View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by SHAREOK repository

PROACTIVE APPROACHES FOR SYSTEM DESIGN UNDER UNCERTAINTY APPLIED TO NETWORK SYNTHESIS AND CAPACITY PLANNING

By

JUAN MA
Bachelor of Engineering in Logistics Systems
Engineering
Huazhong University of Science and Technology
Wuhan, Hubei, China
2010

Submitted to the Faculty of the
Graduate College of
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2015

COPYRIGHT ©

By

JUAN MA

December, 2015

PROACTIVE APPROACHES FOR SYSTEM DESIGN UNDER UNCERTAINTY APPLIED TO NETWORK SYNTHESIS AND CAPACITY PLANNING

Dissertation Approved:

Dr. Balabhaskar Balasundaram

Dissertation co-advisor

Dr. Manjunath Kamath

Dissertation co-advisor

Dr. Ying Tat Leung

Dr. Tieming Liu

Dr. Goutam Chakraborty

Name: JUAN MA

Date of Degree: DECEMBER, 2015

Title of Study: PROACTIVE APPROACHES FOR SYSTEM DESIGN UNDER UNCERTAINTY APPLIED TO NETWORK SYNTHESIS AND CAPACITY PLANNING

Major Field: INDUSTRIAL ENGINEERING AND MANAGEMENT

Abstract:

The need to design systems under uncertainty arises frequently in applications such as telecommunication network configuration, airline hub-and-spoke/inter-hub network design, power grid design, transportation system design, call center staffing, and distribution center design. Such problems are very challenging because: (1) design problems with sophisticated configuration requirements for medium to large scale systems often yield large-sized linear/nonlinear mathematical models with both continuous and discrete decision variables, and (2) in most cases input parameters such as demand arrival rates are subject to uncertainty, whereas engineers have to make a design decision "today," before the outcomes of the uncertain parameters can be observed. The purpose of this study was to develop proactive modeling methodologies and effective solution techniques for such system design problems. Particular emphasis was placed on a network design problem with connectivity and diameter requirements under probabilistic edge failures and a service system capacity planning problem under uncertain demand rates.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

The need to design systems under uncertainty arises frequently in applications such as telecommunication network configuration, airline hub-and-spoke/inter-hub network design, power grid design, transportation system design, call center staffing, and distribution center design. System design under uncertainty, therefore, constitutes a significant and challenging research area of operations research and management science. On the one hand, certain system design problems with sophisticated configuration requirements for medium to large scale systems often yield large-sized linear/nonlinear mathematical models with both continuous and discrete decision variables. The resulting formulations are thus very challenging to solve and require advanced optimization techniques. On the other hand, in most cases if not all, input parameters such as demand arrival rates are subject to uncertainty. Yet, in those cases, system engineers have to make a design decision "today," before the outcomes of the uncertain parameters can be observed. Illustrative examples of system design under uncertainty of this kind are introduced next.

Consider a battlefield wireless communication network design problem. It is indisputable that ensuring effective communication is vital for the command and control of ground forces in any battlefield. Strategic level communication is often achieved via certain established wireless communication networks. The topological features of the underlying wireless networks that enable reliable and fast information exchange include: 1) redundant disjoint data delivery channels, and 2) fewer intermediary nodes and links. Ensuring a certain level of communication reliability by maintaining redundant communication channels is extremely important due to changeable mission contexts and targeted attacks during which network infrastructure might easily malfunction or fail. Some examples of the wireless network infrastructure are radios, gateways, line-of-sight and beyond-line-of-sight communication links (Kerivin and Mahjoub, 2005). Although factors such as battlefield terrain and weather could also affect communication reliability at the operational level, the design of the underlying telecommunication network topology is critical in the sense that it can limit or enable efficient and reliable communication at the strategic level. In addition, the feature of fewer intermediary nodes and links is especially significant for wireless networks where a short communication distance is desirable to limit possible signal loss between transmitters. Now in this scenario, a design decision has to be made simultaneously satisfying certain reliability and efficiency requirements, often also under a given budget constraint. Due to the adversarial nature of the environments, parameters such as the survival of a gateway are uncertain until a mission ends and personnel are available to check the equipment statuses. Therefore an appropriate design decision has to be made before uncertainty reveals itself.

