Dynamic and Robust Network Resource Allocation

by

Peter Yun Zhang

B.A.Sc. with Honours, University of Toronto (2011)
M.A.Sc., University of Toronto (2013)

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Engineering Systems

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

Author .....
Institute for Data, Systems, and Society
August 9, 2019

Certified by .....
David Simchi-Levi
Professor of Engineering Systems
Thesis Supervisor

Certified by .....
Stephen C. Graves
Abraham J. Siegel Professor of Management
Doctoral Committee Member

Certified by .....
Nikolaos Trichakis
Zenon Zannetos (1955) Career Development Professor
Associate Professor of Operations Management
Doctoral Committee Member

Accepted by .....
Stephen C. Graves
Graduate Chair, Institute for Data, Systems, and Society

Dynamic and Robust Network Resource Allocation

by

Peter Yun Zhang

Submitted to the Institute for Data, Systems, and Society
on August 9, 2019, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Engineering Systems

Abstract

Networks are essential modeling tools in engineering, business, and public policy research. They can represent physical connections, such as manufacturing processes. They can be relationships among people, such as patient treatment in healthcare. They can also represent abstract interactions, such as the biological reaction between a certain vaccine and a certain virus. In this work, we bring several seemingly disparate problems under the same modeling framework, and show their thematic coherence via the angle of dynamic optimization on networks.

Our research problems are drawn from business risk management, public health security, and public policy on vaccine selection. A common theme is the integrative design of (1) strategic resource placement on a network, and (2) operational deployment of such resources. We outline the research questions, challenges, and contributions as follows.

Modern automotive manufacturing networks are complex and global, comprising tens of thousands of parts and thousands of plants and suppliers. Such interconnection leaves the network vulnerable to disruptive events. A good risk mitigation decision support system should be data-driven, interpretable, and computational efficient. We devise such a tool via a linear optimization model, and integrate the model into the native information technology system at Ford Motor Company.

In public security, policymakers face decisions regarding the placement of medical resources and training of healthcare personnel, to minimize the social and economic impact of potential large scale bio-terrorism attacks. Such decisions have to integrate the strategic positioning of medical inventories, understanding of adversary's behavior, and operational decisions that involve the deployment and dispensing of medicines. We formulate a dynamic robust optimization model that addresses this decision question, apply a tractable solution heuristic, and prove theoretical guarantees of the heuristic's performance. Our model is calibrated with publicly available data to generate insights on how the policymakers should balance investment between medical inventory and personnel training.

The World Health Organization and regional public health authorities decide on the influenza (flu) vaccine type ahead of flu season every year. Vaccine effectiveness

has been limited by the long lead time of vaccine production – during the production period, flu viruses may evolve and vaccines may become less effective. New vaccine technologies, with much shorter production lead times, have gone through clinical trials in recent years. We analyze the question of optimal vaccine selection under both fast and slow production technologies. We formulate the problem as a dynamic distributionally robust optimization model. Exploiting the network structure and using tools from discrete convex analysis, we prove some structural properties, which leads to informative comparative statics and tractable solution methods. With publicly available data, we quantify the societal benefit of current and future vaccine production technologies. We also explore the reduction in disease burden if WHO expand vaccine portfolio to include more than one vaccine strain per virus subtype.

In each of the applications, our main contributions are four-fold. First, we develop mathematical models that capture the decision process. Second, we provide computational technology that can efficiently process these models and generate solutions. Third, we develop theoretical tools that guarantee the performance of these computational technology. Last, we calibrate our models with real data to generate quantitative and implementable insights.

Thesis Supervisor: David Simchi-Levi
Title: Professor of Engineering Systems

Doctoral Committee Member: Stephen C. Graves
Title: Abraham J. Siegel Professor of Management

Doctoral Committee Member: Nikolaos Trichakis
Title: Zenon Zannetos (1955) Career Development Professor
Associate Professor of Operations Management

Acknowledgments

I have been fortunate to learn from and collaborate with some extraordinary researchers prior to and during my PhD. My PhD advisor, Professor David Simchi-Levi, demonstrates how one's academic career can keep growing through focus, rigor, vision, and relentless effort to get to the bottom of research questions. Professor Nikos Trichakis is the best junior advisor that one can hope for, leading the way with vision, intuition, and rigor that fostered growth, and with patience and care that gave me room to experiment. Professor Steve Graves has continuously contributed to the field of operations management and supply chain research, and led our conversations with sharp intuition. Professor Dick Larson brings compassion to research and teaching, and has inspired me to stay on a path that I believe to be meaningful. His textbook *Urban Operations Research*, co-authored with Professor Amedeo Odoni, was the textbook that opened the world of operations research to me. Many other faculty members at MIT have invested heavily in teaching. The late Professor Joe Sussman's teaching was filled with a rare combination of humor and interrogation, built on his care for the engineering systems students. Professor Robert Freund taught us how to discern conceptual innovation from technical dexterity, and showed us the beauty and structure of optimization theory. Being a teaching assistant to Professor Deishin Lee and Professor Sean Willems has broadened my understanding for teaching. Sometimes learning happens without extensive dialogue. For example, I learned by closely following the works of Professor Dan Iancu, Professor Larry Wein, Professor Margaret Brandeau, Professor Aharon Ben-Tal, Professor Arkadi Nemirovski, Professor Laurent El Ghaoui, and Professor Dimitri Bertsekas. I am also deeply indebted to my mentors from University of Toronto, without them I would not have pursued this research path: Professor Cristina Amon, Dr David Romero, Professor Tim Chan, Professor Chris Beck, and Professor Yu-Ling Cheng. Apart from these mentors and their teachings, I also grew from working with many talented researchers in academia and industry, including Yehua Wei, Bill Schmidt, Oleg Gusikhin, and Don Zhang on our Ford project; Raj Subbu and Ana Muriel from our UTC project; James Poblete

from our Boeing project; Matthew Cook and Raj Raghatate from our BMW project; and Ashish Jagmohan, Pavithra Harsha, Carlo Siebenschuh, and Baruch Schieber during my internship at IBM.

My colleagues often offered their minds, hands, and hearts without asking for anything in return. I would not have been able to endure this PhD process without their support. This includes those from the ESD and IDSS family (Alex Jacquillat, Sarah Fletcher, Lita Das, Yin Jin Lee, Paul Kishimoto, Amanda Giang, Michael Davidson, Morgan Edwards, Henry Fingerhut, Florian Metzler, Marco Miotti, Yan Jin, Qi Yang, and many others that I might have missed), the DSL family (Yehua Wei, He Wang, Mila Nambiar, Krish Johnson, Clark Pixton, Zach Owen, Arzum Akkas, Wang Chi Cheung, Will Ma, Louis Chen, Michael Beeler, Yan Zhao, Xiao Fang, Zhenzhen Yan, Jinglong Zhao, Hanwei Li, Hanzhang Qin, Li Wang, Yiqun Hu, Yunzong Xu, Rui Sun, Elaheh Fata, Kirby Ledvina, Ruihao Zhu, and Xiaoyue Gong). I am especially thankful for the support of Janet Kerrigan and Beth Milnes, who did most of the background heavy lifting to shield us from more stress.

I am deeply indebted to my friends, who have tolerated my frequent disappearances and still accepted me when I reached out: Jim Kuo, Kimia Ghobadi, Sami Yamani Douzi Sorkhabi, Taewoo Lee, Jane Ip, Kevin Liu, Nancy Zhang, Yehua Wei, Lei Zhang, Auyon Siddiq, Velibor Mišić, Tout Wang, Yichuan Liu, Stephen Fang, Yeting Du, Shu Wei, Jean Li, Jamie Wong, Clare Kim, Chi-Yang and Shufen Cheung, Hanwen Li, Wei Luo, Yifan Wang, Wei Zang, Nalu and Qingqing, Yingjun Su, Tom Xiong, Bo Xu, Hanfu Li, Zhifang Cheng, Joseph Yan, Lucy Liu, Zhihao Li, Eva Shan, Mingquan Zhang, Philip Gao, Polly Han, Yansha Jin, Jing Leng, Thomse He, Cecilia Liu, Daizhuo Chen, and many others that I have missed here.

Most importantly, I owe everything to my parents, my grandparents, my girlfriend Mengqi, my aunts, and my cousins – I would not have been who I am if not for their unconditional love and care.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Network Design and Resource Allocation

1.1.1 Motivation

In automotive manufacturing, firms are exposed to a variety of low-probability, high-impact risks that can disrupt their operations and supply chains. These risks are difficult to predict and quantify; therefore, they are difficult to manage. As a result, managers may suboptimally deploy countermeasures, leaving their firms exposed to some risks, while wasting resources to mitigate other risks that would not cause significant damage. For example, Ford Motor Company has a global footprint of more than 50 plants. These plants turn billions of parts into millions of cars and trucks every year. A lengthy disruption in the complex supply and assembly network could lead to significant repercussions at Ford and in the automotive industry: during the 2011 Thailand floods, disruption at Ford's second-tier suppliers idled global production for one of its most profitable product lines. At Ford, how could supply chain managers utilize existing information technology infrastructure to consolidate data and provide analytical decisions to prepare against such high impact disruptions?

In public health, medical resource prepositioning and allocation are crucial in preparing against low-probability and high-impact health emergencies. For example, to defend against bio-terrorism attacks, policymakers face intertwining decisions that

involve inventory prepositioning, dispensing capacity, predicting attacker's move, and resource allocation. In order to be prepared to deliver adequate medication in a timely manner in response to a bioattack, medical countermeasures (MCM) inventories need to be positioned across the country. Some inventories can be stored in central locations to take advantage of pooling effects, while others need to be positioned near densely populated areas in order to improve deployment responsiveness. At the same time, training of emergency healthcare personnel is also critical, to help with the dispensing of medical resources such as antibiotics and vaccines. Where and how much MCM inventory should then be prepositioned in order to cost-effectively defend against bioattacks? How much dispensing capacity should be installed at target areas? These are the questions that require the integration of multiple facets of the medical supply chains.

Influenza (flu) vaccine preparation is another resource-intensive activity that public health policymakers address. In contrast to the low-probability emergency events, influenza epidemic happens every year – October to May in the northern hemisphere, and May to October in the southern hemisphere. Globally, influenza-related respiratory diseases lead to hundreds of thousands of deaths each year. In the United States, the Centers for Disease Control and Prevention (CDC) estimates that influenza causes millions of illnesses and tens of thousands of deaths annually. Influenza viruses evolve continuously, and vaccine strains are selected annually by the World Health Organization (WHO) – along with each country's own public health agency such as the CDC in the United States – to match them. Due to long production lead time of the traditional egg-based vaccine manufacturing technology, vaccine strains have to be selected at least half a year before the upcoming epidemic season, resulting in low vaccine effectiveness. In 2013 and 2016, the Food and Drug Administration in the United States approved new vaccine manufacturing technologies, with significantly shorter production lead times. With the arrival of new vaccine manufacturing technologies, how should the WHO and CDC select vaccine strains and plan productions to reduce disease burden?

1.1.2 Problem Description and Research Objective

Our goal in this thesis is to address the three above mentioned problems in a unified way. We provide analytical tools that can give quantitative and structural answers to these questions. New theories are developed to address the novel problems, for both model tractability and comparative statics. Drawn from seemingly disparate applications, these questions can in fact be modeled under one analytical decision framework. We point out three key ingredients for the framework.

First, these questions all deal with decisions on *networks*. For the case of Ford manufacturing, the network is the inbound supply and manufacturing network that connect components, facilities, and final products. Nodes represent parts, suppliers, products, and manufacturing processes that transform several parts into a larger component. Arcs represent the potential resource allocation pathways in this assembly network. In the medical countermeasure supply chain, the network nodes comprise medical stockpiles, dispensing facilities, and neighborhoods of potential populations under protection. Arcs are the shipment routes and allocation capabilities. For the case of vaccine selection and production, the network is more abstract. In this case, nodes represent the vaccine candidates and potential types of epidemic viruses. Arcs are used to carry the efficacy of using one type of vaccine to one type of virus. Throughout the thesis, we use (V, E) (or its variations) to represent a network comprising node set V and arc set E.

Second, these questions all deal with multiple levels of decision making. More specifically, there are a set *strategic* decisions x and a set of *operational* decisions y. For the case of supply chain risk management at Ford, strategic decisions refer to the network inventory positioning decisions prior to any disruption events, and operational decisions refer to the parts allocation and production decisions if the network is disrupted. Public health planning for biodefense also naturally separates into decisions that preposition inventories at various locations and regularly train healthcare volunteers, and decisions that govern the mobilization of resources and people if an attack happens. For the application in influenza vaccine preparation,

strategic decisions are made months ahead of the flu season to fix the types of vaccines to be produced, and start manufacturing process that has long lead times. The operational decisions are the ramping up of fast production process during the flu season, and administration of vaccines to population susceptible to influenza.

The third ingredient of the framework is uncertainty modeling. On a high level, we include random vector ξ and its governing set Γ. The first two applications deal with low-probability events, for which it is difficult to create probability measures from historical data. The third application of influenza vaccine planning has the flavor of probabilistic decision making. In all three cases, the uncertainty realizes between strategic decision and operational decision epochs. This makes the overall problem into a so-called multi-stage decision problem.

The framework is anchored on a cost function $f(x, y, \xi)$ (*e.g.*, manufacturing cost, disruption loss, healthcare disease burden, economic cost, or combinations of them), subject to constraints that naturally arise from the network setting (*e.g.*, flow conservation, capacity limits):

$$\min_{x \in X} \max_{\mu \in \Gamma} \quad \mathbb{E}_{\xi \sim \mu}[\min_{y \in Y} f(x, y, \xi)] \tag{1.1a}$$

$$\text{subject to} \quad \text{Network constraints.} \tag{1.1b}$$

The choice of uncertainty modeling $\max_{\mu \in \Gamma} \mathbb{E}_{\xi \sim \mu}$ follows from the distributionally robust optimization literature, which has received intensive attention from both the operations research and computer science communities due to its modeling flexibility and computational tractability. For our modeling purpose, we note that it is a framework that generalizes the stochastic programming approach (letting Γ be a singleton set) and classical robust optimization approach (setting Γ to include Dirac delta functions on the support). This flexibilitiy gives us a unified view of worst-case approach in a low-information setting and probabilistic approach in a high-information setting. Throughout this thesis, formulation (1.1) manifests into different instances, which are increasingly more general as we progress from the beginning to the end of the thesis.

1.2 Literature Review

We first review papers in the optimization literature that focus on the modeling and optimization theory related to problem formulation (1.1). We then review different streams of literature related to our three applications: *Supply Chain Risk Management* for Ford manufacturing network design; *Resource Allocation for Military Applications*, *Biodefense*, and *Inventory Prepositioning* for the biodefense supply chain design; and *Vaccine Selection* for the influenza vaccine strain selection and production problem.

Robust Network Flow. Atamtürk and Zhang [12] study the computational complexity of two-stage robust network flow under budgeted uncertainty sets and find it to be intractable in general. For a tree structure, they use a dynamic programming (DP) approach to show that it can be solved in polynomial time. Unfortunately, their DP approach is not suitable for our problem. Being developed within a constructive proof to show that robust tree network flow problems are polynomially solvable, the DP approach breaks down and cannot be applied for general, non-tree network flow problems. Furthermore, it is also tailored to deal only with uncapacitated network flow problems and it is unclear if it can be applied to tackle dispensing capacity issues. From a practical standpoint, because the SNS network need not be a tree, and because dispensing capacity is a major concern [142], the DP approach by Atamtürk and Zhang cannot be applied to tackle the SNS design problem we study in our work. They also provided a cutting plane algorithm that can be applied to solve general (capacitated) networks, but this exact solution algorithm has no polynomial time guarantee and is not practical for large instances, *e.g.*, the ones we solved in Section 3.6. Our goal is different. Instead of focusing on computational complexity, we study a tractable solution method for general networks involving a rich set of decisions (inventory, capacity, shipment, and dispensing over time), and provide strong theoretical and numerical backing for its performance. Our solution method enables us to apply it to a practically relevant policy design problem, demonstrated in a case study involving millions of nodes and considering important practical decisions and constraints.

Affinely Adjustable Robust Optimization. An introduction to this topic is included in Ben-tal *et al.* [17]. There have only been a handful of results that show AARC heuristics to be optimal under special circumstances [9, 20, 21, 72]. Our work contributes to this stream by providing an optimality proof under a new context. Bertsimas *et al.* [21] studies one-dimensional problems, *e.g.*, inventory problems that involve a single stocking level, while in Delage and Ardestani-Jaafari's work [9], the authors deal with newsvendor type problems under demand uncertainty. We differ from these papers by studying multidimensional problems, *i.e.*, shipment decisions in a supply chain network. The papers by Iancu *et al.* [72] and Bertsimas *et al.* [20] deal with families of problems, where for AARC optimality (or "good" performance) they require the uncertainty sets to be lattices, or possess a "symmetric" structure. In our model, in order to preclude overly conservative solutions, we rely on simplex-type budget constraints that invalidate both the lattice and the symmetric structures. Put differently, we prove optimality for a problem that is according to [20] in some sense the "least conducive" to the AARC optimality conditions set forth in the literature so far.

Supply Chain Risk Management. We follow two streams of research in our work for the supply chain risk management. The first area of scholarship pertains to supply chain network modeling and optimization, which broadly consider the optimal network structure under steady state operations [48, 65] or under the possibility of a disruption [140, 116, 95]. Closely related is research that evaluates coordination strategies between buyers and suppliers in the presence of disruption risk [149, 38, 150]. Less attention has been given to evaluating the impact of a disruption based on the optimal response of an existing network once that disruption has occurred. A recent exception is [128], which evaluates response strategies that minimize the service-level impact when disruption occurs on a multiechelon network for a random duration. Another is [94], which evaluates the interaction between the supplier and buyer response strategies under a random-duration disruption.

We make three important contributions to this literature. First, we develop our model for practical applications using large-scale supply chain data from Ford. Sec-

ond, we evaluate the optimal contingency plans for settings in which the disruption duration is either known exactly or described by an uncertainty set. Finally, our model quantifies the disruption exposure across all the nodes in the company's supply chain based on company-level performance impacts.

The second stream of research seek to classify operational disruptions and quantify their impact. Scholars and practitioners generally agree that operational disruptions materially and negatively impact company performance on average [129, 70, 160]. There is less agreement, however, on how we should classify and forecast such disruptions [82, 144, 156, 141]. Researchers are only beginning to understand which disruptions have the greatest impact on firm performance. Answering this research question is important because it informs firms on which disruptions warrant mitigation investments. [41] propose that supply chain density, complexity, and node criticality contribute to the severity of disruptions. [144] theorizes that a firm's vulnerability to disruption depends on its supply chain strategies, including postponement strategies and inventory placement. [23] identify that a firm's organizational integration practices are associated with the firm's ability to mitigate the consequences of disruptions. [82] provide evidence that changes to risk-assessment and risk-mitigation practices reduce the impact of disruptions in the chemical industry.

Our research generally aligns with concepts applied in other disciplines, including estimating maximum foreseeable loss (*i.e.*, the maximum loss if all safeguards in a system break) in the insurance industry and conducting failure analysis (*i.e.*, assessing the structural resilience when a critical member of a system is removed) in structural design. Until now, however, the field of operational risk management has not given these principles much attention.

Resource Allocation for Military Applications. For many military defense problems, such as ours, stochastic and priority-list approaches are inappropriate for capturing the strategic interaction between intelligent defender and attacker [4, 63, 102]. Therefore, it is desirable to use a sequential game approach, such as DAD. A main challenge for this approach is tractability. In the past decade, there has been concentrated and valuable effort in developing scalable solution approaches, especially for

the interdiction type of problems [4, 6, 13, 28, 79, 88]. Atkinson [13] and Lazzaro [88] show that the state-of-the-art algorithms can solve networks with thousands of nodes. Alderson *et al.* [6] mentions that for the significantly simpler bilevel games, problems with thousands of nodes can be readily solved by decomposition algorithms with commercial solvers.

Our work in the biodefense supply chain design differs from the existing DAD papers in two ways. First, we deal with problem sizes on the order of *millions* of nodes and present for them scalable approximate solution approaches backed by performance guarantees. Such drastic increase in problem size would render existing decomposition-based solution approaches insufficient. To deal with high dimensionality, we use the AARC heuristic from RO and contribute back to the theoretical RO literature by proving a novel performance guarantee of AARC (see related discussion below). Second, we provide an integrated modeling of cost, health deprivation, and antibiotic efficacy over time for our application. This in part requires a precise analysis of time-expanded multi-commodity network flow, and how the AARC heuristic is analytically suited for it. We consider these insights to be generalizable since multi-commodity network flow models are considered essential primitives for DAD problems [5].

Another set of military operations research papers, such as [18, 69, 122, 163], also study Stackelberg game problems. For example, Berman and Gavious [18] study the defender-attacker problem as a min-max facility location problem. Zhuang and Bier [163] study the difference between random and fully strategic attackers. While they focus on equilibrium analysis, our paper formulates and solves a large-scale, high-fidelity prescriptive model to support decisions.

Biodefense. Existing biodefense literature does not focus on the interaction between defender and attacker. It generally assumes an abundance of MCM inventories and exogenous demand scenarios [24, 40, 71, 80, 81, 89, 90, 91, 93, 157]. They study, for example, the optimal number and layout of points of dispense [71, 89, 90, 91, 93], the effectiveness of predetermined, exogenous inventory levels on the treatment of patients after bioattacks and influenza outbreaks [24, 80, 81], and a small number of

response strategies for airborne anthrax attacks [40, 157]. Taking MCM supply for granted and assuming attack scenarios to be known is limiting, because in practice, MCMs are shipped from different locations with different lead times (from hours to days) and shipment sizes, resulting in time-dependent flows, and terrorists' decisions are unknown in advance. In contrast, we take an integrated approach and study the entire supply chain within the DAD framework.

Inventory Prepositioning. Beyond the biodefense literature, other streams have studied prepositioning problems, including humanitarian logistics [30, 31, 47] and manufacturing [131, 133]. The latter two papers provide a decision support tool to evaluate performance under disruptive risks for a given inventory configuration and a given disruption scenario. Unlike the *descriptive* nature of these models, ours derives *prescriptive* suggestions that incorporate two stages of decision making.

Uichanco [152] and Simchi-Levi *et al.* [135] study inventory prepositioning to hedge against capacity disruption and demand uncertainties. Uichanco provides a solution method that scales with the number of vertices in the polytope describing the network structure. Such a method would lead to an intractable formulation in our case because a national biodefense network typically comprises millions of nodes and edges. In contrast, we provide a tractable model that can deal with problems of the size faced by U.S. policymakers. By incorporating dispensing capacity optimization, our model is more general too. In [135], the authors also focus on providing an exact solution method. Not surprisingly, this leads to an intractable formulation too, namely an LP with an exponential number of constraints. They propose a constraint generation algorithm, which they find to work well in practice for bi-partite graphs involving 200 nodes. Their constraint generation algorithm still involves the solution of mixed-integer linear programs in its subroutines. Such approaches could not possibly be employed to tackle networks with millions of nodes, like ours. In contrast, our focus is to provide tractable and scalable (polynomial-time) algorithms, which require the solution only of linear optimization problems. Although not being exact for general network topologies, we back our approach with optimality guarantees as discussed above.

Vaccine Selection. The influenza vaccine selection problem is not entirely new to the operations community. This current work is most closely related to that of Wu, Wein, and Perelson (2005) [161]. The authors addressed the vaccine strain selection problem in a multi-year case, focusing on the year-by-year antigenic drift and cancellation effect between multiple vaccinations. They modeled the vaccine effectiveness by quantifying the similarity between vaccine strains and circulating virus strains with a theoretical biology model. Other influenza vaccine research have focused on various aspects of the vaccine lifecycle. For example, Lee *et al.* (2016) studied the vaccine effectiveness prediction problem with machine learning methods [92]. Özaltın *et al.* (2011) focused on the problem of optimal timings of the strain selection and production decisions within one influenza season [115]. Cho (2009) studied the yield uncertainty of vaccine production [37]. Teytelman and Larson (2013) studied vaccine allocation policies, considering regional differences in influenza timings [146]. Several papers also studied supply chain coordination and competition problems among vaccine manufacturers or between manufacturers and policymakers [35, 36, 42, 43, 84, 114].

1.3 Contribution and Organization of the Thesis

In a three-year research engagement with Ford Motor Company, we addressed the practical need of supply chain risk mitigation by developing a novel risk-exposure model that assesses the impact of a disruption originating anywhere in a firm's supply chain. Our approach defers the need for a company to estimate the probability associated with any specific disruption risk until after it has learned the effect such a disruption will have on its operations. As a result, the company can make more informed decisions about where to focus its limited risk-management resources. We demonstrate how Ford applied this model to identify previously unrecognized risk exposures, evaluate predisruption risk-mitigation actions, and develop optimal post-disruption contingency plans, including circumstances in which the duration of the disruption is unknown.

For biodefense supply chain design, we formulated the problem as a two-stage

robust optimization model, and applied a tractable solution heuristic to the problem. We prove that the heuristic is actually optimal in important cases of the problem, and use extensive computational studies to demonstrate its near-optimal performance in general settings. One key feature of this solution heuristic is its ability to solve network design problems with millions of nodes, which is at least two orders of magnitude larger than what is normally used in the literature. This high-performance solution technology enables us to calibrate the model with publicly available data to work through a high fidelity case study. Our integrated framework enables us to derive prescriptive recommendations to the CDC. For example, we find that if CDC wanted to ensure 85% survivability for attacks that simultaneously affected (any) two states, each with at most two cities being attacked, with a detection time of under 48 hours, then the minimum required annual inventory and dispensing capacity budget would be about $210 million; for survivability target of 92%, the budget would increase to $553 million. This increasing marginal cost phenomenon, along with other cost/policy implications, can be explored quantitatively within our model.

In the vaccine selection problem, we incorporate different production technologies in the vaccine selection and production problem to address the arrival of new vaccine manufacturing technologies in the past few years. To analyze this problem, we bring together several technical tools from a few areas – theoretical biology, discrete convex analysis, and distributionally robust optimization – to create a vehicle for our model and theory. The model and theory are general, and can be used to analyze other network newsvendor type and facility location type of problems as mentioned in the literature review. For example, we show tractability of our model under popular choices of ambiguity sets, such as those defined by Wasserstein distance, or consistent with a set of marginal distributions. We also provide comparative statics on the problem, providing structural insights into the policy and technology decisions involving vaccine selection and production. We show that the newly FDA-approved manufacturing technologies have the potential to impact vaccine policies well beyond increasing vaccine availability. In particular, we show that selecting more than one vaccine strain for each virus subtype is *complementary* to adding a fast production

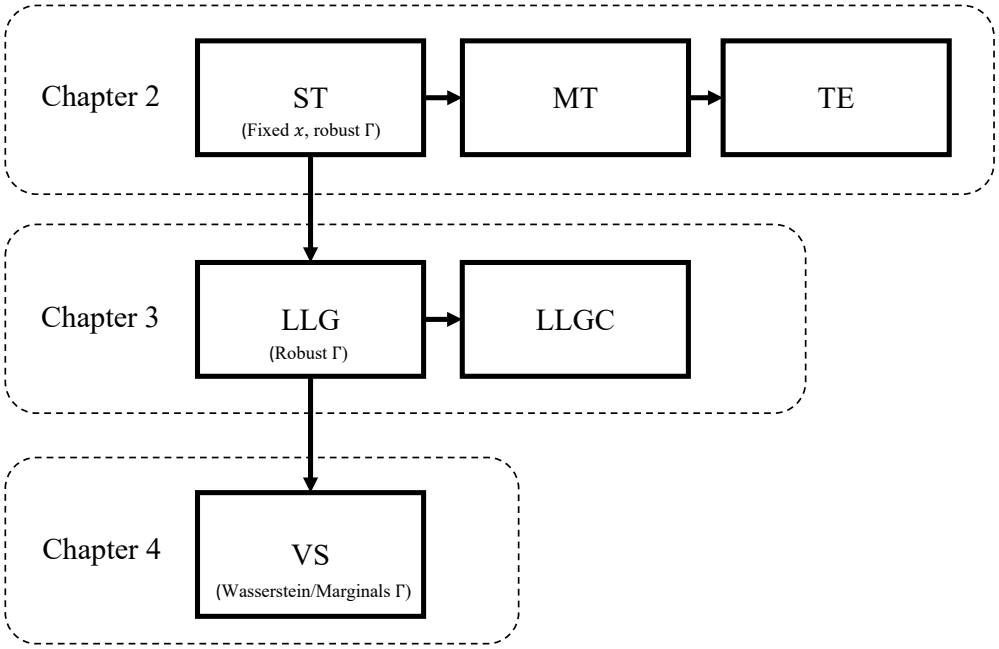

Figure 1-1: Models in the thesis and their relationship. Each box represents one model. Each arc points from a less general model to a more general model. Descriptions in the brackets show how exactly this particular model is a specific implementation of formulation (1.1).

technology, and sometimes has a much higher potential in reducing disease burden. As a result, the policymakers should re-examine the current practice of selecting just one vaccine strain per subtype per flu season, because the benefit may outweigh the cost of selecting an additional strain.

Main models in the content chapters of the thesis are roughly in increasing order of generality (Figure 1-1). Our theoretical contributions cover tractability and optimality results on all models. Furthermore, we also develop algorithms and theoretical characterizations to highlight the behavior of some of these models.

Chapter 2

Mitigating Manufacturing Risk

2.1 Ford Manufacturing Supply Chain

Many companies face considerable operational and supply chain risks that can materially impact company performance. Given the complexity and scope of Ford Motor Company's operations, this is certainly its situation. Ford maintains over 50 plants worldwide, which annually utilize 35 billion parts to produce six million cars and trucks. It has up to 10 tiers of suppliers between itself and its raw materials. Its Tier 1 suppliers number 1,400 companies across 4,400 manufacturing sites. A lengthy disruption anywhere in this extended supply chain can have significant financial repercussions for Ford. A disruption to one of its second-tier suppliers during the 2011 Thailand floods elevated the importance of this issue. As a result of this disruption, Ford idled global production for one of its most profitable product lines.

Ford is one of many companies exposed to such disruptions. For example, the 2011 flooding in Thailand led Intel to cut its quarterly revenue target by $1 billion [147]. Driven in part by greater global trade and the adoption of lean operating principles, many companies now operate with globally dispersed manufacturing facilities and extended supply chains. Normal accident theory holds that because major disruptions are an inherent property of such complex and tightly coupled systems, they should be considered unavoidable or normal [118]. It falls to operations and supply chain managers to navigate this new normal. Traditional operational-disruption

risk-assessment methods oblige firms to identify the probability and magnitude of disruption risks early in the analysis process [124, 83]; however, managers face a number of challenges in implementing such a solution. First, it is difficult and often impossible for managers to accurately estimate the likelihood of low-probability, high-impact disruptive events [14, 143]. Second, managers tend to misallocate resources when facing low-probability events [77, 76], ignore risks regardless of their potential significance [97], and distrust or disregard precise probability estimates [85, 97]. This can lead to inaction; [99] found that most firms do little to proactively prepare for such low-probability, high-impact disruptive events.

In this paper, we apply a new model, proposed by Simchi-Levi in March 2012 [62] and described in [132], for analyzing operational-disruption risk and detail the development and implementation of this model at Ford. Throughout the paper, we share the primary results of our analysis using masked versions of Ford's operational and supply chain data.

Limitations of the Legacy Risk-Analysis Approach

For many companies, even those that have world-class operations and supply chain management systems, proactively managing high-impact, low-probability disruption risks is challenging. One obstacle to conducting a more insightful analysis of disruption risks is that operational disruptions are both difficult to predict and have a highly uncertain impact on performance. In Ford's case, the scale and dynamic nature of its supply chain further complicate this problem. These factors increase both the number of disruption scenarios to consider and the frequency at which we should evaluate those scenarios. A second obstacle is data availability, particularly on suppliers at lower tiers within the supply chain. Supply chain transparency is a challenge for the entire automotive industry. Suppliers to the industry have historically been reluctant to provide the automobile manufacturers with detailed information about their suppliers and their suppliers' suppliers. As a result, although manufacturers typically have good information on Tier 1 suppliers (*i.e.*, companies that supply directly to the manufacturer), they have considerably less information on lower-tier suppliers in the

supply chain.

Given these limitations, legacy risk-management processes often focus on tracking the status of only a handful of suppliers and part numbers. These tend to be suppliers that provide major assembly components and represent a large portion of the total component costs. Many large manufacturers recognize that material exposures are likely to be hidden among the suppliers who are not included in this regular review process. Because of the difficulties in predicting disruptions, the data limitations, and the size of their supply chains, companies often cannot identify where these exposures are, much less quantify their impact. For example, managers at Ford estimate that conducting a traditional risk analysis for all of Ford's more-than-4,000 Tier 1 supplier sites would likely take two or three years, at which time the analysis would be obsolete.

2.2 Data

The major data columns collected in the analysis include:

1. Total annual part volume (part / supplier / Ford factory / car model)
2. Number of parts per vehicle (part / car model)
3. Production volume (Ford factory / car model)
4. TTR (part / supplier)
5. Profit margin (car model)
6. Part-to-subcomponent mapping (part)
7. Inventory (part)

TTR is the number of days that a supplier estimates it requires to return to full production after a major disruption. Ford collected TTR for two types of disruptions-one in which the supplier's production facility is lost and another in which both the supplier's production facility and tooling are lost. Part-to-subcomponent mapping

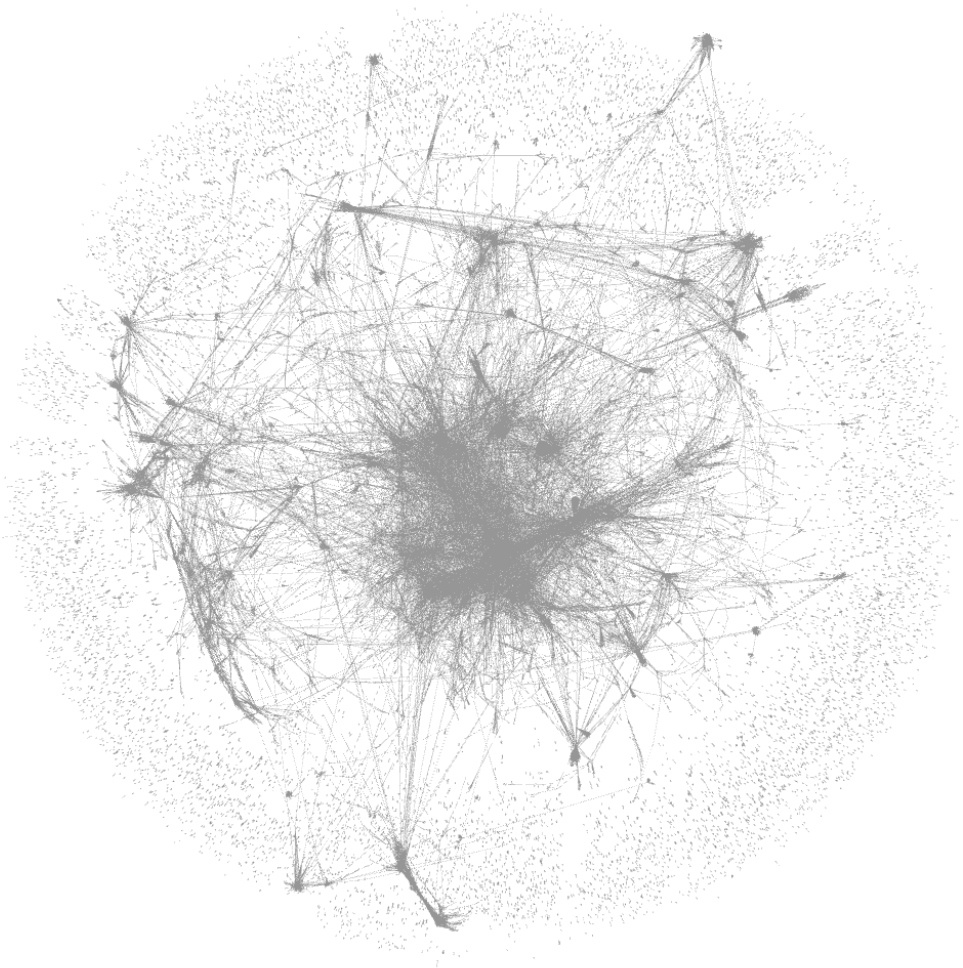

Figure 2-1: Overview of the bill-of-materials network with 74680 nodes and 97479 edges. Each node is a part or component in the manufacturing process, and edges represent the assembly requirements.

identifies how a part from a second- or third-tier supplier maps to a part or subcomponent in the tier above. Inventory measures the number of parts at the supplier site, in transit or at the Ford site. For second- or third-tier parts, this includes the number of parts held at those echelons.

