of last year, the following financial constraint needs to be satisfied in scenario a.

$$wq_t + es_t + us_t^2 \le f'_{t-1} \tag{4.1}$$

where $f'_{t-1} = f_{t-1} + es_{t-1}$. In scenario b, since purchase cost can be delayed, we have a looser constraint instead.

$$es_t + us_t^2 \le f'_{t-1} \tag{4.2}$$

The retailer's problem in year t is then modeled as

$$\begin{array}{ll} \max & \pi_t(s_t, q_t) \\ s.t. & \text{(4.1) in scenario } a\text{, or (4.2) in scenario } b \\ & s_t \geq s_{t-1}, q_t \geq 0 \end{array}$$

This optimization model is infeasible if $f_{t-1} < us_{t-1}^2$, i.e. the retailer's fund is not enough to pay for the total cost even if he neither expands his size nor orders anything from the supplier (in scenario a). In this case, the retailer has already gone bankrupt at the end of last period. Therefore, we always assume $f_{t-1} \geq us_{t-1}^2$ when solving the retailer's problem.

Lemma 13. *Financial constraint (4.2) can be replaced by a linear constraint*

$$s_t \leq \frac{\sqrt{e^2 + 4uf'_{t-1}} - e}{2u} := \bar{s}_t.$$

Thus, (4.2) is equivalent to imposing an upper bound $\bar{s}_t$ on s_t. Because (4.1) implies (4.2), $\bar{s}_t$ is also an upper bound on s_t in scenario b. In addition, $s_t \leq \bar{s}_t$ ensures the non-negativity of q_t.s

Lemma 14. *The size upper bound $\bar{s}_t$ is always greater than or equal to the retailer's starting size s_{t-1} if $f_{t-1} \geq us_{t-1}^2$.*

Lemma (14) guarantees the feasibility of the retailer's problem. In the rest of this section, we derive the retailer's expansion and order policies in both scenario a and b, and for both additive and multiplicative demand models.

Additive Demand Model

In the additive demand model, $D_{t,i}(s_t) = \alpha s_t + \epsilon_{t,i}$, where $\epsilon_{t,i} \sim^{iid} F$ and $E[\epsilon_{t,i}] = 0$, for $i = 1, \ldots, n$. The support of $\epsilon_{t,i}$ is $[-\alpha s_0, +\infty)$ to make demand non-negative at any size no less than the initial size. To make life easier, define $z_t = q_t - \alpha s_t$. Then the retailer's expected profit in period t can be conveniently expressed as a function of s_t and z_t.

$$\begin{aligned} & \pi_t(s_t, z_t) && (4.3) \\ = \; & -\left[np + (n-1)(h - w^{(k)})\right] \mathbb{E}[\epsilon_{t,i} - z_t]^+ - \left[w^{(k)} + (n-1)h\right] z_t \\ & -us_t^2 + [n\alpha(p - w^{(k)}) - e]s_t + es_{t-1} \\ = \; & \begin{cases} -\left[np + (n-1)(h - w^{(k)})\right] F^1(z_t) - \left[w^{(k)} + (n-1)h\right] z_t & \\ -us_t^2 + [n\alpha(p - w^{(k)}) - e]s_t + es_{t-1} & \text{if } z_t \geq -\alpha s_0 \\ n(p - w^{(k)})z_t - us_t^2 + \left[n\alpha(p - w^{(k)}) - e\right] s_t + es_{t-1} & \text{otherwise} \end{cases} && (4.4) \end{aligned}$$

Suppose s_t is given, and there is no financial constraint, then the optimal value of z_t is

$$z_\beta^{(k)} = F^{-1}(\beta^{(k)})$$

where $\beta^{(k)} = \frac{n(p-w^{(k)})}{np+(n-1)(h-w^{(k)})}$. $z_\beta^{(k)}$ is the unconstrained optimal value of z_t for the original problem. Later we will obtain the constrained optimal value of z_t and solve for the optimal expanded size for both scenarios.

Scenario *b*: Trade Credit on Terms of Net k When trade credit on terms of net $k(\geq 1)$ is offered, the retailer does not have to pay at the time of ordering, so z_t is not involved in the financial constraint. The optimal value of z_t is $z_t^* = z_\beta^{(k)}$ for any s_t. Now the retailer's expected profit in period t is only a function of s_t.

$$\begin{aligned}\pi_t(s_t, z_\beta^{(k)}) &= -us_t^2 + [n\alpha(p-w^{(k)}) - e]s_t - \left[np + (n-1)(h-w^{(k)})\right] F^1(z_\beta^{(k)}) \\ &\quad - \left[w^{(k)} + (n-1)h\right] z_\beta^{(k)} + es_{t-1}\end{aligned}$$

$\pi_t(s_t, z_\beta^{(k)})$ has a concave quadratic form. It is maximized at

$$\hat{s}_t^{(k)} = \frac{n\alpha(p-w^{(k)}) - e}{2u}.$$

It is notable that $\hat{s}_t^{(k)}$ has the same value in any period, so the subscript t is omitted. $\hat{s}^{(k)}$ is the optimal size of the retailer in scenario b. We assume that $n\alpha(p-w^{(k)}) - e > 0$ and $s_0 < \hat{s}^{(k)}$, otherwise there is no incentive for the retailer to expand.

The optimal values of s_t and z_t are

$$\begin{aligned} s_t^* &= \begin{cases} s_{t-1} & \text{if } \hat{s}^{(k)} < s_{t-1} \\ \hat{s}^{(k)} & \text{if } s_{t-1} \leq \hat{s}^{(k)} \leq \bar{s}_t \\ \bar{s}_t & \text{otherwise} \end{cases} \\ z_t^* &= z_\beta^{(k)} \end{aligned} \tag{4.5}$$

for $t = 1, \ldots, T$. The retailer's expansion policy in scenario b with additive demand is to stay at s_{t-1} if $s_{t-1} > \hat{s}^{(k)}$, to expand up to $\bar{s}_t$ if $\bar{s}_t < \hat{s}^{(k)}$, and to expand to the optimal size $\hat{s}^{(k)}$ otherwise. His order policy is to order $q_t^* = \alpha s_t^* + z_\beta^{(k)}$, for $t = 1, \ldots, T$.

Scenario *a*: No Trade Credit Financial constraint (4.1) needs to be satisfied when trade credit is not available. With z_t defined, (4.1) can be rewritten as

$$wz_t + (\alpha w + e)s_t + us_t^2 \leq f'_{t-1} \qquad (1)$$

It involves both s_t and z_t, thus the retailer needs to balance between size and base-stock level. Let $z_t^*(s_t)$ denote the optimal value of z_t for any given s_t. $z_t^*(s_t)$ takes value $z_\beta^{(0)}$ if s_t is small enough, but is restricted by (4.1) otherwise. Let s_t^f be the maximum value of s_t such that $z_t^*(s_t) = z_\beta^{(0)}$. Its value is given by

$$s_t^f = \begin{cases} \frac{\sqrt{(\alpha w+e)^2 + 4uf'_{t-1} - 4z_\beta^{(0)}wu} - (\alpha w+e)}{2u} & \text{if } f'_{t-1} \geq z_\beta^{(0)}w - (\alpha w+e)^2/(4u) \\ -\infty & \text{otherwise.} \end{cases}$$

The constrained optimal value of z_t is

$$z_t^*(s_t) = \begin{cases} z_\beta^{(0)} & \text{if } s_t \le s_t^f \\ \left[f'_{t-1} - (\alpha w + e)s_t - u s_t^2\right]/w =: z_t(s_t) & \text{otherwise.} \end{cases} \tag{4.6}$$

Plugging (4.6) back into (4.4), we can solve for the optimal expanded size. Because $\pi_t(s_t, z_t)$ takes different forms when z_t is in different regions, we need to check whether $z_t(s_t) \ge -\alpha s_0$. $z_t(s_t) \ge -\alpha s_0$ if and only if

$$s_t \le \frac{\sqrt{(\alpha w + e)^2 + 4u f'_{t-1} + 4\alpha w u s_0} - (\alpha w + e)}{2u} =: s_t^s$$

Hence, s_t^s is the maximum value of s_t such that $z_t(s_t) \ge -\alpha s_0$.

So far we have obtained the two threshold values of s_t, s_t^f and s_t^s. The relations among s_t^f, s_t^s, and $\bar{s}_t$ is described by the proposition below.

Proposition 15. *In the additive demand case, $s_t^f < s_t^s \le \bar{s}_t$, for any t.*

Therefore, the retailer's expected profit when $z_t = z_t^*(s_t)$ is

$$\pi_t(s_t, z_t^*(s_t)) = \begin{cases} g_1(s_t) & \text{if } s_t \le s_t^f \\ g_2(s_t) & \text{if } s_t^f \le s_t \le s_t^s \\ g_3(s_t) & \text{if } s_t \ge s_t^s \end{cases}$$

where

$$\begin{aligned} g_1(s_t) &= -u s_t^2 + [n\alpha(p-w) - e]s_t + \Big[e s_{t-1} - [np + (n-1)(h-w)]\, F^1(z_\beta^{(0)}) \\ &\quad - [w + (n-1)h]\, z_\beta^{(0)}\Big] \\ g_2(s_t) &= -[np + (n-1)(h-w)]\, F^1(z_t(s_t)) - [w + (n-1)h]\, z_t(s_t) - u s_t^2 \\ &\quad + [n\alpha(p-w) - e]\, s_t + e s_{t-1} \\ g_3(s_t) &= n(p-w) z_t(s_t) - u s_t^2 + [n\alpha(p-w) - e]\, s_t + e s_{t-1}. \end{aligned}$$

Lemma 16. *In the additive demand case, $g_1(\cdot)$ is strictly concave over $\mathbb{R}$, $g_2(\cdot)$ is strictly concave over $[s_t^f, \infty)$, and $g_3(\cdot)$ is strictly concave and decreasing over $[0, \infty)$.*

g_1 has exactly the same form as the expected profit in scenario b, so it is maximized at

$$\hat{s}^{(0)} = \frac{n\alpha(p-w) - e}{2u}.$$

Let $\check{s}_t$ be the unique value of s_t such that $g_2'(s_t) = 0$. There is no close-form solution for $\check{s}_t$ in general, but it can be efficiently obtained by (binary search).

Lemma 17. *In the additive demand case, $\pi_t(s_t, z_t^*(s_t))$ is continuously differentiable, for any t.*

Let $\tilde{s}_t$ be the unconstrained optimal expanded size of the retailer in period t with lower bound s_{t-1} and upper bound $\bar{s}_t$ omitted, then Lemma 16 and Lemma 17 lead to the following proposition.

Proposition 18. *In the additive demand case,*

$$\tilde{s}_t = \begin{cases} \check{s}_t & \text{if } s_t^f < \hat{s}^{(0)} \\ \hat{s}^{(0)} & \text{otherwise.} \end{cases}$$

for any t.

Corollary 19. *In the additive demand case, $\tilde{s}_t < \bar{s}_t$, for any t.*

The optimal values of s_t and z_t are

$$\begin{aligned} s_t^* &= \max\{s_{t-1}, \tilde{s}_t\} \\ z_t^* &= \begin{cases} z_\beta^{(0)} & \text{if } s_t^* \le s_t^f \\ z_t(s_t^*) & \text{otherwise.} \end{cases} \end{aligned} \tag{4.7}$$

Multiplicative Demand Model

In this section, we derive the retailer's expansion and order policies in both scenarios in the multiplicative demand case. Recall that the multiplicative demand model assumes that $D_{t,i}(s_t) = \alpha s_t \epsilon_{t,i}$, with $\epsilon_{t,i} \sim^{iid} F$ and $E[\epsilon_{t,i}] = 1$, for $i = 1, \ldots, n$. The support of $\epsilon_{t,i}$ is assumed to be $[0, \infty)$. As for the additive demand model, we define $z_t = q_t/(\alpha s_t)$ for mathematical convenience, and rewrite the retailer's expected profit in period t as a function of s_t and z_t.

$$\begin{aligned} \pi_t(s_t, z_t) &= -\alpha \left[np + (n-1)(h - w^{(k)})\right] s_t \mathbb{E}[\epsilon_{t,i} - z_t]^+ - \alpha \left[w^{(k)} + (n-1)h\right] s_t z_t - u s_t^2 \\ &\quad + \left[n\alpha p + (n-1)\alpha(h - w^{(k)}) - e\right] s_t + e s_{t-1} \\ &= -\alpha \left[np + (n-1)(h - w^{(k)})\right] s_t F^1(z_t) - \alpha \left[w^{(k)} + (n-1)h\right] s_t z_t - u s_t^2 \\ &\quad + \left[n\alpha p + (n-1)\alpha(h - w^{(k)}) - e\right] s_t + e s_{t-1} \end{aligned} \tag{4.8}$$

The unconstrained optimal value of z_t is also equal to $z_\beta^{(k)}$. We will show that the retailer's expansion and order policies in the multiplicative demand case also have similar structures as those in the additive demand case.

Scenario *b*: Trade Credit on Terms of Net k z_t is always unconstrained when trade credit is available, so $z_t^* = z_\beta^{(k)}$ for all t. With z_t set at optimal, the expected profit of the retailer in period t is

$$\begin{aligned}&\pi_t(s_t, z_\beta^{(k)}) = -us_t^2 \\ &+\left[\alpha\left[np+(n-1)(h-w^{(k)})\right]\left(1-F^1(z_\beta^{(k)})\right) - z_\beta^{(k)}\alpha\left[w^{(k)}+(n-1)h\right] - e\right]s_t \\ &+es_{t-1}.\end{aligned}$$

It is maximized at

$$\hat{s}^{(k)} = \frac{\alpha\left[np+(n-1)(h-w^{(k)})\right]\left(1-F^1(z_\beta^{(k)})\right) - z_\beta^{(k)}\alpha\left[w^{(k)}+(n-1)h\right] - e}{2u}.$$

Without loss of generality, we assume that $\hat{s}^{(k)} > s_0$.