Overall, data on bill-of-materials and potential disruption durations are collected for assemblies occuring in North America (Figures 2-1 and 2-2.

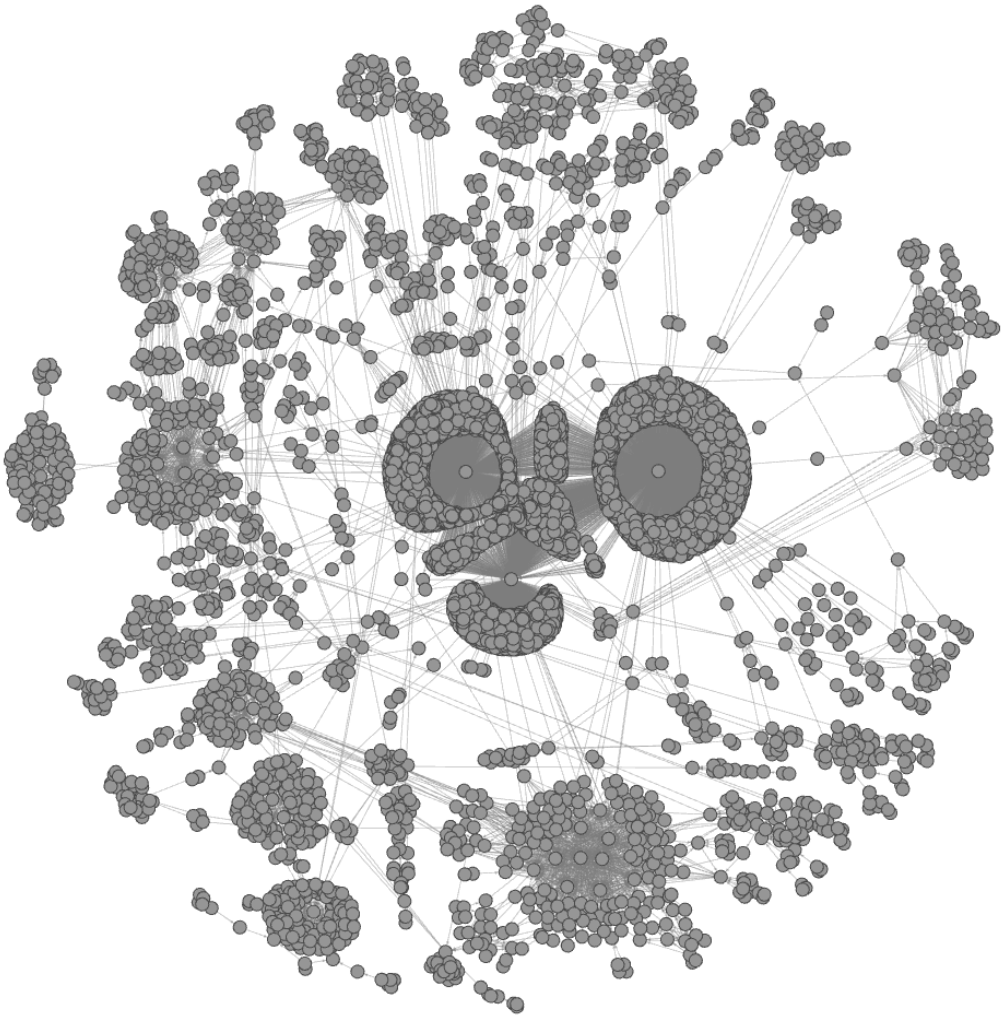

Figure 2-2: Zooming into Figure 2-1: bill-of-materials corresponding to three assemblies, with 8172 nodes and 11678 edges.

2.3 Model and Theory

Recognizing that managers have limited ability to predict low-probability, high-impact risks or collect detailed data on lower levels of their supply chain, our approach, initially described in [132], advocates integrating a vulnerability-based analysis into supply chain risk assessments. In such an analysis, the focus is on understanding the impact of a disruption, regardless of its source. This defers the need to estimate the probability associated with any specific risk and collect detailed information from sub-tier suppliers until after Ford has determined the impact a disruption will have on its operations. At that point, Ford can make a more informed decision about where to focus its limited risk-assessment resources. Our approach suits the goal of analyzing supply chain disruptions because the impact of a disruption often does not depend on the cause of the disruption but rather on its duration. In addition, the potential mitigation actions that a company can practically employ in response to a supply chain disruption are often the same regardless of the specific causes of the disruption. Finally, our approach implicitly recognizes that supply chains are in a continuous state of flux. In the face of such constant change, maintaining up-to-date predictions of the likelihood of specific risks is nontrivial; however, given that a disruption does occur, estimating a firm's vulnerability is more tractable.

2.3.1 Time-To-Recover Model

The model considers the supply chain as a graph representing the movements of supplier parts from each supplier facility to each of a firm's facilities and product lines. A node, also referred to as a stage, in the graph is equivalent to a part or manufacturing process at a particular supplier or Ford facility. Inputs to the model include operational and (or) financial measures (*e.g.*, unit profitability) and in-transit and on-site inventory levels for each node. Our model incorporates the time-to-recover (TTR) of each node in the supply chain network, which represents the time it takes for a node to recover to full functionality after a disruption [98, 132]. This value can be unique at each node in the firm's supply chain.

The model iterates over each node in the graph, disrupting the node for the duration of its TTR and calculating the corresponding impact on the firm's performance. It determines the performance impact assuming the firm responds optimally to the disruption scenarios, where the model simulates the optimal responses by solving an associated linear optimization problem. The model can accommodate different performance measures as the objective for this optimization, including minimizing the lost units of production, lost sales, or lost profit margin. For each disruption scenario, the model searches on how to reallocate existing inventory, redirect supply alternatives, and idle downstream plants such that the disruption has the smallest impact. The resulting performance impact (PI) is the impact of that disruption scenario on the firm's chosen performance measure during the TTR. To simplify cross-scenario comparisons, the model can also calculate a risk-exposure index (REI) [132], which normalizes the PI for each scenario by the maximum PI over all scenarios considered in the analysis.

The model can accommodate simultaneous disruptions in multiple supply chain nodes. This allows management to analyze complex disruption scenarios, including disruptions that affect all the parts from one supplier plant or disruptions that affect all the same part regardless of the supplier. We can extend the model to account for alternative sources of supply and supplier capacity commitments. This facilitates an explicit examination of interactive effects, which may occur when multiple firms try to adjust to supply disruptions at the same time. For example, if a supplier fails to deliver to one firm, it may have gone down for multiple firms. Such an event makes other potentially compensating nodes (*e.g.*, backup suppliers) more congested.

2.3.2 Time-To-Survive Model

In many cases, accurate TTR information may not be available. More importantly, a supplier may be optimistic when assessing its TTR; that is, a supplier may underestimate the time required to recover and hence may underestimate Ford's exposure to a disruption. Therefore, Ford is interested in identifying suppliers whose disruption impact is sensitive to the given TTR information. For this purpose, we introduce the

time-to-survive (TTS) concept, which we define as the maximum amount of time the system can function without performance loss if a particular node is disrupted [134]. As we will show, we determine the TTS associated with a specific node by solving an optimization problem that takes into account the disruption of this node, inventory levels, and alternative sources of supply; see below for the model formulation. The firm can determine whether a more accurate measure of TTR is necessary by comparing the TTS value associated with a specific node with the TTR estimate of that node. If the TTS far exceeds the TTR, it implies that a large change in TTR will have little impact on the firm's risk exposure; however, nodes with short TTS values require Ford to engage these suppliers in a detailed discussion about their TTRs.

2.3.3 Mathematical Models

We first present a single-tier (ST) supply chain model to illustrate some of the main concepts, and then extend it to a multiple-tier (MT) model that encompasses more components. The basic premise of both models is that, given a supply chain structure (a graph) and a disruption scenario (interrupted nodes and edges), we determine how to allocate the firm's remaining resources to optimize its ability to satisfy exogenous demand. A node (or stage) in the graph is a component or manufacturing process at a particular supplier or assembly site; an edge is a directed flow of materials from an upstream stage to a downstream stage. We formulate both models as linear optimization programs. We summarize our notation for the single-tier model in Table 1 and for the multiple-tier model in Table 2.

In the ST model, the firm has a set of plants ($\mathcal{A}$), which produce a set of products ($\mathcal{V}$). The firm's objective for each disruption scenario is to minimize the impact of the disruption on its chosen performance metric. We capture this through the following

Symbol	Explanation
Superscript $^{(n)}$	Disruption scenario n.
$\mathcal{A}$	Set of all suppliers sites (plants).
$\mathcal{V}$	Set of all final nodes (vehicles).
$\mathcal{F}^{(n)}$	Set of production edges under disruption scenario n.
$t^{(n)}$	TTR for disruption scenario n.
c_i	Total production capacity of node i per unit time.
s_i	Finished goods inventory of node i.
f_j	Profit margin of product j.
d_j	Demand for j (per time unit).
l_j	Lost production volume of vehicle type j.
y_{ij}	Amount of product j produced at plant i.

Table 2.1: This table lists the parameters and variables of the single-tier model and their explanations.

linear program.

$$
\begin{aligned}
\text{Minimize} \quad & \textstyle\sum_{j\in\mathcal{V}} f_j l_j^{(n)} \\
\text{s.t.} \quad & \textstyle\sum_{i:(i,j)\in\mathcal{F}^{(n)}} y_{ij}^{(n)} + l_j^{(n)} \;\geq\; d_j t^{(n)} - s_j, \quad \forall j \in \mathcal{V} \\
& \textstyle\sum_{j:(i,j)\in\mathcal{F}^{(n)}} y_{ij}^{(n)} \;\leq\; c_i t^{(n)}, \quad \forall i \in \mathcal{A} \setminus n \\
& y_{ij}^{(n)}, l_j^{(n)} \;\geq\; 0, \quad \forall i \in \mathcal{A}, j \in \mathcal{V}
\end{aligned}
$$

In this model, decision variable $y_{ij}^{(n)}$ is the cumulative production of j at plant i in disruption scenario n. Variable $l_j^{(n)}$ is the amount of lost demand for product j in disruption scenario n. Parameter $f_j^{(n)}$ refers to the impact of one unit of loss in sales for product j, for example, the profit margin; $t^{(n)}$ is the TTR for this disruption scenario. d_j and s_j are the demand and inventory for product j, respectively. Flexibility design $\mathcal{F}^{(n)}$ is the set of edges that are still alive during disruption scenario n.

The objective function is the minimization of the total weighted loss as a result of the disruption. The first constraint is a lower-bound constraint for the number of units lost for product j, given the production and inventory conditions. The second constraint is a total capacity constraint on the assembly plant i. We can replace

the linear objective function with a convex one in a more general case, for example, accounting for lost market share if the loss exceeds a specific threshold.

Solving one instance of this linear program measures the impact of one disruption scenario. A crucial step of using this model is the construction of the *set* of disruption scenarios of interest. The identification of this set is a self-contained step that can be performed by the business executives and risk managers. For example, when the firm aims to identify the most crucial node of the system, the disruption scenarios are constructed as all events that relate to the removal of a single node from the graph. This is the paradigm adopted for the analysis at Ford.

Although the ST model explicitly captures only the last tier of the production system, it can be used to analyze a disruption at a supplier in an upstream tier. To do so, we disrupt the nodes in the final tier that depend on the upstream supplier, and solve ST. This is reasonable if the firm has little control over the nodes prior to the last tier and if the firm knows which final-tier nodes will be affected by the disruption. These assumptions may be too conservative, for example, in situations in which the firm has control over upstream resource allocation and routing. We present a multiple-tier model that addresses this more general case.

The MT model is similar to the ST model. We include the concept of parts, which refers to the set of nodes that are functionally equivalent in the manufacturing process, but potentially processed at a different plant or supplier site.

$$
\begin{array}{rlcll}
\text{Minimize} & \sum_{j\in\mathcal{V}} f_j l_j & & & \\
\text{s.t.} & u_j - \sum_{i\in\mathcal{P}_{jk}} y_{ij}/r_{kj} & \leq & 0, & \forall k \in \mathcal{N}^-(j), \forall j \in \mathcal{D} \\
& \sum_{j\in\mathcal{N}^+(i)} y_{ij} - u_i & \leq & s_i, & \forall i \in \mathcal{U} \\
& u_j & = & 0, & \forall j \in \mathcal{S}^{(n)} \\
& l_j + \sum_{k\in V_j} u_k & \geq & d_j t^{(n)}, & \forall j \in \mathcal{V} \\
& \sum_{k\in A_\alpha} u_k & \leq & c_\alpha t^{(n)}, & \forall \alpha \in \mathcal{A} \\
& l_j, u_j, y_{ij} & \geq & 0. &
\end{array}
$$

The first constraint is a bill-of-materials constraint; for every node j, we limit the production of node j (denoted by u_j) by the most-scarce parent part. More

specifically, for this node j (*e.g.*, an engine), there are multiple parent nodes (*e.g.*, components of an engine). Variable y_{ij} represents the material flow from node i to node j. If two parent nodes, i and i', represent the same physical and (or) functional part (*e.g.*, the same type of bolts from two different suppliers), we say that i and i' are of the same *part type*. We invoke an additional index k to denote the part type of a node, and use r_{kj} to represent the amount of type k parts required to produce one unit of node j. The ratio y_{ij}/r_{kj} is then the units of node j that can be produced with y_{ij} units of type k parts from node i. We use $\mathcal{P}_{jk}$ to represent the set of all nodes that are (1) upstream of j, and (2) of part type k. Hence, $\sum_{i\in\mathcal{P}_{jk}} y_{ij}/r_{kj}$ represents the maximum amount of j that can be produced, given the aggregated supply of type k materials from upstream nodes in $\mathcal{P}_{jk}$.

The second constraint is also a bill-of-materials constraint, which limits the total outflow of node i ($\sum_{j\in\mathcal{N}^+(i)} y_{ij}$) to be less than the sum of production (u_i) and inventory (s_i) at the current location.

The third constraint is the disruption constraint, which limits the production of disrupted node j (*i.e.*, u_j) to be zero. The fourth and fifth constraints are similar to the first and second constraints in the ST model.

In both the ST and MT models, we make the simplifying assumption that processing lead times are not significant relative to the impact of the disruption. In the MT model, we also assume that the costs of rerouting materials and manufacturing changeovers are not significant relative to the impact of the disruption. These are often reasonable assumptions in the context of high-impact disruptions, the effect of which dwarfs the impact of these other issues.

The ST and MT linear programs generate prescriptive contingency plans that minimize the impact of the disruption on the firm's chosen performance metric. Under each disruption scenario, the optimization model generates a corresponding set of optimal values for the decision variables, denoting the best routing and resource-allocation plans for that disruption.

Symbol	Explanation
$\mathcal{D}$	Set of all but the first-tier nodes.
$\mathcal{U}$	Set of all but the final nodes (vehicles).
$\mathcal{S}^{(n)}$	Set of all disrupted nodes for disruption scenario n.
$\mathcal{A}$	Set of all suppliers sites (plants).
A_α	Set of all nodes produced at supplier and (or) plant α.
$\mathcal{V}$	Set of all final nodes (vehicles).
V_j	Set of all final nodes (vehicles) that are of the same type (j).
$\mathcal{N}^-(i)$	Set of parts required to produce node i.
$\mathcal{N}^+(i)$	Set of nodes that require node i.
$\mathcal{P}_{jk}$	Set of all nodes that are in the upstream of node j and of part type k.
$t^{(n)}$	TTR for disruption scenario n.
u_i	Total production quantity of nodes i during time $t^{(n)}$.
l_j	Lost production volume of vehicle type j.
y_{ij}	Allocation of upstream node i to downstream node j during time $t^{(n)}$.
s_i	Finished goods inventory of node i.
r_{kj}	Number of type k parts required to make one unit of node j.
f_j	Performance impact (*e.g.*, profit margin) of one unit of product j.
d_j	Demand for j per time unit.
c_i	Production capacity of node i per unit time.

Table 2.2: This table lists the parameters and variables of the multiple-tier model and their explanations.

2.3.4 Procedure for Finding Pareto Efficient Solutions Under TTR Uncertainty

Given a finite set of n TTR values and an ordering of their importance (given by a manager, for example), we can find a resource-allocation strategy that is Pareto efficient (*i.e.*, not dominated by any other strategy) on this set of TTR values. This is in spirit the same as finding a lexicographically optimal solution in multiobjective optimization [45], where the n objectives correspond to the performance impact under these n TTR values. Using $\mathbf{x}$ to represent the resource-allocation strategy, and $f(\mathbf{x}, t)$ and $\{\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}(t)\}$ as the objective function and feasible region of the TTR model, respectively, we provide the procedure for finding a Pareto-efficient solution as follows:

Algorithm 1 Pareto-Efficient Resource-Allocation Strategy Algorithm

1: Solve the original TTR linear optimization model with $t = t_1$, and obtain resource-allocation strategy $\mathbf{x}^1$, which minimizes $f(\mathbf{x}, t_1)$ over the set $\{\mathbf{x} | \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$.

2: Determine the strategy $\mathbf{x}^2$, which minimizes $f(\mathbf{x}, t_2)$ over the set $\{\mathbf{x} | f(\mathbf{x}, t_1) = f(\mathbf{x}^1, t_1), \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$.

3: For $3 \leq k \leq n$, determine the strategy $\mathbf{x}^k$, which minimizes $f(\mathbf{x}, t_k)$ over the set $\{\mathbf{x} | f(\mathbf{x}, t_i) = f(\mathbf{x}^i, t_i)$ for each $1 \leq i \leq k-1, \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$.

2.3.5 Time-To-Survive Model

We define time-to-survive (TTS) to be the longest time that the firm can last without losing demand after a disruption happens. TTS for the disruption scenario n can be calculated by solving the following linear program. This model is a special case of the TTR model in the sense that we can find the TTS of the network by solving a number of TTR models with different TTR values, and look for the smallest TTR value corresponding to the financial impact being strictly positive. This TTS formulation is more efficient because we can find the TTS by solving a single linear program.

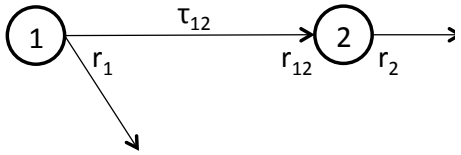

Figure 2-3: Basic setup: τ_{ij} is the lead time between nodes i and j; r_i is the amount of inventory i at node i, r_{ij} is inventory of i at node j (including the pipeline inventory on (i,j)). We assume $r_{ij} \geq \tau_{ij}$.

$$
\begin{array}{lll}
\text{Maximize} & t^{(n)} & \\
\text{s.t.} & u_j - \sum_{i \in \mathcal{P}_{jk}} y_{ij}/r_{kj} \leq 0, & \forall k \in \mathcal{N}^-(j), \forall j \in \mathcal{D} \\
& \sum_{j \in \mathcal{N}^+(i)} y_{ij} - u_i \leq s_i, & \forall i \in \mathcal{U} \\
& u_j = 0, & \forall j \in \mathcal{S}^{(n)} \\
& \sum_{k \in V_j} u_k \geq d_j t^{(n)}, & \forall j \in \mathcal{V} \\
& \sum_{k \in A_\alpha} u_k \leq c_\alpha t^{(n)}, & \forall \alpha \in \mathcal{A} \\
& u_j, y_{ij}, t^{(n)} \geq 0, &
\end{array}
$$

where the constraints and variables are similar to the TTR models, except that (1) $t^{(n)}$ is now a decision variable (TTS), and (2) we do not allow any loss (demand is strictly satisfied in the fourth constraint). The objective value of each optimization instance reveals the TTS of the underlying disruption scenario.

2.3.6 Model with Lead Time

For a general supply chain system (directed acyclic graph) with unique parts (no dual sourcing), lead times can be simplified away by the following transformation:

$$\bar{r}_{ij} = r_{ij} - \tau_{ij} \quad \forall \text{ edge } (i,j) \tag{2.1a}$$

$$\bar{\tau}_{ij} = 0 \quad \forall \text{ edge } (i,j) \tag{2.1b}$$

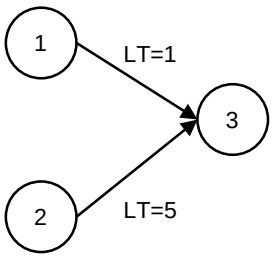

Figure 2-4: Dual Sourcing: nodes 1 and 2 are the same parts.

where i is the upstream node, j is the downstream node, r_{ij} is the nominal inventory level, τ_{ij} is the nominal lead time between nodes i and j. And we know how to solve zero lead time models. Based on these assumptions:

1. Inventory holding cost is negligible.
2. Setup time at each node is insignificant compared with the whole duration of lead time.
3. No demand backlogging.
4. Each node is unique.

2.3.7 Multitree Systems With Dual Sourcing

Dual Sourcing at Top Tier

In the simple setup of Figure 2-4, nodes 1 and 2 are the same parts. If node 1 is down for 1 week, then due to the long lead time of arc (2,3), node 2 does not do anything and the effective TTR is 1 week.

On the other hand, if node 1 is down for 10 weeks, since node 2 has enough time to react, then we can ramp up production at node 2. The effective TTR is 5 weeks.

In general, if one of the dual sources is down at the top tier, then we can transform the inventory levels similar to Equations (2.1), and in addition, change the TTR to the shortest lead time of the alternative sources. Based on the assumptions:

1. Dual sources are uncapacitated.

2. Cost of increased production at alternative source is insignificant compared with the final product margins.

2.3.8 Multitree

Assuming multitree networks – there exists at most one directed path between any pair of nodes – we present an optimization model with size complexity polynomial in the number of nodes and edges, independent from the time horizon.

Definition The supply chain network is defined as the bill-of-materials network, including the current flow of materials and the substituable edges (may carry zero flow during normal operations).

We let $G = (V, E)$ be the network we work with, denote the set of directed paths as $\mathcal{P}$. The set of sink nodes (final products) is S. Denote the rate of flow with function $f : \mathcal{P} \times [0, T] \to \mathbb{R}$, then the flow entering path P at time θ is written as $f_P(\theta)$. Let $\delta^+(i)$ and $\delta^-(i)$ be the set of *paths* starting and ending at node i. The lead time on an path p is $\tau_p \geq 0$. The flow generation (dissipation) at node i is $b_i(t)$ for time period t. Then for a single material, we have the (weak) flow conservation equation:

$$\sum_{\tau=0}^{t} b_\tau + \sum_{P \in \delta^-(i)} \sum_{\theta=0}^{t-\tau_P} f_P(\theta) + r_i \geq \sum_{P \in \delta^+(i)} \sum_{\theta=0}^{t} f_P(\theta), \quad \forall i \in V, t \in T. \tag{2.2}$$

The left hand side is the sum of (i) generation of flow, (ii) incoming flow from upstream nodes that have arrived, and (iii) inventory of materials at this node. The right hand side is the total outflow of flow.

In the case of a supply chain network, where a node may combine multiple upstream parts into one type of downtream part, we have the following modified flow conservation:

$$\sum_{\tau=0}^{t} b_\tau + \min_{q \in q^-(i)} \Big\{ \sum_{P \in \delta_q^-(i)} \sum_{\theta=0}^{t-\tau_P} f_P(\theta) \Big\} + r_i \geq \sum_{P \in \delta^+(i)} \sum_{\theta=0}^{t} f_P(\theta), \quad \forall i \in V, t \in T. \tag{2.3}$$

where $q^-(i)$ is the set of upstream *parts* of i, and $\delta_q^-(i)$ are the paths that enter node i as part type q. Without loss of generality, we assume the part usage ratios are always 1:1.

Directly incorporating these flow conservation constraints would result in an optimization model pseudo-polynomial in size (polynomial in the size of T).

$$\text{TE}: \min \quad \sum_{s\in S} c_s y_s \tag{2.4a}$$

$$\text{s.t.} \sum_{\tau=0}^{t} b_\tau + \min_{q\in q^-(i)} \Big\{ \sum_{P\in\delta_q^-(i)} \sum_{\theta=0}^{t-\tau_P} f_P(\theta) \Big\} + r_i \geq \sum_{P\in\delta^+(i)} \sum_{\theta=0}^{t} f_P(\theta), \quad \forall i \in V\setminus S, t \in T. \tag{2.4b}$$

$$y_i + \sum_{\tau=0}^{t} b_\tau t + \min_{q\in q^-(i)} \Big\{ \sum_{P\in\delta_q^-(i)} \sum_{\theta=0}^{t-\tau_P} f_P(\theta) \Big\} + r_i \geq \sum_{P\in\delta^+(i)} \sum_{\theta=0}^{t} f_P(\theta), \quad \forall i \in S, t \in T. \tag{2.4c}$$

$$f_P(t), y_i \geq 0, \quad \forall t \in T, P \in \mathcal{P}, i \in V. \tag{2.4d}$$

where c_s is the profit margin for product s. Constraints (2.4b) provides the weak flow conservation at each intermediate node. Constraints (2.4c) records the amount of lost sales (y_s) at each sink node. At optimality, either y_s is zero, or the constraints (2.4c) are tight.

Due the following observation, we can actually change the formulation into polynomial in size.

Observation 2.3.1 *The time-expanded network of multitree system G is a set of disconnected multitrees with the same topology as G. The separability comes from the fact that lead time is unique between any two nodes.*

Since the separability relies on the uniqueness of travel time between any two nodes, we can also apply all following arguments to a graph where the travel time between any two nodes is unique, regardless of the path taken.

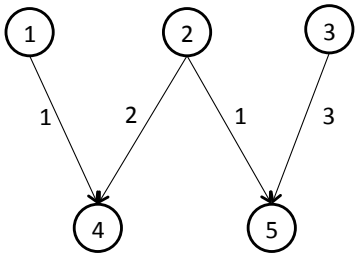

Figure 2-5: A simple multitree network with lead times on the arcs.

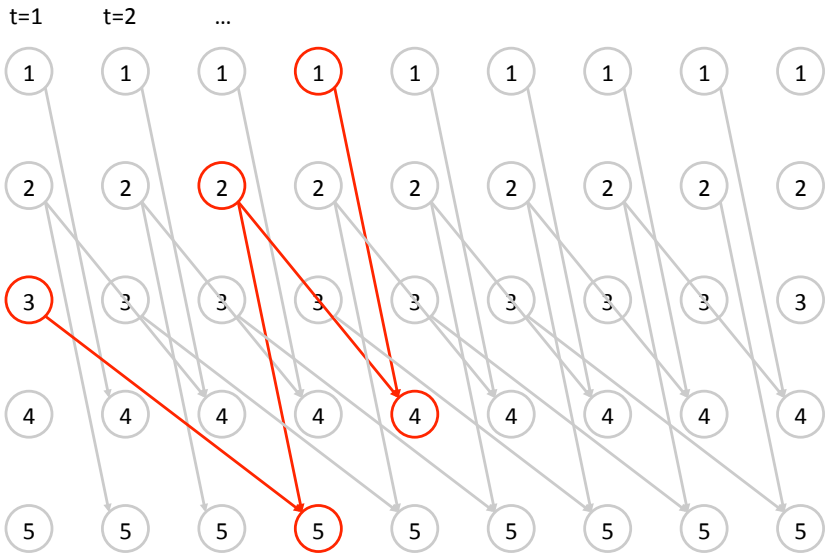

Figure 2-6: The time-expanded version of the network in Figure 2-5, showing its decomposability.

Example We can illustrate this observation by an example (Figures 2-5 and 2-6).

Assuming f is an optimal solution to the time expanded network formulation TE. Then let f^g denote the flow in one of the separable multitree G of the time expanded network (for example, the one highlighted in red in Figure 2-6).

Lemma 2.3.2 *(Smoothing.) For a set of k adjacent neighbor sub-graphs $\{g_i\}$, if each f^{g_i} is feasible on g_i, then flow $\bar{f} = \frac{1}{k}\sum_i f^{g_i}$ is also feasible on g_i, $\forall i$.*

Proof For any two vectors f^{g_1} and f^{g_2} satisfying linear constraints $Af^{g_1} \geq b$ and $Af^{g_2} \geq b$, their average also satisfies the same system of inequalities, *i.e.*, $A(f^{g_1} + f^{g_2})/2 = b$. This argument applies to an arbitrary (finite) number of vectors f^{g_i}. $\square$

$$\text{TE2}: \min \quad \sum_{s\in S}\sum_{t\in\mathcal{T}} c_s y_s(t) \tag{2.5a}$$

$$\text{s.t.} \sum_{\tau=0}^{t} b_\tau + \min_{q\in q^-(i)} \Big\{ \sum_{P\in\delta_q^-(i)} \theta_j f_P(j) \Big\} + r_i \geq \sum_{P\in\delta^+(i)} \theta_j f_P(j), \quad \forall i \in V\setminus S, t\in\mathcal{T}. \tag{2.5b}$$

$$\theta_i y_i(t) + \sum_{\tau=0}^{t} b_\tau + \min_{q\in q^-(i)} \Big\{ \sum_{P\in\delta_q^-(i)} \theta_j f_P(j) \Big\} + r_i \geq \sum_{P\in\delta^+(i)} \theta_j f_P(j), \quad \forall i \in S, t\in\mathcal{T}. \tag{2.5c}$$

$$f_P(t), y_i(t) \geq 0, \quad \forall t\in\mathcal{T}, P\in\mathcal{P}, i\in V. \tag{2.5d}$$

for each P and j, j' is the index such that $\sum_{z=0}^{j}\theta_z - \tau_P = \sum_{z=0}^{j'}\theta_z$. We basically divided the time horizon into $|\mathcal{T}|$ events, the ith interval has length θ_i. Here an event is a sequence of consecutive time periods in which the parameter $b_i t$ remain constant for all i.

Theorem 2.3.3 *Given an assembly network G that is a multitree, we can solve the multi-disruption problem in polynomial time. the objective value of TE2 is the same as TE, and TE2 is polynomial in the number of events and network size.*

Proof TE is the optimization problem on the time expanded network $G^{|T|}$. By Observation 2.3.1, the constraints are decomposable across the events t. The subset of optimal solution corresponding to a particular event may not be identical. But within each event, the parameter values $b_i(t)$ remain constant for all i in V, and for all t in this event. Thus by Lemma 2.3.2, we can smooth these flow and shortage variables to be identical, leading to the formulation TE2. This formulation is polynomial in the number of events and original network size. $\square$

2.4 Implementation

We implemented our model as a decision support system during a three-year research engagement between MIT and Ford. The first phase of the project included the assessment of existing risk-management approaches. In the second phase, we worked with the Ford optimization and information technology (IT) teams to focus on model design and implementation, and the integration of the optimization model and Ford's IT system. The modeler and optimization specialists communicated weekly, and received help from Ford's procurement team to validate the model's output.

Ford's procurement staff used the decision support system in three ways: (1) strategically, to identify exposure to risk associated with parts and suppliers, effectively prioritize and allocate resources, segment suppliers, and develop mitigation strategies; (2) tactically, to track daily changes in risk exposure to alert procurement executives to changes in their risk position; and (3) operationally, to identify effective ways to allocate resources after a disruption. Using the model to conduct a comprehensive analysis of its risk exposures (*i.e.*, the strategic level), Ford identified several supply chain nodes that would have a large impact on its operations if disrupted. These large exposures lie in unlikely places, such as nonstrategic suppliers or parts that the company spends relatively little money to procure. Armed with this information, Ford can make more informed decisions on how to deploy its risk-assessment resources and mitigate the effects of a disruption to these nodes.

In this section, we describe the insights our model provides at the strategic, tactical, and operational levels for Ford's risk-analysis, procurement, and management teams.

2.4.1 Evaluation of Node Criticality with the TTS Model

As we discussed in the previous section, TTR information is not known accurately in many practical situations because of information uncertainty and optimistic supplier assessments. Therefore, the first step of our risk-analysis process is to identify the disruption scenarios that would lead to immediate performance deterioration, namely,

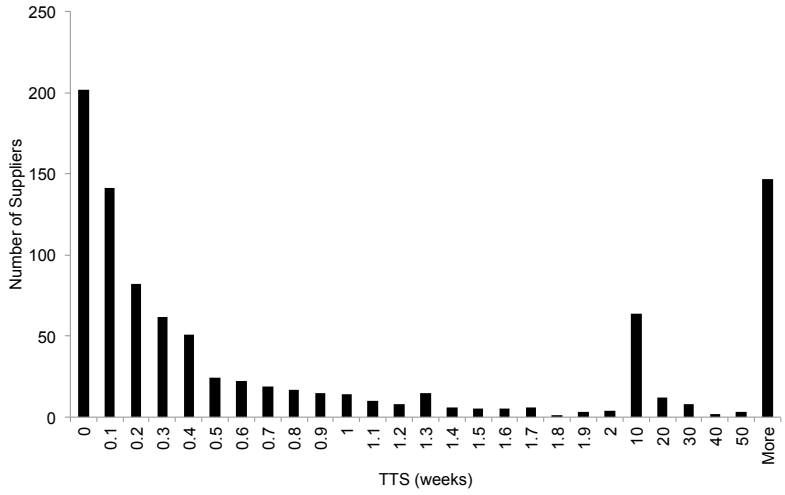

Figure 2-7: A significant portion of the suppliers have very low TTS values, thus requiring more accurate TTR evaluation and closer monitoring for risk-exposure assessment. In addition, some suppliers have very high TTS values, possibly because of redundant inventory buffers.

to find nodes with small TTS values. Nodes that represent higher exposure levels will have a TTS value that is lower than a threshold value, for example, TTR plus a safety allowance.

Figure 2-7 shows that the suppliers included in the analysis have a range of TTS values. Many suppliers have TTS values of less than a week. Ford's management can use this information to concentrate on the PI of low-TTS suppliers and acquire corresponding TTR information. In addition, by identifying the nodes with high TTS values, this analysis can identify potential waste, caused by excessive protection (strategic inventory), within the system. For such nodes, a firm may reduce (strategic) inventory, thus providing significant cost savings.

2.4.2 Application to Strategic Decisions

Strategically, Ford utilizes the TTR model to identify the risk exposure of parts and suppliers, allowing it to prioritize resource allocation. Furthermore, by combining the risk exposure of suppliers with other information, such as the total spend at various supplier sites, Ford gains insights about possible mitigation strategies it could adopt toward various types of suppliers. Below, we describe these applications of the model

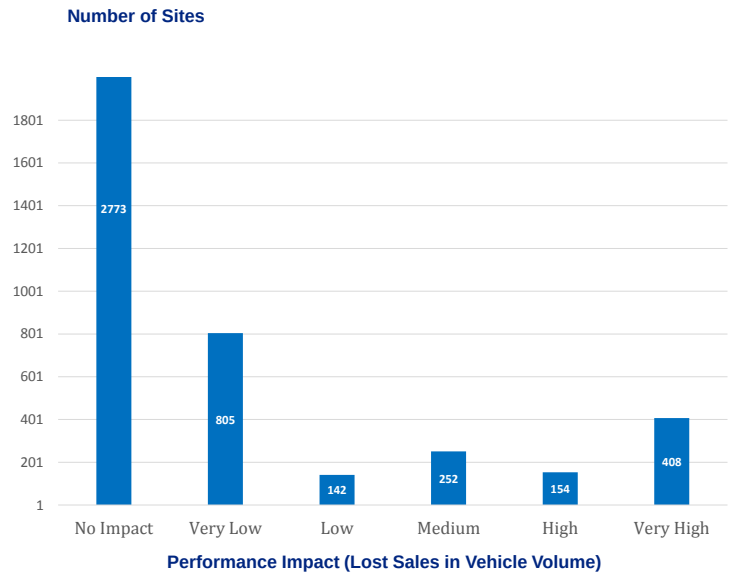

Figure 2-8: Among 4,534 sites examined, 2,773 sites have zero impact at the time of analysis and 408 have very high impact.

to Ford.

Figure 2-8 is based on the PI output (in this case, lost sales measured by the impact of vehicle-production volume) from a model run, including all the critical suppliers and Ford plants that support Ford's North American assembly plants. As the figure indicates, a significant portion of the suppliers do not expose Ford to any risk; however, more than 400 sites have very high PIs.

In Figure 2-9, we take a closer look at these suppliers with very high PIs and see that some of the largest exposures reside in unlikely places, such as the production and (or) procurement of low-cost, commoditized parts. Therefore, some of the traditional risk-mitigation strategies (*e.g.*, focusing on high-spend suppliers) may lead to wasteful resource allocation at low-exposure sites and insufficient protection at high-exposure sites.

Figure 2-9 suggests that Ford should reduce its exposure to risk by segmenting suppliers into three categories depending on the supplier's PI and total spend. Each segment presents a different set of challenges; therefore, Ford should use different

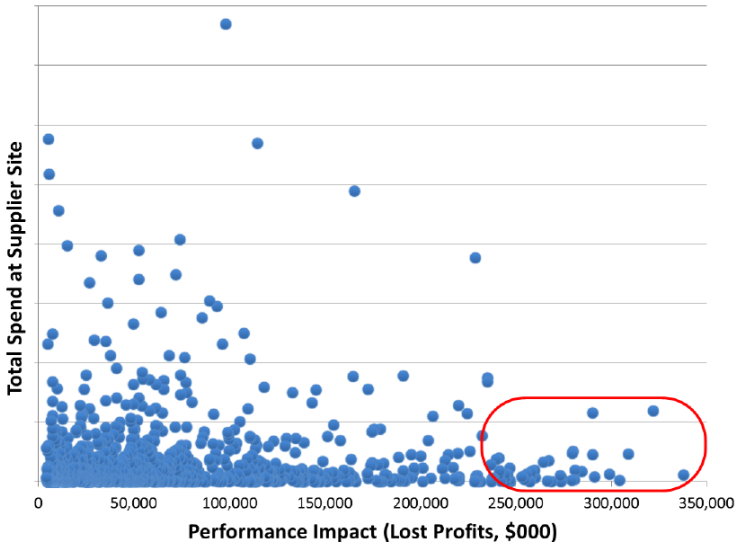

Figure 2-9: Impact of a disruption at a supplier site (node) on Ford's lost profits is not correlated with the amount Ford spends at the supplier. Each circle represents a unique supplier site.

mitigation strategies, as Figure 2-10 illustrates. First, suppliers on the left side of the chart have low exposure; therefore, Ford's primary actions in many of these cases should involve signing long-term supply contracts and tracking inventory.

Supplier sites with high total spend and high PI are at the top right side of the chart. This segment includes, for example, suppliers of seats and instrument panels. These items strongly affect the customer experience, and their prices represent a large portion of the total manufacturing cost. We typically refer to them as strategic components and their corresponding suppliers as strategic suppliers. For many companies, this segment represents 20 percent of their suppliers, which accounts for about 80 percent of total spend. Typically, each of these components has a single strategic supplier. An appropriate supply strategy for these items is to focus on long-term partnerships with suppliers and implement effective supply contracts where Ford can share risks with suppliers and track performance. Importantly, because of the high total spend with these suppliers, Ford may be able to compel some of these suppliers to have backup supply sites in different regions.

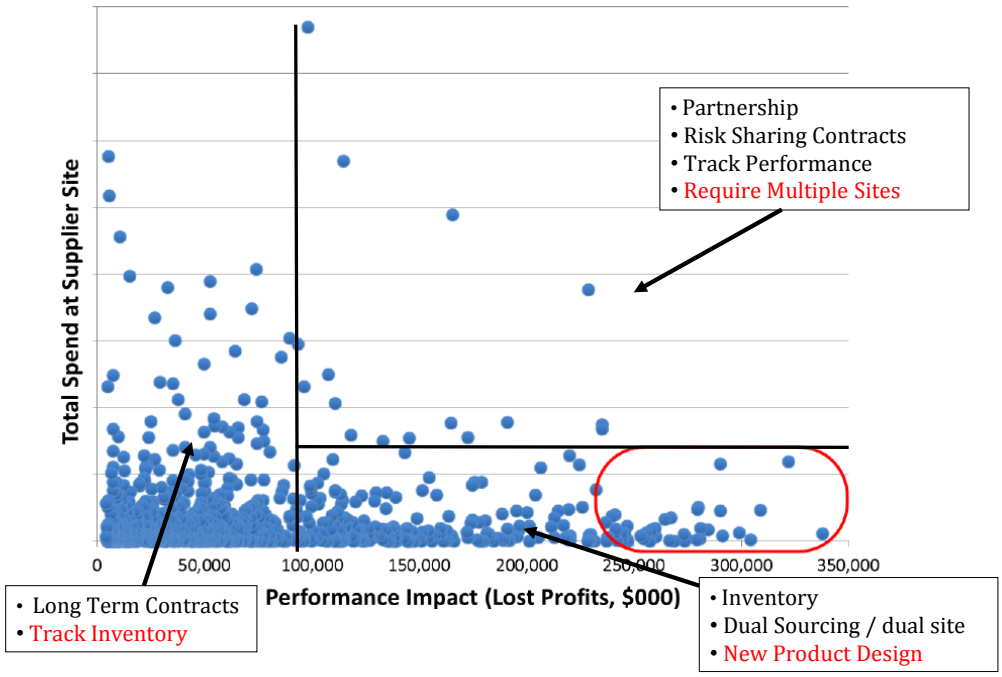

Figure 2-10: This graph suggests supplier segmentation and different risk-mitigation strategies for different groups of suppliers.

The most challenging suppliers are those whose total spend is low and PI is high (*i.e.*, suppliers at the bottom right side of the chart). To ensure supply, a firm may invest in inventory, require the supplier to have dual sites in different regions, or apply a dual-sourcing strategy. Unfortunately, each of these mitigation strategies may cause a problem. Investing in inventory may not be consistent with the lean strategy the company is applying. Low total spend implies that the firm is not in a good position to require the supplier to have multiple sites. In addition, some of these suppliers are associated with high-volume, low-cost, and low-margin components. For these components, competition typically shifts to a few manufacturers that dominate the market because of their lower costs and superior quality; as a result, engaging in a dual-sourcing strategy is difficult. In our experience, one possible mitigation strategy involves a new product design in which components are standardized, allowing the firm to shift more volume and more spend to the supplier; hence, the firm would be in a good position to require dual sites.

2.4.3 Application to Tactical Decisions

Recall that for some components, risk exposure is directly proportional to the level of inventory of that component in Ford's supply chain. To identify risk exposure, pipeline inventory information is uploaded to the system on a regular basis, and the system determines the performance impact by component anywhere in the supply chain. When performance impact is above a specific level, procurement specialists initiate a process to understand the reason and take corrective action. In that respect, our system serves as a control tower that allows the firm to monitor suppliers' performance and inventory trends to take action before problems occur. Because the company takes actions in anticipation of a potential adverse event, it can minimize the financial impact if such events happen.

2.4.4 Application to Operational Decisions

Operationally, Ford supply-risk specialists use the model to respond to a disruption event. For example, a few months ago, political problems in one region motivated the procurement department to identify the high-exposure suppliers in that region and find alternative sources of supply for these components.

In such situations, our TTR model optimizes inventory and capacity allocation decisions when a disruption occurs, assuming that accurate TTR information is available immediately after a disruption occurs. Unfortunately, TTR may be different for different modes of disruptions (*e.g.*, process disruption versus tooling damage), and the firm may not know the exact TTR value when a disruption occurs. Therefore, identifying robust allocations of inventory and capacities under such uncertainty in TTR values is important.

Figure 2-11 provides a stylized example that compares the impact of different resource-allocation strategies when the length of the disruption varies. In this figure, each curve represents the financial impact of one resource-allocation strategy. For example, the solid curve corresponds to the optimal resource-allocation strategy for TTR=1; we evaluate the performance of this resource-allocation strategy for all TTR

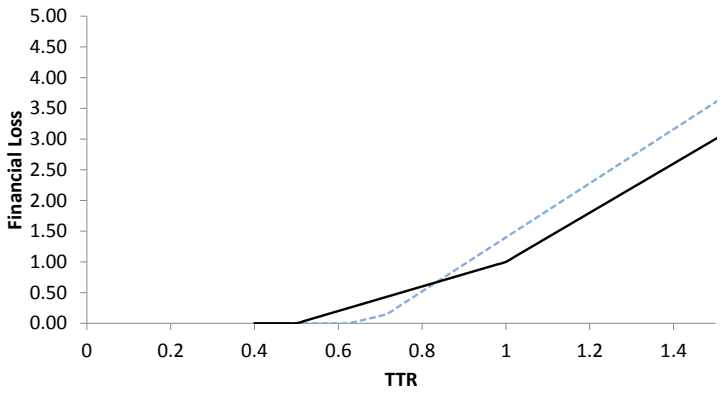

Figure 2-11: Each curve represents the financial impact of one resource-allocation decision. The solid curve is optimal for TTR=1, but suboptimal for TTR=0.7; the reverse is true for the dotted line.

values between 0 and 2. Similarly, the dotted curve is associated with the optimal resource-allocation strategy when TTR=0.7. Figure 2-11 suggests that neither of the two strategies dominates; that is, neither strategy outperforms the other on all TTR values between 0.7 and 1. This is not always the case. Another stylized example (Figure 2-12) shows that the strategy associated with the solid line outperforms the strategy associated with the dotted line. The former strategy outperforms all other strategies for TTR values between 0 and 2 (Figure 2-12 does not show other strategies); that is, the solid line either matches or dominates the performance of any other resource-allocation strategy determined by using a single TTR value between 0 and 2.

Motivated by these different cases, we developed an algorithm that can (1) find a dominating strategy if it exists, or (2) find a Pareto-optimal strategy, which always exists. That is, managers can specify the ranking of potential TTR values, and the algorithm provides a strategy that is not dominated by any other strategy.

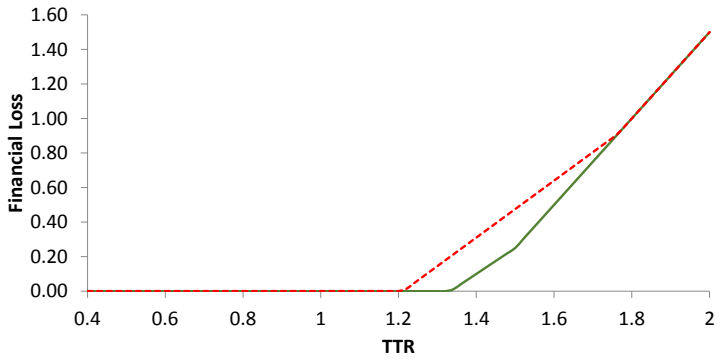

Figure 2-12: Each curve represents the financial impact of one resource-allocation decision. The solid curve is optimal for all nonnegative TTR values.

2.4.5 System Architecture

To allow procurement and risk specialists to take advantage of our model, Ford developed a decision support system that integrates various databases, the TTR and TTS models, and a data-visualization software package. The data sources include Ford's material requirements planning (MRP) system, its purchasing database, and sales-volume planning information. Figure 2-13 describes the system architecture in which various data sets, including bill of material, part routing, inventory levels, and plant vehicle volumes are used to map Ford's supply chain. [67] describe the basic methodology of mapping the Ford supply chain. Our optimization engine uses the results to generate the various performance impacts. These performance measures are then presented to the users by Tableau data visualization, which includes a geographic mapping capability. Thus, users can view results both in tabular form and in various graphical formats. Figure 2-14 provides a screenshot of our interface; the size of the circles identifies the performance impact of a disruption to the supplier in that geographic location. The two tables at the bottom of Figure 2-14 provide detailed information on suppliers and parts. For each supplier, the table on the left provides the vehicle affected, total spend at that supplier, financial impact, and production-volume impact if that supplier is disrupted for the duration of its TTR. The table on

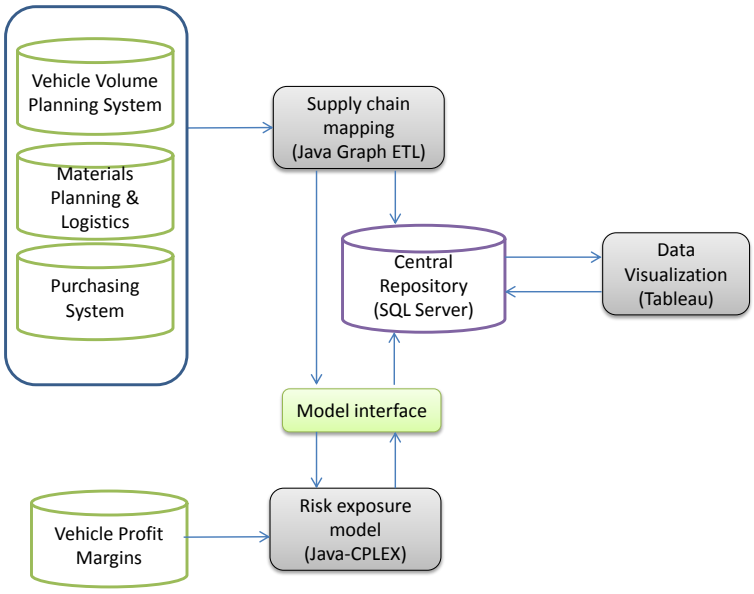

Figure 2-13: Ford's risk-analysis framework integrates databases, our risk-exposure model, and an output visualization tool.

the right provides all affected parts associated with each supplier.

Procurement and risk specialists regularly use the system to track risk exposures in real time as inventory levels fluctuate and the supply chain structure evolves. The frequency of updates relies on the efficient data-integration technology developed by Ford and the computational tractability of our linear programming models.

2.4.6 Realized Benefits for Ford

Ford spends several million dollars per year to proactively manage its operational and supply chain risk. Two points make clear why Ford must deploy its risk-management resources in the most effective manner possible. First, it must spread these resources across a huge operational footprint. Ford's operations and supply chain include over 4,400 Tier 1 supplier sites, hundreds of thousands of lower-tier suppliers (Tier 2 and lower), over 50 Ford-owned facilities, 130,000 unique parts, 35 billion total parts

Supplier	Vehicle Impacted	Total Part Cost	Financial Impact	Volume Impact
x11	cc1	$$$	$$$	vv
x12	cc2	$$$	$$$	vv
x13	cc3	$$$	$$$	vv
x14	cc4	$$$	$$$	vv
x15	cc5	$$$	$$$	vv
x16	cc6	$$$	$$$	vv
x17	cc7	$$$	$$$	vv
x18	cc8	$$$	$$$	vv
x19	cc9	$$$	$$$	vv
x20	cc10	$$$	$$$	vv

Supplier	Part Names
x11	y11
x11	y12
x11	y13
x12	y21
x12	y22
x13	y31
x13	y32
x13	y33
x13	y34
x14	y41

Figure 2-14: Critical suppliers are mapped to geographical location. The size of a circle indicates the magnitude of the impact on Ford's performance if a supplier is disrupted. The table view gives detailed information regarding the financial and vehicle-volume impact associated with these suppliers.

annually, and over $80 billion annually in external procurement costs. Second, the cost of failure can be huge because supply chain disruptions can have a significant impact on Ford's ability to match supply with demand. Indeed, Ford estimates that the lost revenue associated with a disruption can be significant. To illustrate this point, recall that in 2011 Toyota lost 800,000 units of production volume as a result of the Japan earthquake and more than 240,000 units of production volume as a result of the flood in Thailand. Honda faced similar challenges [127].

The risk-exposure model produces important and tangible benefits for Ford to help it effectively identify and manage its risks. First, Ford has identified supplier sites that have a material impact if disrupted, but that it did not recognize as high-exposure sites. Based on the model results, 2,600 Tier 1 supplier sites have nonzero vehicle-volume impact that, if disrupted, would adversely impact Ford's revenue by up to $2.5

billion. Ford now classifies these exposures as high-priority issues that require a formal remediation analysis. Identifying these suppliers is particularly compelling because they represent 1,500 additional supplier sites that will now receive a larger share of Ford's risk-management resources. Identifying and addressing such risk exposures directly supports Ford's corporate strategy.

A second benefit is identifying low-exposure supplier sites that are currently receiving an excessive allocation of Ford's risk-management resources. The model has identified over 400 supplier sites that Ford includes in its risk-monitoring program, but which pose insignificant exposure to the company if disrupted. This information has allowed Ford to more efficiently allocate its supplier risk-management resources.

By reallocating these resources, Ford is better able to protect itself from the highest-impact exposures in its operations and supply chain. For example, the lost revenue associated with a two-week disruption to the newly classified high-impact supplier sites ranges from several hundred thousand dollars to $2.5 billion; in contrast, the lost-revenue impact associated with a two-week disruption to each of the formerly classified high-impact supplier sites is minimal. In the words of Ford manager Michael Sanders, "This has been one of the key game changers for us. This enables us to focus on the supplier sites which would have a high or very high impact on performance if disrupted, and lets us put all our resources and all our knowledge into making sure we have robust plans to protect us in the event that something happens with any one of those sites" [130].

Finally, our model detects hidden risks in Ford's supply chain. For example, it identified a low-cost sensor that has high vehicle exposure; however, because of the low total spend, Ford's procurement group was not paying much attention to this component. Following the risk analysis, the commodity team acknowledged the sourcing concentration and associated risk and developed a mitigation strategy.

Chapter 3

Optimizating Biodefense Supply Chain

3.1 Introduction

In this paper, we study the problem of designing a medical countermeasures (MCM) supply chain in preparation against bioattacks, *i.e.*, the intentional release of pathogens or biotoxins against humans to cause serious illness and death. Bioattacks have been a rising first-order concern to many countries worldwide in the last 15 years; the U.S. alone, for instance, has poured approximately $60 billion into biodefense preparedness since the 9/11 attacks [78]. Bioattacks are considered a major threat because a minute quantity of pathogens is sufficient to infect humans; furthermore, the appropriate MCMs need to be administered within a short time window to effectively reduce casualty. For example, if bioterrorists were to release *Bacillus anthracis*, *i.e.*, anthrax, over a large city, hundreds of thousands of people could be at risk, and MCMs would need to be administered to them within a few days to have the intended effect [103, 142]. Distributing such large MCM quantities to the general public within such short time periods poses a considerable operational challenge, which is often compounded by attack detection delays.

In order to be prepared to deliver adequate medication in a timely manner in response to a bioattack, first, MCMs need to be stockpiled at appropriate locations.

Some inventories can be stored in central locations to take advantage of deployment flexibility, pooling effects, and scale economies in holding costs, while others can be positioned closer to populous areas, or even predispensed to the intended users in the form of home medical kits, in order to reduce transportation time. Second, for end users to have prompt access to deployed MCMs, target areas need to have adequate dispensing capacity in place, *e.g.*, in the form of publicly accessible dispensing facilities or dedicated courier delivery services. Too little capacity would result in inefficient usage of available inventory and devastating delays; too much would result in wasted resources. Given candidate storage locations, where and how much MCM inventory should then be prepositioned in order to cost-effectively defend against bioattacks? How much dispensing capacity should be installed at target areas? Our goal in this paper is to address these research questions and develop a decision support framework that could guide policy design.

MCM supply chain in the U.S. The Centers for Disease Control and Prevention (CDC) maintains MCM inventories within the Strategic National Stockpile (SNS), which is a critical component of the national health security programs managed by the CDC Office of Public Health Preparedness and Response (PHPR) [106, 107]. The inventory storage and delivery infrastructure of SNS form a crucial line of defense for the nation's public health security.

For many bioattack threats, SNS currently stores MCM inventories in central repositories, while its delivery infrastructure enables deployment of virtually any amount of stored MCMs to almost any major city in the U.S. within 12 to 36 hours, at negligible shipping costs. But centrally stored inventories may not be responsive enough for events such as anthrax attacks. In recent years, CDC has been increasingly aware of alternatives, namely the prepositioning of MCMs to locations closer to the population. Prepositioning has been raised as a first-order issue for SNS, and PHPR policymakers have commissioned several studies to examine different strategies [103, 142]. To date, several facets of the problem have been analyzed in isolation, including location and transportation options, cost of each system component, and health effects of delayed treatment.

However, a comprehensive cost-benefit analysis at the *system level* that *integrates* individual considerations into a practical quantitative decision support tool has been elusive. Our work helps fill in this gap, the importance of which is indeed highlighted in the following excerpt from a report prepared by the Board of Scientific Counselors for PHPR [73, 105]:

> "Both analytic and simulation (experimental) modeling activities should be increased substantially. Modeling will allow [the Division of] SNS to make quantitatively-based decisions on how much inventory to hold and where to hold it. An end-to-end model capturing the flow of materials in the SNS, as well as costs and logistical and health measures, should begin at the SNS-managed inventory site and go all the way to the point of dispensing to the public. Using such models will reveal bottlenecks, provide cost estimates, and help SNS properly evaluate the costs and consequences of alternative Response Supply Chain configurations."

Herein, we focus on an end-to-end analytic model as described in the quote above. The absence of such integrative models so far attests to the underlying technical challenges: (a) the multiplicity of decisions involved, both strategic and operational; (b) the subtle way prepositioning could influence bioattackers' actions; and (c) the scale of the problem. In particular, on the strategic level, policymakers need to decide how much inventory to store and how much capacity to install among thousands of locations, in anticipation of an attack. On the operational level, in response to an attack, appropriate shipment decisions need to be made to efficiently dispense MCMs based on the specific demand and supply conditions. The *static* inventory positioning and capacity installation decisions need to be made in conjunction with myriads of *contingent* shipment policies. Furthermore, some information about planned responses is already available in the public domain, and prepositioning MCMs, particularly predispensing medical kits, is likely to grant to bioattackers even greater visibility to planned responses. Consequently, bioattackers, because they usually act in a pre-planned way to inflict the greatest damage possible, are more likely to tar-

get under-served populated areas, as opposed to ones that have access to abundant stockpiles. That attackers' actions are likely to be adversarial is well recognized by biodefense policymakers, see, *e.g.*, [142], and the intelligence and military operations research literature [7, 27, 63, 64, 79].

Our methodology. We model the SNS as a network: inventory stockpiles as nodes, and shipment routes as edges. Bioattacks correspond to demand surges at some of the nodes. To model interactions between bioattackers' and policymakers' decisions, we adopt the defender-attacker-defender model (DAD), which can be viewed as a *multistage* robust optimization (RO) model [17, 27]. Specifically, we consider MCM prepositioning as a first-stage static defender decision, bioattacks as second-stage adjustable attacker decisions, and MCM shipments as third-stage adjustable defender decisions. We calibrate the objective and uncertainty set of possible demand scenarios in ways that reflect policymakers' considerations. In particular, we deal with minimizing either inventory holding costs under survivability guarantees for all possible attack scenarios, or worst-case inventory and health costs, where each scenario comprises simultaneous attacks in multiple locations. To preclude overly pessimistic scenarios in which too many locations are attacked, our uncertainty sets include attack budget constraints, which are both standard and crucial in the DAD framework [5].

Unfortunately, there are no scalable solution approaches that perform provably well even for the inventory prepositioning problem, let alone joint optimization of inventory and dispensing capacity decisions. In particular, DAD models and multistage RO problems of the type we study here are generally intractable [5, 17, 27, 28]. There has been valuable effort in solving reasonably sized problems [6], but they are orders of magnitude smaller than what our problem entails. In the RO literature, a popular heuristic is the so-called Affinely Adjustable Robust Counterpart (AARC), which restricts adjustable decisions to be affine in the uncertainty variables [17]. Although AARCs often lead to tractable formulations, they can perform rather poorly in high-dimensional problems [19]. Worse, all known performance guarantees for AARCs apply to problems with "symmetric" uncertainty sets. The aforementioned attack budget constraints that are necessary in our problem lead to sets that are the

"least-symmetric," casting considerable doubt over the performance of AARCs in our setting (see section 1.2 and discussion in §3.3.2 for more details).

Our contributions. First, for the inventory prepositioning problem, we provide an approximate solution approach, with theoretical backing for its performance, that can deal with network sizes on the order of millions of nodes. In particular, we consider the Affinely Adjustable Robust Counterpart, which reduces to a linear optimization problem (LP) by restricting the adjustable shipment policies to be affine in the demand shocks. We provide theoretical backing to its performance by proving that AARC is in fact optimal under certain conditions. Our work contributes to the RO literature by proving optimality of AARC in a new and important context—see our discussion in Literature Review and in §3.4. Numerical studies we conducted revealed that AARCs provide near-optimal performance for the inventory prepositioning problem under general settings as well.

Second, we consider the joint problem of inventory prepositioning and dispensing capacities optimization. We show how dispensing capacity, which is costly, can be captured as a first-stage decision and how the third-stage shipment decisions can be reformulated so as to reflect the underlying capacitated dispensing operations. We derive the AARC for this joint problem as an LP and conduct numerical studies to confirm that it still produces near-optimal solutions.

Third, successfully applying AARC to our problem speaks to the potential of its broader impact for other DAD problems. In particular, the AARC approach is applicable to tri-level programs that are nonlinear and with integer variables. In the sense of [5], AARC is suitable for eight out of the nine types of operator models (the only exception is simulation-based operator model that does not have a closed form description at all).

Fourth, we use our model in a thorough, large-scale case study of MCM supply chain design for the SNS to defend against anthrax attacks. Tractability of our approach enables us to perform the first study we are aware of in the literature that is at a realistic, nationwide scale and deals with networks with millions of nodes. We calibrate our model using multiple literature sources that studied different facets of

SNS design in isolation. Our integrated framework enables us to derive prescriptive recommendations to the CDC. For example, we find that if CDC wanted to ensure 85% survivability for attacks that simultaneously affected (any) two states, each with at most two cities being attacked, with a detection time of under 48 hours, then the minimum required annual inventory and dispensing capacity budget would be about $210 million; for survivability target of 92%, the budget would increase to $553 million. This increasing marginal cost phenomenon, along with other cost/policy implications, can be explored quantitatively within our model.

3.2 The Inventory Prepositioning Problem

We begin by considering the MCM inventory prepositioning problem, assuming for now that sufficient dispensing capacity is installed. For exposition purposes, we use SNS to exemplify our discussion, but our model and solution heuristic generalize to any network. The SNS network is designed to protect the public susceptible to bioattacks in a set of geographic locations, *e.g.*, densely populated towns or neighborhoods, which we shall refer to as *demand* locations. The demand locations are split into administrative divisions according to a hierarchical structure, *e.g.*, boroughs, municipalities, provinces, states, etc. There are $L+1$ levels in the hierarchy, indexed by $\ell = 0, \ldots, L$, whereby each division at the lowest (Lth) level comprises precisely one demand location. Divisions at some intermediate ℓth level comprise subsets of divisions at the lower $(\ell+1)$th level in a nested fashion; the highest (0th) level includes a single division, *e.g.*, federal/national level. For a division i at the ℓth level, let $\mathcal{P}(i)$ be its parent division at the $(\ell-1)$th level, $\mathcal{C}(i)$ the set of (children) divisions that it includes at the $(\ell+1)$th level, and $\mathcal{D}(i)$ the set of demand locations that it includes, *i.e.*,

$$\mathcal{D}(i) = \{\text{demand location } j \,:\, \underbrace{\mathcal{P}(\mathcal{P}(\ldots\mathcal{P}(\mathcal{P}}_{L-\ell \text{ times}}(j))\ldots)) = i\}.$$

Figure 3-1 includes an illustrative example, where the demand locations are cities, split under state and federal divisions.

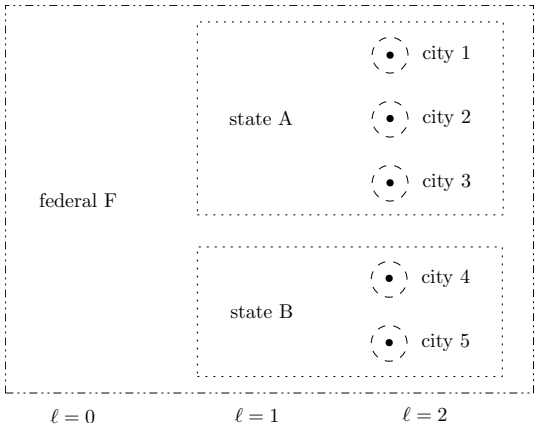

Figure 3-1: Five demand locations (cities) split into three administrative divisions, federal $\ell = 0$, state $\ell = 1$ and city $\ell = 2$. Note that $\mathcal{P}(1) = \mathcal{P}(2) = \mathcal{P}(3) = \text{A}$ and $\mathcal{P}(4) = \mathcal{P}(5) = \text{B}$; $\mathcal{C}(\text{A}) = \mathcal{D}(\text{A}) = \{1, 2, 3\}$ and $\mathcal{C}(\text{B}) = \mathcal{D}(\text{B}) = \{4, 5\}$; $\mathcal{D}(\text{F}) = \{1, 2, 3, 4, 5\}$.

To serve the demand locations, each administrative division at the $0, \ldots, L-1$ levels maintains a *stockpile* of appropriate medical countermeasures. In particular, in case of a bioattack, MCMs would typically need to be shipped from the—usually remotely located—stockpiles to the demand locations, where they are subsequently dispensed to the affected population.

Let the demand locations and stockpiles correspond to nodes in a directed graph (V, E), which we refer to as demand and stockpile nodes, respectively. The subset of demand nodes is denoted with V_D. We index the root node (stockpile node at the highest level division) with 0. The edges in E connect stockpile nodes with demand nodes. Specifically, $(i, j) \in E$ if and only if inventory can be shipped from i to $j \in V_D$ once an attack occurs. Shipments have negligible costs and are not subject to capacity constraints for the purposes of the SNS network.[1] However, there is a fixed lead time τ_{ij} for shipping inventory from i to j for all $(i, j) \in E$. Note that even though a "hierarchical inventory network" like the SNS resembles a tree structure, and the demand and stockpile node sets are usually disjoint, our model and solution approach are more general and can be applied to any network (V, E), accommodating demand and stockpile on any node.

We focus on a particular bioattack threat, in anticipation of which a single MCM

[1]Unlike shipping MCMs, dispensing is often capacitated—a situation we address in §3.5.

type is stored. Inventory of this MCM can be prepositioned at any node $i \in V$. If i is a demand node, the stored inventory corresponds to predispensed medical kits. Let $\mathbf{x} \in \mathbb{R}^{|V|}$ denote the amounts of stored inventories allocated at the nodes. The set of all feasible inventory allocations, $X \subseteq \mathbb{R}^{|V|}$, is assumed to be a polytope and could include inventory related constraints, *e.g.*, non-negativity constraints, budget constraints subject to inventory setup cost and variable cost. A unit of inventory is adequate to treat precisely one individual. Associated with storing inventory at any node are purchasing, replenishment, maintenance costs, etc. We refer to these costs simply as *holding* costs, and denote the per-unit costs with $\mathbf{h} \in \mathbb{R}^{|V|}$. That is, storing x_i MCM units at node $i \in V$ costs $h_i x_i$.

When an attack takes place, a subset of demand nodes are affected, and parts of their populations are in need of treatment with the MCMs. We also say that division i is affected, if any of its children demand nodes $\mathcal{D}(i)$ is affected. For example, if city 1 is attacked in Figure 3-1, we say that it is affected and so is state A. Let $\mathbf{d} \in \mathbb{R}^{|V_D|}$ be the (a priori unknown) vector with the realized number of affected individuals in each of the demand nodes. Once $\mathbf{d}$ is revealed, inventory stored at the affected demand nodes is available immediately. In case of shortages, inventory from other stockpile nodes can be shipped to satisfy the demands, as long as there exist edges between them. For each edge $(i, j) \in E$, let f_{ij} denote the amount of inventory shipped from i to satisfy demand at node j. Note that shipment decisions are contingent on the realized demand, which we sometimes make explicit by writing them as $f_{ij}(\mathbf{d})$. Inventory shipped from i would be made available for treatment at node j only after τ_{ij} time units. Delays in treatment of affected population could lead to lower probability of survival. We use $\rho_{ij} \in [0, 1]$ to denote the *survival probability* (or survivability) of an individual in node j if treated with inventory shipped from node i, and $\overline{\rho} \in [0, 1]$ if left untreated. For example, suppose that node j is affected and 50% of the affected individuals are treated with inventory prepositioned at that node, 30% are treated with inventory shipped from node i, and 20% are left untreated. The *average* survivability would then be $0.5\rho_{jj} + 0.3\rho_{ij} + 0.2\overline{\rho}$.

With respect to the possible attack scenarios, CDC considers ones involving si-

multaneous attacks to multiple locations across the country [73]. In view of this, we consider attack scenarios (or demand vectors) where, for each administrative division i, at most Γ_i of its children divisions $\mathcal{C}(i)$ are affected. We assume Γ_i's to be non-negative integers and refer to them as the *attack scale* parameters. For example, if the administration levels are states and cities, $\Gamma_i = 2$ for state i would mean that we consider attack scenarios with at most 2 cities in state i being affected.

In case demand location i is affected, the maximum number of individuals in need for treatment at that location is $\hat{d}_i$.[2] A further probabilistic characterization of the possible demand vectors $\mathbf{d}$ appears to be, unfortunately, prohibitive for a variety of reasons. One relates to limited historical data, given that bioattacks have so far been rarely encountered in practice. A second, and more important, reason is the nature of terrorist attacks being *adversarial* and *endogenous*, instead of purely random and exogenous. In particular, unlike natural disasters, for instance, bioattacks are premeditated and often carefully planned in order to maximize the damage inflicted. The choice of which locations to attack then, which essentially drives the realization of $\mathbf{d}$, is likely to be influenced by the inventory decisions, given that prepositioned inventory (or lack thereof) is partially visible to the public. Both issues of endogeneity and the adversarial choice of the attacked areas are indeed considered and acknowledged by policymakers [142] and academics studying the DAD framework [4, 6, 7, 28, 32, 79, 88]. In the next section, we consider a suitable model for $\mathbf{d}$ that addresses these issues, instead of postulating a probabilistic description.

In general, more (less) inventory in the system leads, on the one hand, to higher (lower) holding costs and, on the other hand, to fewer (more) potential casualties. Therefore, CDC needs to balance inventory costs and the costs of life loss. We consider two possible ways to navigate this tradeoff, leading to two problem formulations:

1. Policymakers explicitly quantify the cost of life lost to be b monetary units. The

[2]The maximum number of affected individuals for a bioattack is estimated by policymakers based on census considerations, *e.g.*, population density, transmission and contagion characteristics, *e.g.*, the required quantity of spores to carry out an attack and their ability to remain aloft and travel further for aerosol attacks, and others. The estimation usually relies on experimental data, see for example our discussion in Section 3.6.

problem is then to select an inventory allocation so as to minimize holding costs plus worst-case life loss costs under all possible attack scenarios. We refer to this formulation as *Life Loss Cost* (LLC).

2. Policymakers specify survivability targets: let $1 - \epsilon_i$ be the target for average survivability in demand node i, for some $\epsilon_i \in [0, 1]$, which we refer to as the *survivability target* parameters. The problem is then to select an inventory allocation so as to minimize holding costs, while providing the survivability guarantees implied by the specified targets in all demand locations, under all possible attack scenarios. We refer to this formulation as *Life Loss Guarantee* (LLG).

Our model is well-suited for policy decision-making purposes. In particular, the attack scale parameters $\{\Gamma_i : i \in V\}$ enable policymakers to specify the severity of attacks they want to hedge against, both in terms of *magnitude*, since higher values of Γ for lower level nodes translate into more areas affected, *e.g.*, cities, and in terms of *complexity*, since higher values of Γ for higher level nodes translate into the affected areas being dispersed among more divisions, *e.g.*, states. Furthermore, the life loss cost b in the (LLC) formulation, or the survivability target parameters $\{\epsilon_i : i \in V_D\}$ in the (LLG) formulation, reflect the policymakers' aversion to casualties and insufficient coverage. Based on the policymakers' selections, the model prescribes appropriate inventory prepositioning strategies and elicits their minimum required costs, which allows for tradeoff analyses.

3.3 Formulation and Solution Approach

For both (LLC) and (LLG), we formulate the inventory prepositioning problem as a DAD problem, or multistage robust optimization problem. In particular, at the first stage, the defender decides on the prepositioned inventory allocation $\mathbf{x} \in X$. At the second stage, the attacker chooses an attack scenario or demand vector $\mathbf{d}$ from the set of scenarios compatible with the specified attack scale parameters, denoted

by $U \subset \mathbb{R}^{|V_D|}$, so as to minimize expected survivors. We also refer to U as the demand uncertainty set. At the third stage, after the realized demand vector $\mathbf{d} \in U$ is revealed, the defender decides on the shipment strategy $\{f_{ij}(\mathbf{d}) : (i,j) \in E\}$. Formally, we model (LLC) as

$$\text{(LLC):} \quad \min_{\mathbf{x},\, \mathbf{f}(\cdot),\, \mathbf{s}(\cdot)} \quad \mathbf{h}^T\mathbf{x} \quad + \quad \max_{\mathbf{d} \in U} \quad b \sum_{i \in V_D} \left((1-\overline{\rho}) s_i(\mathbf{d}) + \sum_{j:(j,i) \in E} (1-\rho_{ji}) f_{ji}(\mathbf{d}) \right) \tag{3.1a}$$

$$\text{subject to} \quad x_i \geq \sum_{j:(i,j) \in E} f_{ij}(\mathbf{d}), \quad \forall i \in V, \forall \mathbf{d} \in U \tag{3.1b}$$

$$s_i(\mathbf{d}) + \sum_{j:(j,i) \in E} f_{ji}(\mathbf{d}) = d_i, \quad \forall i \in V_D, \forall \mathbf{d} \in U \tag{3.1c}$$

$$s_i(\mathbf{d}) \geq 0, \quad \forall i \in V_D, \forall \mathbf{d} \in U \tag{3.1d}$$

$$f_{ij}(\mathbf{d}) \geq 0, \quad \forall (i,j) \in E, \forall \mathbf{d} \in U \tag{3.1e}$$

$$\mathbf{x} \in X. \tag{3.1f}$$

The auxiliary variables $\{s_i(\cdot) : i \in V_D\}$ capture demand shortages, *i.e.*, the number of affected individuals left untreated at each node. Note that both the shipment decision variables $\mathbf{f}(\cdot)$ and the shortage variables $\mathbf{s}(\cdot)$ are adjustable, contingent on the realized demand vector $\mathbf{d}$. The objective is to minimize inventory holding costs plus worst-case life loss costs. Constraint (3.1b) is a node capacity constraint: the total amount of inventory shipped from node i should be less than the amount stockpiled at i. Constraint (3.1c) defines the demand shortage variable s_i for each node i, and d_i is the ith element of $\mathbf{d}$.