Then the retailer's constrained optimal expanded size in period t is

$$s_t^* = \begin{cases} s_{t-1} & \text{if } \hat{s}^{(k)} < s_{t-1} \\ \hat{s}^{(k)} & \text{if } s_{t-1} \le \hat{s}^{(k)} \le \bar{s}_t \\ \bar{s}_t & \text{otherwise} \end{cases} \tag{4.9}$$

To sum up, the retailer's expansion policy in scenario b with multiplicative demand is the same as that with additive demand. His order policy in this case is to order $q_t^* = z_\beta^{(k)}\alpha s_t^*$, for $t = 1, \ldots, T$.

Scenario *a*: No Trade Credit When no trade credit is provided, the retailer's base-stock level, or equivalently, z_t, might be restricted by financial constraint (4.1). It is required that

$$w\alpha s_t z_t + us_t^2 + es_t \le f'_{t-1}.$$

It follows that

$$z_t^*(s_t) = \begin{cases} z_\beta^{(0)} & \text{if } s_t \le s_t^f \\ z_t(s_t) = \frac{f'_{t-1} - us_t^2 - es_t}{\alpha w s_t} & \text{otherwise} \end{cases}$$

where

$$s_t^f = \frac{\sqrt{(z_\beta^{(0)}\alpha w + e)^2 + 4uf'_{t-1}} - (z_\beta^{(0)}\alpha w + e)}{2u}$$

is the maximum value of s_t such that $z_t^*(s_t)$ is not restricted by the financial constraint. Note that $z_t^*(s_t)$ is non-negative whenever $s_t \le \bar{s}_t$, so it is always feasible and in the support of ϵ_t.

Analogous to the additive demand case, we have the following proposition.

Proposition 20. *In the multiplicative demand case, $s_t^f < \bar{s}_t$, for any t.*

When $z_t = z_t^*(s_t)$, the retailer's expected profit is

$$\pi(s_t, z_t^*(s_t)) = \begin{cases} g_1(s_t) & \text{if } s_t \le s_t^f \\ g_2(s_t) & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} g_1(s_t) &= -us_t^2 \\ &\quad + \left[\alpha\left[np + (n-1)(h-w)\right]\left(1 - F^1(z_\beta^{(0)})\right) - z_\beta^{(0)}\alpha\left[w + (n-1)h\right] - e\right] s_t \\ &\quad + es_{t-1} \\ g_2(s_t) &= -\alpha\left[np + (n-1)(h-w)\right] s_t F^1(z_t(s_t)) + \frac{(n-1)hu}{w}s_t^2 \\ &\quad + \left[\alpha\left[np + (n-1)(h-w)\right] + \frac{(n-1)h}{w}e\right] s_t - \left[f_{t-1} + \frac{(n-1)h}{w}f'_{t-1}\right] \end{aligned}$$

Lemma 21. *In the multiplicative demand case, $g_1(\cdot)$ is strictly concave on $\mathbb{R}$ and $g_2(\cdot)$ is strictly concave on $[s_t^f, \infty)$.*

It has been solved in scenario a that g_1 is maximized at

$$\hat{s}^{(0)} = \frac{\alpha\left[np + (n-1)(h-w)\right]\left(1 - F^1(z_\beta^{(0)})\right) - z_\beta^{(0)}\alpha\left[w + (n-1)h\right] - e}{2u}.$$

Let $\check{s}_t$ be the unique value of s_t such that $g_2'(s_t) = 0$, then g_2 is maximized at $\check{s}_t$.

Lemma 22. *In the multiplicative demand case, $\pi_t(s_t, z_t^*(s_t))$ is continuously differentiable, for any t..*

By Lemma 21 and Lemma 22, $\pi(s_t, z_t^*(z_t))$ is overall concave in s_t and has a unique maximum. Let $\tilde{s}_t$ be the unconstrained optimal expanded size of the retailer in period t with lower bound s_{t-1} and upper bound $\bar{s}_t$ omitted, then we have the following result.

Proposition 23. *In the multiplicative demand case,*

$$\tilde{s}_t = \begin{cases} \check{s}_t & \textit{if } s_t^f < \hat{s}^{(0)} \\ \hat{s}^{(0)} & \textit{otherwise.} \end{cases}$$

for any t.

Corollary 24. *In the multiplicative demand case, $\tilde{s}_t < \bar{s}_t$, for any t.*

The optimal values of s_t and z_t are

$$\begin{aligned} s_t^* &= \max\{s_{t-1}, \tilde{s}_t\} \\ z_t^* &= \begin{cases} z_\beta^{(0)} & \text{if } s_t^* \le s_t^f \\ z_t(s_t^*) & \text{otherwise.} \end{cases} \end{aligned} \tag{4.10}$$

Discussion on Retailer's Growing Process

In this section, we summarize and compare the retailer's growing processes in scenario a and scenario b from several aspects.

Expansion and inventory policies. The retailer essentially follows the same expansion policy in all cases: he expands to an unconstrained optimal size $\tilde{s}_t$ if $\tilde{s}_t$ is between his starting size s_{t-1} and the size upper bound $\bar{s}_t$, expands to $\bar{s}_t$ if $\tilde{s}_t$ exceeds the size upper bound, and stays at s_{t-1} otherwise. We call this a *bounded expand-up-to policy*. However, $\tilde{s}_t$ is different from scenario to scenario. In scenario a, $\tilde{s}_t = \check{s}_t$ in all periods except the last period of his growing process in which $\tilde{s}_t = \hat{s}^{(k)}$. While in scenario b, $\tilde{s}_t$ always equals to $\hat{s}^{(0)}$.

The retailer has different base-stock levels in scenario a and scenario b. In scenario b, he always orders up to the unconstrained base-stock level $q_t^* = \alpha s_t^* + z_\beta^{(k)}$ (in the additive case) or $q_t^* = \alpha s_t^* z_\beta^{(k)}$ (in the multiplicative case). In scenario a, he orders up to the restricted base-stock level $q_t^* = \alpha s_t^* + z_t(s_t^*)$ (in the additive case) or $q_t^* = \alpha s_t^* z_t(s_t^*)$ (in the multiplicative case) when his expanded size s_t^* exceeds s_t^f, and orders up to the unconstrained base-stock level $q_t^* = \alpha s_t^* + z_\beta^{(0)}$ (in the additive case) or $q_t^* = \alpha s_t^* z_\beta^{(0)}$ (in the multiplicative case) otherwise.

Growing manner and speed. We first look at the retailer's growing manner in scenario b. By (4.5) and (4.9), we see that s_t^* never exceeds $\hat{s}^{(k)}$ as long as $s_0 < \hat{s}^{(k)}$. The retailer expands up to $\bar{s}_t$ in every period when $\bar{s}_t$ is less than $\hat{s}^{(k)}$. Once $\bar{s}_t \geq \hat{s}^{(k)}$ is satisfied in some period, the retailer expands to $\hat{s}^{(k)}$ and ends his growing process.

In scenario a, the retailer may expand to $\check{s}_t$ in some periods. One immediate question is whether it is possible for the retailer to grow over $\hat{s}^{(0)}$. The answer is no. To see this, we start from the following lemma.

Lemma 25. *In scenario* a, $\check{s}_t < \hat{s}^{(0)}$ *if* $s_t^f < \hat{s}^{(0)}$.

Lemma 25 shows that the retailer's size can not exceeds $\hat{s}^{(0)}$ in any year in which he expands to $\check{s}_t$. We define the first year in which $s_t^f \geq \hat{s}^{(0)}$ is satisfied as the *mature year* and all the previous periods the *developing years*. We will see from the next proposition that these names accurately describe the two stages of the retailer's growing process.

Proposition 26. *In scenario* a, *the retailer expands to* $\max\{s_{t-1}, \check{s}_t\}$ *in every developing year. His size is always less than* $\hat{s}^{(0)}$ *during these years. In the mature year, he expands to* $\hat{s}^{(0)}$ *and finishes his growing process.*

By Proposition 26, we have the following two corollaries.

Corollary 27. *The retailer expands more aggressively in scenario* b *than in scenario* a*. That is, in a certain year, he expands more in scenario* b *than in scenario* a *given that the starting size and fund are the same.*

Corollary 28. *In scenario* a*, the retailer orders up to the restricted base-stock level* $q_t^* = \alpha s_t^* + z_t(s_t^*)$ *in the additive case or* $q_t^* = \alpha s_t^* z_t(s_t^*)$ *in the multiplicative case in developing years, and orders up to the unconstrained base-stock level* $q_t^* = \alpha s_t^* + z_\beta^{(0)}$ *in the additive case or* $q_t^* = \alpha s_t^* z_\beta^{(0)}$ *in the multiplicative case in the mature year.*

Corollary 28 greatly simplifies the retailer's order policy in scenario *a*.

Target size. No matter under additive or multiplicative demand model, there exists a constant size $\hat{s}^{(k)}$ that has been shown to be the retailer's target size, i.e. the size that the retailer grows towards and stays at once reaching it. The target size takes different values under different demand models. More specifically,

$$\hat{s}^{(k)} = \frac{n\alpha(p - w^{(k)}) - e}{2u}$$

in the additive demand case, while

$$\begin{aligned}\hat{s}^{(k)} \;=\; & \frac{1}{2u}\Big\{\alpha\left[np + (n-1)(h - w^{(k)})\right] - \alpha\left[np + (n-1)(h - w^{(k)})\right] F^1(z_\beta^{(k)}) \\ & - z_\beta^{(k)}\alpha\left[w^{(k)} + (n-1)h\right] - e\Big\}\end{aligned}$$

in the multiplicative demand case, for $k = 0, 1, \ldots$. It is notable that $\hat{s}^{(k)}$ is determined only by α, n, p, $w^{(k)}$, e and u in the additive demand case, while it also depends on h and the demand distribution in the multiplicative demand case. We also conduct a sensitivity analysis on $\hat{s}^{(k)}$ and obtain the following proposition.

Proposition 29. $\hat{s}^{(k)}$ *is increasing in* α *and* p*, and is decreasing in* $w^{(k)}$*,* h*,* e*, and* u*. for* $k = 0, 1, 2, \ldots$.

Proposition 29 is consistent with our intuition: it is better for the retailer to have bigger size if his size has more impact on the demand or if the product is more profitable, and it is better to have a smaller size if it is more expensive to expand or to maintain his size. Because $w^{(k+1)} > w^{(k)}$ for $k = 0, 1, 2, \ldots$, the retailer's target size in scenario b is always larger than that in scenario a, and the longer the payment can be delayed, the larger the target size is.

4.3 Numerical Studies

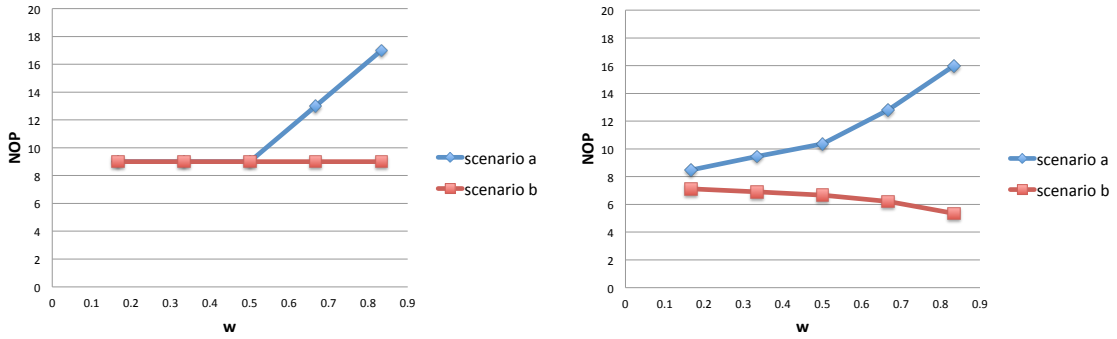
It has been proved in last section that the retailer grows faster with trade credit, so trade credit is favorable for the retailer in this sense. We further examine the retailer's growing speed in this section through numerical experiments. On the other hand, since demand is random, it is possible for the retailer to go bankrupt before he expands to his target size. Thus, it is also necessary to check whether trade credit affects the chance that the retailer reaches his target size. This issue is also addressed in this section. In these experiments, we also simulate the behavior of the supplier in order to evaluate trade credit's impact on her profitability.

Experiment Settings

We test the system in various settings. Each setting is characterized by: (1) availability of trade credit (no trade credit or trade credit on terms of Net 1), (2) demand model (additive or multiplicative), and (3) other parameters (costs, prices, initial size and fund, etc.). For each setting, we generate 5000 sample paths. Each sample path consists of 500 years which is long enough for the retailer to reach his target size. $\epsilon_{t,i}$'s are generated independently according to a uniform distribution (on $[-\alpha s_0, \alpha s_0]$ in the additive case and on $[0, 2]$ in the multiplicative case).

For each sample path, the simulation terminates if $T > 5000$ or if the retailer goes bankrupt. If the retailer is operating until the end of simulation horizon, he pays the supplier all unpaid amount if there is any. If the retailer goes bankrupt, we assume that he ends up with nothing. If his end fund is not enough to pay for all the unpaid amount, he can liquify his asset by selling it at unit price $e_0 (< e)$. We detect bankruptcy as follows. The retailer goes bankrupt at the end of a year if his ending fund is not enough to pay the operational cost of next year even if he would not expand. He goes bankrupt at the end of a period in the middle of a year if his ending fund is not enough to pay the purchasing cost of next period (in scenario a) or if his ending fund is negative (in scenario b).