Similarly, we formulate (LLG) as follows:

$$\text{(LLG):} \quad \min_{\mathbf{x},\, \mathbf{f}(\cdot),\, \mathbf{s}(\cdot)} \quad \mathbf{h}^T\mathbf{x} \tag{3.2a}$$

$$\text{subject to} \quad \overline{\rho} s_i(\mathbf{d}) + \sum_{j:(j,i)\in E} \rho_{ji} f_{ji}(\mathbf{d}) \geq (1-\epsilon_i) d_i, \quad \forall i \in V_D,\, \forall \mathbf{d} \in U \tag{3.2b}$$

$$\text{(3.1b), (3.1c), (3.1d), (3.1e), (3.1f).} \tag{3.2c}$$

Compared to (LLC), the (LLG) formulation has an objective of minimizing holding costs, while ensuring that the average survivability at each node is higher than the target set, as reflected by the added constraint (3.2b). Throughout this paper we assume that $\rho_{ji} \geq \overline{\rho}$ for all $(j,i) \in E$, *i.e.*, treatment increases survivability; and $\overline{\rho} \leq 1 - \epsilon_i$ for all $i \in V_D$, otherwise (LLG) reduces to a trivial problem with optimal cost of zero. Note that we impose a survivability target on a node level instead of in aggregation—this provides better control and could ensure equitable population protection across regions. We note later in Theorem 3.4.3 that (LLG) is a special case of (LLC).

3.3.1 Set of Attack Scenarios

We provide a formulation for the set of attack scenarios under consideration. In particular, recall that, given some attack scale parameters $\{\Gamma_i : i \in V\}$, a possible scenario involves demand nodes affected so that, for each administrative division i, no more than Γ_i of its children divisions $\mathcal{C}(i)$ are affected. Consider

$$U := \Bigg\{ \mathbf{d} \in \mathbb{R}^{|V_D|} : \boldsymbol{\xi} \in \mathbb{R}^{|V|},\, d_i = \hat{d}_i \xi_i \,\forall i \in V_D,\, 0 \leq \xi_i \leq \xi_{\mathcal{P}(i)} \,\forall i \in V \setminus \{0\},\, \xi_0 = 1, \\ \sum_{j \in \mathcal{C}(i)} \xi_j \leq \Gamma_i \xi_i \,\forall i \in V \Bigg\}, \tag{3.3}$$

where ξ_i indicates whether location i is attacked, and $\hat{d}_i$ is the maximum number of people affected when location i is attacked. As we will show, only the extreme points of U will be relevant for the optimal solution of (LLC) and (LLG). Therefore, to

argue that the polytope U represents the set of attack scenarios we are interested in, it suffices to show that its extreme points precisely correspond to the attack scenarios under consideration. Indeed, as we will also show, the auxiliary variables $\boldsymbol{\xi}$ are binary at the extreme points of U. The constraint $\xi_i \leq \xi_{\mathcal{P}(i)}$ enforces all parent divisions of an affected node to also be affected, *i.e.*, $\xi_j = 1$ for all j such that $i \in \mathcal{D}(j)$ and i is affected. Finally, constraint $\sum_{j \in \mathcal{C}(i)} \xi_j \leq \Gamma_i \xi_i$ ensures that, if i is affected, at most Γ_i of its children divisions are affected, as we required. As a side note, it can be readily seen that the set U also includes all remaining attack scenarios that are compatible with the attack scale parameters, but have a number of affected individuals $d_i \leq \hat{d}_i$. To facilitate exposition, we also let

$$\Xi := \left\{ \boldsymbol{\xi} \in \mathbb{R}^{|V|} : \, 0 \leq \xi_i \leq \xi_{\mathcal{P}(i)} \, \forall i \in V \setminus \{0\}, \, \xi_0 = 1, \, \sum_{j \in \mathcal{C}(i)} \xi_j \leq \Gamma_i \xi_i \, \forall i \in V \right\}.$$

We can then simply express $U = \left\{ \mathbf{d} \in \mathbb{R}^{|V_D|} \, : \, d_i = \hat{d}_i \xi_i \, \forall i \in V_D, \, \boldsymbol{\xi} \in \Xi \right\}$.

3.3.2 Solution Approach

By involving both static, $\mathbf{x}$, and recourse decisions, $\mathbf{f}(\cdot)$, $\mathbf{s}(\cdot)$, formulations (LLC) and (LLG) fall into the class of so-called multi-stage adjustable robust optimization problems (ARO). Problems in that class are, in general, computationally intractable [17], because they require the optimization over functions, or policies, instead of vectors, and this makes them infinite-dimensional problems. Specifically, $\mathbf{f}(\cdot)$ and $\mathbf{s}(\cdot)$ are policies that could take different values contingent on the uncertain parameters' realization (namely, the demand $\mathbf{d}$).

Most popular techniques to deal with ARO problems in the literature are heuristics and constraint/column generation methods. One popular heuristic, which we also adopt to tackle our problem, is to limit attention to policies restricted to depend affinely on the uncertain parameters, which are often referred to as *Affine Policies (AP)*, as opposed to *Fully-Adjustable Policies (FP)*. This restriction often enables tractability, see, *e.g.*, [17]. In our setting, the affinely adjustable robust counterpart

of (LLG), for example, is

$$\text{(ALLG):} \min_{\mathbf{x},\mathbf{F},\mathbf{S}} \quad \mathbf{h}^T\mathbf{x} \tag{3.4a}$$

$$\text{subject to} \quad \overline{\rho}(\mathbf{S}_i^T\mathbf{d} + S_i^0) + \sum_{j:(j,i)\in E} \rho_{ji}(\mathbf{F}_{ji}^T\mathbf{d} + F_{ji}^0) \geq (1-\epsilon_i)d_i, \quad \forall i \in V_D, \forall \mathbf{d} \in U \tag{3.4b}$$

$$x_i \geq \sum_{j:(i,j)\in E} (\mathbf{F}_{ij}^T\mathbf{d} + F_{ij}^0), \quad \forall i \in V, \forall \mathbf{d} \in U \tag{3.4c}$$

$$(\mathbf{S}_i^T\mathbf{d} + S_i^0) + \sum_{j:(j,i)\in E} (\mathbf{F}_{ji}^T\mathbf{d} + F_{ji}^0) = d_i, \quad \forall i \in V_D, \forall \mathbf{d} \in U \tag{3.4d}$$

$$\sum_{j:(j,i)\in E} (\mathbf{F}_{ji}^T\mathbf{d} + F_{ji}^0) \leq d_i, \quad \forall i \in V_D, \forall \mathbf{d} \in U \tag{3.4e}$$

$$\mathbf{F}_{ji}^T\mathbf{d} + F_{ji}^0 \geq 0, \quad \forall (i,j) \in E, \forall \mathbf{d} \in U \tag{3.4f}$$

$$\mathbf{x} \in X, \tag{3.4g}$$

where $\mathbf{F}_{ij} \in \mathbb{R}^{|V_D|}$, $F_{ij}^0 \in \mathbb{R}$ for all $(i,j) \in E$, and $\mathbf{S}_i \in \mathbb{R}^{|V_D|}$, $S_i^0 \in \mathbb{R}$ for all $i \in V_D$ are vectors of decision variables corresponding to the affine policies' coefficients. The affinely adjustable version of (LLC) is of similar form. We can reformulate (ALLC) and (ALLG) using standard robust optimization techniques (see Corollary 1.3.5. in [17]) to obtain linear optimization problems. Importantly, the resulting linear optimization problems are polynomial in size of the original inputs, enabling tractability and scalability.

On the flipside, tractability of the AP heuristic to deal with ARO problems often comes at the cost of suboptimal solutions. In fact, for some AROs, the suboptimality gap between the objective value under AP and the objective value of the original formulation under FP can grow indefinitely with the dimension of the problem [19]. As we pointed out in the literature review, a handful of papers have recently identified conditions under which AP are indeed optimal. These conditions require the absence of simplex-type constraints in the uncertainty set. In our model, however, the simplex-type constraints $\sum_j \xi_j \leq \Gamma_i \xi_i$ are essential to preclude excessively conservative demand scenarios, where an arbitrary number of nodes are affected. Con-

sequently, given the current state of affairs in the robust optimization literature, the performance of AP heuristics (ALLC) and (ALLG) remains questionable. The next section is devoted to providing evidence that (ALLC) and (ALLG) are indeed likely to produce near-optimal solutions for the original (LLC) and (LLG) formulations.

3.4 Optimality of Affine Shipment Policies

We provide evidence that affine policies are near-optimal for our problem formulations (LLC) and (LLG). First, and more importantly, we analytically show that under additional assumptions (on the survivability parameters and the network structure), AP are indeed optimal. Second, we conduct numerical studies illustrating that the suboptimality gap remains small for instances that violate the additional assumptions guaranteeing optimality.

3.4.1 Optimality Result

Consider the following assumptions.

Assumption 3.4.1 *The survival probabilities under treatment are all equal, i.e.,* $\rho_{ij} = \rho \ \forall (i,j) \in E$.

Assumption 3.4.2 *Stockpiles serve all demand nodes in their division, and only demand nodes in their division, i.e.,* $(i,j) \in E \iff j \in \mathcal{D}(i)$.

Assumption 3.4.1 requires that the difference in shipment times between stockpile and demand nodes bears no effect on the survivability of treated individuals. Note that this assumption will be violated for biothreats with incubation periods shorter than the shipment times, *e.g.*, for nerve-agents that require treatment within minutes or hours after an attack [142]. However, it will be satisfied for biothreats with longer incubation periods, *e.g.*, for anthrax attacks that are detected early (see also our discussion in Section 3.6). Assumption 3.4.2 requires that stockpiles, which are maintained by administrative divisions, serve affected locations only within their division.

In the U.S., this would mean that the stockpiles maintained by states are reserved for usage by their residents. In emergency situations, however, neighboring states, for example, could also provide assistance and this assumption might not hold—we explore this further in the next section.

Let $z^\star_{\text{LLC}}$, $z^\star_{\text{LLG}}$, $z^\star_{\text{ALLC}}$, $z^\star_{\text{ALLG}}$ be the optimal values of formulations (LLC), (LLG), (ALLC), (ALLG) respectively. We have the following result.

Theorem 3.4.3 *Under Assumptions 3.4.1 and 3.4.2, affine policies are optimal for (LLG), i.e.,*

$$z^\star_{\text{LLG}} = z^\star_{\text{ALLG}}.$$

Furthermore, if $\rho = 1$, affine policies are optimal for (LLC), i.e.,

$$z^\star_{\text{LLC}} = z^\star_{\text{ALLC}}.$$

To illustrate the applicability of our methodology beyond the scope of MCM inventory prepositioning, we prove our result in a more general model. In particular, for any node we consider shipments to any of its children nodes, not just its demand (leaf) nodes. Furthermore, we allow demand to occur at any node in V, not just the (leaf) nodes in V_D. We formalize this generalization within the proof of Theorem 3.4.3, which can be found in Appendix 3.8, and show that it subsumes problems (LLC) and (LLG) as special cases. We discuss alternative applications in Section 3.7.

3.4.2 Performance of Affine Policies for the General Case

To quantify the performance of the AP heuristic for more general cases, we conducted two numerical studies, in which we sequentially relaxed the optimality-guaranteeing Assumptions 3.4.1 and 3.4.2. For brevity, we present our studies for the (LLC) formulation—our studies on the (LLG) formulation yielded quantitatively similar results.

We measured the AP heuristic's performance via its suboptimality gap, defined as the relative difference between the heuristic's optimal cost, $z^\star_{\text{ALLC}}$, and the true

optimal cost, $z^\star_{\text{LLC}}$, that is $(z^\star_{\text{ALLC}} - z^\star_{\text{LLC}})/z^\star_{\text{LLC}}$. Because the underlying problem is intractable, it is impossible to compute $z^\star_{\text{LLC}}$ for large instances. To quantify the gap then, we rely on the method in [136] to generate a valid lower bound on $z^\star_{\text{LLC}}$. Using this bound instead of the true optimal cost enables us to obtain an *upper bound* on the suboptimality gap. Despite such conservativeness, we found the performance of the AP heuristic to be strong across a wide range of parameter values: median suboptimality gaps were less than 1.5%, and, importantly, they did not grow with the problem's scale.

In Study 1, we relax Assumption 3.4.1 and allow the survival probabilities to vary. At a high level, we created 5000 instances of (LLC). For each instance, we generated a random tree-style graph with varying numbers of nodes. Inventory cost, antibiotic efficacy, and attack scale parameters were also randomly sampled—see Table 3.1 for the sampling ranges, and Appendix 3.9 for more details.

Parameter	Sampling Range
Approximate number of nodes	$\{100, 200, 500, 1000, 2000, 5000, 10000\}$
Number of levels	$\{2, 3, 4, 5\}$
Inventory cost	$[0, 0.2] \times b$
Efficacy	$[0, 1]$
Attack scale	$[0, 0.2] \times \lvert C_i \rvert$ for each node i
Demand ($\hat{d}_i$)	[0,1000] for each node i

Table 3.1: Parameter setup for Study 1: b is the demand loss cost, and $|C_i|$ is the number of children for node i.

In Study 2, we relax Assumption 3.4.2 and introduce edges that violate the tree structure. In particular, we considered the same setup as in Study 1, and then for every pair of nodes that were not already connected and resided in adjacent levels in the graph, we added an edge between them with some probability, p_{arc}, so as to obtain "non-tree" graphs. We sampled p_{arc} in $[0, 0.01]$, resulting in graphs that had up to about 200% more edges than the tree-style graphs in Study 1.

Using a high-performance computing cluster, we ran the AP heuristic and the exact solution method in [136] for each instance. Each run was given a 4-hour time limit, 16 GB of memory, and 2 CPUs (2.1 GHz each). Across both Studies 1 & 2,

AP successfully solved all 10000 instances, with an average run time of 6 seconds. Although the exact solution method managed to solve only 2% of the instances, it generated valid lower bounds for 91% of them, with an average run time of 14116 seconds. The generated lower bounds were then used for our comparisons.

Results. Table 3.2 provides statistics of the upper bounds on the suboptimality gaps we obtained by using the generated valid lower bounds on the optimal costs. Qualitatively, our analysis reveals that the suboptimality gap is small and does not

Min	1st Quantile	Median	Mean	3rd Quantile	Max
0.0%	0.4%	1.4%	1.9%	3.0%	9.9%

Table 3.2: Statistics of upper bounds on suboptimality gaps for AP heuristic for Studies 1 & 2.

have strong correlation with any of the parameters we varied, including, among others, the number of nodes and the attack severity. In particular, the studies suggest that the suboptimality gap does not grow with the network size. This is a key consideration given that we cannot tractably characterize the gap for even larger networks due to limitations of the exact solution method. More details are included in Appendix 3.9.

3.5 Optimizing Dispensing Capacity

In this section, we relax the assumption of sufficient dispensing capacity being pre-installed, and jointly optimize inventory prepositioning and dispensing capacity decisions. By dispensing capacity, we refer to the ability of the appropriate local authorities to distribute MCMs to the general public at demand nodes subsequent to a bioattack—technically, it corresponds to the MCM delivery rate from so-called Receive, Stage and Store facilities to end users. Such joint optimization is essential: too much inventory and not enough dispensing capacity would result in inventory buildup while the population suffers from dispensing delays; too little inventory and too much capacity would result in idle workers and under-utilized resources.

Limited dispensing capacity could introduce further delays before the affected

population has access to MCMs, which in turn could affect their efficacy. Unlike delays due to transportation lead times, which are independent of the MCM quantity shipped, delays due to limited dispensing capacity are dependent on the MCM quantity being dispensed. This creates the need to introduce a time dimension explicitly in our model.

We illustrate our approach using the (LLG) formulation. Subsequent to an attack, consider T discrete time periods, each with duration δ, indexed by $t = 1, 2, \ldots, T$. Let u_j be the dispensing capacity in MCM units per time period at demand node j. In view of this capacity, a shipment from some node i, although it would still arrive at j after τ_{ij} time periods, might not be immediately available to the population at its entirety. To capture this, let f_{ijt} be the MCM amount shipped from i and dispensed at j at time t. To measure antibiotic efficacy, let ρ_t be the survival probability if treatment is received t time periods after the attack. We use δ small enough so that the probability of survival remains approximately constant and equal to ρ_t within time period $[(t-1)\delta, t\delta]$.

Let c_j be the cost per unit capacity installation (*e.g.*, training of staff, preparation and maintenance of dispensing facilities), and let p_j be the cost per unit of MCM dispensed after an attack. The joint inventory prepositioning and dispensing capacity optimization problem can be cast as:

$$\text{(LLGC):} \quad \min_{\mathbf{x},\,\mathbf{u},\,\mathbf{f}(\cdot),\,\mathbf{s}(\cdot)} \max_{\mathbf{d}\in U} \quad \mathbf{h}^T\mathbf{x} + \mathbf{c}^T\mathbf{u} + \mathbf{p}^T\mathbf{f}(\mathbf{d}) \tag{3.5a}$$

$$\text{subject to} \quad x_i \geq \sum_{j:(i,j)\in E} \sum_{t=1}^{T} f_{ijt}(\mathbf{d}), \quad \forall i \in V,\, \forall \mathbf{d} \in U \tag{3.5b}$$

$$s_j(\mathbf{d}) + \sum_{i:(i,j)\in E} \sum_{t=1}^{T} f_{ijt}(\mathbf{d}) = d_j, \quad \forall j \in V_D,\, \forall \mathbf{d} \in U \tag{3.5c}$$

$$\overline{\rho} s_j(\mathbf{d}) + \sum_{i:(i,j)\in E} \sum_{t=1}^{T} \rho_t f_{ijt}(\mathbf{d}) \geq (1-\epsilon_j) d_j, \quad \forall j \in V_D,\, \forall \mathbf{d} \in U \tag{3.5d}$$

$$\sum_{i:(i,j)\in E} f_{ijt}(\mathbf{d}) \leq u_j, \quad \forall j \in V_D,\, \forall \mathbf{d} \in U,\, \forall t \in [T] \tag{3.5e}$$

$$f_{ijt}(\mathbf{d}) = 0, \quad \forall (i,j) \in E,\, \forall \mathbf{d} \in U,\, \forall t \in [\tau_{ij} - 1] \tag{3.5f}$$

$$f_{ij}(\mathbf{d}) \geq 0, \quad \forall (i,j) \in E,\, \forall \mathbf{d} \in U \tag{3.5g}$$

$$s_i(\mathbf{d}) \geq 0, \quad \forall i \in V_D,\, \mathbf{d} \in U \tag{3.5h}$$

$$\mathbf{x} \in X, \mathbf{u} \geq 0. \tag{3.5i}$$

Constraint (3.5e) ensures that no more than u_j MCM units are dispensed at demand node j across all origin nodes i, at each time period t. Constraint (3.5f) ensures that no MCMs shipped from node i can be dispensed at node j before the associated shipment lead time τ_{ij}. All other constraints are similar in spirit to the ones we have in (LLG).

By restricting the flow and demand shortage variables to be affine in demand vector $\mathbf{d}$, the robust counterpart of (LLGC) remains a linear optimization problem. Thus, we are able to solve large problem instances that match the fidelity and scale of a national biodefense network.

Finally, note that introducing a time dimension here can be thought of as simply introducing additional edges in the set E. Therefore, beyond adding capacity variables and accounting for costs, the only essential way that (LLGC) differs from (LLG)

is the inclusion of capacity constraints (3.5e). In the Appendix 3.9, we present a numerical study similar to the ones in §3.4.2, in which we quantify the AP heuristic's suboptimality gap for problems including such capacity variables, costs, and constraints. As before, we find performance to be strong, with a median suboptimality gap less than 1%.

3.6 Strategic National Stockpile Supply Chain

In this section, we apply our work in a case study on SNS design for aerosolized *Bacillus anthracis* (anthrax) attacks in the U.S., which are of particular interest to public health experts owing to their relatively high probability of occurrence and potentially devastating impact [142]. Anthrax spores spread easily in the air to affect a large number of people through inhalation, with high fatality rates if left untreated for even just a few days. More precisely, if t is the time between infection and treatment, then, according to the studies of [25, 26, 158, 159], the survival probability $\rho(t)$ of anthrax infected population can be approximated, for $t \leq 200$ hours, by

$$\rho(t) = e^{-(0.004t)^2}. \tag{3.6}$$

Considering the time it would take to detect an anthrax attack and the time to ship and dispense MCMs from warehouses to individuals, which is likely to be on the order of days [142], CDC is exploring different options in prepositioning MCMs and coordinating the SNS network to achieve better antibiotic efficacy. We now discuss how to calibrate our model to tackle this problems.

3.6.1 Model calibration

For the purposes of this study, we consider the (LLGC) formulation, that is, policymakers specify coverage targets and the model optimizes the required inventory, capacity, and dispensing costs.

Network. There are 12 federally-managed SNS warehouses for MCM storage, with

their exact locations being classified for security reasons. It is known that the locations have been chosen such that MCMs can be transported from them to any state within 24-36 hours of a deployment call [107]. In our model, it is appropriate then to model these warehouses as a single virtual federal stockpile node. In case of an attack, inventories can be shipped from the federal inventory node to state-managed warehouses, where additional inventory is usually stored. We therefore model the SNS stockpile network with 1 federal stockpile node, 52 state and special district nodes. Inventory from the state warehouses can then be forwarded to city, county, and local authorities for dispensing to the general public. Using census data to obtain precise information about major cities or so-called Metropolitan Statistical Areas (MSAs), counties, we include 377 MSA nodes, 3221 county nodes, and 77072 neighborhood nodes. In addition, we also include 77072 household nodes, each denoting the aggregate demand in a neighborhood. The federal stockpile node serves all downstream state nodes. State nodes serve not only their downstream MSAs, but also all MSA nodes in neighboring states. MSAs serve predominantly their downstream counties, but also neighboring MSAs and counties, and so forth. In total, the network comprises 157795 nodes and 335010 edges.

Transportation times. As we remarked above, the time to ship MCMs from the federal node to any state level node varies between 24 to 36 hours. From state level to MSA/county level, we assume the transportation time to be within 2 to 6 hours, *e.g.*, if shipped via ground transportation, depending on the geography of the state. According to field experiments, including, for example, a study in Philadelphia in 2005 [1] and another one in Minneapolis in 2008 [11], policymakers have estimated the shipment of MCMs from local authorities to dispensing points to take 10 to 12 hours.

Holding costs. We consider the SNS design for one MCM type, *e.g.*, *ciprofloxacin* or *doxycycline*, to treat individuals affected by anthrax attacks in the U.S. Inventory costs include the cost of purchasing, storage, management, replenishment, and shipment from manufacturer to warehouse. According to a commissioned paper by the Institute of Medicine [68], the costs for anthrax antibiotics stored at regional ware-

houses amount to $2.10 per person, while home kits are more expensive at $10 per person, due to higher packaging and delivery costs. We estimate the federal, state, MSA, county, and neighborhood holding costs to be between $1 and $2 per person. In general, upstream holding costs are cheaper than downstream ones due to anticipated scale economies. Annual management and replenishment costs are 85% of purchase cost according to [68].

Demand. To estimate the demand volume in a neighborhood at each MSA in case of an attack, we assume that an airborne attack has the same spread radius in every geographic region, and we use a value of 3000 square miles as in the field study on the Minneapolis MSA [11]. The affected population at node i can then be calculated as (population density at i) $\times$ min{3000 sq miles, area of neighborhood}. We obtained population density and area data of neighborhood from statistics of the Census Bureau [153].

Dispensing modes and costs. Two dispensing modes are currently available: (1) Point-of-Dispenses (PODs), and (2) door-to-door delivery by the United States Postal Service (USPS) [157]. PODs are what the supply chain literature would consider a "pull" strategy, in which people commute to pre-specified locations in their neighorhoods to pick up MCMs for their families. The USPS option is a "push" strategy, in which the staff members of USPS deliver MCMs into the mailboxes of each household. Dispensing capacity is shared among neighborhoods, so that there are approximately 2 PODs per county.

Primary costs associated with the PODs setup are: medical staff training costs and wages, supervisor training costs and wages, security wages, and public facility rental costs. Using the recommendation in [162], we estimate a labor wage of $18.64 per hour, $2785 security cost per POD per day, and $5000 administrative overhead cost per POD per day; for each POD, the total number of staff is 300 and can support a dispensing service rate of 1000 patients per hour. The resulting cost of operating a POD during dispensing phase is $3.39 per hour for a one-person/hour increase in dispensing capacity. For capacity installation costs, we assume that a two-day drill is performed annually, resulting in $162.72 per unit capacity per year.

The main costs associated with the USPS option are the wages for the delivery staff and wages for the accompanying security officers. Based on the Minneapolis-St. Paul study [68] with USPS delivery involving 179 staff, wage of $23.72 per person per hour, covering 205,000 households in 8.5 hours (estimate of 6.9 hours spent on transportation and 1.6 hours on material handling) for a population density of 1000 households per square mile, we derive a population-density-dependent USPS dispensing cost function: \$ $(356/\sqrt{\text{density}} + 19)$ per hour for every one-person/hour increase in delivery capacity. For capacity installation costs, we also assume a two-day drill every year. In addition to these costs, the USPS option also has an upper-bound on how many workers can take up the delivery job, as each delivery worker has to be accompanied by one security officer, and the total number of security officers is likely a bottleneck during such an emergency situation. We assume that each county can support at most 10 delivery workers at any given time.

Survival parameters. We choose a time step of $\delta = 8$ hours, in accordance with the standard work shift length of 8 hours per group of staff for dispensing and delivery [142]. Then based on the survival probability calculation aforementioned in (3.6), for a detection time of 48 hours, we calculate $\{\rho_1, \rho_2, \rho_3, \rho_4, \dots\}$ to be $\{0.957, 0.944, 0.929, 0.912, \dots\}$. If a person inhales anthrax spores and is left untreated, his/her survival probability is estimated to be 0.2 [154].

3.6.2 Results

Having calibrated the model, we compute the optimal costs for different policy parameters: survivability target, attack scale, detection time, and dispensing mode. Table 3.3 is a summary of the model inputs and outputs. Table 3.4 provides solution times and information about the size of the resulting optimization problems.

The default input parameters we consider are attack scale parameters $\Gamma_i = 2$, $\forall i \in V$, and detection time $\tau_0 = 48$ hours.[3] Below, we perform several sensitivity analyses to illustrate the flexibility of our framework and how it can be used to guide policymaking.

[3]For a discussion of anticipated detection times in case of an anthrax attack, see [142].

Input	Output
Survivability target: ϵ_i	Total cost (inventory+capacity+dispensing)
Attack scale: Γ_i	Optimal inventory prepositioning strategy: $\mathbf{x}^\star$
Detection time: τ_0	Optimal capacity installation: $\mathbf{u}^\star$

Table 3.3: Input/output for our case study.

Min	1st Quantile	Median	Mean	3rd Quantile	Max
192	719	798	671	839	954

Table 3.4: Solution time (seconds). The resulting optimization problems for the case study instances have 1100000 variables and 115000 constraints, approximately, after pre-processing using Gurobi.

Varying survivability target. Figure 3-2 depicts the minimum required annual cost as a function of the survivability target. For example, for an 80% survivability target the annual budget needs to be about $194 million dollars; if the survivability target is set to 92%, the budget would need to be increased to $553 million, approximately. The curve in Figure 3-2 also illustrates that increasing the survivability target beyond 90% requires a rather steep cost increase. We can also interpret the curve as a Pareto frontier associated with the cost-survivability tradeoff: the region left to the curve represents the achievable outcomes.

Figure 3-3 shows how much inventory and at which level it needs to be prepositioned for different survivability targets. It can be seen that as the target increases, more inventory is required. Notably, for a target less than 90%, predispensed medi-

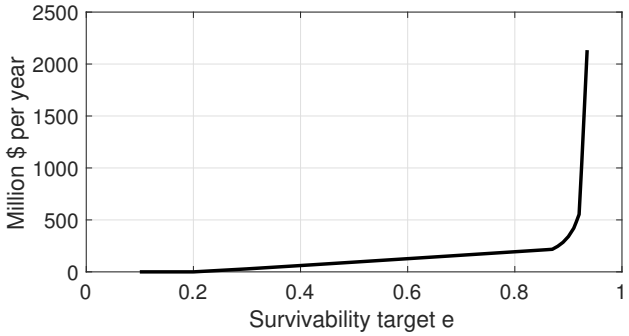

Figure 3-2: Minimum required annual cost for different survivability targets.

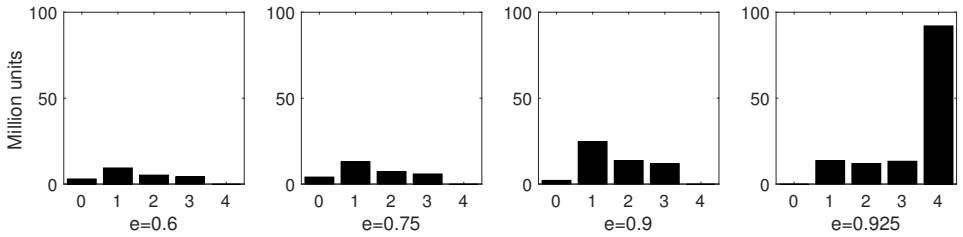

Figure 3-3: Amount (y-axis) and location (x-axis) of prepositioned inventory for different survivability targets $\{0.6, 0.75, 0.9, 0.925\}$; at the x-axis, level 0 is federal, level 1 is state, level 2 is MSA, level 3 is county, and level 4 is predispensed medical kits in households.

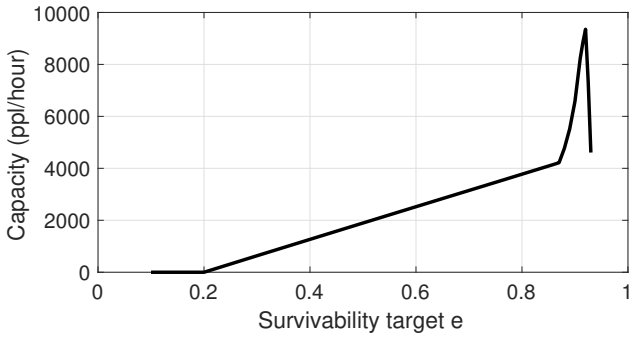

Figure 3-4: Average capacity installation per county (number of people served per hour) for different survivability targets.

cal kits are redundant, whereas for a higher target a rather large number of them is required.

Figure 3-4 shows county-level capacity decisions for different survivability targets. Although required capacity naturally increases with the desired target, interestingly, it drops at a steep rate beyond a certain point.

Put together, the analysis of varying survivability target illustrates a phase change that takes place around the target of 90%, approximately. To better understand this phenomenon, recall first that survivability targets translate into certain responsiveness requirements, through relationship (3.6), for example, with higher targets calling for higher responsiveness. For targets lower than 90% then, storing inventory upstream at federal/state levels provides acceptable responsiveness, and because it is so much more cost-efficient per unit and also enjoys pooling benefits (in the sense that one kit could serve multiple downstream recipient households), it makes predispensed

kits redundant. For targets higher than 90%, upstream storage no longer provides adequate responsiveness for the entire SNS network, and predispensed kits become necessary, which explains their introduction to the prepositioning strategy. The high rate at which these kits are introduced then, compared with kits stored upstream, can be explained by the lack of pooling benefits (one predispensed kit could only serve a single household). This in turn explains the observed steep increase in inventory costs and the decrease in required dispensing capacity.

Varying attack scale. We now explore different attack scale parameters. Figure 3-5 reports the minimum required annual costs and optimal inventory amounts at each level, under nine different cases: $\{1, 2, 3\}$ states being attacked and each state having $\{1, 2, 3\}$ MSAs or counties affected. Recall that the first parameter reflects the geographic complexity of the attack, and the second parameter the magnitude of an attack within a state.

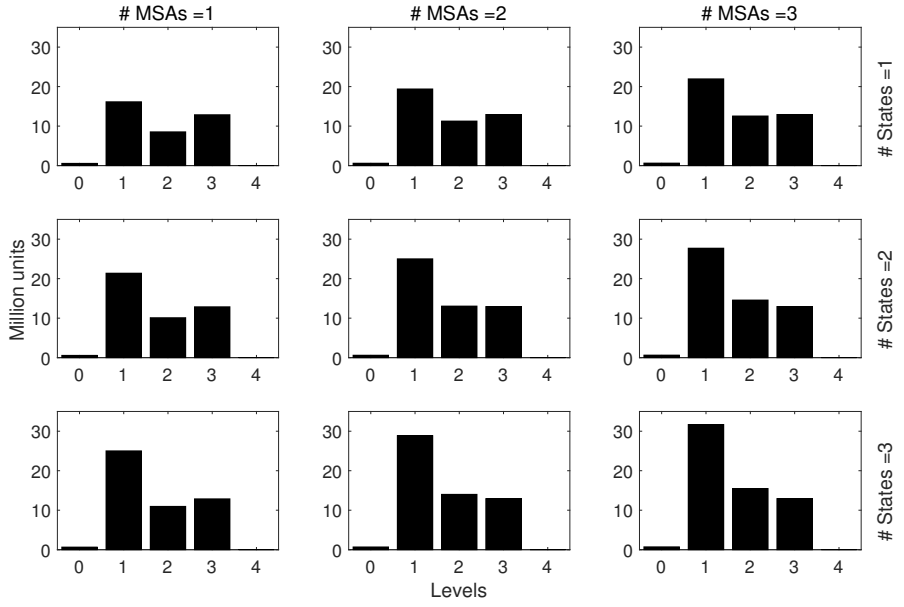

Figure 3-5: Amount (y-axis) and location (x-axis) of prepositioned inventory for different attack scale parameters; at the x-axis, level 0 is federal, level 1 is state, level 2 is MSA, level 3 is county, and level 4 is predispensed medical kits in households.

Varying detection time. Figure 3-6 shows the cost-survivability trade-off for different detection times τ_0, highlighting the importance of timely response mechanisms.

In particular, if the detection mechanism is able to identify an attack and start deployment within 60 hours, approximately $788 million USD is needed for a 90% survivability target annually. For enhanced detection mechanism capable of detecting attacks within 24 hours, the corresponding annual costs reduce to just over $210 million.

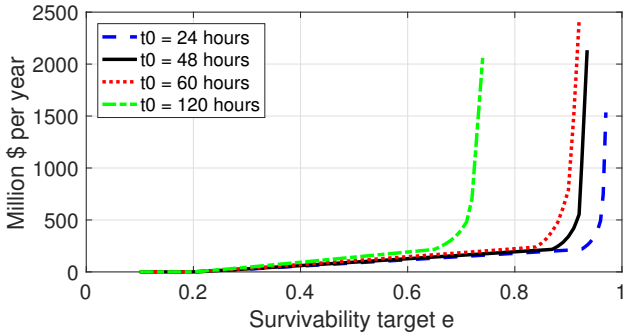

Figure 3-6: Minimum required annual costs for different survivability targets and detection times.

Choosing between different dispensing modes. Figure 3-7 illustrates the cost-survivability trade-offs for different dispensing modes. Using USPS dispensing mode alone is generally more costly as compared with using PODs, due to the upper bound on USPS capacity and reliance on home medical kits. Overall, we find that even though USPS is empirically efficient in high density neighborhoods, PODs should still remain as the primary dispensing channel due to their scalability.