Our model involves a large number of parameters, but it is neither practical nor necessary to change all of them in simulation. First, we fix α, p, and s_0 at 1, because every system can be transformed into a system with $\alpha = 1$, $p = 1$, and $s_0 = 1$ by re-scaling parameters properly. For analytical simplicity, we also set $n = 4$, $h = 0.1$, $e_0 = e/2$, and $c = p/3$. w changes between 0 and $p - h$. e and u are set at values such that $\hat{s}^{(0)} = 100$. After all the above parameters are determined, f_0 takes 10 values evenly spaced in the interval $[f_{0,min}, f_{0,max}]$, where $f_{0,min} = us_0^2$, the minimum value of f_0 such that the retailer does not go bankrupt at the beginning, and $f_{0,max} = u(\hat{s}^{(k)})^2 + e(\hat{s}^{(k)} - s_0)$, the minimum value of f_0 such that the retailer is able to extend to his target size in the first period in scenario b. For



(a) Additive demand model (b) Multiplicative demand model

Figure 4.2: NOP of the Retailer with $f_{0,index} = 4$

each f_0, we define

$$f_{0,index} = \frac{10(f_0 - f_{0,min})}{f_{0,max} - f_{0,min}} + 1$$

as its index, so $f_{0,min}$ has index 1, and $f_{0,max}$ has index 10. By using the index of f_0, we are able to refer to the same level of initial fund under different parameters.

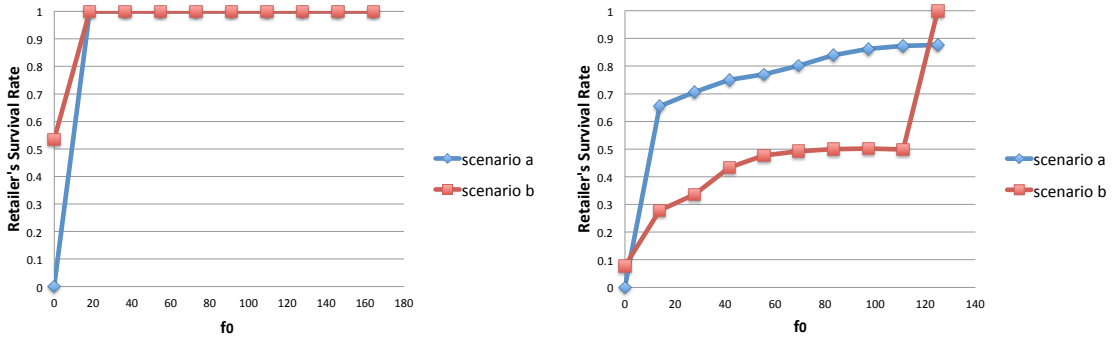
Retailer's Growing Speed

We already know that the retailer grows faster in scenario b than in scenario a. We are also interested in by how much faster he grows with trade credit. Since $\hat{s}^{(0)}$ is fixed at 100, we measure the growing speed of the retailer through the number of periods needed for him to reach his target size (NOP). The average NOP in scenario a and that in scenario b with $f_{0,index} = 4$ are plotted in Figure 4.2 as functions of w. As expected, the retailer always needs less periods to reach his target size in scenario b than in scenario a no matter which demand model is used.

We also observe that NOP changes as w increases. However, while NOP increases convexly in scenario a, it stays constant or decreases slightly in scenario b. Hence, the retailer's growing speed is very sensitive to the change in the wholesale price in scenario a, but is quite robust in scenario b. Moreover, although $\hat{s}^{(0)}$ is maintained at 100, $\hat{s}^{(1)}$ actually increases as w increases. It increases from 103 to 175 in the additive demand case and from 104.1 to 248.3 in the multiplicative demand case when w increases from 1/6 to 5/6. Therefore, the retailer grows much faster in scenario a than in scenario b.

Retailer's Survival Rate

We define the *survival rate* of the retailer as the probability that he reaches his target size before going bankrupt. In this section, we investigate the impacts of trade credit on the



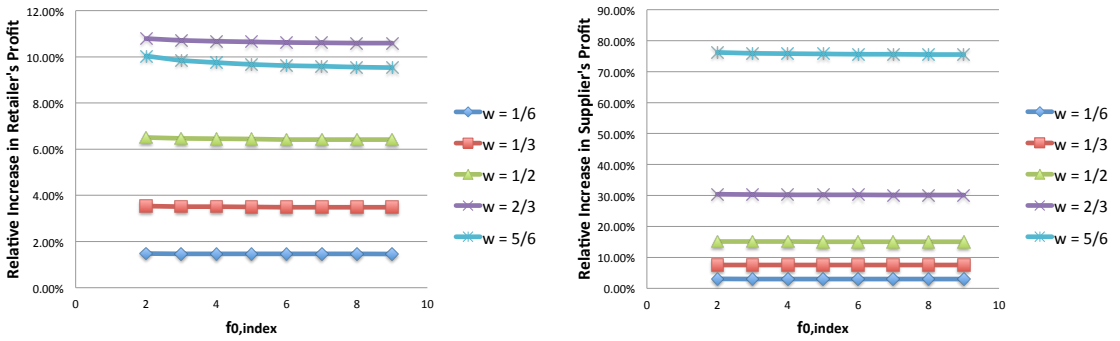
(a) Additive demand model (b) Multiplicative demand model

Figure 4.3: Retailer's Survival Rate with $w = 1/2$

retailer's survival rate. Figure 4.3 displays the retailer's survival rate when $w = 1/2$. In the additive demand case, the survival rate of the retailer is always 1 except for the case that $f_0 = f_{0,min}$. This means that there is little chance for the retailer to fail when demand is negatively correlated no matter trade credit is used or not. In the multiplicative demand case, survival rate is much lower, and the retailer is more likely to survive in scenario a than in scenario b in most cases. The opposite happens only when f_0 equals $f_{0,min}$ or $f_{0,max}$. Ignoring the two extremes cases of f_0, we find that the retailer's survival rate when he uses trade credit is within 0.5. That is, with more than a half chance he fails to reach his target size. Because only failures before target size is reached is counted here, we expect that the retailer's overall survival rate over the simulation horizon is even lower. We also observe an increasing trend in the retailer's survival rate with respect to f_0, which implies that the retailer is more likely to survive if he has more initial fund.

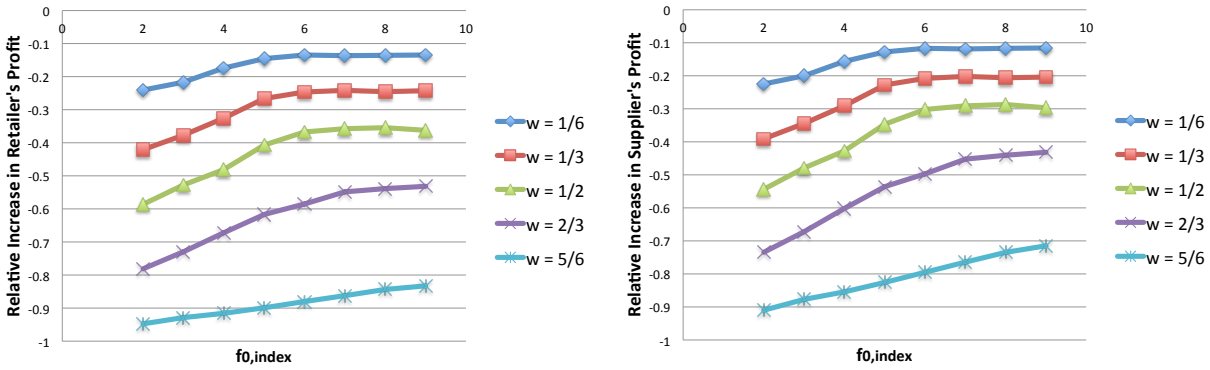
Retailer's and Supplier's Profits

It is shown that the retailer grows towards a larger target size with a faster speed when trade credit is available, so he should be able to achieve a higher profit over the simulation horizon if he survives. Nevertheless, it is also shown in the last study that his survival rate might be lower at the same time, so his overall profit is not necessarily higher when trade credit is available. Similar things happen with the supplier. If the retailer survives, because the retailer has a bigger size and always orders the unconstrained order quantity in scenario b, the supplier sells more in this scenario, and thereby makes more profit. However, she may also lose profit due to trade credit. One risk of trade credit comes directly from nonpayment – the retailer fails to pay all the unpaid amount when he goes bankrupt. Another potential risk is that the retailer might be more likely to go bankrupt in an early stage so that the supplier can not make any profit after that.



(a) Retailer (b) Supplier

Figure 4.4: Relative Increase in Retailer's and Supplier's Profit in the Additive Demand Case



(a) Retailer (b) Supplier

Figure 4.5: Relative Increase in Retailer's and Supplier's Profit in the Multiplicative Demand Case

We estimate the retailer's and the supplier's expected profits in scenario a and scenario b with both additive and multiplicative demand models, and calculate the relative increases in their profits from scenario a to scenario b. The relative increase in profit with additive demand model is displayed in Figure 4.4 and that with multiplicative demand model in Figure 4.5. In the additive demand case, because the retailer survives with probability 1, both the retailer and the supplier have higher profit in scenario b than in scenario a. On the contrary, we observe decreases in retailer's and supplier's profits in the multiplicative demand case. This implies that, when demand is positively correlated, trade credit may lower the profits of both the retailer and the supplier. Another interesting result is, the supplier's and the retailer's profit always change in the same direction when trade credit is extended. More specifically, the supplier gets more profit with trade credit if and only if the retailer also gets more profit. So there is opportunity for them to reach a win-win agreement on the terms of

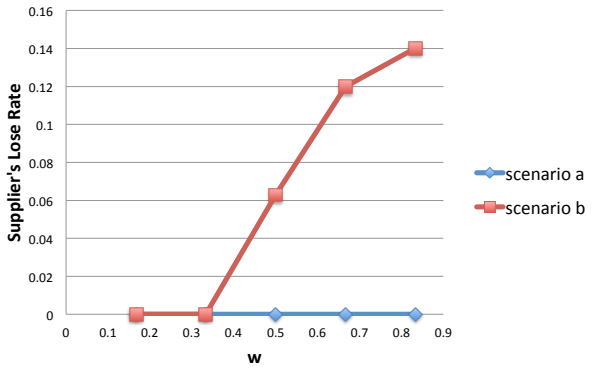


Figure 4.6: Supplier's Lose Rate with $f_{0,index} = 4$ in the Multiplicative Demand Case

trade credit. At last, figure 4.4 also shows that the value of f_0 has little effect on the relative increase in profit in the additive demand case, but can mitigate the risk of trade credit in the multiplicative demand case.

We also investigate whether trade credit may cause the bankruptcy of the supplier when she is a small business herself. Our model verifies the findings in [10]. We define the *lose rate* of the supplier as the probability that she gets negative profit over the simulation horizon. Her lose rate is always zero in the additive demand case because the retailer almost never fails in that case. However, positive lose rate is observed under multiplicative demand model. Figure 4.6 plots the supplier's lose rate with $f_{0,index} = 4$ in the multiplicative demand case. It is notable that the supplier may lose money with positive probability in scenario b. Therefore, when the supplier is also financially constrained, she may go bankrupt because of the failure of the retailer if she extends trade credit to him. Moreover, if the supplier also uses trade credit, she may further cause the failure of her credit provider.

4.4 Summary

In this chapter, we study the impact of trade credit on the growth of small businesses and their suppliers through a mathematical model. In the model, we consider a supply chain in which a single retailer seeks to expand his size and a single supplier seeks to create new business and gain long-term profit. The retailer is financially constrained and has no access to bank loans. The supplier may choose to extend trade credit to the retailer in order to help him expand his size and potentially increase her own sales. Demand from end users is assume to be random and depend on the retailer's size. The retailer's problem is solved analytically under both the additive and the multiplicative demand models. It is proved that the retailer grows faster and orders more with trade credit. Numerical studies show that both the supplier and the retailer benefit from trade credit in the additive demand

case. They achieve higher profit when trade credit is extended. On the contrary, in the multiplicative demand case, the retailer's survival rate is lower when he uses trade credit, and thereby his overall profit is also lower. At the same time, the supplier loses profit by extending trade credit, and some times her profit is even negative.

We get the following major insights from our model. First of all, trade credit facilitates the growth of small businesses and can stabilize their growth speed when wholesale price changes a lot. Second, using or providing trade credit is more risky if demand is positively correlated. It not only makes buyer more likely to go bankrupt, but can also cause the failure of suppliers. As last, the effect of trade credit is dependent on demand correlation. Thus, suppliers need to carefully investigate the market before making decisions.

There are several directions for future work. For example, we only compare scenario b when $k = 1$ with scenario a in numerical studies. It is not clear how the impact of trade credit changes as k increases. By carrying out numerical studies with more general k values, researchers will be able to get some insight on the design of trade credit terms, and to explain why net 30 and net 10 are commonly used in practice. It is also interesting to study how business failures diffuse in a more complex supply chain where trade credit is extended by every node to its downstream.

Appendix A

Proofs of Chapter 2

The proofs for all lemmas, theorems and important statements are given in this section. Since some of them may be used in multiple settings, we prove each of them in the most general setting. That is, we prove Claim 2 and Theorem 3 for (P_t^I), Lemma 4 for (P_t^{II}), Lemma 1, Theorem 6, and Theorem 7 for (PS_t), Proposition 8 and Proposition 9 for problem (PG), and Proposition 10 and Proposition 11 for uncapacitated problems. The proofs of Theorem 5 is not given here, because it is a special version of Theorem 6.

Proof of Lemma 1. Suppose Y is an optimal solution to (PS_t), and $k_j = \sum_{i \in S_j} Y_{ij}$ is the number of suppliers connected with retailer j in this solution. So $\sum_{j=1}^{M} k_j = t$, and $0 \leq k_j \leq u_j$, for any j. Construct another solution Y^*, s.t. $Y^*_{ij} = 1$ for $j = 1, \ldots, M$, $i = i^j_{(1)}, \ldots, i^j_{(k_j)}$, and all other Y^*_{ij}'s equal to 0. Then Y^* is a feasible solution to (PS_t) because $\sum_{j=1}^{M} \sum_{i \in S_j} Y^*_{ij} = \sum_{j=1}^{M} k_j = t$. In addition, because the suppliers in each supplier set is ranked in increasing order of failure probability, we have $\prod_{i \in S_j} q_i^{Y^*_{ij}} \leq \prod_{i \in S_j} q_i^{Y_{ij}}$ for any j. Thus, Y^* is at least as good as Y, and thereby optimal. Therefore, we can always find an optimal solution to (PS_t) in which, if a retailer is supplied by k suppliers, then it must be supplied by the first k suppliers in its supplier set. □

Proof of Claim 2. Consider the function $g(q_1) = q_1^k - kq_1 + k - 1$. g is convex when $q_1 \geq 0$, and $g'(q_1) = kq_1^{k-1} - k$. By first order condition, g achieves its minimum 0 at $q_1 = 1$. Thus, $q_1^k - kq_1 + k - 1 \geq 0$, for any $q_1 \geq 0$. Also, by the assumption $q_1^k \leq q_1 q_2 \ldots q_k$, we have

$$q_1 q_2 \ldots q_k - kq_1 + k - 1 \geq q_1^k - kq_1 + k - 1 \geq 0$$

and equivalently

$$1 - q_1 \geq \frac{1 - q_1 q_2 \ldots q_k}{k}.$$

□

Proof of Theorem 3. To prove Theorem 3, we need to prove the following lemma.

Lemma 30. *There exists an optimal solution to (PS_t^I) in which only one or two successive L_k's are positive.*

Proof of Lemma 30. Suppose L^A is an optimal solution to (PS_t^I). Let $k_1^A = \min\{k : L_k^A > 0\}$, $k_2^A = \max\{k : L_k^A > 0\}$. If $k_2^A - k_1^A \leq 1$, then we are done. Otherwise, define $l_{min} = \min\{L_{k_1^A}^A,\ L_{k_2^A}^A\}$. If $k_1^A + k_2^A$ is even, let $c = (k_1^A + k_2^A)/2$. Now we construct another solution L^B s.t. $L_{k_1^A}^B = L_{k_1^A}^A - l_{min}$, $L_c^B = L_c^A + 2l_{min}$, $L_{k_2^A}^B = L_{k_2^A}^A - l_{min}$, and $L_k^B = L_k^A$, for any $k \neq k_1^A, c, k_2^A$. Then L^B satisfies

$$\sum_{k=1}^{N} L_k^B = \sum_{k=1}^{N} L_k^A \leq M$$

$$\sum_{k=1}^{N} k \cdot L_k^B = \sum_{k=1}^{N} k \cdot L_k^A - k_1^A \cdot l_{min} + c \cdot 2l_{min} - k_2^A \cdot l_{min} = \sum_{k=1}^{N} k \cdot L_k^A = t.$$

So L^B is a feasible solution to (PS_t^I). The increase in objective function value is

$$\begin{aligned}
Obj(L^B) - Obj(L^A) &= -\left(1 - \prod_{i=1}^{k_1^A} q_i\right) l_{min} + \left(1 - \prod_{i=1}^{c} q_i\right) 2l_{min} - \left(1 - \prod_{i=1}^{k_2^A} q_i\right) l_{min} \\
&= l_{min} \left(\prod_{i=1}^{k_1^A} q_i\right) \left(1 + \prod_{i=k_1^A+1}^{k_2^A} q_i - 2 \prod_{k_1^A+1}^{c} q_i\right) \\
&\geq l_{min} \left(\prod_{i=1}^{k_1^A} q_i\right) \left(1 + \left(\prod_{k_1^A+1}^{c} q_i\right)^2 - 2 \prod_{k_1^A+1}^{c} q_i\right) \\
&= l_{min} \left(\prod_{i=1}^{k_1^A} q_i\right) \left(1 - \prod_{k_1^A+1}^{c} q_i\right)^2 \\
&\geq 0
\end{aligned}$$

If $k_1^A + k_2^A$ is odd, let $c_1 = \lfloor (k_1^A + k_2^A)/2 \rfloor$, $c_2 = \lceil (k_1^A + k_2^A)/2 \rceil$. Construct L^B s.t. $L_{k_1^A}^B = L_{k_1^A}^A - l_{min}$, $L_{c_1}^B = L_{c_1}^A + l_{min}$, $L_{c_2}^B = L_{c_2}^A + l_{min}$, $L_{k_2^A}^B = L_{k_2^A}^A - l_{min}$, and $L_k^B = L_k^A$, for any $k \neq k_1^A, c_1, c_2, k_2^A$. Then L^B satisfies

$$\sum_{k=1}^{N} L_k^B = \sum_{k=1}^{N} L_k^A \leq M$$

$$\sum_{k=1}^{N} k \cdot L_k^B = \sum_{k=1}^{N} k \cdot L_k^A - k_1^A \cdot l_{min} + c_1 \cdot l_{min} + c_2 \cdot l_{min} - k_2^A \cdot l_{min} = \sum_{k=1}^{N} k \cdot L_k^A = t.$$

So L^B is a feasible solution to (PS_t^I). The increase in objective function value is

$$\begin{aligned}
Obj(L^B) - Obj(L^A) &= -\left(1-\prod_{i=1}^{k_1^A} q_i\right) l_{min} + \left(1-\prod_{i=1}^{c_1} q_i\right) l_{min} + \left(1-\prod_{i=1}^{c_2} q_i\right) l_{min} \\
&\quad - \left(1-\prod_{i=1}^{k_2^A} q_i\right) l_{min} \\
&= l_{min}\left(\prod_{i=1}^{k_1^A} q_i\right)\left(1+\prod_{i=k_1^A+1}^{k_2^A} q_i - \prod_{k_1^A+1}^{c_1} q_i - \prod_{k_1^A+1}^{c_2} q_i\right) \\
&\geq l_{min}\left(\prod_{i=1}^{k_1^A} q_i\right)\left(1+\left(\prod_{k_1^A+1}^{c_1} q_i\right)^2 - 2\prod_{k_1^A+1}^{c_1} q_i\right) \\
&= l_{min}\left(\prod_{i=1}^{k_1^A} q_i\right)\left(1-\prod_{k_1^A+1}^{c_1} q_i\right)^2 \\
&\geq 0
\end{aligned}$$

To sum up, we can always find another optimal solution L^B. What is more, if we define $k_1^B = \min\{k : L_k^B > 0\}$, $k_2^B = \max\{k : L_k^B > 0\}$, then $k_2^B - k_1^B < k_2^A - k_1^A$. By repeating this procedure for finitely many times, we can find an optimal solution in which at most two successive L_k's are positive. This completes the proof of Lemma 30.
It is easy to show that, if $a = \max\{k \in \mathbb{Z} : kM \leq t\}$, $b = t - aM$, then $L_a = M - b$, $L_{a+1} = b$, and $L_k = 0$, for any $k \neq a, a+1$ is the only feasible solution that satisfies the condition of Lemma 30. Therefore, it must be optimal for (PS_t^I). □

Proof of Lemma 4. Let K^A be an optimal solution to (PS_t^{II}). If for any l, j s.t. $\mu_l \leq \mu_j$ and $K_j^A < u_j$, we have $K_l^A \leq K_j^A$, then we are done. Otherwise, suppose there exist j_1, j_2, s.t. $K_{j_1}^A < u_{j_1}$, $\mu_{j_1} \geq \mu_{j_2}$, and $K_{j_1}^A < K_{j_2}^A$. We construct K^B s.t. $K_{j_1}^B = K_{j_1}^A + 1$, $K_{j_2}^B = K_{j_2}^A - 1$, and $K_j^B = K_j^A$, for any $j \neq j_1, j_2$. Then K^B satisfies

$$\begin{aligned}
&\sum_{j=1}^{M} K_j^B = \sum_{j=1}^{M} K_j^A = t \\
&0 \leq K_j^B \leq u_j, integer \qquad j = 1, \ldots, M
\end{aligned}$$

So K^B is a feasible solution to (PS_t^{II}). The increase in objective function value is

$$\begin{aligned} Obj(K^B) - Obj(K^A) &= \sum_{j=1}^{M} \mu_j \left(1 - \bar{q}^{K_j^B}\right) - \sum_{j=1}^{M} \mu_j \left(1 - \bar{q}^{K_j^A}\right) \\ &= \left[\mu_{j_1} \left(1 - \bar{q}^{K_{j_1}^B}\right) + \mu_{j_2} \left(1 - \bar{q}^{K_{j_2}^B}\right)\right] \\ &\quad - \left[\mu_{j_1} \left(1 - \bar{q}^{K_{j_1}^A}\right) + \mu_{j_2} \left(1 - \bar{q}^{K_{j_2}^A}\right)\right] \\ &= (1 - \bar{q}) \left(\mu_{j_1} \bar{q}^{K_{j_1}^A} - \mu_{j_2} \bar{q}^{K_{j_2}^A - 1}\right) \end{aligned}$$

Given $K_{j_1}^A < K_{j_2}^A$, we have $K_{j_1}^A \leq K_{j_2}^A - 1$, and thus $\bar{q}^{K_{j_1}^A} \geq \bar{q}^{K_{j_2}^A - 1}$. We also have $\mu_{j_1} \geq \mu_{j_2}$. Hence, the increase in objective function value is nonnegative. K^B is also an optimal solution to (PS_t^{II}). By repeating this procedure for finitely many times, we can always find an optimal solution K^*, s.t., if $K_j^* < u_j$, $\mu_l \leq \mu_j$, then $K_l^* \leq K_j^*$, for any l, j. □

Proof of Theorem 6. Suppose K is an optimal solution to (PS_t) with $t > 0$. If $K_{\hat{j}} \geq 1$, then we are done. Otherwise, there exists $\tilde{j} \neq \hat{j}$ s.t. $K_{\tilde{j}} \geq 1$, because K needs to satisfy $\sum_{j=1}^{M} K_j = t \geq 1$. We construct another solution K^* s.t. $K_{\hat{j}}^* = 1$, $K_{\tilde{j}}^* = K_{\tilde{j}} - 1$, and $K_j^* = K_j$ for all other j's. Obviously, K^* is a feasible solution to PS_t. The increase in objective function value is

$$\begin{aligned} Obj(K^*) - Obj(K) &= \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j^*} q_{i_{(k)}^j}\right) - \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j} q_{i_{(k)}^j}\right) \\ &= \mu_{\hat{j}} \left(1 - q_{i_{(1)}^{\hat{j}}}\right) - \mu_{\hat{j}}(1-1) + \mu_{\tilde{j}} \left(1 - \prod_{k=1}^{K_{\tilde{j}}-1} q_{i_{(k)}^{\tilde{j}}}\right) - \mu_{\tilde{j}} \left(1 - \prod_{k=1}^{K_{\tilde{j}}} q_{i_{(k)}^{\tilde{j}}}\right) \\ &= \mu_{\hat{j}} \left(1 - q_{i_{(1)}^{\hat{j}}}\right) - \mu_{\tilde{j}} \left(1 - q_{i_{(K_{\tilde{j}})}^{\tilde{j}}}\right) \prod_{k=1}^{K_{\tilde{j}}-1} q_{i_{(k)}^{\tilde{j}}} \end{aligned}$$

By the definition of $\hat{j}$, and the order of suppliers in supplier sets, we have

$$\mu_{\hat{j}} \left(1 - q_{i_{(1)}^{\hat{j}}}\right) \geq \mu_{\tilde{j}} \left(1 - q_{i_{(1)}^{\tilde{j}}}\right) \geq \mu_{\tilde{j}} \left(1 - q_{i_{(K_{\tilde{j}})}^{\tilde{j}}}\right) \geq \mu_{\tilde{j}} \left(1 - q_{i_{(K_{\tilde{j}})}^{\tilde{j}}}\right) \prod_{k=1}^{K_{\tilde{j}}-1} q_{i_{(k)}^{\tilde{j}}}$$

Thus, the increase in objective function value is non-negative. K^* is also an optimal solution to (PS_t). □

Proof of Theorem 7. Let K^* be an optimal solution to (PS_t). For any $K \leq K^*$, we say K is part of K^*, and construct a remaining problem

$$\begin{aligned}(P^R) \qquad \max \quad & \sum_{j=1}^{M} \mu_j^R \left(1 - \prod_{k=K_j+1}^{K_j+L_j} q_{i_{(k)}^j}\right) \\ s.t. \quad & \sum_{j=1}^{M} L_j = t^R \\ & 0 \leq L_j \leq u_j^R, integer, \qquad j = 1, \ldots, M.\end{aligned}$$

in which $t^R = t - \sum_{j=1}^{M} K_j$, $\mu_j^R = \mu_j \prod_{k=1}^{K_j} q_{i_{(k)}^j}$, and $u_j^R = u_j - K_j$, for any j. To prove Theorem 7, we first introduce and prove the following lemma.