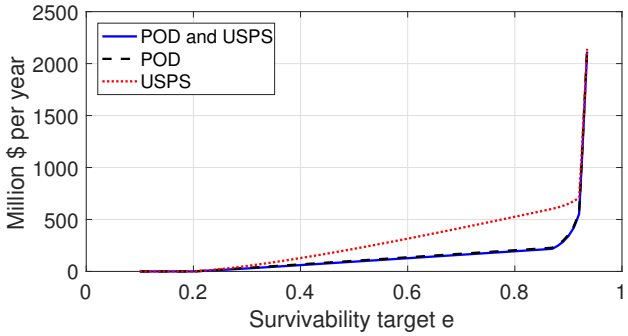

Figure 3-7: Minimum required annual costs for different survivability targets and dispensing modes.

3.7 Extensions and Concluding Remarks

In this paper, we considered a framework to tackle the problem of designing a cost-effective and responsive public health stockpile supply chain for protection against bioterrorism attacks. Our framework captured many of the key drivers facing CDC in maintaining the Strategic National Stockpile, for example, holding costs, pooling, capacity, and responsiveness. Our methodology can be extended to capture additional SNS design considerations, or used in other supply chain applications. For example, an interesting direction for future research could be to allow for multiple types of MCMs, for the same or different biothreats. In that case, our model could be used to guide the design of CDC's overall portfolio and mix of MCMs in view of budgetary constraints.

3.8 Proofs

Proof Proof of Theorem 3.4.3. Under Assumption 3.4.2, our model can be re-formulated to one where, instead of inventory shipments over the edges of (V, E), $\{f_e : e \in E\}$, we consider inventory shipments over the edges of the directed out-tree (V, T) with root node 0 and $T := \{(i, j) : i \in V, j \in \mathcal{C}(i)\}$. This is because, under Assumption 3.4.2, any inventory shipment to node j occurs from some "parent node" i such that $j \in \mathcal{D}(i)$, *i.e.*, $\mathcal{P}^{L-\ell}(j) = i$ for some ℓ, where $\mathcal{P}^n(k) := \mathcal{P}(\mathcal{P}^{n-1}(k))$ for n positive integer, $k \in V$ and $\mathcal{P}^0(k) := k$. Thus, any f_{ij} can be thought of as a flow along the unique path $i \to \mathcal{P}^{L-\ell-1}(j) \to \ldots \mathcal{P}(j) \to j$. It can then be readily seen that our model entails a *path-based* network flow formulation for inventory shipments in the graph (V, T). In the rest of the proof, we consider the associated *edge-based* formulation, and denote the shipments over the edges with $\mathbf{y} = \{y_t : t \in T\}$. Clearly, this is without loss, since for every feasible $(\mathbf{f}, \mathbf{s})$ in the path-based formulation, there exists $\mathbf{y}$ such that $(\mathbf{y}, \mathbf{s})$ is feasible for the edge-based formulation, and vice versa (for more details see Chapter 3.5 in [3]). Furthemore, we assume that demand occurs at every node; specifically, demand at node i is $d_i = \hat{d}_i \xi_i$ for all $i \in V$ and some $\boldsymbol{\xi} \in \Xi$.

Unmet demand at each node in $V \setminus V_D$ is treated in the same way as at nodes in V_D: either penalized at a cost rate b under the (LLC), or subject to a guarantee under the (LLG).

To prove optimality of affine policies, it suffices to show that for any static inventory allocation decision $\mathbf{x} \in X$, there exist policies that are affine in the uncertain demand for the adjustable decisions and achieve the same worst-case cost under fully-adjustable policies. Thus, we henceforth consider the static inventory allocation decision $\mathbf{x}$ as fixed.

We first deal with (LLC). At the end of the proof, we argue how (LLG) can be cast as a special case of (LLC). Using the edge-based formulation, the more general demand model, and a fixed inventory allocation $\mathbf{x} \in X$, it can be readily seen that when $\rho = 1$, the (LLC) problem is equivalent to (in the sense that they have the same optimal set)[4]

$$
\begin{aligned}
\min_{\mathbf{y}(\cdot),\mathbf{s}(\cdot)} \max_{\boldsymbol{\xi}\in\Xi} \quad & \mathbf{1}^T \mathbf{s}(\boldsymbol{\xi}) \\
\text{subject to} \quad & s_i(\boldsymbol{\xi}) + y_{\mathcal{P}(i)i}(\boldsymbol{\xi}) + x_i \geq \sum_{j\in\mathcal{C}(i)} y_{ij}(\boldsymbol{\xi}) + \hat{d}_i \xi_i, \ \forall i \in V, \forall \boldsymbol{\xi} \in \Xi \\
& y_{\mathcal{P}(i)i}(\boldsymbol{\xi}) + x_i \geq \sum_{j\in\mathcal{C}(i)} y_{ij}(\boldsymbol{\xi}), \ \forall i \in V, \forall \boldsymbol{\xi} \in \Xi \\
& \mathbf{y}(\cdot), \mathbf{s}(\cdot) \geq 0,
\end{aligned}
$$

where $\mathbf{1}$ is the vector of all ones. Let z_{F} be the optimal value of this fully-adjustable formulation. Correspondingly, let z_{A} be the optimal value of its affinely adjustable counterpart, *i.e.*, when we restrict $\mathbf{y}(\cdot)$ and $\mathbf{s}(\cdot)$ to be affine in $\boldsymbol{\xi}$. It suffices then to show that $z_{\mathrm{F}} = z_{\mathrm{A}}$.

We now introduce some useful notation:

- For some index set I, let $\boldsymbol{\xi}_I := \{\xi_i : i \in I\}$ and $\mathrm{proj}_I \Xi := \{\boldsymbol{\xi}_I : \boldsymbol{\xi} \in \Xi\}$.
- Let V^l be the set of nodes at the lth level, *i.e.*, $V^l := \{j \in V : \mathcal{P}^l(j) = 0\}$.

[4]Technically, this means that if $(\mathbf{f}, \mathbf{s})$ is optimal for (LLC), then there exists $\mathbf{y}$ such that $(\mathbf{y}, \mathbf{s})$ is optimal for the corresponding edge-based formulation, and vice versa.

- Let A_i be the set of ancestor nodes of i, *i.e.*, $A_i := \{i, \mathcal{P}(i), \mathcal{P}^2(i), \ldots, 0\}$.

- We will frequently look into subgraphs of (V, T), specifically out-trees rooted at some node $i \in V^l$, denoted by (V_i, T_i), where $V_i := \{j \in V : \mathcal{P}^k(j) = i \text{ for some } k = 0, 1, 2, \ldots, L - l\}$, and $T_i := \{(k, l) : k \in V_i, l \in \mathcal{C}(k)\}$.

- Let $\mathcal{Q}_i(\xi_i', x_i, \mathbf{x}_{-i})$ by the set of feasible policies for (V_i, T_i), with inventory at node i being x_i, the inventory on $V_i \setminus \{i\}$ being $\mathbf{x}_{-i}$, and the uncertain parameter being ξ_i' at node i, *i.e.*,

$$\begin{aligned}
\mathcal{Q}_i(\xi_i', x_i, \mathbf{x}_{-i}) :=& \{\{\mathbf{y}(\cdot), \mathbf{s}(\cdot)\} : \\
& x_i \geq \sum_{j \in \mathcal{C}(i)} y_{ij}(\boldsymbol{\xi}),\ \forall \boldsymbol{\xi} \in \Xi_i \\
& s_i(\boldsymbol{\xi}) + x_i \geq \sum_{j \in \mathcal{C}(i)} y_{ij}(\boldsymbol{\xi}) + \hat{d}_i \xi_i',\ \forall \boldsymbol{\xi} \in \Xi_i \\
& y_{\mathcal{P}(k)k}(\boldsymbol{\xi}) + x_k \geq \sum_{j \in \mathcal{C}(k)} y_{kj}(\boldsymbol{\xi}),\ \forall k \in V_i \setminus \{i\}, \forall \boldsymbol{\xi} \in \Xi_i \\
& s_k(\boldsymbol{\xi}) + y_{\mathcal{P}(k)k}(\boldsymbol{\xi}) + x_k \geq \sum_{j \in \mathcal{C}(k)} y_{kj}(\boldsymbol{\xi}) + \hat{d}_k \xi_k,\ \forall k \in V_i \setminus \{i\}, \forall \boldsymbol{\xi} \in \Xi_i \\
& \mathbf{y}(\boldsymbol{\xi}), \mathbf{s}(\boldsymbol{\xi}) \geq 0,\ \forall \boldsymbol{\xi} \in \Xi_i\}.
\end{aligned}$$

  where $\Xi_i(\xi_i') := \{\boldsymbol{\xi} \mid \boldsymbol{\xi} \in \Xi, \xi_i = \xi_i'\}$. We use shorthand notation Ξ_i when there is no ambiguity in ξ_i'.

- Let $\Theta_i(\xi_i', \mathbf{s})$ be the worst-case demand loss in (V_i, T_i) under $\xi_i = \xi_i'$ and (feasible) resource allocation policy $\{\mathbf{y}(\cdot), \mathbf{s}(\cdot)\}$, $\Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i})$ be the worst-case demand loss in V_i under $\xi_i = \xi_i'$ with an optimal fully-adjustable policy, and $\mathcal{Q}_i^\star(\xi_i', x_i, \mathbf{x}_{-i})$ be the set of all optimal policies:

$$\begin{aligned}
\Theta_i(\xi_i', \mathbf{s}) &:= \max_{\boldsymbol{\xi} \in \Xi_i(\xi_i')} \sum_{k \in V_i} s_k(\boldsymbol{\xi}), \\
\Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i}) &:= \min_{\mathbf{y}(\cdot), \mathbf{s}(\cdot)} \Theta_i(\xi_i', \mathbf{s}) \text{ subject to } \{\mathbf{y}(\cdot), \mathbf{s}(\cdot)\} \in \mathcal{Q}_i(\xi_i', x_i, \mathbf{x}_{-i}), \\
\mathcal{Q}_i^\star(\xi_i', x_i, \mathbf{x}_{-i}) &:= \operatorname{argmin}_{\mathbf{y}(\cdot), \mathbf{s}(\cdot)} \Theta_i(\xi_i', \mathbf{s}) \text{ subject to } \{\mathbf{y}(\cdot), \mathbf{s}(\cdot)\} \in \mathcal{Q}_i(\xi_i', x_i, \mathbf{x}_{-i}).
\end{aligned}$$

Similarly, we define the affinely-adjustable counterpart $\Omega_i^A(\xi_i', x_i, \mathbf{x}_{-i})$ as the worst-case demand loss in V_i, restricted to $\{\mathbf{y}(\cdot), \mathbf{s}(\cdot)\} \in \mathcal{Q}_i(\xi_i', x_i, \mathbf{x}_{-i})$ and affine.

Using our notation, $z_{\mathrm{F}} = \Omega_0^F(1, x_0, \mathbf{x}_{-0})$ and $z_{\mathrm{A}} = \Omega_0^A(1, x_0, \mathbf{x}_{-0})$. It thus suffices to show that

$$\Omega_0^F(1, x_0, \mathbf{x}_{-0}) = \Omega_0^A(1, x_0, \mathbf{x}_{-0}). \tag{3.10}$$

We now present two useful intermediate results. The first provides a recursive expression for Ω_i^F, and the second shows that Ξ has binary vertices.

Proposition 3.8.1 *For all $i \in V$, $\xi_i' \in [0,1]$, $\mathbf{x} \geq 0$,*

$$\Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i}) = \left(\hat{d}_i \xi_i' - x_i + \max_{\boldsymbol{\xi}_{\mathcal{C}(i)} \in \mathrm{proj}_{\mathcal{C}(i)} \Xi_i(\xi_i')} \sum_{j \in \mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}) \right)^+.$$

Moreover, $\Omega_i^F(\xi_i, x_i, \mathbf{x}_{-i})$ is convex in ξ_i over $\xi_i \in [0,1]$.

Proposition 3.8.2 *Ξ is a polytope with binary vertices,* i.e., *$\boldsymbol{\xi} \in ext(\Xi) \implies \boldsymbol{\xi} \in \{0,1\}^n$.*

Proposition 3.8.1 shows the intuitive fact that given a fixed inventory decision and under the optimal policy, the demand shortage in V_i should be the sum of shortages in $\{V_j\}_{j \in \mathcal{C}(i)}$ plus the demand on node i, subtracting away inventory x_i. Based on such recursive definition, we can view the tree as a collection of hub-and-spoke clusters, each consisting of one node and several edges pointing away from this node (Figure 3-8).

Now we prove (3.10) by induction. For a given tree (V_i, T_i) and uncertainty set Ξ_i, we limit our attention to $\{\mathbf{y}(\cdot), \mathbf{s}(\cdot)\}$ that satisfies the following properties:

(I) $\forall m \in V_i,\ j \in \mathcal{C}(m),\ \boldsymbol{\xi} \in \Xi_i : s_m(\boldsymbol{\xi}) = s_m(\xi_m),\ y_{mj}(\boldsymbol{\xi}) = y_{mj}(\xi_j)$, *i.e.*, the policies are univariate in ξ_m, ξ_j respectively.

(II) $\forall m \in V_i$: (a) $y_{\mathcal{P}(m)m}(0) = 0$, (b) $y_{jk}(0) = \Omega_k^A(0, x_k, \mathbf{x}_{-k}) = 0\ \forall (j,k) \in T_m$, and (c) $s_j(0) = 0\ \forall j \in V_m$, *i.e.*, these policies are linear.

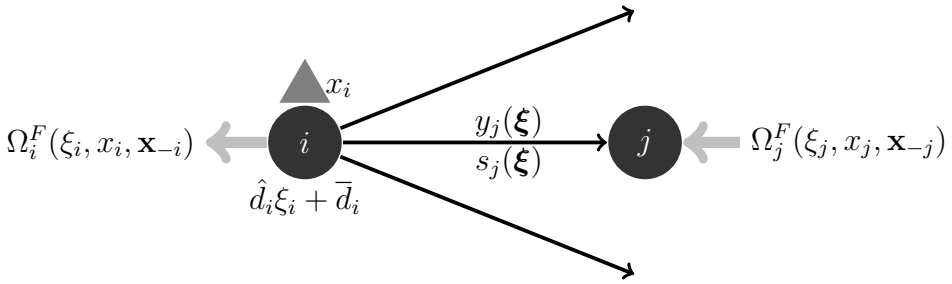

Figure 3-8: Recursive representation of Ω_i^F.

Let

$$\overline{\mathcal{Q}}_i(\xi_i, x_i, \mathbf{x}_{-i}) := \{\{\mathbf{y}, \mathbf{s}\} \in \mathcal{Q}_i(\xi_i, x_i, \mathbf{x}_{-i}) \ : \ \{\mathbf{y}, \mathbf{s}\} \text{ satisfies (I)-(II)}\}$$
$$\overline{\mathcal{Q}}_i^\star(\xi_i, x_i, \mathbf{x}_{-i}) := \overline{\mathcal{Q}}_i(\xi_i, x_i, \mathbf{x}_{-i}) \cap \mathcal{Q}_i^\star(\xi_i, x_i, \mathbf{x}_{-i}).$$

We are now ready to formalize the induction process.

Induction Hypothesis. If for some $l \in \{0, 1, \ldots, L-1\}$, $\forall j \in V^{l+1}, \mathbf{x} \geq 0$, $\exists\{\bar{\mathbf{y}}^j(\cdot), \bar{\mathbf{s}}^j(\cdot)\}$ such that

$$\text{(Feasibility) } \forall \xi_j \in [0,1], \{\bar{\mathbf{y}}^j(\cdot), \bar{\mathbf{s}}^j(\cdot)\} \in \overline{\mathcal{Q}}_j(\xi_j, x_j, \mathbf{x}_{-j}), \tag{3.11}$$
$$\text{(Optimality) } \forall \xi_j \in \{0,1\}, \Theta_j(\xi_j, \bar{\mathbf{s}}^j) = \Omega_j^A(\xi_j, x_j, \mathbf{x}_{-j}) = \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}), \tag{3.12}$$

then, $\forall i \in V^l, \mathbf{x} \geq 0$, $\exists\{\bar{\mathbf{y}}^i(\cdot), \bar{\mathbf{s}}^i(\cdot)\}$ such that

$$\text{(Feasibility) } \forall \xi_i \in [0,1], \{\bar{\mathbf{y}}^i(\cdot), \bar{\mathbf{s}}^i(\cdot)\} \in \overline{\mathcal{Q}}_i(\xi_i, x_i, \mathbf{x}_{-i}), \tag{3.13}$$
$$\text{(Optimality) } \forall \xi_i \in \{0,1\}, \Theta_i(\xi_i, \bar{\mathbf{s}}^i) = \Omega_i^A(\xi_i, x_i, \mathbf{x}_{-i}) = \Omega_i^F(\xi_i, x_i, \mathbf{x}_{-i}). \tag{3.14}$$

Base Case ($l = L-1$). Given some $\mathbf{x} \geq 0$, for each $j \in V^L$ we define $\bar{s}_j^j(\boldsymbol{\xi}) = \bar{s}_j^j(\xi_j) = \xi_j(\hat{d}_j - x_j)^+$. This policy satisfies (I) and (II) by construction, and it is straightforward to check feasibility and optimality.

General Step. Now we prove the general induction step: suppose (3.11)-(3.12) hold for some $l \in \{0, 1, \ldots, L-1\}$ and $\forall j \in V^{l+1}, \mathbf{x} \geq 0$. To construct $\{\bar{\mathbf{y}}^i(\cdot), \bar{\mathbf{s}}^i(\cdot)\}$, we define the following. Since $\Omega_j^A(\xi_j, x_j, \mathbf{x}_{-j})$ is nondecreasing in ξ_j and nonnegative, there necessarily exists $g_j \geq 0$ such that $\Omega_j^A(\xi_j, x_j, \mathbf{x}_{-j}) = g_j\xi_j$ for $\xi_j \in \{0,1\}$. Let

$\beta_i := \min\left(1, \frac{x_i}{\sum_{j=1}^{\Gamma_i} g_{(j)} + \hat{d}_i}\right)$ if $\sum_{j=1}^{\Gamma_i} g_{(j)} + \hat{d}_i > 0$, and $\beta_i := 1$ otherwise, where $g_{(j)}$ is the jth largest element of $\{g_j\}_{j \in \mathcal{C}(i)}$. ($\beta_i$ is the surge demand coverage ratio)

We are now ready to construct affine $\{\bar{\mathbf{y}}^i, \bar{\mathbf{s}}^i\}$:

$$\{\bar{\mathbf{y}}^i, \bar{\mathbf{s}}^i\} := \begin{cases} \bar{y}^i_{ij}(\xi_j) = \beta_i g_j \xi_j & \forall j \in \mathcal{C}(i) \\ \bar{s}^i_i(\xi_i) = (1-\beta_i)\hat{d}_i \xi_i & \\ \bar{y}^i_{km}(\xi_m) = \bar{y}^{j1}_{km}(\xi_m) & \forall (k,m) \in T_j, \forall j \in \mathcal{C}(i) \\ \bar{s}^i_k(\xi_k) = \bar{s}^{j1}_k(1)\xi_k & \forall k \in V_i \setminus \{i\}. \end{cases} \tag{3.15}$$

Let $\overline{\mathcal{Q}}^\star_{j1} := \overline{\mathcal{Q}}^\star_j(1, x_j + \bar{y}^i_{ij}(1), \mathbf{x}_{-j})$. By the induction hypothesis, $\overline{\mathcal{Q}}^\star_{j1} \neq \emptyset$, since $x_j + \bar{y}^i_{ij}(\xi_j) \geq 0$ for $\xi_j \in [0,1]$. We now check the feasibility and optimality of $\{\bar{\mathbf{y}}^i, \bar{\mathbf{s}}^i\}$.

For (Feasibility), note that Property (I-II) hold by construction. To show $\{\bar{\mathbf{y}}^i, \bar{\mathbf{s}}^i\}$ satisfies the constraints in $\mathcal{Q}_i(\xi_i', x_i, \mathbf{x}_{-i})$, note that the constraints involving node i are satisfied by construction of $\{\bar{\mathbf{y}}^i, \bar{\mathbf{s}}^i\}$. For nodes downstream from i, note that $\bar{y}^i_{km}(\xi_m) = \bar{y}^{j1}_{km}(1)\xi_m$ and $\bar{s}^i_k(\xi_k) = \bar{s}^{j1}_k(1)\xi_k$, for all $(k,m) \in T_j, j \in \mathcal{C}(i)$ and $k \in V_i \setminus \{i\}$. Since $\{\bar{\mathbf{y}}^{j0}(\cdot), \bar{\mathbf{s}}^{j0}(\cdot)\}$ and $\{\bar{\mathbf{y}}^{j1}(\cdot), \bar{\mathbf{s}}^{j1}(\cdot)\}$ satisfy the constraints of $\mathcal{Q}_j(0, x_j + \bar{y}^i_{ij}(0), \mathbf{x}_{-j})$ and $\mathcal{Q}_j(1, x_j + \bar{y}^i_{ij}(1), \mathbf{x}_{-j})$ respectively, it is straightforward to check the constraint satisfaction of $\{\mathbf{y}^i, \mathbf{s}^i\}$ for $\mathcal{Q}_i(\xi_i', x_i, \mathbf{x}_{-i})$.

For (Optimality), note that for $\xi_i' \in \{0,1\}$ we have

$$\Theta_i(\xi_i', \bar{\mathbf{s}}^i) = \bar{s}_i^i(\xi_i') + \max_{\boldsymbol{\xi} \in \Xi_i} \sum_{j \in \mathcal{C}(i)} \sum_{k \in V_j} \bar{s}_k^i(\xi_k) \tag{3.16a}$$

$$= \bar{s}_i^i(\xi_i') + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in [0, \xi_i'], j \in \mathcal{C}(i)}} \sum_{j \in \mathcal{C}(i)} \max_{\boldsymbol{\xi} \in \Xi_j(\xi_j)} \sum_{k \in V_j} \bar{s}_k^{j1}(1) \xi_k \tag{3.16b}$$

$$= \bar{s}_i^i(\xi_i') + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in \{0, \xi_i'\}, j \in \mathcal{C}(i)}} \sum_{j \in \mathcal{C}(i)} \max_{\boldsymbol{\xi} \in \Xi_j(\xi_j)} \sum_{k \in V_j} \bar{s}_k^{j1}(1) \xi_k \tag{3.16c}$$

$$= \bar{s}_i^i(\xi_i') + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in \{0, \xi_i'\}, j \in \mathcal{C}(i)}} \sum_{j \in \mathcal{C}(i)} \Omega_j^A(\xi_j, x_j + \bar{y}_{ij}^i(\xi_j), \mathbf{x}_{-j}) \tag{3.16d}$$

$$= \bar{s}_i^i(\xi_i') + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in \{0, \xi_i'\}, j \in \mathcal{C}(i)}} \sum_{j \in \mathcal{C}(i)} \Omega_j^F(\xi_j, x_j + \bar{y}_{ij}^i(\xi_j), \mathbf{x}_{-j}) \tag{3.16e}$$

$$= \bar{s}_i^i(\xi_i') + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in \{0, \xi_i'\}, j \in \mathcal{C}(i)}} \sum_{j \in \mathcal{C}(i)} \left(\Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}) - \bar{y}_{ij}^i(\xi_j)\right)^+ \tag{3.16f}$$

$$= \bar{s}_i^i(\xi_i') + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in \{0, \xi_i'\}, j \in \mathcal{C}(i)}} \left(\sum_{j \in \mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}) - \sum_{j \in \mathcal{C}(i)} \bar{y}_{ij}^i(\xi_j) \right) \tag{3.16g}$$

$$= \left(\hat{d}_i \xi_i' - x_i + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in \{0, \xi_i'\}, j \in \mathcal{C}(i)}} \sum_{j \in \mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}) \right)^+ \tag{3.16h}$$

$$= \left(\hat{d}_i \xi_i' - x_i + \max_{\substack{\boldsymbol{\xi}_{\mathcal{C}(i)}: \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \le \Gamma_i \xi_i' \\ \xi_j \in [0, \xi_i'], j \in \mathcal{C}(i)}} \sum_{j \in \mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}) \right)^+ \tag{3.16i}$$

$$= \Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i}) = \Omega_i^A(\xi_i', x_i, \mathbf{x}_{-i}). \tag{3.16j}$$

Equalities (3.16a-3.16b) are straightforward applications of definitions of Θ_i and $\bar{\mathbf{s}}^i$.

To show equality (3.16c), we argue that is it sufficient to check $\xi_j = 1$ and $\xi_j = 0$ for $j \in \mathcal{C}(i)$. Suppose that $\Xi_j(\xi_j)$ has n extreme points $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^n$, and $W_m = \{k \in V_j \ : \ \xi_k^m = \xi_j\}$, $\overline{W}_m = V_j \setminus W_m$. Since for each extreme point, ξ_k takes value of either

ξ_j or 0 (*cf.* Proposition 3.8.2), we can write

$$\max_{\boldsymbol{\xi}\in\Xi_j(\xi_j)} \sum_{k\in V_j} \bar{s}_k^{j1}(1)\xi_k = \max_{m\in\{1,\dots,n\}} \sum_{k\in V_j} \left\{ \bar{s}_k^{j1}(1)\xi_k \mathbb{1}_{k\in W_m} \right\},$$

which is a convex function of ξ_j ($\mathbb{1}_{i\in I} = 1$ if $i \in I$ and 0 otherwise). By this token, we can replace $\boldsymbol{\xi}_{\mathcal{C}(i)} \in \text{proj}_{\mathcal{C}(i)}\Xi_i$ with $\boldsymbol{\xi}_{\mathcal{C}(i)} \in \text{proj}_{\mathcal{C}(i)}\Xi_i \cap \{0,1\}^{|\mathcal{C}(i)|}$ for the argument of the outer maximization, since $\Xi_i(0)$ and $\Xi_i(1)$ have binary vertices.

To show equality (3.16d), note that for $\xi_j = 1$, $\Omega_j^A(1, x_j + \bar{y}_{ij}^i(1), \mathbf{x}_{-j}) = \Theta_j(1, \bar{\mathbf{s}}^{j1}) = \bar{s}_j^{j1}(1) + \max_{\boldsymbol{\xi}\in\Xi_j(1)} \sum_{k\in V_j\setminus\{j\}} \bar{s}_k^{j1}(\xi_k)$, which is equal to $\max_{\boldsymbol{\xi}\in\Xi_j(1)} \sum_{k\in V_j} \bar{s}_k^{j1}(1)\xi_k$. For $\xi_j = 0$, $\Omega_j^A(0, x_j + \bar{y}_{ij}^i(0), \mathbf{x}_{-j}) = 0$. Equality (3.16e) holds by the induction hypothesis. Equality (3.16f) follows from the definition of Ω_j^F and (3.20). For equality (3.16g), by the definitions of β_i, we have $(\Omega_j^A(\xi_j, x_j, \mathbf{x}_{-j}) - \bar{y}_{ij}^i(\xi_j)) \geq 0$ for $j \in \mathcal{C}(i)$, therefore we can rewrite (3.16f) into (3.16g). For the transition into (3.16h), it is straightforward to consider $\xi_i' = 0$ and $\xi_i' = 1$ separately. Lastly, (3.16i) holds by the convexity property of $\sum_{j\in\mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j})$ and (3.16j) holds by the recursive property, both shown in Proposition 3.8.1. We have now completed the proof for the inductive step, thus showing $z_{\text{F}} = z_{\text{A}}$.

To complete the proof, we now turn our attention to the (LLG) formulation. By solving for $s_i(\mathbf{d})$ from (3.1c) and substituting in (3.2b), we can eliminate (3.1c) and re-write (3.2b) as

$$\sum_{j:(j,i)\in E} f_{ji}(\mathbf{d}) \geq \frac{1 - \epsilon_i - \overline{\rho}}{\rho - \overline{\rho}} d_i, \quad \forall i \in V, \forall \mathbf{d} \in U.$$

At optimality, the above constraint can be taken to be active without loss (if it is not, we can scale down the associated flows into i so that it becomes active). Thus,

the (LLG) is equivalent to

$$
\begin{aligned}
\min_{\mathbf{x},\,\mathbf{f}(\cdot),\,\mathbf{s}(\cdot)} \quad & \mathbf{h}^T\mathbf{x} \\
\text{subject to} \quad & x_i \geq \sum_{j:(i,j)\in E} f_{ij}(\mathbf{d}), \quad \forall i \in V,\, \forall \mathbf{d} \in U' \\
& s_i(\mathbf{d}) + \sum_{j:(j,i)\in E} f_{ji}(\mathbf{d}) = d_i, \quad \forall i \in V,\, \forall \mathbf{d} \in U' \\
& s_i(\mathbf{d}) = 0, \quad \forall i \in V,\, \forall \mathbf{d} \in U' \\
& f_{ij}(\mathbf{d}) \geq 0, \quad \forall (i,j) \in E,\, \forall \mathbf{d} \in U' \\
& \mathbf{x} \in X,
\end{aligned}
$$

where $U' := \left\{\mathbf{d}' : d'_i = \frac{1-\epsilon_i-\overline{\rho}}{\rho-\overline{\rho}} d_i \,\forall i \in V,\, \mathbf{d} \in U\right\}$. Then, this is in turn equivalent to

$$
\begin{aligned}
\min_{\mathbf{x},\,\mathbf{f}(\cdot),\,\mathbf{s}(\cdot)} \quad \mathbf{h}^T\mathbf{x} \quad + \quad \max_{\mathbf{d}\in U'} \quad & b' \sum_{i\in V} (1-\overline{\rho}) s_i(\mathbf{d}) \\
\text{subject to} \quad & x_i \geq \sum_{j:(i,j)\in E} f_{ij}(\mathbf{d}), \quad \forall i \in V,\, \forall \mathbf{d} \in U' \\
& s_i(\mathbf{d}) + \sum_{j:(j,i)\in E} f_{ji}(\mathbf{d}) = d_i, \quad \forall i \in V,\, \forall \mathbf{d} \in U' \\
& \sum_{j:(j,i)\in E} f_{ji}(\mathbf{d}) \leq d_i, \quad \forall i \in V,\, \forall \mathbf{d} \in U' \\
& f_{ij}(\mathbf{d}) \geq 0, \quad \forall (i,j) \in E,\, \forall \mathbf{d} \in U' \\
& \mathbf{x} \in X
\end{aligned}
$$

for b' large enough. However, this corresponds to an (LLC) instance that satisfies our assumptions and therefore admits an optimal affine adjustable policy. ■

Proof Proof of Proposition 3.8.1.

Note that we can express $\Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i})$ as $\max_{\boldsymbol{\xi} \in \Xi_i(\xi_i')} r_i(\boldsymbol{\xi}, x_i, \mathbf{x}_{-i})$, where

$$
\begin{aligned}
r_i(\boldsymbol{\xi}, x_i, \mathbf{x}_{-i}) := \min_{\mathbf{y},\mathbf{s}} \quad & \sum_{k \in V_i} s_k \\
\text{subject to} \quad & x_i \geq \sum_{j \in \mathcal{C}(i)} y_{ij} \\
& s_i + x_i \geq \sum_{j \in \mathcal{C}(i)} y_{ij} + \hat{d}_i \xi_i \\
& y_{\mathcal{P}(k)k} + x_k \geq \sum_{j \in \mathcal{C}(k)} y_{kj}, \ \forall k \in V_i \setminus \{i\} \\
& y_{\mathcal{P}(k)k} + s_k + x_k \geq \sum_{j \in \mathcal{C}(k)} y_{kj} + \hat{d}_k \xi_k, \ \forall k \in V_i \setminus \{i\} \\
& \mathbf{y}, \mathbf{s} \geq 0.
\end{aligned}
$$

It is straightforward to check that the above optimization problem admits an optimal solution such that $s_i = (\hat{d}_i \xi_i' - x_i)^+$ and $\sum_{j \in \mathcal{C}(i)} y_{ij} = (x_i - \hat{d}_i \xi_i')^+$. Also,

$$
r_i(\boldsymbol{\xi}, x_i + \Delta, \mathbf{x}_{-i}) = (r_i(\boldsymbol{\xi}, x_i, \mathbf{x}_{-i}) - \Delta)^+, \quad \forall \Delta \geq 0. \tag{3.20}
$$

Another fact we will use is that

$$
\min_{\substack{\mathbf{x} \geq 0 \\ \mathbf{1}^T \mathbf{x} = K}} \sum_{i=1}^{n} (q_i - x_i)^+ = (\mathbf{1}^T \mathbf{q} - K)^+, \quad \forall \mathbf{q} \in \mathbb{R}^n, K \geq 0. \tag{3.21}
$$

Using all these properties we have

$$\Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i}) = (\hat{d}_i\xi_i' - x_i)^+ + \max_{\boldsymbol{\xi}\in\Xi_i} \min_{\substack{\{y_{ij}\}_{j\in\mathcal{C}(i)} : y_{ij}\geq 0, \\ \sum_{j\in\mathcal{C}(i)} y_{ij} = (x_i - \hat{d}_i\xi_i')^+}} \sum_{j\in\mathcal{C}(i)} r_j(\boldsymbol{\xi}, x_j + y_{ij}, \mathbf{x}_{-j}) \tag{3.22a}$$

$$= (\hat{d}_i\xi_i' - x_i)^+ + \max_{\boldsymbol{\xi}\in\Xi_i} \min_{\substack{\{y_{ij}\}_{j\in\mathcal{C}(i)} : y_{ij}\geq 0, \\ \sum_{j\in\mathcal{C}(i)} y_{ij} = (x_i - \hat{d}_i\xi_i')^+}} \sum_{j\in\mathcal{C}(i)} (r_j(\boldsymbol{\xi}, x_j, \mathbf{x}_{-j}) - y_{ij})^+ \tag{3.22b}$$

$$= (\hat{d}_i\xi_i' - x_i)^+ + \left(\max_{\boldsymbol{\xi}\in\Xi_i} \sum_{j\in\mathcal{C}(i)} r_j(\boldsymbol{\xi}, x_j, \mathbf{x}_{-j}) - (x_i - \hat{d}_i\xi_i')^+ \right)^+ \tag{3.22c}$$

$$= (\hat{d}_i\xi_i' - x_i)^+ + \left(\max_{\boldsymbol{\xi}_{\mathcal{C}(i)}\in \mathrm{proj}_{\mathcal{C}(i)}\Xi_i} \max_{\substack{\boldsymbol{\xi}\in\Xi_j(\xi_j), \\ j\in\mathcal{C}(i)}} \sum_{j\in\mathcal{C}(i)} r_j(\boldsymbol{\xi}, x_j, \mathbf{x}_{-j}) - (x_i - \hat{d}_i\xi_i')^+ \right)^+ \tag{3.22d}$$

$$= (\hat{d}_i\xi_i' - x_i)^+ + \left(\max_{\boldsymbol{\xi}_{\mathcal{C}(i)}\in \mathrm{proj}_{\mathcal{C}(i)}\Xi_i} \sum_{j\in\mathcal{C}(i)} \max_{\boldsymbol{\xi}\in\Xi_j(\xi_j)} r_j(\boldsymbol{\xi}, x_j, \mathbf{x}_{-j}) - (x_i - \hat{d}_i\xi_i')^+ \right)^+ \tag{3.22e}$$

$$= (\hat{d}_i\xi_i' - x_i)^+ + \left(\max_{\boldsymbol{\xi}_{\mathcal{C}(i)}\in \mathrm{proj}_{\mathcal{C}(i)}\Xi_i} \sum_{j\in\mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}) - (x_i - \hat{d}_i\xi_i')^+ \right)^+ \tag{3.22f}$$

$$= \left(\hat{d}_i\xi_i' - x_i + \max_{\boldsymbol{\xi}_{\mathcal{C}(i)}\in \mathrm{proj}_{\mathcal{C}(i)}\Xi_i} \sum_{j\in\mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j}) \right)^+ . \tag{3.22g}$$

Equality (3.22a) follows from the definition of r_i; (3.22b) follows from (3.20); (3.22c) follows from (3.21); (3.22d) is an equivalent way of writing the maximization operator; exchanging the inner maximization and summation operators is allowed since r_j only depends on $\{\xi_k, k \in V_j\}$, leading to (3.22e); (3.22f) follows from the definition of r_j. Equality (3.22g) holds by considering the fact that $\Omega_j^F \geq 0$ for any $j \in \mathcal{C}(i)$, and checking two cases: one with $\hat{d}_i\xi_i' - x_i < 0$, and the other with $\hat{d}_i\xi_i' - x_i \geq 0$.