Lemma 31. *If L^* is an optimal solution to (P^R), then $(K + L^*)$ is an optimal solution to (PS_t).*

Proof of Lemma 31. It is easy to show that $(K + L^*)$ is feasible for (PS_t). We prove its optimality by contradiction. Suppose L^* is optimal for (P^R), but $(K + L^*)$ is not optimal for (PS_t). Let $L' = K^* - K$. Then L' is a feasible solution to (P^R), and

$$\sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j+L'_j} q_{i_{(k)}^j}\right) > \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j+L_j^*} q_{i_{(k)}^j}\right) \tag{A.1}$$

The objective function value of (P^R) at L' is

$$\begin{aligned} Obj(L') &= \sum_{j=1}^{M} \mu_j^R \left(1 - \prod_{k=K_j+1}^{K_j+L'_j} q_{i_{(k)}^j}\right) \\ &= \sum_{j=1}^{M} \mu_j \left(\prod_{k=1}^{K_j} q_{i_{(k)}^j}\right) \left(1 - \prod_{k=K_j+1}^{K_j+L'_j} q_{i_{(k)}^j}\right) \\ &= \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j+L'_j} q_{i_{(k)}^j}\right) - \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j} q_{i_{(k)}^j}\right). \end{aligned}$$

Similarly, the objective function value of (P^R) at L is

$$Obj(L^*) = \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j+L_j^*} q_{i_{(k)}^j}\right) - \sum_{j=1}^{M} \mu_j \left(1 - \prod_{k=1}^{K_j} q_{i_{(k)}^j}\right).$$

By (A.1), we have $Obj(L') > Obj(L^*)$, which contradicts that L is optimal for (P^R). This completes the proof of Lemma 31.

Let K^ν be the value of K at the end of iteration ν of Algorithm 2. We then prove the following proposition.

Proposition 32. *K^ν is part of an optimal solution to (PS_t) for any $1 \le \nu \le t$.*

Proof of Proposition 32. We prove by mathematical induction.
When $\nu = 1$, $K^1 = e_{\hat{j}}$, where $\hat{j} = \text{argmax}_{j:u_j>0}\left\{\mu_j\left(1 - q_{i^j_{(1)}}\right)\right\}$. By Theorem 3, there exists an optimal solution K^* s.t. $K^*_{\hat{j}} \ge 1$, so K^1 is part of K^*.
Assume K^ν is part of an optimal solution to (PS_t), $\nu < t$. We construct a remaining problem with respect to K^ν.

$$
\begin{aligned}
(P^\nu) \qquad \max \quad & \sum_{j=1}^{M} \mu_j^\nu \left(1 - \prod_{k=K_j^\nu+1}^{K_j^\nu+L_j} q_{i^j_{(k)}}\right) \\
s.t. \quad & \sum_{j=1}^{M} L_j = t^\nu \\
& 0 \le L_j \le u_j^\nu, integer, \qquad j = 1, \ldots, M.
\end{aligned}
$$

where t^ν, μ_j^ν, and u_j^ν are as defined in Algorithm 2. Let $\hat{j} = \text{argmax}_{j:u_j^\nu>0}\left\{\mu_j^\nu\left(1 - q_{i^j_{(K_j^\nu)}} + 1\right)\right\}$.
By Theorem 3, there exists an optimal solution L^* to (P^ν) s.t. $L^*_{\hat{j}} \ge 1$. And by Lemma 31, $(K^\nu + L^*)$ is an optimal solution to (PS_t). $K^{\nu+1} = K^\nu + e_{\hat{j}}$ is part of $(K^\nu + L^*)$, and thereby part of an optimal solution to (PS_t).
By mathematical induction, we conclude that K^ν is part of an optimal solution to (PS_t) for any $1 \le \nu \le t$. This completes the proof of Proposition 32.
Algorithm 2 terminates when $\nu = t$, and returns K^t. By Proposition 32, K^t is part of an optimal solution to (PS_t). In addition, it satisfies $\sum_{j=1}^{M} K_j^t = t$, so it is an optimal solution to (PS_t). □

Proof of Proposition 8. Let $K^{t,\nu}$ be the number of suppliers that have been assigned to retailers at the end of iteration ν of Algorithm 3 with t links to be built. K^ν does not change as the value of t changes as long as $1 \le \nu \le t$. Then for any t_1, t_2 s.t. $1 \le t_1 < t_2 \le \sum_{j=1}^{M} u_j$, we have $K^{t_1,\nu} = K^{t_2,\nu}$ for any $1 \le \nu \le t_1$. Hence, $K^{t_1,t_1} = K^{t_2,t_1}$. The solution when $T = t_1$ is part of that when $T = t_2$. Since t_1 and t_2 are chosen arbitrarily, we conclude that the solutions given by Algorithm 2 are nested. □

Proof of Proposition 9. Algorithm 2 adds links in non-increasing order of marginal benefit, then by Proposition 8, the marginal increment in $S(T)$ is non-decreasing in T. □

Proof of Proposition 10. Let $\mathcal{S}$ be the set of all the candidate links. Define function $F(\cdot)$ as the expected sales of the supply chain given a subset of links are built. Let A be an arbitrary subset of $\mathcal{S}$ which contains no more than $\sum_{j=1}^{M} u_j - 2$ links, and

$B = A \cup \{a_1\}$, where $a_1 \in \mathcal{S} \backslash A$, and $F(B) \ge F(B')\ \forall\ B'$ s.t. $A \subset B' \subset \mathcal{S}$, $|B'| = |B|$,

$C = B \cup \{a_2\}$, where $a_2 \in \mathcal{S} \backslash B$, and $F(C) \geq F(C')\ \forall\ C'$ s.t. $B \subset C' \subset \mathcal{S}$, $|C'| = |C|$.
Let $\widetilde{B} = A \cup \{a_2\}$. Then $A \subset \widetilde{B} \subset \mathcal{S}$, and $|\widetilde{B}| = |B|$, thereby $F(B) \geq F(\widetilde{B})$.
By Corollary 1 in [1], $F(B \cup \widetilde{B}) + F(B \cap \widetilde{B}) \leq F(B) + F(\widetilde{B})$. It follows that $F(C) + F(A) \leq 2F(B)$, or equivalently, $F(C) - F(B) \leq F(B) - F(A)$. □

Proof of Proposition 11. First we show that the sales of the supply chain is concave in C when suppliers are reliable and demands are deterministic.
As mentioned before, the sales of a supply chain is equal to the max flow in a network in which the source node is connected to suppliers well the sink node is connected to retailers. By adding a link from the sink to the source and assigning weight 1 to it and weight 0 to all other links, we convert the max-flow problem into a max-weight circulation problem, and the max weight is equal to the max flow. By Proposition 1.2 in [**murota**], the max flow is concave in C.
Let C^1 and C^2 be any two capacity vectors. Define $\bar{C} = tC^1 + (1-t)C^2$ for $t \in [0,1]$. Let d be the demands, and r the states of suppliers. Then the realized capacity vector is equal to $C \cdot r$ for any capacity vector C (here $(\cdot)$ represents element-wise product). Let $F(C)$ denote the max flow with realized capacity C. Then

$$F(\bar{C} \cdot r) = F\left([tC^1 + (1-t)C^2] \cdot r\right) = F\left(tC^1 \cdot r + (1-t)C^2 \cdot r\right) \geq tF(C^1 \cdot r) + (1-t)F(C^2 \cdot r)$$

Since r and d are arbitrarily chosen,

$$\mathbb{E}\left[F\left(\bar{C} \cdot R\right)\right] \geq t\mathbb{E}\left[F(C^1 \cdot R)\right] + (1-t)\mathbb{E}\left[F(C^2 \cdot R)\right].$$

Therefore, the expected sales is concave in supply capacity. □

Appendix B

Proofs of Chapter 3

Proof of Theorem 12. Let Ω be the feasible region of (P). By Theorem 6.2 in [48], optimizing the expected total cost over Ω is equivalent to optimizing over $conv(\Omega)$. In addition, by Theorem 5.3 in [8], the minimum expected total cost is a concave function of the cost vector, and thus is concave in f, w, and q respectively. □

Appendix C

Proofs of Chapter 4

Proof of Lemma 13. (4.2) is a quadratic inequality of s_t. By solving it, we get

$$\frac{-\sqrt{e^2+euf'_{t-1}}-e}{2u} \le s_t \le \frac{\sqrt{e^2+euf'_{t-1}}-e}{2u} = \bar{s}_t.$$

Because constraints $s_t \ge s_{t-1}$ for $t = 1, 2, \ldots$ implies $s_t > s_0 > 0$, and $\frac{-\sqrt{e^2+euf'_{t-1}}-e}{2u} < 0$, we only need to bound s_t above by $\bar{s}_t$. □

Proof of Lemma 14. Given $f_{t-1} \ge us_{t-1}^2$, then

$$\bar{s}_t \ge \frac{\sqrt{e^2+4u(us_{t-1}^2+es_{t-1})}-e}{2u} = \frac{\sqrt{(e+2us_{t-1})^2}-e}{2u} = s_{t-1}.$$

□

Proof of Proposition 15. First we prove $s_t^f < s_t^s$.

$$\begin{aligned}
\beta^{(0)} = \frac{n(p-w)}{np+(n-1)(h-w)} \in (0,1) \quad &\Leftrightarrow \quad z_\beta^{(0)} > -\alpha s_0 \\
&\Leftrightarrow \quad -4z_\beta^{(0)} wu < 4\alpha wus_0 \\
&\Leftrightarrow \quad s_t^f < s_t^s
\end{aligned}$$

Next we prove $s_t^s \le \bar{s}_t$.

$$\begin{aligned}
s_t^s \le \bar{s}_t \quad &\Leftrightarrow \quad \sqrt{(\alpha w+e)^2+4uf'_{t-1}+4\alpha wus_0} \le \sqrt{e^2+4uf'_{t-1}}+\alpha w \\
&\Leftrightarrow \quad (\alpha w+e)^2+4uf'_{t-1}+4\alpha wus_0 \le e^2+4uf'_{t-1}+\alpha^2w^2+2\alpha w\sqrt{e^2+4uf'_{t-1}} \\
&\Leftrightarrow \quad \sqrt{e^2+4uf'_{t-1}} \ge e+2us_0 \\
&\Leftrightarrow \quad f'_{t-1} \ge us_0^2+es_0
\end{aligned}$$

Because $s_{t-1} \geq s_0$ and $f_{t-1} \geq us_{t-1}^2$, we have

$$f'_{t-1} = f_{t-1} + es_{t-1} \geq us_{t-1}^2 + es_{t-1} \geq us_0^2 + es_0$$

and thereby $s_t^s \leq \bar{s}_t$ also holds. □

Proof of Lemma 16. $g_1(\cdot)$ has a quadratic form with negative coefficient at the second order term, so it is strictly concave on $\mathbb{R}$.

$$\begin{aligned} g_2''(s_t) &= -\left[np + (n-1)(h-w)\right] f\left(z_t(s_t)\right)\left(z_t'(s_t)\right)^2 + \left[np + (n-1)(h-w)\right] F^0(z_t(s_t)) z_t''(s_t) \\ &\quad -\left[w + (n-1)h\right] z_t''(s_t) - 2u \end{aligned}$$

Given

$$z_t(s_t) = \frac{1}{w}\left[f'_{t-1} - (\alpha w + e)s_t - us_t^2\right]$$

we have

$$\begin{aligned} z_t'(s_t) &= -\frac{\alpha w + e}{w} - \frac{2u}{w}s_t \\ z_t''(s_t) &= -\frac{2u}{w} \end{aligned}$$

Thus,

$$\begin{aligned} g_2''(s_t) &= -\left[np + (n-1)(h-w)\right] f\left(z_t(s_t)\right)\left(z_t'(s_t)\right)^2 - 2u \\ &\quad -\frac{2u}{w}\left\{\left[np + (n-1)(h-w)\right] F^0\left(z_t(s_t)\right) - \left[w + (n-1)h\right]\right\}. \end{aligned}$$

If $s_t \geq s_t^f$, then $z_t(s_t) \leq z_\beta^{(0)}$, and thereby

$$F^0\left(z_t(s_t)\right) \geq F^0(z_{beta}^{(0)}) = 1 - \beta^{(0)} = \frac{w + (n-1)h}{np + (n-1)(h-w)}.$$

It follows that

$$g_2''(s_t) \leq -\left[np + (n-1)(h-w)\right] f\left(z_t(s_t)\right)\left(z_t'(s_t)\right)^2 - 2u < 0.$$

$g_2(\cdot)$ is strictly concave on $[s_t^f, +\infty)$.

$$g_3''(s_t) = -\frac{2u}{w}\left[np - (n-1)w\right] < 0$$

so $g_3(\cdot)$ is strictly concave over $\mathbb{R}$.

$$\begin{aligned} g_3'(s_t) &= n(p-w)z_t'(s_t) - 2us_t + n\alpha(p-w) - e \\ &= -\frac{np - (n-1)w}{w}(2us_t + e) \end{aligned}$$

If $s_t \geq 0$, then $2us_t + e > 0$, and $g_3'(s_t) < 0$. Thus, $g_3(\cdot)$ is strictly decreasing on $[0, +\infty)$. □

Proof of Lemma 17. We only need to show the the first derivative of $\pi_t(s_t, z_t^*(s_t))$ is continuous. Because $g_1(\cdot)$, $g_2(\cdot)$, and $g_3(\cdot)$ are continuously differentiable respectively, we only need to show that the first derivative of $\pi_t(s_t, z_t^*(s_t))$ is continuous at s_t^f and s_t^s.