To show convexity of $\Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i})$, we use an induction argument on the level of

node i. First note that for $i \in V^L$, we have $\Omega_i^F(\xi_i', x_i, \mathbf{x}_{-i}) = (\hat{d}_i \xi_i' - x_i)^+$, which is convex and non-decreasing in ξ_i'. For i in any other level of the network, given the result we just proved (3.22g), it suffices to show $\max_{\boldsymbol{\xi}_{\mathcal{C}(i)} \in \text{proj}_{\mathcal{C}(i)} \Xi_i(\xi_i')} \sum_{j \in \mathcal{C}(i)} \Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j})$ is convex in ξ_i'. Since $\Omega_j^F(\xi_j, x_j, \mathbf{x}_{-j})$ is convex in ξ_j (by the induction hypothesis), maximum is obtained at the extreme points. Given that $\text{proj}_{\mathcal{C}(i)} \Xi_i(\xi_i') = \{\boldsymbol{\xi}_{\mathcal{C}(i)} : \xi_j \in [0, \xi_i'], \mathbf{1}^T \boldsymbol{\xi}_{\mathcal{C}(i)} \leq \Gamma_i \xi_i'\}$ and Ω_j^F is non-decreasing in ξ_j, we can equivalently write: $\max_{|S|=\Gamma_i, S \subseteq \mathcal{C}(i)} \left\{ \sum_{j \in S} \Omega_j^F(\xi_i', x_j, \mathbf{x}_{-j}) + \sum_{j \in \mathcal{C}(i) \setminus S} \Omega_j^F(0, x_j, \mathbf{x}_{-j}) \right\}$, which is convex, non-decreasing in ξ_i'.■

Proof Proof of Proposition 3.8.2.

For the sake of reaching a contradiction, suppose there exists some extreme point $\boldsymbol{\xi} \in \text{ext}(\Xi)$, such that its kth entry is fractional, $\xi_k \in (0, 1)$. Note that by definition of Ξ, $\xi_0 = 1$, and thus $k \neq 0$. Let $p = \mathcal{P}(k)$ be its parent, *i.e.*, $k \in \mathcal{C}(p)$. Then it must hold that $0 < \sum_{i \in \mathcal{C}(p)} \xi_i \leq \Gamma_p \xi_p$, and thus $\Gamma_p > 0$ and $\xi_p > 0$.

Suppose $\xi_p = 1$. Then the above condition becomes $0 < \sum_{i \in \mathcal{C}(p)} \xi_i \leq \Gamma_p$. Furthermore, $\xi_i \in [0, 1]$ $\forall i \in \mathcal{C}(p)$. Now consider the $|\mathcal{C}(p)|$-dimensional polytope $\{\boldsymbol{\gamma} : \sum_{i \in \mathcal{C}(p)} \gamma_i \leq \Gamma_p, \gamma_i \in [0, 1] \, \forall i \in \mathcal{C}(p)\}$, which can be readily seen to have binary extreme points. Let these extreme points be $\{\boldsymbol{\gamma}^j, j = 1, \ldots, N\}$. Since $\boldsymbol{\xi}_{\mathcal{C}(p)}$ belongs to this polytope, we can express it as

$$\boldsymbol{\xi}_{\mathcal{C}(p)} = \sum_{j=1}^{N} \alpha_j \boldsymbol{\gamma}^j, \text{ where } \sum_{j=1}^{N} \alpha_j = 1, \alpha_j \geq 0, \forall j \in \{1, \ldots, N\}.$$

In particular, for those $i \in \mathcal{C}(p)$ that $\xi_i > 0$, we have:

$$\sum_{j=1}^{N} \alpha_j \gamma_i^j = \xi_i \iff \sum_{j=1}^{N} \alpha_j \frac{1}{\xi_i} \gamma_i^j = 1. \tag{3.23}$$

Now we use each $\boldsymbol{\gamma}^j$ to construct a vector $\boldsymbol{\xi}^j \in \Xi$ and show that we can write $\boldsymbol{\xi}$ as a convex combination of $\{\boldsymbol{\xi}^j, j = 1, \ldots, N\}$, and $\boldsymbol{\xi} \neq \boldsymbol{\xi}^j, \forall j \in \{1, \ldots, N\}$—which will contradict our assumptions.

We construct $\boldsymbol{\xi}^j$, $j = 1, \ldots, N$, as follows:

$$\begin{aligned}
\boldsymbol{\xi}^j_{V \setminus V_p} &= \boldsymbol{\xi}_{V \setminus V_p}, \\
\xi^j_p &= \xi_p, \\
\boldsymbol{\xi}^j_{\mathcal{C}(p)} &= \boldsymbol{\gamma}^j, \\
\boldsymbol{\xi}^j_{V_i \setminus \{i\}} &= \begin{cases} \frac{1}{\xi_i} \boldsymbol{\xi}_{V_i \setminus \{i\}} & \text{if } \xi_i > 0 \text{ and } \xi^j_i = 1 \\ 0 & \text{otherwise} \end{cases}, \quad \forall i \in \mathcal{C}(p).
\end{aligned}$$

It is straightforward to verify that for all $j = 1, \ldots, N$, $\boldsymbol{\xi}^j \in \Xi$ and $\boldsymbol{\xi}^j \neq \boldsymbol{\xi}$. We next argue that $\sum_{j=1,\ldots,N} \alpha_j \boldsymbol{\xi}^j = \boldsymbol{\xi}$ by checking component-wise. This is straightforward for components $V \setminus V_p$, p, and $\mathcal{C}(p)$. For any $l \in V_i \setminus \{i\}$ for $i \in \mathcal{C}(p)$ such that $\xi_i = 0$, we have that $\xi^j_l = \xi_l = 0$ for all $j = 1, \ldots, N$. Finally, for any $l \in V_i \setminus \{i\}$ for $i \in \mathcal{C}(p)$ such that $\xi_i > 0$, we have that

$$\sum_{j=1}^{N} \alpha_j \xi^j_l = \sum_{j:\xi^j_i = 1} \alpha_j \xi_l / \xi_i + \sum_{j:\xi^j_i = 0} 0 = \sum_{j=1}^{N} \alpha_j \xi_l / \xi_i \xi^j_i = \sum_{j=1}^{N} \left(\alpha_j \xi^j_i / \xi_i\right) \xi_l = \sum_{j=1}^{N} \left(\alpha_j \gamma^j_i / \xi_i\right) \xi_l \overset{(3.23)}{=} \xi_l.$$

Consequently, $\boldsymbol{\xi}$ cannot be an extreme point if $\xi_p = 1$. Since $\xi_p > 0$, we must have $\xi_p \in (0, 1)$. We can propagate this argument upstream to eventually show that $\xi_0 \in (0, 1)$, which contradicts that $\xi_0 = 1$ for all $\boldsymbol{\xi} \in \Xi$, and the proof is complete.■

3.9 Numerical Studies on AP Heuristic

3.9.1 Relaxing Assumptions 1 and 2

We provide details on how problem instances are generated for Studies 1 and 2 in §3.4.2, alongside a quantitative analysis of our results. For each problem instance, we generate a tree-style graph as follows:

1. Uniformly sample the target total number of nodes, $n \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$.
2. Uniformly sample the number of levels, $l + 1 \in \{2, 3, 4, 5\}$.

3. For each non-leaf level $k = 0, 1, \ldots, l-1$, sample a random integer f_k between 5 and 15.

4. Find F (*e.g.*, by line search) such that $\texttt{round}\left(1 + f_0/F + f_0 f_1/F^2 + \cdots + f_0 \cdots f_{l-1}/F^{l-1}\right) = n$, where $\texttt{round}(\cdot)$ is the usual rounding operation.

5. Grow the tree from root in a breadth-first way: for each node in level $k = 0, 1, \ldots, l-1$, assign to it c_i children, where c_i is randomly picked from $\{(\prod_{j=0}^{k} f_j/F^k - 1)^+, \prod_{j=0}^{k} f_j/F^k, \prod_{j=0}^{k} f_j/F^k + 1\}$.

With this procedure, we generate 5000 trees. For Study 2, for each of the 5000 trees, we create a non-tree counterpart by introducing additional edges into the graph. Specifically, for each pair of nodes that are in adjacent levels in the previously constructed tree, we assign them an edge with probability p_{arc}, which is randomly sampled from $[0, 0.01]$ for each instance. The total number of edges in the non-tree counterpart comes out to have 0% to 200% more edges than its base tree.

To generate cost, efficacy, demand, and attack severity parameter values, we first normalize the per unit demand loss cost, $b = 100$. Then for each instance, we do the following:

1. Generate (float-valued) inventory cost upper and lower bounds uniformly: $\underline{h} \in [0, 10]$, and $\bar{h} \in [10, 20]$, and $\delta h = \bar{h} - \underline{h}$.

2. For each node i, assign an inventory cost $h_i \in [\underline{h}, \bar{h}]$ to it. More specifically, h_i is uniformly sampled from $[\underline{h} + \frac{l(i)+0.5}{l+1} \times \delta h,\ \underline{h} + \frac{l(i)+1}{l+1} \times \delta h]$, where $l(i) \in \{0, 1, \ldots, l\}$ is the level in which node i resides.

3. Uniformly sample the average efficacy $\bar{\rho} \in [0, 1]$. For each path p, generate $\rho_p \in [0.5, 1] \times \bar{\rho}$. Assign a flow cost efficacy $100\rho_p$ to the path.

4. For each demand node i, the maximum demand is $\hat{d}_i$, which is uniformly sampled from $[0, 1000]$. (This value is then fed into the optmization formulation $d_i(\xi_i) = \hat{d}_i \xi_i$.)

5. To determine the attack scale for each node, Γ_i, first generate $\bar{\Gamma} \in [0, 0.2]$. Then for each node i in this network instance, uniformly sample an integer Γ_i between $\texttt{round}(0.8\bar{\Gamma}c_i)$ and $\texttt{round}(1.2\bar{\Gamma}c_i)$.

Figures 3-9-3-14 depict, either through scatter plots or standard box plots, the upper bounds on the AP heuristic suboptimality gaps we obtained for Studies 1 & 2 for varying number of nodes, tree depth, inventory cost parameters, attack severity parameters, antibiotic efficacy, and tree-violating edges. It can be seen that suboptimality gaps are small and do not appear to have strong correlation with any of the parameter values tested, including, the number of nodes and the attack severity.

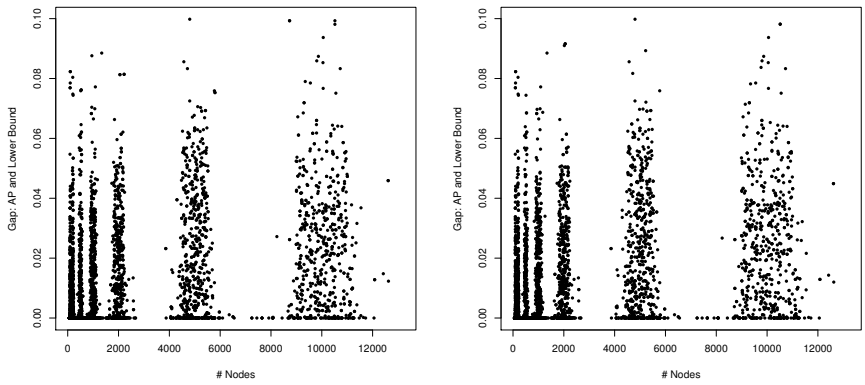

Figure 3-9: Upper bounds on AP suboptimality gap for varying number of nodes (Left: tree; right: non-tree).

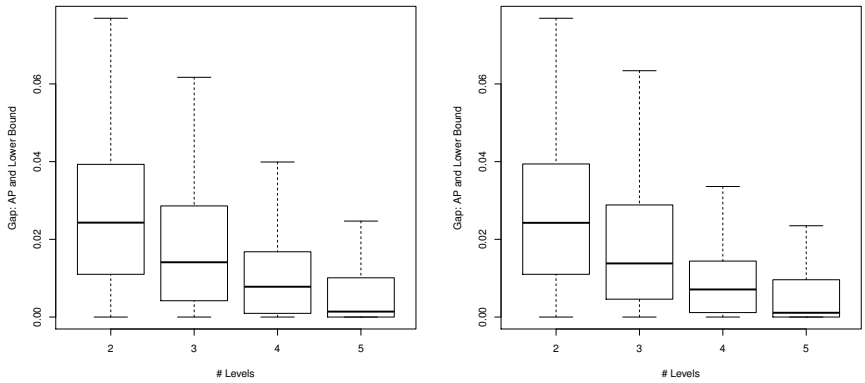

Figure 3-10: Upper bounds on AP suboptimality gap for varying number of levels (Left: tree; right: non-tree).

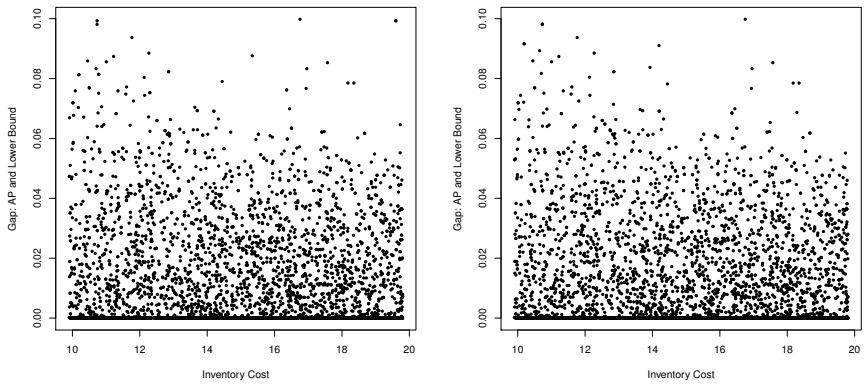

Figure 3-11: Upper bounds on AP suboptimality gap for varying inventory cost, $\bar{h}$ (Left: tree; right: non-tree).

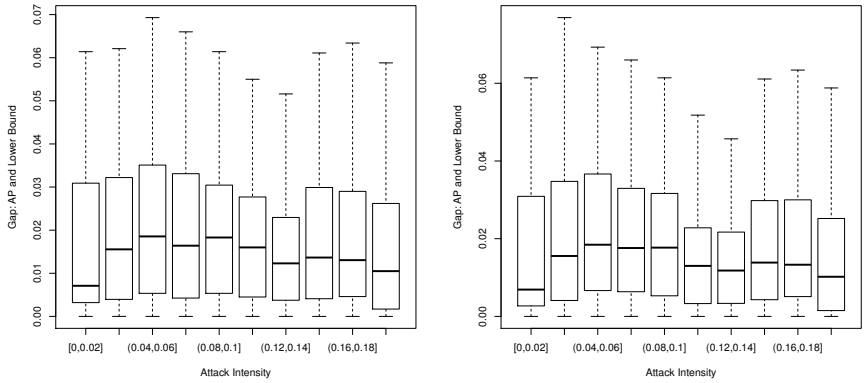

Figure 3-12: Upper bounds on AP suboptimality gap for varying attack severity, $\bar{\Gamma}$ (Left: tree; right: non-tree).

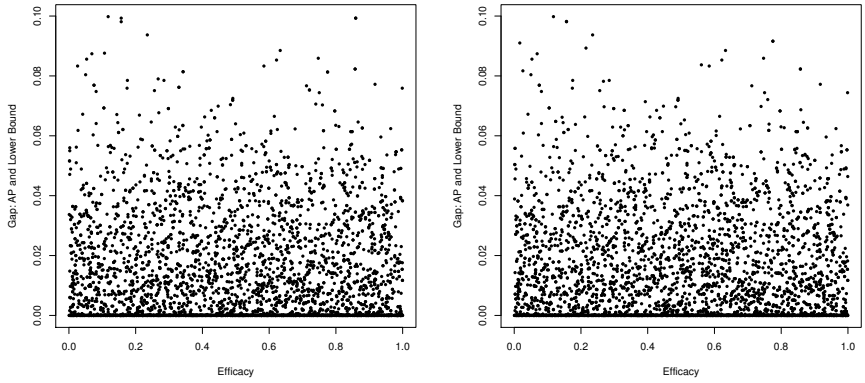

Figure 3-13: Upper bounds on AP suboptimality gap for varying antibiotic efficacy, $\bar{\rho}$ (Left: tree; right: non-tree).

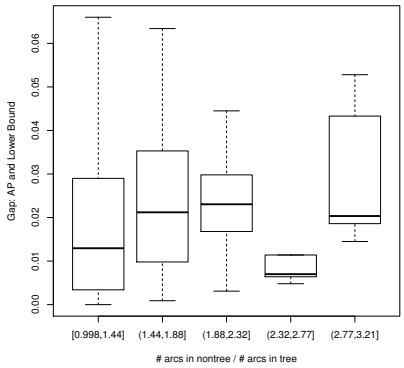

Figure 3-14: Upper bounds on AP suboptimality gap for varying network complexity.

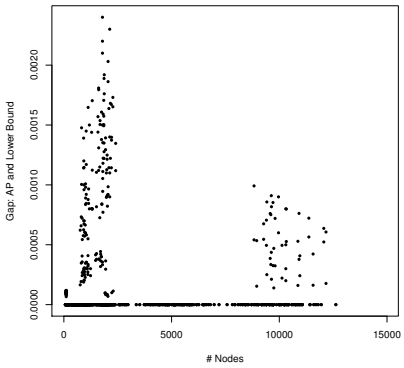

Figure 3-15: Upper bounds on AP suboptimality gap for varying number of nodes, capacitated networks.

3.9.2 Relaxing Infinite Capacity Assumption

To explore the AP heuristic's performance for the joint inventory and dispensing capacity optimization problem, (LLGC) in §3.5, we conducted numerical studies similar to Studies 1 & 2 described above.

Recall that (LLGC) differed from (LLG) by introducing capacity constraints (3.5e) for subsets of edges in the graph, and the associated capacity variables and costs. We conducted two studies, in both of which we considered the same setup as in Study 1. In the first of these two new studies, we generated these subsets of edges in a random fashion. In the second, we considered subsets of edges that had similar structure as in (3.5e)—specifically, sets of edges which all shared the same demand node. Both studies yielded similar results; we focus only on the latter in the remainder.

To introduce dispensing capacity, for each demand node, we partitioned its incoming edges to k equi-sized sets; k was sampled from $\{2, 4, 6, 8\}$. For each resulting set of edges, we introduced a capacity variable and a constraint, as in (3.5e). The associated capacity cost coefficient was sampled as a fraction of the life loss cost.

Table 3.5 provides statistics of the upper bounds on the suboptimality gaps we obtained. Figure 3-15 depicts the suboptimality gap bounds for varying number of

Min	1st Quantile	Median	Mean	3rd Quantile	Max
0.00%	0.00%	0.00%	0.15%	0.00%	0.22%

Table 3.5: Statistics of upper bounds on AP heuristic's suboptimality gaps for the numerical studies on the joint inventory and dispensing capacity optimization problem.

nodes. As before, our analysis reveals that the suboptimality gap is small and does not grow with the size of the network.

Chapter 4

Optimizing Influenza Vaccine Selection

4.1 Introduction

Seasonal influenza (flu) imposes high cost on health and economic activities every year. Globally, influenza-related respiratory diseases led to 290,000 to 650,000 deaths each year, between years 1999 and 2015 [75]. In the United States, the Centers for Disease Control and Prevention (CDC) estimates that influenza causes 9.3 million to 49.0 million illnesses, 140,000 to 960,000 hospitalizations, and 12,000 to 79,000 deaths annually since 2010 [58]. Since influenza is a communicable disease, one strategy to relieve the disease burden is to provide large quantity of preventative vaccines to the public in a timely manner [56]. In the 2017-2018 flu season, influenza vaccination prevented 7.1 million illnesses, 3.7 million medical visits, 109,000 hospitalizations, and 8,000 deaths [123].

Improving vaccine effectiveness (VE) is a focal point of current public health and policy studies. In the 2017-2018 season, VE was estimated to be 38% in the United States. A main reason behind the relatively low vaccine effectiveness is the mismatch between vaccine strains and the circulating virus strains – the actual virus strains that are causing an epidemic [138]. To understand such (in)effectiveness, we briefly summarize how influenza vaccines work here. Influenza vaccination works by exposing

human body's immune system to inactivated or attenuated virus strain. The immune system responds to virus intrusion by identifying which type of protective protein (antibody) from its repertoire works well against such virus, and proliferating such antibodies. The immune system then memorizes the production mechanism of this antibody type, resulting in more effective response to any future virus that is *similar* to the previous one. If the circulating viruses are similar to the vaccine virus, a vaccinated person stands a high chance of staying healthy and non-contagious. If the circulating viruses are not similar to the vaccine virus, the vaccinated person might still be protected (with a lower probability) due to so-called *cross-immunity*.

Ideally, the vaccine strains should be selected to precisely match future circulating virus strains. But this match quality is limited by the current vaccine manufacturing technology [16]. More specifically, three concurring factors cause mismatch between vaccine virus strains and circulating virus strains. First, the viruses' active components (antigens) undergo genetic changes constantly, a phenomenon known as antigenic drift and shift [55]. Drifting refers to relatively small genetic changes, resulting in reduced vaccine effectiveness. Shifting refers to abrupt, major changes of the virus protein structure, usually leading to a pandemic [145]. Second, due to long production lead time of vaccines, the strains need to be selected months ahead of flu season [57]. Third, the strains in the final, delivered vaccines may differ from the original seed strains, due to the strains adapting to their biological environments in the manufacturing process [138]. Therefore, limitations in the traditional vaccine manufacturing technology, along with constant antigenicity changes, reduce matching accuracy between vaccine strains and flu season circulating strains.

Indeed, improving manufacturing technologies has been a first-order priority from the policymakers' point of view, as evidenced from a testimony in a 2018 U.S. House of Representatives meeting [109]:

> "Continuous manufacturing holds great promise for both cell-based and recombinant vaccines because supply could be more easily ramped up on short notice. This would allow us to more rapidly address newly emerging strains or strain drift. Getting all the necessary preparatory work done is

one limiting step of the egg-based processes."

Several new vaccine technologies and manufacturing processes have been approved by the Food and Drug Administration (FDA) in 2013 and 2016 [59], which are more consistent in yield quality and quantity and have shorter production lead times [15, 29].

The implications of better vaccine manufacturing technologies are multi-faceted. A more effective vaccine is not only beneficial for the person being vaccinated, her social network, but also conducive to higher public confidence in vaccinations, all of which contribute to a virtuous cycle of more vaccine benefit and less disease burden for the society. From the international and national policymakers' point of view, this invites examination and update of the current influenza monitoring effort and vaccine strain selection criteria to better suit the new vaccine manufacturing technologies. From the vaccine manufacturing companies' point of view, the new vaccine technologies expand the space of business strategies, introducing opportunities for cooperation and areas of competition.

Research Question and Approach.

In this work, we address the question of how to select strains and produce vaccines, in light of new vaccine technologies. We support this quantitative policy analysis with analytical tools and publicly available data. The overall framework is not drastically different from existing operations management and operations research models. One can imagine the conceptual mapping between traditional operations management models (such as the newsvendor problem and the facility location problem) and the strain selection problem as follows: product / supply site selection decisions – vaccine strain selection, demand type – circulating virus strain type, demand changes – antigenic drift & shift, process flexibility – cross–immunity. And the list goes on. We will make the model more concrete in the sections to come.

Contributions.

We differ from these above mentioned works by incorporating different production technologies in the vaccine selection and production problem. We show that the newly FDA-approved manufacturing technologies have the potential to impact vac-

cine policies well beyond increasing vaccine availability. In particular, we show that selecting more than one vaccine strain for each virus subtype is *complementary* to adding a fast production technology. As a result, the policymakers should re-examine the current practice of selecting just one vaccine strain per subtype per flu season, because the benefit may outweigh the cost of selecting an additional strain.

To analyze this problem, we bring together several technical tools from a few areas – theoretical biology, discrete convex analysis, and distributionally robust optimization – to create a vehicle for our model and theory. The model and theory are general, and can be used to analyze other network newsvendor type and facility location type of problems.

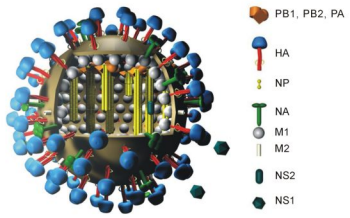

Figure 4-1: (Adopted from [74]) Structure of a virus: the HA and NA are characteristic proteins on a virus that can be recognized by human immune system. These proteins are called antigens – substance that causes immune system's antibodies to be generated.

4.2 Immunology Background

There are four types of influenza viruses: A, B, C, and D. Types A and B cause seasonal epidemics. The sudden changes in influenza A viruses can also cause a pandemic. Type C is mild, and type D is not known to affect humans [53].

Type A influenza virus, for example, are characterized by two proteins, hemagglutinin (HA) and neuraminidase (NA) (Figure 4-1). One key feature of a influenza virus is its imperfect polymerase (DNA and RNA formulation), which results in genetic mutations (antigenic drift and shift) that change the chemical properties of such proteins [86]. HA proteins assist viral entry into human cells by identifying host cells and facilitating the cell membrane merging, for the deposit of viral genome

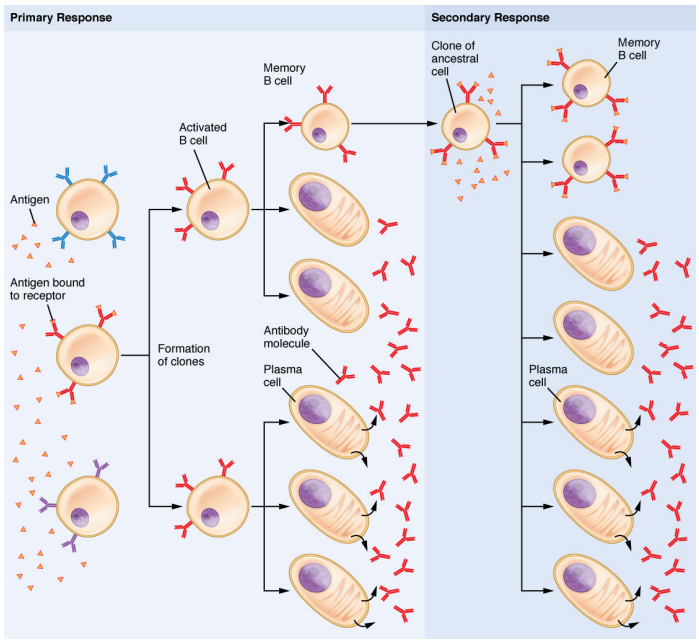

Figure 4-2: Clonal selection of B cells. Figure adopted from [110].

into human cells. NA proteins help the virus move through the human respiratory tract system and also assist the elution of replicated viruses out of host cells. The hemaglutinin has at least 18 subtypes, and the neuraminidase protein has at least 11 subtypes [60].

B cells, or B lympholytes, is a type of white blood cells that produce antibodies against viruses (B refers to the place that produces the cells: bursa of Fabricius in birds and bone marrow in mammals, and is not related to the B in influenza type B) [39]. Each B cell has one type of Y-shaped surface receptors (*i.e.*, antibodies), suitable for binding with specific types of antigens (virus surface proteins such as HA and NA). When a B cell encounters its compatible antigen, the receptor binds with the antigen and is said to be *activated*. After being activated, B cells differentiate into two types. The first type, plasma cells, act as supercharged production facility to generate antibodies, which then inactivate viruses and send out signals to other cells to clean up the bound virus clumps. The second type, memory cells, are primed for future encounters of similar types of antigens. The process of B cell activation, differentiation, and future activation are in summary called the clonal selection process (Figure 4-2).

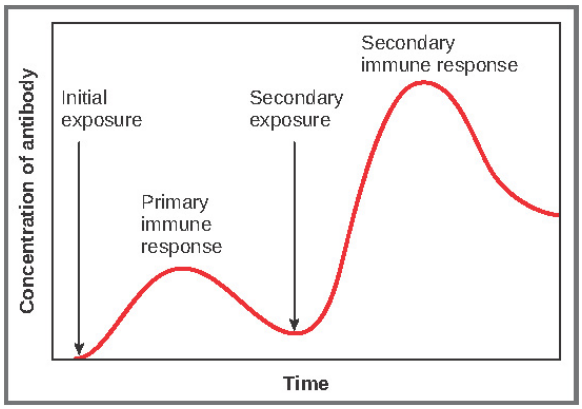

Figure 4-3: More efficient antigen recognition and containment during secondary response after the priming of memory cells. Figure adopted from [110].

Vaccination works by injecting specific types of inactivated or live attenuated viruses into the human body, to induce the production of memory cells that can recognize these types of viruses (B cells undergoing the primary response phase as depicted in Figure 4-2). In the subsequent flu season, if the actual circulating viruses are similar to the vaccine viruses, then memory cells would readily recognize the circulating viruses and more efficiently contain them (Figure 4-3).

As discussed in the previous section, the vaccine effectiveness is largely determined by the antigenic similarity between the vaccine virus strain and circulating virus strain. And this similarity can be measured in so-called *shape space* [117, 161]. In the operations research and computer science terminology, we can think of shape space as a linear feature space in which we can embed virus strains' genetic information. The vaccine effectiveness has been shown to closely follow an exponential function:

$$w(v, s) = e^{-\eta \|v-s\|^2},$$

where v and s are vector representations of the virus and vaccine strains in the feature space, η is the effectiveness coefficient, and norm $\|\cdot\|$ is the Euclidean norm in this feature space. Shape space theory has been studied and validated via extensive lab experiments and empirical data in this biological setting. For the case of influenza virus and antibody interaction, a two-dimensional manifold is sufficient [139].

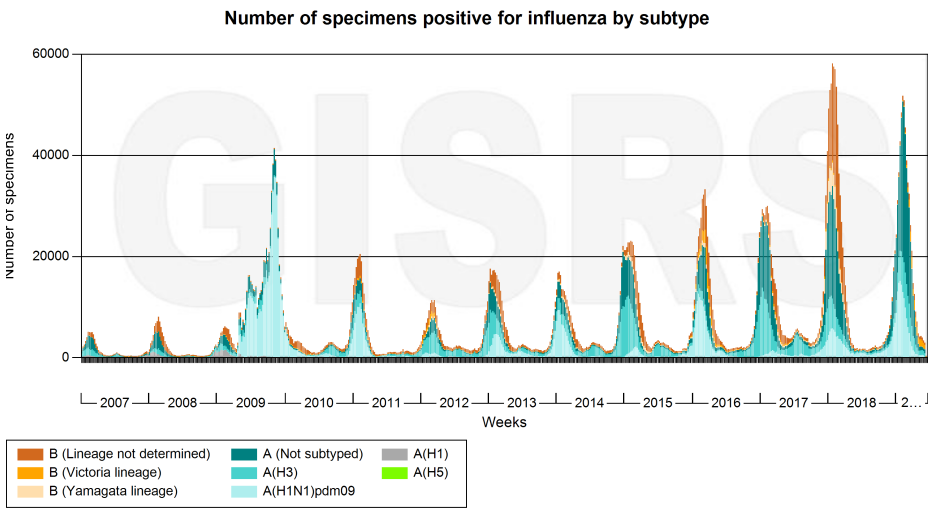

Figure 4-4: Global influenza virus circulation between 2007 and 2019. Data from [111].

4.3 Vaccine Selection and Production

Continuous antigenic shift of influenza viruses makes total immunity difficult. Influenza vaccines have to be updated every year to target newly emerging variants of the type A and B viruses. The World Health Organization (WHO) monitors the circulation of influenza viruses through a global collaboration network, the Global Influenza Surveillance and Response System (GISRS) (Figure 4-4) [112]. In the U.S., the CDC also closely monitors the outpatient illness and viral activities on the federal and local levels [54].

Potential viruses are isolated to undergo lab experiments, and sometimes full gene sequencing, every year [139]. The main method of experiment, *hemaglutination inhibition* (HI) assay, generates a data matrix. Each entry of the matrix HI_{ij} indicates the similarity between the reference virus j and potential circulating virus i. Statistical learning and optimization methods are then used to convert HI data to create antigenic maps, which is a specific implementation of the shape space theory [87].

With the experimental studies and quantitative methods, WHO makes recommendation about the specific virus strains to include the subsequent vaccine production (February/March for the northern hemisphere, and August/September for the southern hemisphere). For example, the WHO recommended the following four strains for

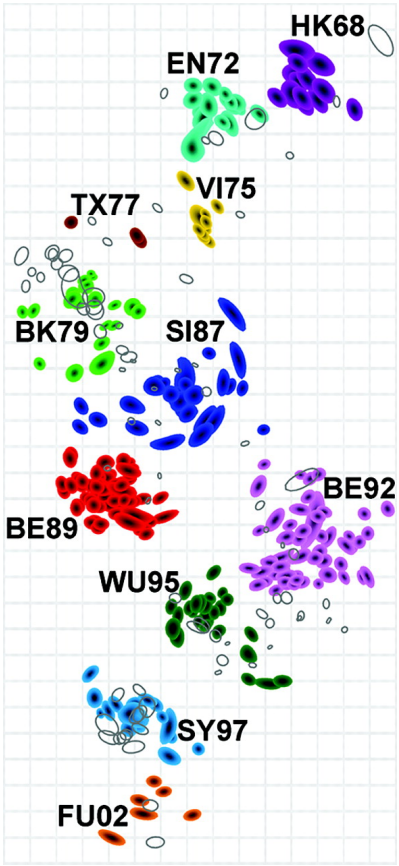

Figure 4-5: The antigenic map of influenza A (H3N2) from year 1968 to 2003. Figure adopted from [139].

the 2019–2020 influenza seasons for northern hemisphere in March 2019 [112]:

- A/Brisbane/02/2018 (H1N1)pdm09-like virus;
- A/Kansas/14/2017 (H3N2)-like virus;
- B/Colorado/06/2017-like virus (B/Victoria/2/87 lineage); and
- B/Phuket/3073/2013-like virus (B/Yamagata/16/88 lineage).

In the first strain, for example, "A" refers to the antigenic type, "Brisbane" refers to the geographical origin, "02" refers to the strain number, "2018" refers to the year of strain isolation, and "pdm09" refers to the pandemic in 2009. For influenza A type viruses, the HA and NA description is included in the parantheses, such as (H1N1) and (H3N2).

Seasonal influenza vaccines are monovalent, trivalent, or quadrivalent – indicating that the vaccines include one, three, or four types of the recommended strains. The traditional vaccine production technology is heavily dependent on hen's eggs [59, 148]. In the United States, the CDC and labs provide candidate vaccine viruses (CVV) to the vaccine manufacturers. Then, the CVVs are incubated in fertilized eggs for several days. Subsequently, the virus-containing fluid is isolated and purified according to FDA-approved processes. The final vaccine yield depends highly on the quantity and quality of healthy eggs. And because quality varies from one lot to the next, each lot has to be separately approved by the FDA. Furthermore, the (random) adaptation of viruses to the biological environment in eggs reduces vaccine effectiveness. Overall, production lead time is at least six to eight months from the time CVVs are provided to manufacturers to delivery [57].

New vaccine technologies recently approved by the FDA can improve both the yield quality and reduce the production lead time of vaccines [96]. For example, the recombinant protein approach completely avoids the reliance on eggs. After strains are selected, the gene encoding for the HA proteins are directly cloned into base viruses, which are injected into insect cells for replication, and later purification. The overall production is more reliable and has a production lead time shorter than

38 days, from gene isolation to delivery [29]. In addition to HA and NA targeting vaccines, new technologies (so-called *universal vaccines*) try to elicit human immune response to target to portions of viruses that do *not* drift or shift much. The first clinical trial for universal vaccines has begun recently [104].

4.4 Decision Model: Vaccine Selection and Production

The main question we hope to address is the impact of vaccine technologies on global strain selection decisions. More precisely, we want to understand how vaccine strain selection policy would change, given the decrease in vaccine production lead times, and increase in the production capacity. In this section, we only model the vaccine strain selection process for one subtype, *e.g.*, influenza A (H1N1), to facilitate exposition and focus on the question of how many strains we should select *within* each virus subtype. However, the model and theory in this and later sections can be readily generalized to include multiple virus subtypes.

We can visualize this optimization problem as a network design problem on a bipartite graph $(V \cup V_D, E)$. On the one side, node $i \in V$ are supply nodes. Let index i represent the ith vaccine strain candidate, and decision variables z_i, x_i, $\tilde{x}_i$ represent the strain selection, production quantity of slow vaccine, and production quantity of fast vaccine respectively. We let $m = |V|$. On the other side of the bipartite graph, let $j \in V_D$ be the jth circulating strain candidate, and d_j be the number of people susceptible to strain j. The support of random vector $d := \{d_1, d_2, \ldots, d_n\}$ is $D := D_1 \times D_2 \times \ldots \times D_n$, where $n = |V_D|$. Decision variable y_{ij} represents the number of vaccine i used to vaccinate people with virus j. The demand is random and takes distribution μ. To hedge against uncertainty, we assume the demand distribution is drawn from a distribution family Γ.

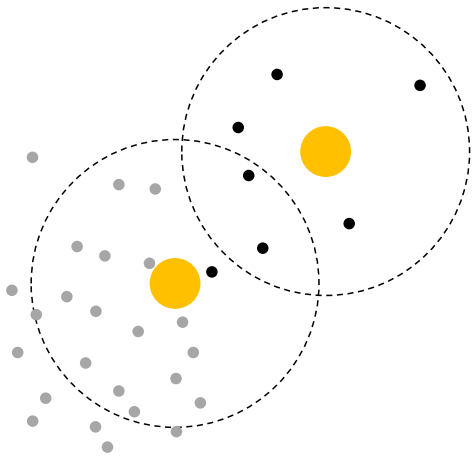

Figure 4-6: Example of vaccine strain selection being misguided by antigenic drift, visualized in a two-dimensional antigenic map. Small black dots: virus strains that show up earlier in the flu season. Small gray dots: virus strains that show up later in the flu season. Upper-right yellow dot: vaccine strain selected to target earlier circulating strains. Lower-left yellow dot: vaccine strain selected later in the season to target all strains. Dotted circle: vaccine's effectiveness range.

$$\text{VS}: \quad \min_{x,z} \max_{\mu \in \Gamma} \quad \mathbb{E}_{d\sim\mu}[\min_{\tilde{x},y} f(x,\tilde{x},y,z,d)] \tag{4.1a}$$

$$\text{subject to} \quad \sum_{j\in N_i^-} y_{ij}(d) \le x_i + \tilde{x}_i(d), \quad \forall i \in V, \forall d \in D \tag{4.1b}$$

$$\sum_{i\in N_j^+} y_{ij}(d) \le d_j, \quad \forall j \in V_D, \forall d \in D \tag{4.1c}$$

$$x_i + \tilde{x}_i(d) \le N z_i, \quad \forall i \in V, \forall d \in D \tag{4.1d}$$

$$\sum_{i\in V} x_i \le c_e, \tag{4.1e}$$

$$\sum_{i\in V} \tilde{x}_i \le c_r, \tag{4.1f}$$

$$\sum_{i\in V} z_i \le \kappa \tag{4.1g}$$

$$\tilde{x}(d), y(d) \ge 0, \quad \forall d \in D \tag{4.1h}$$

$$z \in \mathcal{Z} \subseteq \{0,1\}^m. \tag{4.1i}$$

Constraints (4.1b) ensure that the total amount of vaccine i allocated out is no more than the total amount of vaccine i produced. Constraints (4.1c) show the total number of people vaccinated is no more than the size of susceptible population. Constraints (4.1d) show that the total production quantity of vaccine type i cannot be positive unless we have selected this strain i (N is a large enough constant, *e.g.*, $N = \sum_{j \in V_D} \max\{d_j \,|\, d_j \in D_j\}$). Capacity parameters c_i, c_e, and c_r represent the maximum amount of vaccine i, total amount of eggs available for slow vaccine production, and total amount of cells available for fast vaccine production respectively. Constraint (4.1g) shows that the total number of strains to be selected should not exceed κ. Unless specifically mentioned, coefficients in this work are positive. N_j^+ is the set of supply nodes that can supply j, $N_j^+ := \{i \in V \,|\, (i,j) \in E\}$, N_i^- is the set of demand nodes that are supplied by i, $N_i^- := \{j \in V_D \,|\, (i,j) \in E\}$, $\mathcal{Z}$ is the feasible region for vaccine selection decisions.

Assumption 4.4.1 *Cost function f is linear.*

Linearity on production related variables $z, x, \tilde{x}$ are standard in the operations management literature. Linearity on the fulfillment vector y and demand d are not immediately clear given that we are treating people that potentially can contract communicable disease. Note that this assumption still allows us to incorporate the fact that if one person is vaccinated, her social network also benefits. The only limitation is that we assume there are no "cross" effects when vaccinating the population. A more extensive discussion with simulation results can be found in [46], which numerically shows that the disease burden is very close to a linear function of the number of people vaccinated. More specifically, we assume the cost function to take the form:

$$f(x, \tilde{x}, y, z, d) := \sum_{i \in V} (h_i x_i + \tilde{h}_i \tilde{x}_i + \hat{h}_i z_i) + \sum_{j \in V_D} (b_j d_j - \sum_{i \in N_j^+} v_{ij} y_{ij}),$$

where we assume the disease burden for demand type j is b_j per person. The logistical costs are h_i, $\tilde{h}_i$, and $\hat{h}_i$. The network vaccine effectiveness is v_{ij}, defined as the probability of vaccine preventing illness for the vaccinated person, multiplied by the

expected number of people she protects in her network if she does not carry the flu virus. A consequence of this choice of objective function is that the last term, $(b_j d_j - \sum_{i \in N_j^+} v_{ij} y_{ij})$, can theoretically be negative, if the network vaccination effectiveness is greater than 1 (*e.g.*, VE for the vaccinated person is 50% and that also protects four other people in her social network). But for this specific model and calibration, we observe numerically that this is not a concern due to current low vaccine effectiveness.

We have also made several other assumptions in proposing this model:

Assumption 4.4.2 *(Model Assumptions)*

1. *One person is susceptible to only one type of circulating virus strain.*
2. *Antigenic drift is independent of vaccination,* i.e.*, demand d is not a function of x, y.*
3. *Evaluation of the number of susceptible people is accurate after epidemic influenza has circulated long enough,* i.e.*, $\tilde{x}$ is a decision after realization of d.*
4. *There is only a finite number of strains,* i.e.*, $|V|$ and $|V_D|$ are finite.*
5. *Central planner chooses the strains and production quantity, manufacturers follow truthfully.*

In reality, some people can be affected by multiple strains in the same flu season. But most of these cases are outpatients with exposure within healthcare facilities and constitute on the order of 1% of the overall population [66]. Furthermore, if we insist, our model can be easily extended to incorporate cases where a group of people are susceptible to multiple strains of circulating viruses. We can do so by allowing demand nodes j to represent the jth *composition* of virus strains. The second assumption is a standard assumption in the vaccination literature [161]. The third assumption relies on the fact that new vaccine production technology has a significantly shorter production lead time (on the order of a few weeks), compared with egg-based production technologies (on the order of six to eight months) [29]. The fourth assumption comes from practice – the total number of strains tested

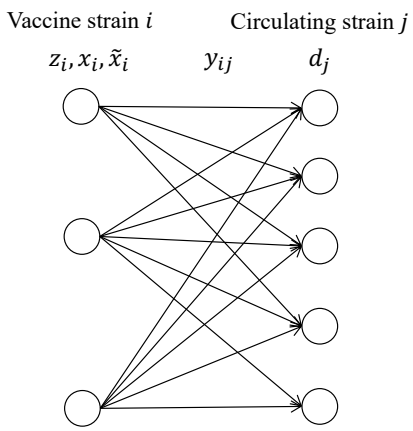

Figure 4-7: Vaccine selection casted as a network design problem.

experimentally is finite. The last assumption is based on current WHO and CDC practices, and the fact that manufacturers have to rely on the vaccine strains selected by national laboratories.

4.5 Theoretical Properties and Solution Technology

In this section, we show that the inner minimization of problem (4.1) can be equivalently formulated as a max weight circulation problem, which in turn allows us to prove several theoretical properties. For example, we show that the inner minimization is supermodular in the uncertain demand d, and as a consequence, the overall problem is efficiently solvable. Similar techniques can also be applied to infer that the model has certain complementarity condition between vaccine strain selection and the new manufacturing capacities.

More precisely, we define the inner minimization problem as

$$
\begin{aligned}
g(z, x, d) := \min_{\tilde{x}, y} \quad & f(x, \tilde{x}, y, z, d) && \text{(4.2a)} \\
\text{subject to} \quad & \sum_{j \in N_i^-} y_{ij} \leq x_i + \tilde{x}_i, \quad \forall i \in V && \text{(4.2b)} \\
& \sum_{i \in N_j^+} y_{ij} \leq d_j, \quad \forall j \in V_D && \text{(4.2c)} \\
& x_i + \tilde{x}_i \leq N z_i, \quad \forall i \in V && \text{(4.2d)} \\
& \sum_{i \in V} \tilde{x}_i \leq c_r, && \text{(4.2e)} \\
& \tilde{x}, y \geq 0 && \text{(4.2f)}
\end{aligned}
$$

For exposition, we sometimes suppress its dependency on x and z and simply write $g(d)$. Sometimes we also elevate the coefficients such as c_r to the arguments of g to emphasize its dependence on these parameters.

We also sometimes refer to the worst-case problem:

$$
g^\star_\Gamma(z, x) := \max_{\mu \in \Gamma} \mathbb{E}_{d \in \mu}\, g(z, x, d).
$$

We suppress the dependency on Γ when there is no ambiguity for the choice of ambiguity set Γ.

Definition An order list $\{v_1, a_1, v_2, a_2, \ldots, v_n, a_n, v_{n+1}\}$ is a *simple cycle* if

1. $v_i \in V$ and $a_i \in E$ for all $i = 1, \ldots, n$,
2. for any $i < i'$, we have that $v_i = v_{i'}$ if and only if $i = 1, i' = n + 1$,
3. for all $i, i' \in \{1, \ldots, n\}$, we have that and $a_i \neq a_{i'}$.

Definition Two arcs are *parallel* if they have opposite orientations in every simple cycle that contains both of them. A set of arcs are parallel arcs if they are pairwise parallel. Similarly, a pair of arcs are in *series* if they have the same orientation in

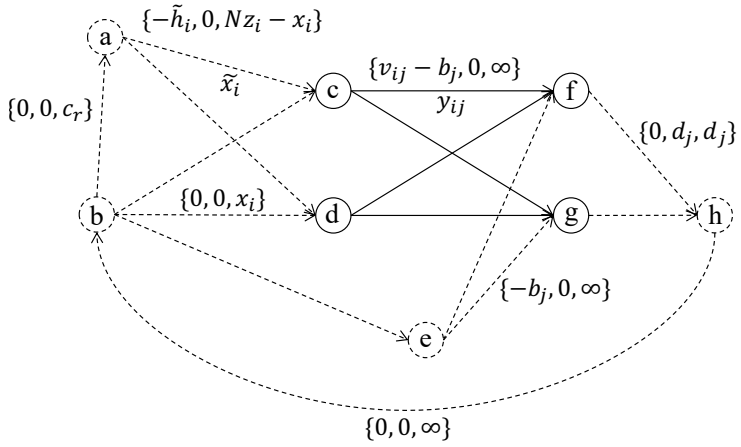

Figure 4-8: Converting the network design problem into an equivalent max weight circulation problem. Solid lines and circles are from the original network, dotted circles and arrows are auxiliary components to construct the circulation. Nodes b and h are source and sink nodes. Node a is added to incorporate fast vaccine capacity $\tilde{x}_i$. Node e is added so that flow on (e, f) and (e, g) represents the number of people that do not receive vaccines on node f and node g. Elements in the tuple on each arc represents its arc cost, flow lower bound, and flow upper bound.

every simple cycle that contains both of them. And a set of arcs are in series if they are pairwise in series.

With '$\wedge$' and '$\vee$' as the conventional meet and join operators for vectors (componentwise min and max), we define the key concept in our analysis:

Definition $g(d) : \mathbb{R}^n \to \mathbb{R}$ is *supermodular* in d if for any d, d', the following condition holds:

$$g(d) + g(d') \leq g(d \wedge d') + g(d \vee d').$$

$g(d)$ is submodular if $-g(d)$ is supermodular.

We also define a weaker notions:

Definition Let $d = (d_1, d_2)$, and $d' = (d'_1, d'_2)$, where d_1 and d'_1 have the same dimensions, and likewise for d_2 and d'_2. Then $g(d) = g(d_1, d_2)$ is *complementary* in its first and second sets of arguments if

$$g(d_1, d_2) + g(d'_1, d'_2) \leq g(d_1 \wedge d'_1, d_2 \wedge d'_2) + g(d_1 \vee d'_2, d_1 \wedge d'_2).$$

We say that d_1 and d_2 are substitutes for $g(d_1, d_2)$ if they are complementary for $-g(d_1, d_2)$.

With these definitions of simple cycles, parallel arcs, and supermodularity, we are ready to present the first result for our model.

Lemma 4.5.1 *$g(d)$ is supermodular in d.*

Proof (Lemma 4.5.1) We first characterize a closely related problem: max weight circulation. A max weight circulation problem is a network optimization model that maximizes the total weighted circulation in a network, subject to lower and upper bounds on the flow on each arc:

$$\min_{y} \quad \sum_{a \in E} w_a y_a \tag{4.3a}$$

$$\text{subject to} \quad \sum_{a \in v^+} y_a = \sum_{a \in v^-} y_a, \quad \forall v \in V \tag{4.3b}$$

$$y_a \in [l_a, u_a], \quad \forall a \in E \tag{4.3c}$$

Here V and E are the node and arc set of the circulation problem, v and a refer to particular node and arc, v^+ and v^- refer to the incoming and outgoing arc sets for v. We associate each arc with a tuple of three parameters: $\{w_a, l_a, u_a\}$. They are the weight of flow, lower bound of flow, and upper bound of flow on this particular arc a.

It is straightforward to see that given a particular d and first stage decision x, z, the network flow problem $g(z, x, d)$ can be represented as a max weight circulation problem. To exemplify, we use Figure 4-8 to represent this specific instance of max weight circulation problem $g'(z, x, d)$, associated with $g(z, x, d)$. In this schematic, we highlight the decision variables $\tilde{x}_i$ and y_{ij} and the coefficient tuples related to $g(x, z, d)$ to show the transformation. According to the way we defined our network for $g'(z, x, d)$, we have that

$$g(z, x, d) = -g'(z, x, d).$$

It is known in the discrete convexity literature that for any max weight circulation problem g', if a set P of arcs are parallel, then g' is submodular in c_P, where c_P is the set of arc capacities of P [61].

But it is obvious that in our network for $g'(d)$, the arcs associated with d_j always connect to the sink node (*e.g.*, node (h) in Figure 4-8), implying that any pair of d_j-capacitated arcs are in parallel.

Therefore, $g'(z, x, d)$ is submodular in d, and $g(z, x, d)$ is supermodular in d. □

Having characterized the supermodularity of $g(d)$, we can then leverage a result in stochastic programming to show that the so-called price of correlation in optimization this problem is large.

Definition (Price of Correlation, Agrawal *et al.* (2012) [2]) We define the price of correlation of any function $g(x, d)$ with regard to some nominal distribution μ_0 and ambiguity set Γ to be:

$$\text{POC} := \sup_{\mu \in \Gamma} \frac{\mathbb{E}_{d\sim\mu}[g(x_S, d)]}{\mathbb{E}_{d\sim\mu}[g(x_D, d)]},$$

where $x_S := \text{argmin}_x \, \mathbb{E}_{d\sim\mu_0} g(x, d)$, and $x_D := \text{argmin}_x \max_{\mu\in\Gamma} \mathbb{E}_{d\sim\mu} g(x, d)$.

In our context, the price of correlation essentially quantifies the risk of solving the optimization model with regard to a fixed demand distribution μ when the true demand distribution could be drawn from the ambiguity set Γ. If POC is close to 1, that means one can simply solve the optimization problem and be "robust" against all realized distributions from Γ. If POC is much larger than 1, then by implementing the solution x_S from stochastic programming, the actual loss could be much larger under some realizations of the actual distribution.

We show in the following statement that, the POC in our case is not bounded above. Let us define: $b := \max\{b_j \,|\, j \in V_D\}$, and $h := \min\{h' \,|\, h' \in \{h_i, \tilde{h}_i\},\, i \in V\}$.

Proposition 4.5.2 *The price of correlation for our optimization problem (4.1) can be as large as b/h.*

Proof (Proposition 4.5.2) The proof is straightforward if we follow the same argument as [2]. □

As a result of Proposition 4.5.2, the price of correlation can be extremely large, given that b represents the per unit disease burden, while the h is the per unit manufacturing cost. This justifies the distributionally robust approach we have taken in problem (4.1), that we included a distribution family Γ to hedge against uncertainty. Popular choices of Γ in the literature include Wasserstein ball, and marginals model.

Definition Let (M, d) be a metric space. The pth Wasserstein distance between two measures μ and ν on M as

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)^p \, \mathrm{d}\gamma(x, y) \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is the measure on $M \times M$ with marginals μ and ν.

Definition The Wasserstein ball (with radius r) around measure μ is defined as

$$B_r^p(\mu) = \{\nu \,|\, W_p(\mu, \nu) \leq r\}.$$

To discuss the tractability of problem (4.1), we have to show that the

Theorem 4.5.3 *(Tractability Under Wasserstein Ambiguity Set) Problem (4.1) is polynomially solvable with regard to a Wasserstein-ball ambiguity set. That is,* $\Gamma = B_r^p(\mu)$*, given some nominal* μ *and radius* r*.*

Proof (Theorem 4.5.3) Formulation (4.1) – with the linear objective – is a special case of the formulation (18b) in Esfahani and Kuhn (2017), Page 30 [100]. $\square$

An alternative formulation for ambiguity set is the marginals model. We assume the given input is a set of marginal distributions μ_j for each demand node j, and we want the problem to be robust against all joint distributions μ that are consistent with such marginals.

Definition Given a set of measures μ_j on M_j for all $j \in V_D$, the set of joint distributions that is consistent with the marginal distributions is

$$\Gamma(\mu_1, \mu_2, \ldots, \mu_n) := \{\mu \,|\, \Pi_j \mu = \mu_j\},$$

where Π_j is the projection onto M_j.

We have assume the marginal distributions are given with a finite number of support points. For example, such marginal μ_j may be an empirical distribution with finitely many data points, or a measure that is absolutely continuous with regard to this empirical distribution.

Assumption 4.5.4 *Each marginal distribution μ_j has at most $l < \infty$ points in the support,* i.e., *$|D_j| \leq l$ for all j in V_D.*

Lemma 4.5.5 *(Monotone Coupling Lemma, Theorem 3.1.2 in Rachev and Rüschendorf (1998) [121], Lemma 1 in Chen* et al. *(2019) [33]) When problem (4.1) is endowed with the marginal ambiguity set $\Gamma(\mu_1, \mu_2, \ldots, \mu_n)$, the worst-case joint distribution $\mu^\star$ can be fully characterized by $\mu_1, \mu_2, \ldots, \mu_n$ (independent of $z, x, \tilde{x}, y$). Furthermore, $\mu^\star$ has at most nl support points and can be computed from $\{\mu_j, j = 1, \ldots, n\}$ in $\mathcal{O}(nl)$ time.*

Proof (Lemma 4.5.5) $g(d)$ is continuous in d since the optimization problem $g(d)$ is always feasible. This allows us to apply the theorem in Rachev and Rüschendorf (1998) and Chen *et al.* (2019) directly to get the desired result. □

This brings us to the tractability result for problem (4.1) under the marginal model.

Theorem 4.5.6 *(Tractability Under Marginal Ambiguity Set) Problem (4.1) can be solved efficiently under the marginal ambiguity set $\Gamma(\mu_1, \mu_2, \ldots, \mu_n)$. More precisely, the problem can be reformulated to a mixed integer linear minimization problem with $\mathcal{O}(mn^2l)$ continuous variables, m binary variables, $\mathcal{O}(mnl + nnl)$ linear constraints, and $\mathcal{O}(ml)$ constraints with binary variables, where m is the number of supply nodes, n is the number of demand nodes, $l = \max\{|D_j| : \ j = 1, \ldots, n\}$ is the maximum number of support points in the discrete marginals.*

Proof (Theorem 4.5.6) With the definition of $g(d)$, problem (4.1) is:

$$\min_{x,z} \max_{\mu \in \Gamma} \mathbb{E}_{d\sim\mu}[g(z, x, d)].$$

Under the marginal model, let $\mu^\star$ be the worst-case distribution in $\Gamma(\mu_1, \mu_2, \ldots, \mu_n)$. Then we get the following reformulation:

$$\min_{x,z} \mathbb{E}_{d\sim\mu^\star}[g(z,x,d)].$$

Therefore, problem (4.1) is equivalent to this optimization problem

$$\min_{x,z,\tilde{x},y} \quad \sum_{i\in V}(h_i x_i + \hat{h}_i z_i + \sum_{s=1}^{S}\tilde{h}_i\tilde{x}_i^s) + \sum_{j\in V_D}\sum_{s=1}^{S}(b_j d_j^s - \sum_{i\in N_j^+} v_{ij}y_{ij}^s)p_s \tag{4.4a}$$

$$\text{subject to} \quad \sum_{j\in N_i^-} y_{ij}^s \le x_i + \tilde{x}_i^s, \quad \forall i \in V,\, s = 1,\ldots,S \tag{4.4b}$$

$$\sum_{i\in N_j^+} y_{ij}^s \le d_j^k, \quad \forall j \in V_D,\, s = 1,\ldots,S \tag{4.4c}$$

$$x_i + \tilde{x}_i^s \le N z_i, \quad \forall i \in V,\, s = 1,\ldots,S \tag{4.4d}$$

$$\sum_{i\in V} x_i \le c_e, \tag{4.4e}$$

$$\sum_{i\in V} \tilde{x}_i^s \le c_r, \quad s = 1,\ldots,S \tag{4.4f}$$

$$\sum_{i\in V} z_i \le \kappa \tag{4.4g}$$

$$x, \tilde{x}, y \ge 0, \tag{4.4h}$$

$$z \in \mathcal{Z} \subseteq \{0,1\}^m, \quad \forall i \in V \tag{4.4i}$$

where the worst-case joint distribution $\mu^\star$ is supported on $\{d_s,\, s = 1,\ldots,S\}$, and $\mu^\star(d_s) = p_s$; $\tilde{x}^s$ and y^s are scenario-based second stage decision variables.

By Lemma 4.5.5, we have $S \le nl$. Therefore problem (4.1) can be solved in form (4.4), with the above mentioned numbers of constraints and variables. $\square$

Up to now, we have demonstrated that our model is tractable for important ambiguity set formats. Next, we hope to generate some insights in the comparative statistics of problem (4.1). Again, we rely on the tools developed in discrete convex analysis by Gale and Politof (1981) [61] and Murota and Shioura (2005) [101].

Recall that with a slight abuse of notation, we can write the inner minimization

problem of (4.1) as $g(c_r, z)$ to emphasize its dependence on the optimization parameters like c_r and z.

Theorem 4.5.7 *(Complementarity in Value) Under the marginal ambiguity model, if $\mathcal{Z} \subseteq \{0,1\}^n$ is a totally ordered lattice, then c_r and z are substitutes for the worst-case cost of problem (4.1), in the sense that:*

$$g^\star(c_r, z) + g^\star(c_r', z') \geq g^\star(c_r \wedge c_r', z \wedge z') + g^\star(c_r \vee c_r', z \vee z').$$

Proof (Theorem 4.5.7)

We first note that c_r and z being substitutes for cost $g^\star(c_r, z)$ is equivalent to saying that c_r and z are complementary for "utility" $-g^\star(c_r, z)$. Thus the naming of the theorem in the direction of "complementarity".

We prove the substitutes statement with the following steps: we show that the inner minimization problem satisfies the substitutes condition, *i.e.*, c_r and z are substitutes for $g(d, c_r, z)$ under totally ordered $\mathcal{Z}$, then we show the condition holds for $g^\star$. To see the first part, note that g corresponds to a max weight circulation problem g' where c_r and z_i (for some fixed i) correspond to flow capacities on two arcs that are always in series (see Figure 4-8). Therefore, by the Main Theorem in Gale and Politof (1981) [61] Theorem 1.3 in Murota and Shioura (2005) [101], c_r and z_i are supermodular for the circulation function g'. Thus, c_r and z_i are submodular for g (provided that setting $z_i = 1$ is within the feasible set set by κ).

The substitution condition of g in c_r and z_i for some fixed i implies the substitution condition of g in c_r and z under the condition that $\mathcal{Z}$ is a totally ordered set. To see this, recall that $\mathcal{Z}$ is a lattice in $\{0,1\}^n$, and its total order implies that for any $\underline{z}, z \in \mathcal{Z}$, $\sum_{i \in V} z_i - \sum_{i \in V} \underline{z}_i = 1$ if and only if the two binary vectors differ in exactly one bit. Therefore, for any fixed $\bar{c}_r \geq \underline{c}_r$, $z > \underline{z}$ differing by 1 bit, substitution of c_r and z_i implies the following condition:

$$g(\underline{c}_r, \underline{z}) + g(\bar{c}_r, z) \leq g(\underline{c}_r, z) + g(\bar{c}_r, \underline{z}).$$

Now suppose $\bar{z} > z$, and $\bar{z}$ differs from z by exactly 1 bit. Then we have that $\bar{z} > z > \underline{z}$, each inequality represents a 1-bit difference. Then repeating the previous argument, we have that:

$$g(\underline{c}_r, z) + g(\bar{c}_r, \bar{z}) \leq g(\underline{c}_r, \bar{z}) + g(\bar{c}_r, z).$$

Adding the two inequalities up, we have that:

$$g(\underline{c}_r, \underline{z}) + g(\bar{c}_r, \bar{z}) \leq g(\underline{c}_r, \bar{z}) + g(\bar{c}_r, \underline{z}).$$

By induction, we can then show that c_r and z are substitutes for g if $\mathcal{Z}$ is totally ordered.

By Proposition 5 in Chen (2017) [34], the expectation operation preserves submodularity. But we know from worst-case cost definition and Lemma 4.5.5 that

$$g^\star(c_r, z) := \max_{\mu \in \Gamma(\mu_1, \ldots, \mu_n)} \mathbb{E}_{d\sim\mu} g(d, c_r, z) = \mathbb{E}_{d\sim\mu^\star} g(d, c_r, z).$$

Thus, c_r and z are substitutes for the worst-case cost $g^\star(c_r, z)$. $\square$

Theorem 4.5.7, roughly speaking, says that the act of adding new vaccine production technology and the act of selecting additional vaccine strains are substitutes in cost, and therefore *complementary* in value: the benefit from doing both is greater than the sum of the benefits of doing just one of them. The assumption on totally ordered $\mathcal{Z}$ is practically sound if we assume that there is a natural preference ordering of the vaccine strain candidates i, *e.g.*, the strain i will always be selected before $i+1$ is selected in the final decision.

Although Theorem 4.5.7 only makes a statement about substitution and complementarity with regard to the inner maximization cost $g^\star$, we observe numerically that the same hold (approximately) for the overall problem (4.1).

4.6 Experiment and Policy Insight

To analyze problem (4.1), we calibrate the model with the following data:

- Total number of vaccine strain candidates: m.
- Total number of circulating virus strain candidates: n.
- Total number of strains to be selected: κ.
- Fixed cost of strain selection: $\hat{h}$.
- Variable cost of vaccine production and delivery for old and new technologies: h, $\tilde{h}$.
- Disease burden: b.
- Value of vaccination: v.
- Estimation of the size of susceptible population: D and Γ.
- Production capacity of old (egg-based) technology: c_e.
- Production capacity of new (recombinant) technology: c_r.

Number of Vaccine Strain Candidates m and Epidemic Strain Candidates n. We could not find the total number of vaccine strain candidates and circulating virus candidates in the literature or public documentations. However, our intuition is that both m and n are on the order of 10, given the size of the experiments in related papers. For example, the hemagglutination inhibition assay analysis tend to have a matrix on the order of 10 strains by 10 strains [8, 108, 125].

Total Number of Strains Selected κ. Currently, the WHO and CDC choose one strain per influenza subtype [113]. For our analysis, we also includes higher numbers to test the impact of including additional vaccine strains.

Fixed Cost of Strain Selection $\hat{h}$. To avoid contamination in the vaccine production process, equipment cannot be re-used for products with different biological or chemical compositions. Therefore, producing more than one type of vaccines would entail the

setup of multiple production facilities. The cost of setting up each additional antigen strain is estimated to be between 50 to 500 million US dollars [119].

Variable Costs h, $\tilde{h}$. The variable cost of vaccine production, transportation, and delivery is estimated to be \$5 to \$15 per person [50].

Disease Burden b. The economic disease burden of a case of influenza illness is estimated to be \$630 to \$2530 per person [120].

Vaccine Effectiveness v. The societal effectiveness of a vaccine is the product of two numbers: individual vaccine effectiveness and network protection (reproduction number R in the public health literature). The individual VE for matched strains is around 54% to 73%, and individual VE for mismatched strains is around 37% to 63% [52, 151]. The reproduction number for seasonal influenza (not pandemic) is estimated to be $R \in [1.19, 1.37]$, meaning that keeping one person healthy can in expectation reduce another 0.19 to 0.37 cases of influenza sickness [22]. Therefore, the overall VE is between 0.64 to 1 for matched vaccine and circulating strains, and between 0.44 and 0.86 for mismatched strains.

Estimation of the Susceptible Population D and Γ. We construct synthetic distribution and ambiguity set for demand distribution based on historical numbers of influenza burden, as summarized in Table 4.1 [49]. Based on the number of circulating strains each year and the total number of influenza illness cases per year, we set the support for demand j to be {0,1,2,3,4,5,6,7,8,9,10} million people, and the marginal probability distribution to be $p = (0.05, 0.05, 0.1, 0.12, 0.18, 0.18, 0.12, 0.1, 0.05, 0.05)$.

Production Capacity of Old (Egg-Based) Technology c_e. The production capacity is somewhere between 130 and 150 million doses per year in the United States [51].

Production Capacity of New (Recombinant) Technology c_r. The current production capacity for recombinant vaccine is estimated to be between 10 and 20 million doses per year [126].

We summarize the parameters used in our experiments in Table 4.2. The computational experiments are run on a MacBook with 1.2 GHz Intel Core M processor and 8 GB 1600 MHz DDR3 memory, consuming about 1 second per instance.

Flu Season	All Ages	0-4 yrs	5-17 yrs	18-49 yrs	50-64 yrs	65+ yrs
2010-2011	21.3	2.8	4.4	7.4	4.8	1.8
2011-2012	9.3	0.9	2.0	3.5	1.9	1.0
2012-2013	33.7	3.6	6.7	11.4	7.9	4.2
2013-2014	29.7	2.5	4.0	13.0	8.5	1.7
2014-2015	30.2	4.7	7.3	8.6	6.4	3.2
2015-2016	25.5	2.3	4.3	10.1	7.1	1.7
2016-2017	29.9	2.3	6.0	8.5	8.5	4.4
2017-2018	48.4	4.0	7.5	14.4	15.6	7.3

Table 4.1: Disease burden in the United States: million people with influenza symptomatic illness [49].

Parameter	Value Range	Increment	Num. of Values
m	10	–	1
n	10	–	1
κ	[1,10]	1	10
$\hat{h}$	[50,500]	200	3
h	[5, 15]	10	2
$\tilde{h}$	[5, 15]	10	2
b	[630,2530]	800	3
v_{ii} matched	[0.64, 1]	0.1	4
v_{ij} mismatched	[0.32,0.6]	0.1	3
c_e	[150,150]	–	1
c_r	[0,100]	20	6

Table 4.2: Values of parameters in our numerical case study. For variation among the different i or j, we added 20% random noise. For cost and people, units are in million people, million dollars, and million dollars per million people.

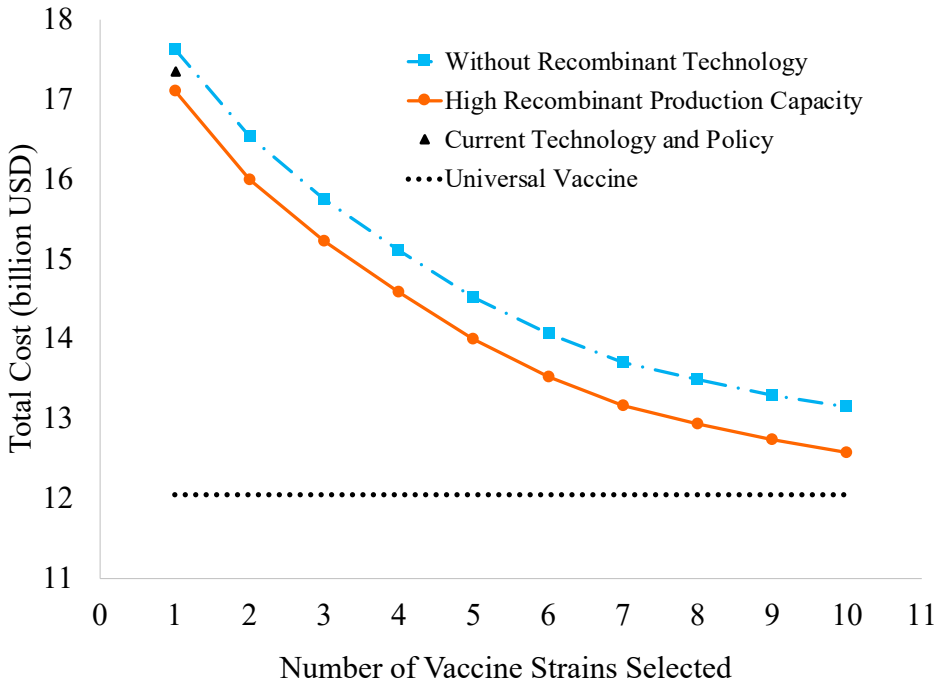

Figure 4-9: Total cost (including disease burden, vaccine selection cost, and manufacturing cost) and the total number of strains selected. High recombinant production capacity refers to $c_r = 40$ million doses per year, and current technology and policy refer to $c_r = 20$ and $\kappa = 1$. Other parameters are fixed: avg($\hat{h}_i$)=450, avg(h_i)=15, avg($\tilde{h}_i$)=5, avg(b_j)=630, avg(v_{ii})=0.64, avg(v_{ij})=0.32, and c_e=150.

We examined the experiment runs and see qualitatively similar results in all experiment runs. We present a subset of the runs to exemplify what we find. We plot such findings in Figure 4-9, and draw three observations.

First, we see that there is a reasonable cost reduction opportunity if we can improve the recombinant technology by simply increasing capacity. This can be done without much technological challenge or policy hurdle, since the recombinant technology is already in production, and selecting one vaccine strain per year (per hemisphere) is business-as-usual.

Second, if policymakers consider selecting more vaccine strains for production, the cost reduction potential is much larger than just ramping up the recombinant vaccine production alone. There are complications too: should we combine all vaccine strains in the same dose? Or should we manufacture different vaccines, each with only one strain per subtype? The former would increase manufacturing complexity and is not readily FDA-approved. The latter introduces myriads of concerns: who should receive a certain type of vaccine? Is this an opportunity to implement personalized medicine in the context of influenza vaccination and improve overall social welfare, or is it going to lead to wider inequality in healthcare?

The third observation is, universal vaccines provide significant potential cost reduction *without* introducing the complication of including multiple strains in the vaccine production and allocation process. The downside is, this technology has not gone through FDA trial, and may take years to come [104].

4.7 Extensions

We discuss briefly the potential to extend our model beyond the current setup.

Modeling Production Quality. New technologies offer the potential to manufacture vaccines with better quality consistency, due to the fact that these new vaccines do not depend on eggs. To model this, we can simply generalize the definition of supply node i from "vaccine strain i" to the ith vaccine strain and technology pair.

Modeling Heterogeneity in Demand Location. Similarly, we can expand the

definition of demand node j, from "virus type j" to the jth virus type and location pair.

Convex Cost Function. Because influenza is communicable, a natural consequence of incorporating location-specific demand is that we would have to consider the interaction between demand nodes. This can be approximated by incorporating nonlinear interaction between the demand nodes and generalize the objective function from linear to nonlinear. From the literature on epidemic simulations [10], we see that the cost function can empirically be approximated as a piecewise affine convex function. This allows us to exploit the convexity and generalize our results to a convex cost network flow question. Results such as the tractability of problem (4.1) under a Wasserstein ambiguity set would still hold.

Improving Vaccination Rate. We did not explicitly model the percentage of people getting vaccinated as a decision variable. However, significant resources are put into encouraging vaccination, especially in the healthcare workforce and those stay near high risk patients. A first step to include such decisions can be adding a decision variable in our model to indicate the flow capacity on certain subsets of arcs.

4.8 Conclusion

In this work, we model the vaccine selection and production problem under a variety of manufacturing technologies and antigenic drift. Our goal is to examine the impact of recently approved manufacturing technologies, and provide some insight on the potential cost-saving directions for policymakers and the scientific community. We provide quantitative evidence that the new technologies hold immediate potential, and can be further leveraged to improve public health conditions with more sophisticated vaccine strain selection policies. Our model is a two-stage distributionally robust optimization problem, and theoretical results can be applied to other operations settings, such as facility location and network newsvendor.

Chapter 5

Summary and Future Outlook

In this thesis, I have presented three distinct applications of optimization theory. In all three cases, the analyses result in practical tools and structural insights for the decision problems. New theories in optimization are also presented, motivated by these applications. I now explicitly tie together these insights and theories on the concept of networks, and on the concept of optimization. The unifying angle will demonstrate thematic coherence of the applications. But more interestingly, it points towards future research directions.