$$\begin{aligned}
g_1'(s_t) &= -2us_t + n\alpha(p-w) - e \\
g_2'(s_t) &= \left\{[np + (n-1)(h-w)]\, F^0(z_t(s_t)) - [w + (n-1)h]\right\} z_t'(s_t) - 2us_t + n\alpha(p-w) \\
&\quad -e \\
g_3'(s_t) &= -\frac{np - (n-1)w}{w}(2us_t + e)
\end{aligned}$$

So we have

$$\begin{aligned}
g_1'(s_t^f) &= -2us_t^f + n\alpha(p-w) - e \\
g_2'(s_t^f) &= \left\{[np + (n-1)(h-w)]\, F^0(z_\beta^{(0)}) - [w + (n-1)h]\right\} z_t'(s_t^f) - 2us_t^f + n\alpha(p-w) \\
&\quad -e \\
&= -2us_t^f + n\alpha(p-w) - e \\
g_2'(s_t^s) &= \left\{[np + (n-1)(h-w)]\, F^0(-\alpha s_0) - [w + (n-1)h]\right\} z_t'(s_t^s) - 2us_t^s + n\alpha(p-w) \\
&\quad -e \\
&= n(p-w)\left(-\frac{\alpha w + e}{w} - \frac{2u}{w}s_t^s\right) - 2us_t^s + n\alpha(p-w) - e \\
&= -\frac{np - (n-1)w}{w}(2us_t^s + e) \\
g_3'(s_t^s) &= -\frac{np - (n-1)w}{w}(2us_t^s + e)
\end{aligned}$$

Therefore, $g_1'(s_t^f) = g_2'(s_t^f)$, and $g_2'(s_t^s) = g_3'(s_t^s)$. □

Proof of Proposition 18. By Lemma 16 and Lemma 17, there exists a unique $\tilde{s}_t$ and it equals to $\hat{s}^{(0)}$ if $\frac{d\pi_t}{ds_t}\big|_{s_t=s_t^f} < 0$, and $\check{s}_t$ otherwise. When $s_t^f > \hat{s}^{(0)}$, we have $\frac{d\pi_t}{ds_t}\big|_{s_t=s_t^f} = g_1'(s_t^f) < 0$, and thereby $\tilde{s}_t = \hat{s}^{(0)}$. When $s_t^f \le \hat{s}^{(0)}$, we have $\frac{d\pi_t}{ds_t}\big|_{s_t=s_t^f} = g_1'(s_t^f) \ge 0$, and thereby $\tilde{s}_t = \check{s}_t$. □

Proof of Corollary 19. Case 1: $s_t^f \ge \hat{s}^{(0)}$.
$\tilde{s}_t = \hat{s}^{(0)} \le s_t^f$. By Proposition 15, $s_t^f < \bar{s}_t$. Therefore, $\tilde{s}_t < \bar{s}_t$.
Case 2: $s_t^f < \hat{s}^{(0)}$.
To prove $\check{s}_t < \bar{s}_t$, only need to prove that $g_2'(\bar{s}_t) < 0$.

$$\begin{aligned}
g_2'(\bar{s}_t) &= [np + (n-1)(h-w)]\, F^0(z_t(\bar{s}_t)) z_t'(\bar{s}_t) - [w + (n-1)h]\, z_t'(\bar{s}_t) - 2u\bar{s}_t + n\alpha(p-w) \\
&= -e \\
&= \left\{[np + (n-1)(h-w)]\, F^0(-\alpha s_0) - [w + (n-1)h]\right\} z_t'(\bar{s}_t) - 2u\bar{s}_t + n\alpha(p-w) \\
&\quad -e
\end{aligned}$$

Because the support of ϵ is $[-\alpha s_0, +\infty)$, $F^0(-\alpha s_0) = 1$. It has also been shown in the proof of Lemma 16 that $z_t'(s_t) = -\frac{\alpha w + e}{w} - \frac{2u}{w} s_t$, then we have

$$\begin{aligned} g_2'(\bar{s}_t) &= -n(p-w)\left(\frac{\alpha w + e}{w} + \frac{2u}{w} s_t\right) - 2u\bar{s}_t + n\alpha(p-w) - e \\ &= -\frac{np - (n-1)w}{w}(2u\bar{s}_t + e) < 0. \end{aligned}$$

□

Proof of Proposition 20.

$$\begin{aligned} s_t^f < \bar{s}_t &\Leftrightarrow \frac{\sqrt{(z_\beta^{(0)}\alpha w + e)^2 + 4uf_{t-1}'} - (z_\beta^{(0)}\alpha w + e)}{2u} < \frac{\sqrt{e^2 + 4uf_{t-1}'} - e}{2u} \\ &\Leftrightarrow \sqrt{(z_\beta^{(0)}\alpha w + e)^2 + 4uf_{t-1}'} < z_\beta^{(0)}\alpha w + \sqrt{e^2 + 4uf_{t-1}'} \\ &\Leftrightarrow (z_\beta^{(0)}\alpha w + e)^2 + 4uf_{t-1}' < (z_\beta^{(0)}\alpha w)^2 + e^2 + 4uf_{t-1}' + 2z_\beta^{(0)}\alpha w\sqrt{e^2 + 4uf_{t-1}'} \\ &\Leftrightarrow 2z_\beta^{(0)}\alpha we < 2z_\beta^{(0)}\alpha w\sqrt{e^2 + 4uf_{t-1}'} \\ &\Leftrightarrow e < \sqrt{e^2 + 4uf_{t-1}'} \end{aligned}$$

Because $f_{t-1}' \geq us_{t-1}^2 > 0$, $e < \sqrt{e^2 + 4uf_{t-1}'}$ always holds, and thereby $s_t^f < \bar{s}_t$ also holds. □

Proof of Lemma 21. $g_1(\cdot)$ has a quadratic form with negative coefficient at the second order term, so it is strictly concave on $\mathbb{R}$.

$$\begin{aligned} g_2''(s_t) &= \alpha\left[np + (n-1)(h-w)\right]\left[\left(2z_t'(s_t) + s_t z_t''(s_t)\right)F^0(z_t(s_t)) - s_t f\left(z_t(s_t)\right)\left(z_t'(s_t)\right)^2\right] \\ &\quad + \frac{2(n-1)hu}{w} \end{aligned}$$

Given

$$z_t(s_t) = \frac{f_{t-1}' - es_t - us_t^2}{\alpha w s_t}$$

we have

$$\begin{aligned} z_t'(s_t) &= -\frac{f_{t-1}'}{\alpha w s_t^2} - \frac{u}{\alpha w} \\ z_t''(s_t) &= \frac{2f_{t-1}'}{\alpha w s_t^3} \end{aligned}$$

Thus,

$$\begin{aligned} g_2''(s_t) &= -\alpha\left[np+(n-1)(h-w)\right]s_t f\left(z_t(s_t)\right)\left(z_t'(s_t)\right)^2 \\ &\quad +\alpha\left[np+(n-1)(h-w)\right]F^0\left(z_t(s_t)\right)\left(-\frac{2u}{\alpha w}\right)+\frac{2(n-1)hu}{w}. \end{aligned}$$

If $s_t \geq s_t^f$, then $z_t(s_t) \leq z_\beta^{(0)}$, and thereby

$$F^0\left(z_t(s_t)\right) \geq F^0(z_{beta}^{(0)}) = 1-\beta^{(0)} = \frac{w+(n-1)h}{np+(n-1)(h-w)}.$$

It follows that

$$\begin{aligned} g_2''(s_t) &\leq -\alpha\left[np+(n-1)(h-w)\right]s_t f\left(z_t(s_t)\right)\left(z_t'(s_t)\right)^2-\frac{2u}{w}\left[w+(n-1)h\right]+\frac{2(n-1)hu}{w} \\ &= -\alpha\left[np+(n-1)(h-w)\right]s_t f\left(z_t(s_t)\right)\left(z_t'(s_t)\right)^2-2u<0 \end{aligned}$$

$g_2(\cdot)$ is strictly concave on $[s_t^f,+\infty)$. □

Proof of Lemma 22. We only need to show the the first derivative of $\pi_t(s_t, z_t^*(s_t))$ is continuous. Because $g_1(\cdot)$ and $g_2(\cdot)$ are continuously differentiable respectively, we only need to show that the first derivative of $\pi_t(s_t, z_t^*(s_t))$ is continuous at s_t^f.

$$\begin{aligned} g_1'(s_t) &= -2us_t+\alpha\left[np+(n-1)(h-w)\right]\left(1-F^1(z_\beta^{(0)})\right) \\ &\quad -z_\beta^{(0)}\alpha\left[w+(n-1)h\right]-e \\ g_2'(s_t) &= \alpha\left[np+(n-1)(h-w)\right]\left[1-F^1(z_t(s_t))+s_tF^0(z_t(s_t))z_t'(s_t)\right]+\frac{(n-1)h}{w}(2us_t+e) \end{aligned}$$

So we have

$$\begin{aligned} g_1'(s_t^f) &= -2us_t^f+\alpha\left[np+(n-1)(h-w)\right]\left(1-F^1(z_\beta^{(0)})\right)-z_\beta^{(0)}\alpha\left[w+(n-1)h\right]-e \\ g_2'(s_t^f) &= \alpha\left[np+(n-1)(h-w)\right]\left[1-F^1(z_\beta^{(0)})+s_t^fF^0(z_\beta^{(0)})z_t'(s_t^f)\right]+\frac{(n-1)h}{w}(2us_t^f+e) \\ &= \alpha\left[np+(n-1)(h-w)\right]\left(1-F^1(z_\beta^{(0)})\right)+\alpha\left[w+(n-1)h\right]s_t^fz_t'(s_t^f) \\ &\quad +\frac{(n-1)h}{w}(2us_t^f+e) \end{aligned}$$

By definition of s_t^f, it satisfies

$$u(s_t^f)^2+(z_\beta^{(0)}\alpha w+e)s_t^f = f_{t-1}'.$$

Then

$$\begin{aligned} z_t'(s_t^f) &= -\frac{f_{t-1}'}{\alpha w s_t^2} - \frac{u}{\alpha w} \\ &= -\frac{1}{\alpha w}\left(2u + \frac{z_\beta^{(0)}\alpha w + e}{s_t^f}\right) \end{aligned}$$

Then

$$\begin{aligned} g_2'(s_t^f) &= \alpha\left[np + (n-1)(h-w)\right]\left(1 - F^1(z_\beta^{(0)})\right) - \frac{w + (n-1)h}{w} s_t^f \left(2u + \frac{z_\beta^{(0)}\alpha w + e}{s_t^f}\right) \\ &\quad + \frac{(n-1)h}{w}(2u s_t^f + e) \\ &= \alpha\left[np + (n-1)(h-w)\right]\left(1 - F^1(z_\beta^{(0)})\right) - 2u s_t^f - z_\beta^{(0)}\alpha\left[w + (n-1)h\right] - e \end{aligned}$$

Therefore, $g_1'(s_t^f) = g_2'(s_t^f)$. □

Proof of Proposition 23. By Lemma 21 and Lemma 22, $\pi_t(s_t, z_t^*(s_t))$ is strictly concave. In addition,

$$\begin{aligned} \lim_{s_t \to +\infty} g_2'(s_t) &= \lim_{s_t \to +\infty} \Big\{ \alpha\left[np + (n-1)(h-w)\right]\left[1 - F^1(z_t(s_t)) + s_t F^0(z_t(s_t)) z_t'(s_t)\right] \\ &\quad + \frac{(n-1)h}{w}(2u s_t + e) \Big\} \\ &= -\infty \end{aligned}$$

Therefore, $\tilde{s}_t$ always exists. $\tilde{s}_t$ equals to $\hat{s}^{(0)}$ if $\frac{d\pi_t}{ds_t}\big|_{s_t = s_t^f} < 0$, and $\check{s}_t$ otherwise. When $s_t^f > \hat{s}^{(0)}$, we have $\frac{d\pi_t}{ds_t}\big|_{s_t = s_t^f} = g_1'(s_t^f) < 0$, and thereby $\tilde{s}_t = \hat{s}^{(0)}$. When $s_t^f \le \hat{s}^{(0)}$, we have $\frac{d\pi_t}{ds_t}\big|_{s_t = s_t^f} = g_1'(s_t^f) \ge 0$, and thereby $\tilde{s}_t = \check{s}_t$. □

Proof of Corollary 24. Case 1: $s_t^f \ge \hat{s}^{(0)}$.
$\tilde{s}_t = \hat{s}^{(0)} \le s_t^f$. By Proposition 20, $s_t^f < \bar{s}_t$. Therefore, $\tilde{s}_t < \bar{s}_t$.
Case 2: $s_t^f < \hat{s}^{(0)}$.
To prove $\check{s}_t < \bar{s}_t$, only need to prove that $g_2'(\bar{s}_t) < 0$.

$$\begin{aligned} g_2'(\bar{s}_t) &= \alpha\left[np + (n-1)(h-w)\right]\left[1 - F^1(z_t(\bar{s}_t)) + \bar{s}_t F^0(z_t(\bar{s}_t)) z_t'(\bar{s}_t)\right] + \frac{(n-1)h}{w}(2u\bar{s}_t + e) \\ &= \alpha\left[np + (n-1)(h-w)\right]\left[1 - F^1(0) + \bar{s}_t F^0(0) z_t'(\bar{s}_t)\right] + \frac{(n-1)h}{w}(2u\bar{s}_t + e) \end{aligned}$$