5.1 Paths, Flowers, and Trees

Real World Networks are Simple. These naturally occurring networks in the three applications exhibit interesting structural properties. For the case of Ford's manufacturing bill-of-materials network, I discovered from data that the network is almost cycle-free: data show that the number of nodes in cycles is on the order of 1% of the total number of nodes. Furthermore, after checking the cycles in greater details, I found that they are all collapsible, in the following sense.

Definition A cycle (V_c, E_c) is *collapsible* in a directed network (V, E), if there are nodes $i, i' \in V_c$, and $j, j' \in V \setminus V_c$, such that (i, j) and (j', i') are in E, and there are no other edges that connect V_c and $V \setminus V_c$.

Practically, these collapsible cycles arise in situations such as: a manufacturing component is used in multiple downstream parts, and these parts are used exclusively make one part in their downstream. From the modeling perspective, this essentially tells us that the network can almost be thought of as an arborescence – the connection path between any pair of upstream / downstream nodes is unique, if we collapse these cycles. In the case of designing biodefense supply chain, the inventory storage & distribution network is largely hierarchical (Figure 3-1), also a special case of arborescence.

Simple Networks and Clean Theory. Theoretically, this simple structure would enable us to generate more algorithmic and analytical tools, due to the desirable combinatorial properties of such networks. These collapsible cycles are reminiscent to what mathematician Jack Edmonds called a flower in his seminal paper on maximum matching [44]. In this thesis, I also showed that simple and efficient heuristics work well on these simple networks: time-expanded version of arborescence flow is polynomially solvable, and affine policy is optimal for two-stage robust optimization on trees.

Simplicity is Natural. It is tempting to dismiss such simplicity as contrived or artificial. Afterall, manufacturing systems are designed by people, and organizational structures are byproducts of governance. However, after examining numerous complex systems from human organizations to human immunological systems, one cannot ignore such commonality, and wonder why these systems exhibit such strong hierarchical, arborescence, tree-like structures. Perhaps the answer was already alluded to by Herbert Simon, who argued that hierarchy leads to decomposability, and decomposibility leads to robustness against disruptions [137]. This is points to an interesting direction for future discussion: what types of simplicity simultaneously lead to systemic robustness and are conducive to computationally efficient design algorithms?

5.2 Data Structure in Optimization

An optimization model comprises the sequence of decisions, objective, and constraints. These components implicitly encode the modeler's knowledge of the underlying decision problem. For example, the objective function encodes certain beliefs about the causality of a system – how it reacts to exogenous decisions imposed upon it. Consequently, optimization models are often used as *prescriptive* decision frameworks.

With the recent availability of large amounts of data, researchers and practitioners have rushed to throw data to "train" such belief models. This is a strong application of optimization tools. However, it also creates risks if haphazardly done.

Problem of Locality. The first danger of fitting regression models to data without being cognizant about the data environment is locality: data are generated from limited ranges of system states. Extrapolating the system behavior beyond certain range of states is risky. But optimization is inherently range-breaking – it explores the potential of all boundary cases of decisions, many of which may not have been present in the data generation processes.

Problem of Correlation. A problem of blindly fitting regression models is the overwhelming emphasis on correlation over causality (at least in the current research styles of machine learning and data-driven optimization). This is not an issue if one does not claim causality. However, in the case of optimization, models are used as prescriptive tools, therefore modelers are implicitly claiming causality. Yet so far there has not been much effort in data-driven optimization that properly encodes the logic of events into data modeling.

Conservativeness as Temporary Solution. In this thesis, I touched on these problems by using worst-case approaches. These approaches address uncertainty in a conservative way – even though we do not know what will happen exactly given certain actions, we want to evaluate the cost in the most conservative way. This might help us mitigate some of the discrepancy between causality and correlation modeling. But I made no attempt to quantify such error. In the best case, a worst-case approach

is a round-about solution to address the problem of correlation.

The only tools that are close to consciously incorporating causality in optimization are Markov decision process and reinforcement learning type of models – the learner interacts with the environment, learns the response of the environment, and incorporates this knowledge into its probabilistic understanding of the world. However, the current state-of-the-art theory in reinforcement learning hints that this is not an efficient solution for many cases, as the number of interactions needed to train simple tasks goes well beyond $\mathcal{O}(n^3)$, n being some characteristic size of the model. This is far from satisfactory in many decision problems for engineering, business, policy, and healthcare, because the number of allowable trial-and-errors is sometimes on the order of $\mathcal{O}(1)$, or even 0.

Structured Uncertainty Models. To provide a tool that can be truly prescriptive, optimizers need to invent new theory and computational machinery to precisely model processes and their uncertainty. This has already been fueled, partially, by the machine learning community, with its recent focus on optimal transport theory and optimization: theories from optimal transport help researchers better describe the relationship between a measure μ and a measure ν, which in turn allows us to model empirical data and potential underlying data generation processes in a more structured and computationally efficient way. This led to new progress in distributionally robust optimization theory, with new types of ambiguity sets Γ (see formulation 1.1).

Analogous to the new uncertainty modeling approach for statistical applications, I think we have much to do for complex systems. One way forward is to directly incorporate the nominal probability distribution of a set of parameters as a graphical model – a graph-based representation that emphasizes dependency structures, and sometimes causal relationships, among random variables and decision variables. We can then construct the ambiguity set Γ based on metrics that are more natural [155] to similarities between instances of graphical models. An approach like this has several potential advantages:

1. Causality of a system is naturally incorporated into the decision problem.

2. It can simultaneously incorporate data, and expert understanding of how the system works. This is especially important for areas requiring deep domain knowledge, such as drug discovery, complex engineering systems, and economic and political policy design.

3. Optimization models are "algorithmizable" by design, therefore the inclusion of structured uncertainty would mechanize many more difficult decision problems, as mentioned in the previous point.

4. It is interpretable by design.

5. It might still be susceptible to mature mathematical theories such as optimal transport.

6. If the prior knowledge is strong, a small data set is sufficient to describe the uncertainty. This relates to the long held debate in artificial intelligence regarding nature versus nurture in learning, and might shed light on how optimizers can contribute to general artificial intelligence.

Overall, a more structured approach to uncertainty modeling could bring optimization theory closer to being robust and precise tools for decision-making in complex systems.

Bibliography

[1] Mary Agócs, Shannon Fitzgerald, Steven Alles, Graham John Sale, Victor Spain, Edward Jasper, Thomas Lee Grace, and Esther Chernak. Field Testing a Head-of-Household Method to Dispense Antibiotics. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 5(3):255–267, sep 2007.

[2] Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Price of Correlations in Stochastic Optimization. *Operations Research*, 60(1):150–162, February 2012.

[3] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. Network flows: theory, algorithms, and applications. *Prentice Hall*, 1993.

[4] David L. Alderson, Gerald G. Brown, Matthew W. Carlyle, and Louis Anthony Cox. Sometimes There Is No "Most-Vital" Arc: Assessing and Improving the Operational Resilience of Systems. *Military Operations Research*, 18(1):21–37, 2013.

[5] David L. Alderson, Gerald G. Brown, and W. Matthew Carlyle. *Assessing and Improving Operational Resilience of Critical Infrastructures and Other Systems*, chapter Chapter 8, pages 180–215. INFORMS, 2014.

[6] David L Alderson, Gerald G Brown, and W Matthew Carlyle. Operational Models of Infrastructure Resilience. *Risk Analysis*, 35(4):562–586, apr 2015.

[7] David L Alderson, Gerald G Brown, W Matthew Carlyle, and R Kevin Wood. Solving Defender-Attacker-Defender Models for Infrastructure Defense. In *12th INFORMS Computing Society Conference*, pages 28–49. INFORMS, apr 2011.

[8] Christopher S. Anderson, Patrick R. McCall, Harry A. Stern, Hongmei Yang, and David J. Topham. Antigenic cartography of H1n1 influenza viruses using sequence-based antigenic distance calculation. *BMC Bioinformatics*, 19(1):51, December 2018.

[9] Amir Ardestani-Jaafari and Erick Delage. Robust Optimization of Sums of Piecewise Linear Functions with Application to Inventory Problems. *Operations Research*, 64(2):474–494, apr 2016.

[10] Nimalan Arinaminpathy, Inkyu Kevin Kim, Paul Gargiullo, Michael Haber, Ivo M. Foppa, Manoj Gambhir, and Joseph Bresee. Estimating Direct and Indirect Protective Effect of Influenza Vaccination in the United States. *American Journal of Epidemiology*, 186(1):92–100, July 2017.

[11] Assistant Secretary for Preparedness and Response. National postal model for the delivery of medical countermeasures. Technical report, Department of Health and Human Services, Washington, D.C., 2010.

[12] Alper Atamtürk and Muhong Zhang. Two-Stage Robust Network Flow and Design Under Demand Uncertainty. *Operations Research*, 55(4):662–673, aug 2007.

[13] Michael Philip Atkinson. *Mathematical models of terror interdiction*. PhD thesis, Stanford University, 2009.

[14] Erik Banks. *Catastrophic risk: Analysis and management*. Wiley Finance Series. J. Wiley and Sons, Chichester; Hoboken, NJ, 2005.

[15] Ian G. Barr, Ruben O. Donis, Jacqueline M. Katz, John W. McCauley, Takato Odagiri, Heidi Trusheim, Theodore F. Tsai, and David E. Wentworth. Cell culture-derived influenza vaccines in the severe 2017âĂŞ2018 epidemic season: a step towards improved influenza vaccine effectiveness. *npj Vaccines*, 3(1):44, December 2018.

[16] Ian G. Barr et al. Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1n1), A(H3n2) and B influenza viruses: Basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009âĂŞ2010 Northern Hemisphere season. *Vaccine*, 28(5):1156–1167, February 2010.

[17] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. Robust Optimization. *Princeton University Press*, 2009.

[18] Oded Berman and Arieh Gavious. Location of terror response facilities: A game between state and terrorist. *European Journal of Operational Research*, 177(2):1113–1133, mar 2007.

[19] Dimitris Bertsimas and Vineet Goyal. On the power and limitations of affine policies in two-stage adaptive optimization. *Mathematical Programming*, 134(2):491–531, sep 2012.

[20] Dimitris Bertsimas, Vineet Goyal, and Xu Andy Sun. A geometric characterization of the power of finite adaptability in multistage stochastic and adaptive optimization. *Mathematics of Operations Research*, 36(1):24–54, 2011.

[21] Dimitris Bertsimas, Dan A. Iancu, and Pablo A. Parrilo. Optimality of Affine Policies in Multistage Robust Optimization. *Mathematics of Operations Research*, 35(2):363–394, may 2010.

[22] Matthew Biggerstaff, Simon Cauchemez, Carrie Reed, Manoj Gambhir, and Lyn Finelli. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infectious Diseases*, 14(1):480, December 2014.

[23] Michael J. Braunscheidel and Nallan C. Suresh. The organizational antecedents of a firm's supply chain agility for risk mitigation and response. *Journal of Operations Management*, 27(2):119–140, 2009.

[24] Dena M. Bravata, Gregory S. Zaric, Jon-Erik C. Holty, Margaret L. Brandeau, Emilee R. Wilhelm, Kathryn M. McDonald, and Douglas K. Owens. Reducing Mortality from Anthrax Bioterrorism: Strategies for Stockpiling and Dispensing Medical and Pharmaceutical Supplies. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 4(3):244–262, sep 2006.

[25] Ron Brookmeyer. The statistical analysis of truncated data: application to the Sverdlovsk anthrax outbreak. *Biostatistics*, 2(2):233–247, jun 2001.

[26] Ron Brookmeyer, Elizabeth Johnson, and Sarah Barry. Modelling the incubation period of anthrax. *Statistics in Medicine*, 24(4):531–542, feb 2005.

[27] Gerald Brown, Matthew Carlyle, Javier Salmer?n, and Kevin Wood. Defending Critical Infrastructure. *Interfaces*, 36(6):530–544, dec 2006.

[28] Gerald G Brown, W Matthew Carlyle, and R Kevin Wood. Optimizing Department of Homeland Security Defense Investments : Applying Defender-Attacker (-Defender) Optimization To Terror Risk Assessment and Mitigation. In *Department of Homeland Security Bioterrorist Risk Assessment: A Call for Change*. National Academies Press, Washington, D.C., 2008.

[29] Barry C Buckland. The development and manufacture of influenza vaccines. *Human Vaccines & Immunotherapeutics*, 11(6):1357–1360, June 2015.

[30] Aakil M. Caunhye, Xiaofeng Nie, and Shaligram Pokharel. Optimization models in emergency logistics: A literature review. *Socio-Economic Planning Sciences*, 46(1):4–13, mar 2012.

[31] Melih Çelik, Özlem Ergun, Ben Johnson, Pinar Keskinocak, Álvaro Lorca, Pelin Pekgün, and Julie Swann. Humanitarian Logistics. In *2012 TutORials in Operations Research*, pages 18–49. INFORMS, oct 2012.

[32] Timothy C. Y. Chan, Zuo-Jun Max Shen, and Auyon Siddiq. Robust defibrillator deployment under cardiac arrest location uncertainty via row-and-column generation. *Operations Research*, 66(2):358–379, 2018.

[33] Louis Chen, Will Ma, Karthik Natarajan, David Simchi-Levi, and Zhenzhen Yan. Distributionally Robust Linear and Discrete Optimization with Marginals. Working paper, Massachusetts Institute of Technology, 2019.

[34] Xin Chen. L$^\natural$-Convexity and Its Applications in Operations. Working paper, University of Illinois Urbana-Champaign, 2017.

[35] Stephen E. Chick, Sameer Hasija, and Javad Nasiry. Information Elicitation and Influenza Vaccine Production. *Operations Research*, 65(1):75–96, February 2017.

[36] Stephen E. Chick, Hamed Mamani, and David Simchi-Levi. Supply Chain Coordination and Influenza Vaccination. *Operations Research*, 56(6):1493–1506, December 2008.

[37] Soo-Haeng Cho. The Optimal Composition of Influenza Vaccines Subject to Random Production Yields. *Manufacturing & Service Operations Management*, 12(2):256–277, April 2010.

[38] Sunil Chopra, Gilles Reinhardt, and Usha Mohan. The importance of decoupling recurrent and disruption risks in a supply chain. *Naval Research Logistics*, 54(5):544–555, 2007.

[39] Max D. Cooper. The early history of B cells. *Nature Reviews Immunology*, 15(3):191–197, March 2015.

[40] David L. Craft, Lawrence M. Wein, and Alexander H. Wilkins. Analyzing Bioterror Response Logistics: The Case of Anthrax. *Management Science*, 51(5):679–694, may 2005.

[41] Christopher W. Craighead, Jennifer Blackhurst, M. Johnny Rungtusanatham, and Robert B. Handfield. The severity of supply chain disruptions: Design characteristics and mitigation capabilities. *Decision Sciences*, 38(1):131–156, 2007.

[42] Tinglong Dai, Soo-Haeng Cho, and Fuqiang Zhang. Contracting for On-Time Delivery in the U.S. Influenza Vaccine Supply Chain. *Manufacturing & Service Operations Management*, 18(3):332–346, July 2016.

[43] Sarang Deo and Charles J. Corbett. Cournot Competition Under Yield Uncertainty: The Case of the U.S. Influenza Vaccine Market. *Manufacturing & Service Operations Management*, 11(4):563–576, October 2009.

[44] Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.

[45] Matthias Ehrgott. *Multicriteria Optimization*. Springer-Verlag, Berlin/Heidelberg, 2005.

[46] Martin Eichner, Markus Schwehm, Linda Eichner, and Laetitia Gerlier. Direct and indirect effects of influenza vaccination. *BMC Infectious Diseases*, 17(1):308, December 2017.

[47] Özlem Ergun, Gonca Karakus, Pinar Keskinocak, Julie Swann, and Monica Villarreal. Operations Research to Improve Disaster Supply Chain Management. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc., Hoboken, NJ, USA, jan 2011.

[48] Marshall Fisher, Janice Hammond, Walter Obermeyer, and Ananth Raman. Configuring a supply chain to reduce the cost of demand uncertainty. *Production and Operations Management*, 6(3):211–225, 1997.

[49] Centers for Disease Control and Prevention. Burden Estimates for the 2010-2011 Influenza Season.

[50] Centers for Disease Control and Prevention. Current CDC Vaccine Price List.

[51] Centers for Disease Control and Prevention. Historical Reference of Seasonal Influenza Vaccine Doses Distributed.

[52] Centers for Disease Control and Prevention. Past Seasons Vaccine Effectiveness Estimates.

[53] Centers for Disease Control and Prevention. Types of Influenza Viruses.

[54] Centers for Disease Control and Prevention. Weekly U.S. Influenza Surveillance Report.

[55] Centers for Disease Control and Prevention. How Flu Viruses Can Change, September 2017.

[56] Centers for Disease Control and Prevention. Key Facts About Seasonal Flu Vaccine, September 2018.

[57] Centers for Disease Control and Prevention. Antigenic Characterization, March 2019.

[58] Centers for Disease Control and Prevention. Disease Burden of Influenza, April 2019.

[59] Centers for Disease Control and Prevention. How Influenza (Flu) Vaccines Are Made, April 2019.

[60] R. A. M. Fouchier, V. Munster, A. Wallensten, T. M. Bestebroer, S. Herfst, D. Smith, G. F. Rimmelzwaan, B. Olsen, and A. D. M. E. Osterhaus. Characterization of a Novel Influenza A Virus Hemagglutinin Subtype (H16) Obtained from Black-Headed Gulls. *Journal of Virology*, 79(5):2814–2822, March 2005.

[61] David Gale and Themistocles Politof. Substitutes and complements in network flow problems. *Discrete Applied Mathematics*, 3(3):175–186, July 1981.

[62] Dan Gilmore. Supply chain: Risky business, 2012. accessed November 28, 2014.

[63] Boaz Golany, Edward H Kaplan, Abraham Marmur, and Uriel G Rothblum. Nature plays with dice ? terrorists do not: Allocating resources to counter strategic versus probabilistic risks. *European Journal of Operational Research*, 192(1):198–208, jan 2009.

[64] Boaz Golany, Moshe Kress, Michal Penn, and Uriel G Rothblum. Network Optimization Models for Resource Allocation in Developing Military Countermeasures. *Operations Research*, 60(1):48–63, feb 2012.

[65] Stephen C Graves and Sean P Willems. Supply chain design: Safety stock placement and supply chain configuration. In A.G. de Kok and Stephen C Graves, editors, *Handbooks in Operations Research and Management Science*, volume 11, chapter 3, pages 95–132. Elsevier, Amsterdam, 2003.

[66] Tatiana SchÃďffer Gregianini, Ivana R Santos Varella, Patricia Fisch, LetÃŋcia Garay Martins, and Ana B G Veiga. Dual and Triple Infections With Influenza A and B Viruses: A Case-Control Study in Southern Brazil. *The Journal of Infectious Diseases*, page jiz221, May 2019.

[67] Oleg Gusikhin and Erica Klampfl. JEDI: Just-in-Time Execution and Distribution Information Support System for Automotive Stamping Operations. In Dieter Armbruster and Karl G. Kempf, editors, *Decision Policies for Production Networks*, pages 119–142. Springer-Verlag, London, 2012.

[68] James Guyton, Robert Kadlec, Chandresh Harjivan, Shabana Farooqi, Sheana Cavitt, and Joseph Buccina. A Cost and Speed Analysis of Strategies for Prepositioning Antibiotics for Anthrax. *Commissioned paper by the Institute of Medicine*, 2011.

[69] Mengran Hao, Shilan Jin, and Jun Zhuang. Robustness of Optimal Defensive Resource Allocations in the Face of Less Fully Rational Attacker. Technical report, Homeland Security Center, 2009.

[70] Kevin B. Hendricks and Vinod R. Singhal. Association between supply chain glitches and operating performance. *Management Science*, 51(5):695–711, 2005.

[71] Nathaniel Hupert, Wei Xiong, Kathleen King, Michelle Castorena, Caitlin Hawkins, Cindie Wu, and John A. Muckstadt. Uncertainty and Operational Considerations in Mass Prophylaxis Workforce Planning. *Disaster Medicine and Public Health Preparedness*, 3(S2):S121–S131, dec 2009.

[72] Dan A Iancu, Maxim Sviridenko, Mayank Sharma, and Maxim Sviridenko. Supermodularity and Affine Policies in Dynamic Robust Optimization. *Operations Research*, 61(4):941–956, aug 2013.

[73] Thomas Inglesby and Barbara Ellis. Division of Strategic National Stockpile (DSNS) Program Review: A Report from the Board of Scientific Counselors (BSC). Technical report, Center for Disease Control and Prevention, Atlanta, 2012.

[74] European Bioinformatics Institute. Hemagglutinin (PF00509).

[75] A Danielle Iuliano et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet*, 391(10127):1285–1300, March 2018.

[76] Eric J Johnson, John Hershey, Jacqeline Meszaros, and H Kunreuther. Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty*, 7(1):35–51, 1993.

[77] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263, March 1979.

[78] Jocelyn Kaiser. Taking Stock of the Biodefense Boom. *Science*, 333(6047):1214–1214, sep 2011.

[79] Edward H. Kaplan. OR Forum - Intelligence Operations Research: The 2010 Philip McCord Morse Lecture. *Operations Research*, 60(6):1297–1309, dec 2012.

[80] Kathleen King. Logistical Models For Planning And Operating Medical Countermeasure Distribution Networks During Public Health Emergencies. *PhD dissertation, Cornell University*, 2012.

[81] Kathleen King and John A. Muckstadt. Evaluating Planned Capacities for Public Health Emergency Supply Chain Models. Technical report, Cornell University, 2009.

[82] P. R. Kleindorfer and G. H. Saad. Managing disruption risks in supply chains. *Production and Operations Management*, 14(1):53–68, 2005.

[83] A Michael Knemeyer, Walter Zinn, and Cuneyt Eroglu. Proactive planning for catastrophic events in supply chains. *Journal of Operations Management*, 27(2):141–153, 2009.

[84] Laura J. Kornish and Ralph L. Keeney. Repeated Commit-or-Defer Decisions with a Deadline: The Influenza Vaccine Composition. *Operations Research*, 56(3):527–541, June 2008.

[85] Howard Kunreuther. Limited knowledge and insurance protection. *Public Policy*, 24(2):227–261, 1976.

[86] Linda C. Lambert and Anthony S. Fauci. Influenza Vaccines for the Future. *New England Journal of Medicine*, 363(21):2036–2044, November 2010.

[87] Alan Lapedes and Robert Farber. The Geometry of Shape Space: Application to Influenza. *Journal of Theoretical Biology*, 212(1):57–69, September 2001.

[88] Gary Lazzaro. *Tri-level optimization algorithms for solving defender-attacker-defender network models*. Phd dissertation, Naval Postgraduate School, 2016.

[89] Eva K. Lee, Chien-Hung Chen, Ferdinand Pietz, and Bernard Benecke. Modeling and Optimizing the Public-Health Infrastructure for Emergency Response. *Interfaces*, 39(5):476–490, oct 2009.

[90] Eva K. Lee, Siddhartha Maheshwary, Jacquelyn Mason, and William Glisson. Decision support system for mass dispensing of medications for infectious disease outbreaks and bioterrorist attacks. *Annals of Operations Research*, 148(1):25–53, nov 2006.

[91] Eva K. Lee, Siddhartha Maheshwary, Jacquelyn Mason, and William Glisson. Large-Scale Dispensing for Emergency Response to Bioterrorism and Infectious-Disease Outbreak. *Interfaces*, 36(6):591–607, dec 2006.

[92] Eva K. Lee, Helder I. Nakaya, Fan Yuan, Troy D. Querec, Greg Burel, Ferdinand H. Pietz, Bernard A. Benecke, and Bali Pulendran. Machine Learning for Predicting Vaccine Immunogenicity. *Interfaces*, 46(5):368–390, October 2016.

[93] Eva K. Lee, Hannah K. Smalley, Yang Zhang, Ferdinand Pietz, and Bernard Benecke. Facility location and multi-modality mass dispensing strategies and emergency response for biodefence and infectious disease outbreaks. *International Journal of Risk Assessment and Management*, 12(2/3/4):311, 2009.

[94] Cameron A. MacKenzie, Kash Barker, and Joost R. Santos. Modeling a severe supply chain disruption and post-disaster decision making with application to the japanese earthquake and tsunami. *IIE Transactions*, 46(12):1243–1260, 2014.

[95] Ho-Yin Mak and Zuo-Jun Shen. Risk diversification and risk pooling in supply chain design. *IIE Transactions*, 44(8):603–621, 2012.

[96] Keiichi Makizumi, Kazuhiko Kimachi, Katsuhiko Fukada, Tomohiro Nishimura, Yasuhiro Kudo, Shuro Goto, Takato Odagiri, Masato Tashiro, and Yoichiro Kino. Timely production of A/Fujian-like influenza vaccine matching the 2003âĂŞ2004 epidemic strain may have been possible using MadinâĂŞDarby canine kidney cells. *Vaccine*, 26(52):6852–6858, December 2008.

[97] James G March and Zur Shapira. Managerial perspectives on risk and risk taking. *Management Science*, 33(11):1404–1418, 1987.

[98] D. Miklovic and R. Witty. Case study: Cisco addresses supply chain risk management. Document of the presentation at Gartner Security and Risk Management Summit 2010 (ID Number: G00206060), September 2010.

[99] Ian I Mitroff and Murat C Alpaslan. Preparing for evil. *Harvard Business Review*, 81(4):109–15, 124, April 2003.

[100] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, September 2018.

[101] Kazuo Murota and Akiyoshi Shioura. Substitutes and complements in network flows viewed as discrete convexity. *Discrete Optimization*, 2(3):256–268, September 2005.

[102] National Research Council. *Department of Homeland Security Bioterrorism Risk Assessment: A Call for Change*. The National Academies Press, Washington DC, 2008.

[103] Anna Nicholson, Scott Wollek, Benjamin Kahn, and Jack Hermann. The Nation's Medical Countermeasure Stockpile. *National Academies Press*, 2016.

[104] National Institutes of Health. NIH begins first-in-human trial of a universal influenza vaccine candidate | National Institutes of Health (NIH), April 2019.

[105] Office of Public Health Preparedness and Response. Board of Scientific Counselors, 2016.

[106] Office of Public Health Preparedness and Response. Public Health Preparedness Capabilities: National Standards for State and Local Planning, 2016.

[107] Office of Public Health Preparedness and Response. Strategic National Stockpile, 2016.

[108] Suzanne E. Ohmit, Joshua G. Petrie, Rachel T. Cross, Emileigh Johnson, and Arnold S. Monto. Influenza Hemagglutination-Inhibition Antibody Titer as a Correlate of Vaccine-Induced Protection. *The Journal of Infectious Diseases*, 204(12):1879–1885, December 2011.

[109] Subcommittee on Oversight and Investigations. Examining U.S. Public Health Preparedness for and Response Efforts to Seasonal Influenza, March 2018.

[110] OpenStax. *Anatomy and Physiology*. Rice University, March 2013.

[111] World Health Organization. FluNet.

[112] World Health Organization. Global Influenza Surveillance and Response System (GISRS).

[113] World Health Organization. WHO recommendations on the composition of influenza virus vaccines.

[114] Osman Y. Özaltın, Oleg A. Prokopyev, and Andrew J. Schaefer. Optimal Design of the Seasonal Influenza Vaccine with Manufacturing Autonomy. *INFORMS Journal on Computing*, 30(2):371–387, May 2018.

[115] Osman Y. Özaltın, Oleg A. Prokopyev, Andrew J. Schaefer, and Mark S. Roberts. Optimizing the Societal Benefits of the Annual Influenza Vaccine: A Stochastic Programming Approach. *Operations Research*, 59(5):1131–1143, October 2011.

[116] Peng Peng, Lawrence V Snyder, Andrew Lim, and Zuli Liu. Reliable logistics networks design with facility disruptions. *Transportation Research*, 45(8):1190–1211, 2011.

[117] Alan S. Perelson and George F. Oster. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of Theoretical Biology*, 81(4):645–670, December 1979.

[118] Charles Perrow. *Normal accidents: Living with high risk technologies*. Princeton University Press, Princeton, NJ, 2011.

[119] Stanley Plotkin, James M. Robinson, Gerard Cunningham, Robyn Iqbal, and Shannon Larsen. The complexity and cost of vaccine manufacturing âĂŞ An overview. *Vaccine*, 35(33):4064–4071, July 2017.

[120] Wayan C.W.S. Putri, David J. Muscatello, Melissa S. Stockwell, and Anthony T. Newall. Economic burden of seasonal influenza in the United States. *Vaccine*, 36(27):3960–3966, June 2018.

[121] S. T. Rachev and Ludger Rüschendorf. *Mass transportation problems: theory*, volume 1 of *Probability and its applications*. Springer, New York, 1998.

[122] Jesus Rios and David Rios Insua. Adversarial Risk Analysis for Counterterrorism Modeling. *Risk Analysis*, 32(5):894–915, may 2012.

[123] Melissa A Rolfes et al. Effects of Influenza Vaccination in the United States During the 2017âĂŞ2018 Influenza Season. *Clinical Infectious Diseases*, February 2019.

[124] Allan R. Sampson and Robert L. Smith. Assessing risks through the determination of rare event probabilities. *Operations Research*, 30(5):839, 1982.

[125] M. R. Sandbulte, K. B. Westgeest, J. Gao, X. Xu, A. I. Klimov, C. A. Russell, D. F. Burke, D. J. Smith, R. A. M. Fouchier, and M. C. Eichelberger. Discordant antigenic drift of neuraminidase and hemagglutinin in H1n1 and H3n2 influenza viruses. *Proceedings of the National Academy of Sciences*, 108(51):20748–20753, December 2011.

[126] Sanofi. Sanofi Ships First Flu Vaccines for 2018-2019 Season, July 2018.

[127] William Schmidt and David Simchi-Levi. Nissan Motor Company: Building Operational Resiliency, 2013. accessed March 2, 2015.

[128] Amanda J. Schmitt. Strategies for customer service level protection under multi-echelon supply chain disruption risk. *Transportation Research Part B: Methodological*, 45(8), 2011.

[129] Yosef Sheffi. *The resilient enterprise: Overcoming vulnerability for competitive advantage*. MIT Press, Cambridge, MA, 2005.

[130] David Simchi-Levi and Michael Sanders. New approach to supplier risk management at ford, 2013. [Video file] accessed March 15, 2015.

[131] David Simchi-Levi, William Schmidt, and Yehua Wei. From superstorms to factory fires: managing unpredictable supply-chain disruptions. *Harvard Business Review*, 92(1-2):96, 2014.

[132] David Simchi-Levi, William Schmidt, and Yehua Wei. From superstorms to factory fires: Managing unpredictable supply-chain disruptions. *Harvard Business Review*, 92(1/2):96–101, 2014.

[133] David Simchi-Levi, William Schmidt, Yehua Wei, Peter Yun Zhang, Keith Combs, Yao Ge, Oleg Gusikhin, Michael Sanders, and Don Zhang. Identifying Risks and Mitigating Disruptions in the Automotive Supply Chain. *Interfaces*, 45(5):375–390, oct 2015.

[134] David Simchi-Levi, He Wang, and Yehua Wei. Increasing supply chain robustness through process flexibility and strategic inventory. Working paper, Massachusetts Institute of Technology, 2013.

[135] David Simchi-Levi, He Wang, and Yehua Wei. Increasing Supply Chain Robustness through Process Flexibility and Inventory. *Working paper, MIT*, 2016.

[136] David Simchi-Levi, He Wang, and Yehua Wei. Constraint Generation for Two-Stage Robust Network Flow Problem. *INFORMS Journal on Optimization, forthcoming*, 2018.

[137] Herbert Alexander Simon. *The sciences of the artificial*. MIT Press, 3. ed., [nachdr.] edition, 2008. OCLC: 552080160.

[138] Danuta M. Skowronski, Naveed Z. Janjua, Gaston De Serres, Suzana Sabaiduc, Alireza Eshaghi, James A. Dickinson, Kevin Fonseca, Anne-Luise Winter, Jonathan B. Gubbay, Mel Krajden, Martin Petric, Hugues Charest, Nathalie Bastien, Trijntje L. Kwindt, Salaheddin M. Mahmud, Paul Van Caeseele, and Yan Li. Low 2012âĂŞ13 Influenza Vaccine Effectiveness Associated with Mutation in the Egg-Adapted H3n2 Vaccine Strain Not Antigenic Drift in Circulating Viruses. *PLoS ONE*, 9(3):e92153, March 2014.

[139] D. J. Smith. Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science*, 305(5682):371–376, July 2004.

[140] Lawrence V Snyder, Maria P Scaparra, Mark S Daskin, and Richard L Church. Planning for disruptions in supply chain networks. *Tutorials in Operations Research*, 2006.

[141] ManMohan S. Sodhi, Byung-Gak Son, and Christopher S. Tang. Researchers' perspectives on supply chain risk management. *Production and Operations Management*, 21(1):1–13, 2012.

[142] Clare Stroud, Kristin Viswanathan, Tia Powell, and Robert R. Bass, editors. *Prepositioning Antibiotics for Anthrax*. The National Academies Press, Washington, DC, 2012.

[143] Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, Cambridge, Mass., 2007.

[144] Christopher Tang. Robust strategies for mitigating supply chain disruptions. *International Journal of Logistics*, 9(1):33–45, March 2006.

[145] Jeffery K. Taubenberger and David M. Morens. 1918 Influenza: the Mother of All Pandemics. *Emerging Infectious Diseases*, 12(1):15–22, January 2006.

[146] Anna Teytelman and Richard C. Larson. Multiregional Dynamic Vaccine Allocation During an Influenza Epidemic. *Service Science*, 5(3):197–215, September 2013.

[147] Shara Tibken. Intel cuts its outlook – Giant blames Thai flood for $1 billion drop in sales goal, 2011. accessed March 2, 2015.

[148] V. Tisa, I. Barberis, V. Faccio, C. Paganino, C. Trucchi, M. Martini, and F. Ansaldi. Quadrivalent influenza vaccine: a new opportunity to reduce the influenza burden. *Journal of Preventive Medicine and Hygiene*, 57(1):E28–33, 2016.

[149] Brian Tomlin. On the value of mitigation and contingency strategies for managing supply chain disruption risks. *Management Science*, 52(5):639–657, May 2006.

[150] Brian Tomlin. Impact of supply learning when suppliers are unreliable. *Manufacturing & Service Operations Management*, 11(2):192–209, 2009.

[151] Andrea C Tricco, Ayman Chit, Charlene Soobiah, David Hallett, Genevieve Meier, Maggie H Chen, Mariam Tashkandi, Chris T Bauch, and Mark Loeb. Comparing influenza vaccine efficacy against mismatched and matched strains: a systematic review and meta-analysis. *BMC Medicine*, 11(1):153, December 2013.

[152] Joline Uichanco. A robust model for pre-positioning emergency relief items before a typhoon with an uncertain trajectory. *Working paper, University of Michigan*, 2015.

[153] United States Census Bureau. Population, Housing Units, Area, and Density: 2000, 2000.

[154] U.S. Food and Drug Administration. Anthrax, 2015.

[155] Titouan Vayer, Laetitia Chapel, Remi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6275–6284, 2019.

[156] Stephan M. Wagner and Christoph Bode. An empirical investigation into supply chain vulnerability. *Journal of Purchasing and Supply Management*, 12(6):301–312, 2006.

[157] L. M. Wein, D. L. Craft, and E. H. Kaplan. Emergency response to an anthrax attack. *Proceedings of the National Academy of Sciences*, 100(7):4346–4351, apr 2003.

[158] D. A. Wilkening. Sverdlovsk revisited: Modeling human inhalation anthrax. *Proceedings of the National Academy of Sciences*, 103(20):7589–7594, may 2006.

[159] D. A. Wilkening. Modeling the Incubation Period of Inhalational Anthrax. *Medical Decision Making*, 28(4):593–605, jul 2008.

[160] World Economic Forum. Building resilience in supply chains, 2013. accessed March 2, 2015.

[161] Joseph T. Wu, Lawrence M. Wein, and Alan S. Perelson. Optimization of Influenza Vaccine Selection. *Operations Research*, 53(3):456–476, June 2005.

[162] Gregory S. Zaric, Dena M. Bravata, Jon-Erick Cleophas Holty, Kathryn M. McDonald, and Margaret L. Brandeau. Modeling the Logistics of Response to Anthrax Bioterrorism. *Medical Decision Making*, 28(3):332–350, 2008.

[163] Jun Zhuang and Vicki M. Bier. Balancing Terrorism and Natural Disasters: Defensive Strategy with Endogenous Attacker Effort. *Operations Research*, 55(5):976–991, oct 2007.