Because the support of ϵ is $[0, +\infty)$ and $\mathbb{E}[\epsilon] = 1$, $F^0(0) = 1$ and $F^1(0) = \mathbb{E}[\epsilon] = 1$. It is also known that $\bar{s}_t$ satisfies $u\bar{s}_t^2 + e\bar{s}_t = f_{t-1}'$, so

$$z_t'(\bar{s}_t) = -\frac{f_{t-1}'}{\alpha w \bar{s}_t^2} - \frac{u}{\alpha w} = -\frac{u\bar{s}_t^2 + e\bar{s}_t}{\alpha w \bar{s}_t^2} - \frac{u}{\alpha w} = -\frac{1}{\alpha w}(2u + \frac{e}{\bar{s}_t}).$$

It follows that

$$\begin{aligned} g_2'(\bar{s}_t) &= -\alpha\left[np+(n-1)(h-w)\right]\bar{s}_t\frac{1}{\alpha w}(2u+\frac{e}{\bar{s}_t})+\frac{(n-1)h}{w}(2u\bar{s}_t+e) \\ &= -\frac{np-(n-1)w}{w}(2u\bar{s}_t+e)<0. \end{aligned}$$

□

Proof of Lemma 25. We only need to check whether $g_2'(s\hat{s}^{(0)})<0$ when $s_t^f<\hat{s}^{(0)}$.
In the additive demand case,

$$g_2'(\hat{s}^{(0)})=\left\{\left[np+(n-1)(h-w)\right]F^0(z_t(\hat{s}))-\left[w+(n-1)h\right]\right\}z_t'(\hat{s}^{(0)})-2u\hat{s}^{(0)}+n\alpha(p-w)-e$$

If $s_t^f<\hat{s}^{(0)}$, then $z_t(\hat{s}^{(0)})<z_\beta^{(0)}$, and

$$F^0(z_t(\hat{s}^{(0)}))>F^0(z_\beta^{(0)})=\frac{w+(n-1)h}{np+(n-1)(h-w)}.$$

We also have

$$z_t'(\hat{s}^{(0)})=-\frac{\alpha w+e}{w}-\frac{2u}{w}\hat{s}^{(0)}<0.$$

Then

$$g_2'(\hat{s}^{(0)})<-2u\hat{s}^{(0)}+n\alpha(p-w)-e=0$$

In the multiplicative demand case,

$$\begin{aligned} g_2'(\hat{s}^{(0)}) &= \alpha\left[np+(n-1)(h-w)\right]\left[1-F^1(z_t(\hat{s}^{(0)}))+\hat{s}^{(0)}F^0(z_t(\hat{s}^{(0)}))z_t'(\hat{s}^{(0)})\right] \\ &\quad +\frac{(n-1)h}{w}(2u\hat{s}^{(0)}+e) \\ &= \alpha\left[np+(n-1)(h-w)\right]\left[1-F^1(z_t(\hat{s}^{(0)}))\right] \\ &\quad +\alpha\left[np+(n-1)(h-w)\right]F^0(z_t(\hat{s}^{(0)}))\hat{s}^{(0)}\left(-\frac{f_{t-1}'}{\alpha w(\hat{s}^{(0)})^2}-\frac{u}{\alpha w}\right) \\ &\quad +\frac{(n-1)h}{w}(2u\hat{s}^{(0)}+e) \end{aligned}$$

We have show that, if $s_t^f<\hat{s}^{(0)}$, then

$$F^0(z_t(\hat{s}^{(0)}))>F^0(z_\beta^{(0)})=\frac{w+(n-1)h}{np+(n-1)(h-w)}.$$

Thus,

$$\begin{aligned}
g_2'(\hat{s}^{(0)}) &< \alpha\left[np+(n-1)(h-w)\right]\left[1-F^1(z_t(\hat{s}^{(0)}))\right] \\
&\quad -\alpha\left[w+(n-1)h\right]\left(\frac{f'_{t-1}}{\alpha w\hat{s}^{(0)}}+\frac{u}{\alpha w}\hat{s}^{(0)}\right) \\
&\quad +\frac{(n-1)h}{w}(2u\hat{s}^{(0)}+e) \\
&= \alpha\left[np+(n-1)(h-w)\right]\left[1-F^1(z_t(\hat{s}^{(0)}))\right] \\
&\quad -\alpha\left[w+(n-1)h\right]\left(z_t(\hat{s}^{(0)})+\frac{e}{\alpha w}+\frac{2u}{\alpha w}\hat{s}^{(0)}\right) \\
&\quad +\frac{(n-1)h}{w}(2u\hat{s}^{(0)}+e) \\
&= \alpha\left[np+(n-1)(h-w)\right]\left[1-F^1(z_t(\hat{s}^{(0)}))\right]-\alpha\left[w+(n-1)h\right]z_t(\hat{s}^{(0)})-2u\hat{s}^{(0)} \\
&\quad -e \\
&= -\alpha\left[np+(n-1)(h-w)\right]\left[F^1(z_t(\hat{s}^{(0)}))-F^1(z_\beta^{(0)})\right] \\
&\quad +\alpha\left[w+(n-1)h\right](z_\beta^{(0)}-z_t(\hat{s}^{(0)})) \\
&= -\alpha\left[np+(n-1)(h-w)\right]\left[F^0(z_\beta^{(0)})(z_\beta^{(0)}-z_t(\hat{s}^{(0)}))-\left(F^1(z_t(\hat{s}^{(0)}))-F^1(z_\beta^{(0)})\right)\right] \\
&= -\alpha\left[np+(n-1)(h-w)\right]\left[\int_{z_t(\hat{s}^{(0)})}^{z_\beta^{(0)}}F^0(z_\beta^{(0)})dx-\int_{z_t(\hat{s}^{(0)})}^{z_\beta^{(0)}}F^0(x)dx\right]
\end{aligned}$$

Because $F^0(x) > F^0(z_\beta^{(0)})$ for any $x < z_\beta^{(0)}$, we have

$$g_2'(\hat{s}^{(0)}) < 0.$$

□

Proof of Proposition 26. Let year $\hat{t}$ be the mature year. By Proposition 18 and Proposition 23,

$$\tilde{s}_t = \begin{cases} \check{s}_t & \text{if } t < \hat{t} \\ \hat{s}^{(0)} & \text{if } t = \hat{t} \end{cases}$$

Then by (4.7) and (4.10), for any $t < \hat{t}$, $s_t^* = \max\{s_{t-1}, \check{s}_t\}$, and Lemma 25 implies that $s_t^* \le \check{s}_t < \hat{s}^{(0)}$ during these developing years. In the mature year, $s_{\hat{t}}^* = \max\{s_{\hat{t}-1}, \hat{s}^{(0)}\}$. Because $s_{\hat{t}-1} = s_{\hat{t}-1}^*$, and it has been shown that $s_{\hat{t}-1}^* < \hat{s}^{(0)}$, we have $s_{\hat{t}}^* = \hat{s}^{(0)}$ □

Proof of Corollary 27. In year t, assume that the retailer starts from the same size s_{t-1} and the same fund f_{t-1} in both scenarios, $s_{t-1} < \min\{\hat{s}^{(0)}, \hat{s}^{(k)}\}$, and $f_{t-1} > us_{t-1}^2$ (i.e. the fund allows expansion). Let s_t^a and s_t^b denote the retailer's expanded sizes in scenario a and scenario b, respectively. Then by Proposition 26, $s_t^a = \max\{s_{t-1}, \check{s}_t\} < \hat{s}^{(0)}$, and by Proposition 29, $\hat{s}^{(0)} < \hat{s}^{(k)}$, so $s_t^a < \hat{s}^{(k)}$. In addition, Corollary 19 and Corollary 24

together with the condition $f_{t-1} > us_{t-1}^2$ implies that $s_t^a =< \bar{s}_t$. We have also shown that $s_t^b = \min\{\hat{s}^{(k)}, \bar{s}_t\}$. Therefore, $s_t^a < s_t^b$. □

Proof of Corollary 28. In any developing year, $s_t^f < \hat{s}^0 \Rightarrow \tilde{s}_t = \check{s}_t \Rightarrow s_t^* = \max\{s_{t-1}, \check{s}_t\}$. By Lemma 17 and Lemma 22, $s_t^f < \hat{s}^0$ also implies $s_t^f < \check{s}_t$. Therefore, $s_t^* \geq \check{s}_t > s_t^f$, and thereby $z_t^* = z_t(s_t^*)$.
In the mature year, $s_t^f \geq \hat{s}^0 \Rightarrow \tilde{s}_t = \hat{s}^0 \Rightarrow s_t^* = \max\{s_{t-1}, \hat{s}^0\}$. By Proposition 26, $s_{t-1} \geq \hat{s}^0$, so $s_t^* \leq \hat{s}^0 \leq s_t^f$. It follows that $z_t^* = z_\beta^{(0)}$. □

Proof of Proposition 29. In the additive demand case,

$$\begin{aligned}
\frac{\partial \hat{s}^{(k)}}{\partial \alpha} &= \frac{n(p - w^{(k)})}{2u} > 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial p} &= \frac{n\alpha}{2u} > 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial w^{(k)}} &= -\frac{n\alpha}{2u} < 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial e} &= -\frac{1}{2u} < 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial u} &= -\frac{n\alpha(p - w^{(k)}) - e}{2u^2} = -\frac{\hat{s}^{(k)}}{u} < 0
\end{aligned}$$

So $\hat{s^{(k)}}$ is increasing in α and p, and decreasing in w^k, e and u.
In the multiplicative demand case, $z_\beta^{(k)}$ is a function of α, p, $w^{(k)}$ and h, so we first derive the partial derivative of $\hat{s}^{(k)}$ with respect to $z_\beta^{(k)}$.

$$\frac{\partial \hat{s}^{(k)}}{\partial z_\beta^{(k)}} = \frac{\alpha\left[np + (n-1)(h - w^{(k))}\right] F^0(z_\beta^{(k)}) - \alpha\left[w^{(k)} + (n-1)h\right]}{2u} = 0.$$

Then we have

$$\begin{aligned}
\frac{\partial \hat{s}^{(k)}}{\partial \alpha} &= \frac{\left[np + (n-1)(h - w^{(k)})\right]\left(1 - F^1(z_\beta^{(k)})\right) - z_\beta^{(k)}\left[w^{(k)} + (n-1)h\right]}{2u} \\
&= \frac{1}{2u}\frac{2u\hat{s}^{(k)} + e}{\alpha} > 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial p} &= \frac{n\alpha\left(1 - F^1(z_\beta^{(k)})\right)}{2u} = \frac{n\alpha}{2u}\left(F^1(0) - F^1(z_\beta^{(k)})\right) > 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial w^{(k)}} &= \frac{-(n-1)\alpha\left(1 - F^1(z_\beta^{(k)})\right) - z_\beta^{(k)}\alpha}{2u} = -\frac{\alpha}{2u}\left[(n-1)\left(F^1(0) - F^1(z_\beta^{(k)})\right) + z_\beta^{(k)}\right] \\
&< 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial h} &= \frac{(n-1)\alpha\left(1 - F^1(z_\beta^{(k)})\right) - z_\beta^{(k)}\alpha(n-1)}{2u} = \frac{(n-1)\alpha}{2u}\left[F^1(0) - F^1(z_\beta^{(k)}) - z_\beta^{(k)}\right] \\
&= \frac{(n-1)\alpha}{2u}\left[\int_0^{z_\beta^{(k)}} F^0(x)dx - \int_0^{z_\beta^{(k)}} 1dx\right] < 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial e} &= -\frac{1}{2u} < 0 \\
\frac{\partial \hat{s}^{(k)}}{\partial u} &= -\frac{1}{2u^2}\Big\{\alpha\left[np + (n-1)(h - w^{(k)})\right] - \alpha\left[np + (n-1)(h - w^{(k)})\right]F^1(z_\beta^{(k)}) \\
&\quad -z_\beta^{(k)}\alpha\left[w^{(k)} + (n-1)h\right] - e\Big\} \\
&= -\frac{\hat{s}^{(k)}}{u} < 0
\end{aligned}$$

So $\hat{s^{(k)}}$ is increasing in α and p, and decreasing in w^k, h, e and u.

□

Bibliography

[1] O. Z. Aksin and F. Karaesmen. "Characterizing the Performance of Process Flexibility Structures". In: *Operations Research Letters* 35.2007 (2007), pp. 477–484.

[2] S. Andradottir, H. Ayhan, and D. G. Down. "Compensating for Failures with Flexible Servers". In: *Operations Research* 55.4 (2007), pp. 753–768.

[3] Groothedde B. and Ruijgrok C. abd Tavasszy L. "Towards Collaborative, Intermodal Hub Networks: A Case Study in the Fast Moving Consumer Goods Market". In: *Transportation Res. Part E* 41.6 (2005), pp. 567–583.

[4] A. Bassamboo, R. S. Randhawa, and J. A. Van Mieghem. "A Little Flexibility is All You Need: On the Asymptotic Value of Flexibility in Parallel Queuing Systems with Linear Capacity Sizing Costs". Working Paper. Kellogg School of Management, Northwestern University. 2009.

[5] A. Bassamboo, R. S. Randhawa, and J. A. Van Mieghem. "Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing". In: *Management Science* 56.8 (2010), pp. 1285–1303.

[6] T. Beck, A. Demirguc-Kunt, and V. Maksimovic. "Financial and Legal Constraints to Growth: Does Firm Size Matter?" In: *The Journal of Finance* 60.1 (2005), pp. 137–177.

[7] J. R. Berge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, 2011.

[8] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[9] M. P. Bitler, A. M. Robb, and J. D. Wolken. "Financial Services Used by Small Businesses: Evidence from the 1998 Survey of Small Business Finances". In: *Federal Reserve Bulletin* (2001).

[10] D. B. Bradley and M. J. Rubach. "Trade Credit and Small Businesses: a Cause of Business Failures?" The Small Business Advancement National Center, University of Central Arkansas. 2002.

[11] L. M. B. Cabral and J. Mata. "On the Evolution of the Firm Size Distribution: Facts and Theory". In: *American Economic Review* 93.4 (2003), pp. 1075–1090.

[12] R. S. de Camargo, G. Miranda Jr., and H. P. Luna. "Benders Decomposition for Hub Location Problems with Economies of Scale". In: *Transportation Science* 43.1 (2009), pp. 86–97.

[13] R. S. de Camargo et al. "Multiple Allocation Hub-and-Spoke Network Design under Hub Congestion". In: *Computers & Operations Research* 36.12 (2009), pp. 3097–3106.

[14] J. F. Campbell. "Hub Location for Time Definite Transportation". In: *Comput. Oper. Res.* 36.12 (2009), pp. 3107–3116.

[15] J. F. Campbell. "Integer Programming Formulations of Discrete Hub Location Problems". In: *Eur. J. Oper. Res.* 72.2 (1994), pp. 387–405.

[16] J. F. Campbell and M. E. O'Kelly. "Twenty-Five Years of Hub Location Research". In: *Transportation Science* 46.2 (2012), pp. 153–169.

[17] R. E. Carpenter and B. C. Petersen. "Is the Growth of Small Firms Constrained by Internal Finance?" In: *The Review of Economics and Statistics* 84.2 (2002), pp. 298–309.

[18] H.-C. Chang et al. "The Optimal Pricing and Ordering Policy for an Integrated Inventory Model When Trade Credit Linked to Order Quantity". In: *Applied Mathematical Modelling* 33.2009 (2009), pp. 2978–2991.

[19] L. Chen, G. Kok, and J. Tong. "The Effect of Payment Schemes on Inventory Decisions: The Role of Mental Accounting". In: (012). To apear in *Management Science*.

[20] Y.-T. Chiu. *How Japan's Earthquake Is Shaking Up Taiwan's High-Tech Sector*. 2011. URL: `http://spectrum.ieee.org/semiconductors/memory/how-japans-earthquake-is-shaking-up-taiwans-hightech-sector`.

[21] M. C. Chou, C. P. Teo G. A. Chua, and H. Zheng. "Design for Processing Flexibility: Efficiency of the Long Chain and Sparse Structure". In: *Operations Research* 58.1 (2010), pp. 43–58.

[22] M. C. Chou, C. P. Teo, and H. Zheng. "Process Flexibility: Design, Evaluation and Applications". In: *Flexible Service and Manufacturing Journal* 20 (2008), pp. 59–94.

[23] Bryan D. "Extensions to the Hub Location Problem: Formulations and Numerical Examples". In: *Geographical Anal.* 30.4 (1998), pp. 315–330.

[24] T. Deng and Z.-J. Shen. "Process Flexibility Design in Unbalanced Networks". To appear in Manufacturing and Service Operations Management. 2012.

[25] *Designing Hub Networks with Connected and Isolated Hubs*. 43rd Hawaii Internat. Conf. System Sci. (HICSS-43). IEEE Computer Society. Koloa, Kauai, 2010.

[26] G. W. Emery. "An Optimal Financial Response to Variable Demand". In: *Journal of Financial and Quantitative Analysis* 22 (1987), pp. 209–225.

[27] A. T. Ernst and M. Krishnamoorthy. "An Exact Solution Approach Based on Shortest-Paths for p-Hub Median Problems". In: *Informs Journal on Computing* 10.2 (1998), pp. 149–162.

[28] Campbell J. F., Ernst A. T., and Krishnamoorthy M. *Facility Location: Applications and Theory, Hub location problems*. Springer, 2002.

[29] J. S. Ferris. "A Transactions Theory of Trade Credit Use". In: *Quarterly Journal of Economics* 94 (1981), pp. 243–270.

[30] M. Giannetti, M. Burkart, and T. Ellingsen. "What You Sell is What You Lend? Explaining Trade Credit Contracts". In: *Review of Financial Studies* 24.4 (2011), pp. 1299–1335.

[31] S. C. Graves and B. T. Tomlin. "Process Flexibility in Supply Chain". In: *Management Science* 49.7 (2003), pp. 907–919.

[32] D. Gupta and L. Wang. "A Stochastic Inventory Model with Trade Credit". In: *Manufacturing and Service Operations Management* 11.1 (2009), pp. 4–18.

[33] J Henry. *BMW Says Flexible, not Lean, is the Next Big Thing in Autos*. 2009. URL: `http://industry.bnet.com/auto/10002978/bmw-says-flexible-not-lean-is-the-next-big-thing-in-autos`.

[34] W. J. Hopp and M. L. Spearman. "To Pull or Not to Pull: What Is the Question". In: *Manufacturing and Service Operations Management* 6.2 (2004), pp. 133–148.

[35] Contreras I, Cordeau J-F, and Laporte G. "The Dynamic Uncapacitated Hub Location Problem". In: *Transportation Sci.* (2011).

[36] S. M. Iravani, M. P. Van Oyen, and K. Sims. "Structural Flexibility: A New Perspective on the Design of Manufacturing and Service Operations". In: *Management Science* 51 (2005), pp. 151–166.

[37] W. C. Jordan and S. C. Graves. "Principles on the Benefits of Manufacturing Process Flexibility". In: *Management Science* 41 (1995), pp. 577–594.

[38] F. Karaesmen, F. van der Duyn Schouten, and L. van Wassenhove. "Dedication vs. Flexibility in Field Service Operations". Working Paper. Center for Economic Research, Tilburg University. 1998.

[39] H. Kim and M. E. O'Kelly. "Reliable p-Hub Location Problems in Telecommunication Networks". In: *Geographical Anal.* 41.3 (2009), pp. 283–306.

[40] M. Lim et al. "Flexibility and Fragility in Supply Chain Network Design". Working Paper. Northwestern University. 2008.

[41] A.-G. Lium, Crainic T. G., and Wallace S. W. "A Study of Demand Stochasticity in Service Network Design". In: *Transportation Sci.* 43.2 (2009), pp. 144–157.

[42] M. S. Long, I. B. Malitz, and S. A. Ravid. "Trade Credit, Quality Guarantees, and Product Marketability". In: *Financial Management* 22 (1994), pp. 117–127.

[43] B. S. Maddah, M. Y. Jaber, and N. E. Abboud. "Periodic Review (s, S) Inventory Model with Permissible Delay in Payments". In: *The Journal of the Operational Research Society* 55.2 (2004), pp. 147–159.

[44] T. L. Magnanti and R. T. Wong. "Accelerating Benders Decomposition: Algorithmic Enhancement and Model Selection Criteria". In: *Operations Research* 39.3 (1981), pp. 464–484.

[45] H.-Y. Mak and Z.-J. Shen. "Stochastic Programming Approach to Process Flexibility Design". In: *Flexible Service and Manufacturing Journal* 21 (2009), pp. 75–91.

[46] A. Martin, Canovas L., and Landete M. "New Formulations for the Uncapacitated Multiple Allocation Hub Location Problem". In: *Eur. J. Oper. Res.* 172.1 (2006), pp. 274–292.

[47] Boland N. et al. "Preprocessing and Cutting for Multiple Allocation Hub Location Problems". In: *Eur. J. Oper. Res.* 155.3 (2004), pp. 638–653.

[48] G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimizaion*. John Wiley & Sons, Inc., 1988.

[49] C. K. Ng, J. K. Smith, and R. L. Smith. "Evidence on the Determinants of Credit Terms Used in Interfirm Tade". In: *American Finance Association* 54.3 (1999), pp. 1109–1129.

[50] Berman O., Krass D., and Menezes M. "Facility Reliability Issues in Network p-Median Problems: Strategic Centralization and Co-Location Effects". In: *Operations Research* 55.2 (2007), pp. 332–350.

[51] M. E. O'Kelly. "A Quadratic Integer Program for the Location of Interacting Hub Facilities". In: *Eur. J. Oper. Res.* 32.3 (1987), pp. 393–404.

[52] M. E. O'Kelly. "Activity Levels at Hub Facilities in Interacting Networks". In: *Geographical Anal.* 18.4 (1986), pp. 343–356.

[53] M. E. O'Kelly. "The Location of Interacting Hub Facilities". In: *Transportation Sci.* 20.2 (1986), pp. 92–106.

[54] M. E. O'Kelly and Bryan D. "Hub Location with Flow Economies of Scale". In: *Transportation Res. Part B* 32.8 (1998), pp. 605–616.

[55] M. A. Peterson and R. G. Rajan. "The Benefits of Lending Relationships: Evidence from Small Business Data". In: *The Journal of Finance* 49.1 (1994), pp. 3–37.

[56] M. A. Peterson and R. G. Rajan. "Trade Credit: Theories and Evidence". In: *The Society for Financial Studies* 10.3 (1997), pp. 661–691.

[57] H. Pirkul and D. A. Schilling. "An Efficient Procedure for Designing Single Allocation Hub and Spoke Systems". In: *Management Science* 44.12 (1998), pp. 235–242.

[58] L. Qi and Z.-J. M. Shen. "A Supply Chain Design Model with Unreliable Supply". In: *Naval Res. Logist.* 54 (2007).

[59] Lian Qi, Zuo-Jun Max Shen, and Lawrence V. Snyder. "The Effect of Supply Disruptions on Supply Chain Design Decisions". In: *Transportation Science* (2010).

[60] Alumur S and Kara B. Y. "A Hub Covering Network Design Problem for Cargo Applications in Turkey". In: *J. Oper. Res. Soc.* 60.10 (2009), pp. 1349–1359.

[61] Alumur S and Kara B. Y. "Network Hub Location Problems: The State of the Art". In: *Eur. J. Oper. Res.* 190.1 (2008), pp. 1–21.

[62] S. Saghafian and M. P. Van Oyen. "The Value of Flexible Backup Suppliers and Disruption Risk Information: Newsvendor Analysis with Recourse". In: *IIE Transactions* 44.10 (2012), pp. 834–867.

[63] S. Saghafian, M. P. Van Oyen, and B. Kolfal. "The "W" Network and the Dynamic Control of Unreliable Flexible Servers". In: *IIE Transactions* 43.12 (2011), pp. 893–907.

[64] S.S. Sana and K.S. Chaudhuri. "A Deterministic EOQ Model with Delays in Payments and Price-Discount Offers". In: *European Journal of Operational Research* 184.2008 (2008), pp. 509–533.

[65] Y. Sheffi. "Supply Chain Management Under the Threat of International Terrorism". In: *International Journal of Logistics Management* 12.1 (2002), pp. 1–11.

[66] Z.-J. M. Shen, R. L. Zhan, and J. Zhang. "The Reliable Facility Location Problem: Formulations, Heuristics, and Approximation Algorithms". In: *Journal on Computing* 23.3 (2010), pp. 470–482.

[67] Skorin-Kapov, Skorin-Kapov J D., and O'Kelly ME. "Tight Linear Programming Relaxations of Uncapacitated p-Hub Median Problems". In: *Eur. J. Oper. Res.* 94.3 (1996), pp. 582–593.

[68] J. Smith. "Trade Credit and Information Asymmetry". In: *Journal of Finance* 4 (1987), 863–869.

[69] L. V. Snyder and Z.-J. M. Shen. "Supply Chain Management under the Threat of Disruptions". In: *The Bridge(National Academy of Engineering)* 36.4 (2006), pp. 39–45.

[70] L.V. Snyder and M.S. Daskin. "Reliability Models for Facility Location: The Expected Failure Cost Case". In: *Transportation Science* (2005).

[71] L.V. Snyder et al. *Planning for Disruptions in Supply Chain Networks*. Tutorial, INFORMS, Pittsburgh, PA. 2006.

[72] J.-S. Song and J. Tong. "Payment Schemes, Financing Costs, and Inventory Management". Under revision. The Fuqua School of Business, Duke University. 2011.

[73] Aykin T. "Lagrangian Relaxation Based Approaches to Capacitated Hub-and-Spoke Network Design Problem". In: *Eur. J. Oper. Res.* 79 (1994), pp. 501–523.

[74] Ernst A. T. and Krishnamoorthy M. "Efficient Algorithms for the Uncapacitated Single Allocation p-Hub Median Problem". In: *Location Sci.* 4.3 (1996), pp. 139–154.

[75] Ernst A. T. and Krishnamoorthy M. "Exact and Heuristic Algorithms for the Uncapacitated Multiple Allocation p-Hub Median Problem". In: *European Journal of Operational Research* 104.1 (1998), pp. 100–112.

[76] C.S. Tang. "Robust Strategies for Mitigating Supply Chain Disruptions". In: *International Journal of Logistics: Research and Applications* 9.1 (2006), pp. 33–45.

[77] J.-T. Teng, C.-T. Chang, and S. K. Goyal. "Optimal Pricing and Ordering Policy under Permissible Delay in Payments". In: *International Journal Production Economics* 97.2005 (2005), pp. 121–129.

[78] Z.-H. Tu. *Japan earthquake brings extra orders for Taiwan chipmakers*. 2011. URL: `http://www.wantchinatimes.com/news-subclass-cnt.aspx?cid=1205&MainCatID=12&id=20110406000158`.

[79] Marianov V., Serra D., and ReVelle C. "Location of Hubs in a Competitive Environment". In: *Eur. J. Oper. Res.* 114.2 (1999), pp. 363–371.

[80] N. Wilson and B. Summers. "Trade Credit Terms Offered by Small Firms: Survey Evidence and Empirical Analysis". In: *Journal of Business Finance & Accounting* 29.3 & 4 (2002).

[81] An Y., Zhang Y., and Zeng B. "The Reliable Hub-and-Spoke Design Problem: Models and Algorithms". Working paper. 2011.