Another example of system design under uncertainty is the capacity planning problem of a facility where customers/transactions arrive to be served. Each service encounter may require the service provider to perform one or more tasks using one or more resources. After the entire service encounter is completed, the customer/transaction leaves the facility. This situation is very common in many service industries such as a distribution center or warehouse, a customer service center (either walk-in or call-in), a repair shop, and a healthcare provider. In planning to establish or renovate such a facility, one needs to decide the service capacity, which requires a one-time fixed investment and preparation time. This capacity decision often has to be made according to the future needs which are often not known in the design phase. Examples include the number of checkout lanes in a retail store, the number of service bays in an automobile repair facility, number of beds in a hospital emergency room, and number of workstations in a customer service center. Due to uncertainty, the capacity decision often has to be obtained by solving a stochastic or robust optimization model to balance the risks of poor system performance and the total system configuration cost over a finite horizon, taking into consideration the fixed but possibly multiple-time real-estate or equipment costs starting from the beginning of the horizon. Other issues may also complicate this problem in a service system where customers are present for co-production, a key one being the appearance of a sparse business volume impacting future demand negatively. An example of this is education (e.g., a class) where a student has to be present (could be remotely) and has to work, together with the teacher, in order to learn. Because of this characteristic, the customer will see the facility and possibly other customers, in which case the appearance of a few customers present may make a negative impression on the customer, undermining customer retention. As a result, at some low demand point, the system may go into a negative feedback loop and the demand may die off because of this effect. In this situation, unlike the third-party logistics provider case, having extra capacity at the beginning may not be all good, even if capacity is affordable.

1.1 Modeling Under Uncertainty

Approaches tackling system design under uncertainty can be broadly classified into two categories: *reactive approaches* and *proactive approaches*. In the reactive approach, an uncertain parameter is usually substituted by its mean or mode. One then proceeds with solving the deterministic design optimization problem, followed by a sensitivity analysis to see how the solution will be impacted by different realizations of the uncertain parameter. If the solution turns out to be rather insensitive to changes to the uncertain parameter, then it would be an adequate solution. If the solution is indeed sensitive to the uncertain parameter, one has to be careful in choosing the final solution. For example, if two solutions result in similar total costs, how sensitive the solution is to the uncertain parameter can be used as a tie breaker. In practice, one is usually faced with a tradeoff between the total costs and the sensitivity of them to the uncertain parameter; this is often resolved by an intuitive and subjective decision, likely based on risk guidelines or preferences of the user.

Although replacing an uncertain parameter with its mean or mode is simple and easy, such an approach is rooted in the assumptions that: 1) the mean or mode is very likely to be observed as an outcome, and 2) the variability in the realized outcomes is not significant. In cases where these two assumptions are violated, the approach is no longer satisfactory. Follow-up sensitivity analysis in the reactive approach may help in some cases where the assumptions are violated, but not when the outcomes of the uncertain parameter can only be observed after the design decision has been made and executed. After all, with the design decision having been executed, the decision maker may only have few opportunities to change the response or performance of the designed system.

Proactive approaches generally employ optimization models using *risk measures* to handle uncertainty. The concept of risk measures is more prevalent in financial risk management. Yet, it can also be used to capture similar ideas in engineering settings to handle uncertainty. A risk measure is basically a function that maps the uncertain parameters modeled as random variables defined on an appropriate probability space onto a scalar which quantifies the magnitude of risk. Examples of risk measures are mean, mean-variance, failure probability, value-at-risk, conditional-value-at-risk, and worst case (Rockafellar and Uryasev, 2013). Different risk measures have different statistical and computational features. The choice of a risk measure generally depends on one's risk preference (i.e., risk-seeking, risk-neutral, or risk-averse) and problem-specific characteristics such as whether the impact of the worst case outcome is overwhelming and whether the impact of design requirement violation is quantifiable.

While we are broadly interested in system design under uncertainty using proactive approaches, this dissertation focuses on two specific problems: *the network design problem with connectivity ("reliability") and diameter ("efficiency") requirements under probabilistic edge failures*, and *service system capacity planning problem under uncertain demand rate*. We focus on single-stage decision models in this research, which are appropriate given the nature of our decisions (i.e., strategic, rather than tactical or operational).

1.2 Network Design Under Topological Uncertainty

Here we seek a network design satisfying given connectivity and diameter requirements under probabilistic edge failures indirectly by using the notion of *k-core*. Given a positive integer k, a k-core is a graph in which each vertex has at least k neighbors. The concept of k-core was introduced in social network analysis to identify denser regions of a social network (Seidman, 1983). They have since been employed to retrieve useful information from complex networks. For instance, in bioinformatics, the k-core structures have been employed to investigate protein interaction networks (Altaf-Ul-Amine et al., 2003; Wuchty and Almaas, 2005) to predict protein functions and to identify inherent layered structures.

On the other hand, the potential applications of k-cores in the domain of network design have been overlooked by researchers. In this dissertation, we exploit the graph-theoretic properties of k-cores to introduce a new approach to hop-constrained survivable network design via spanning k-cores that preserves connectivity and diameter under limited edge failures. In particular, high connectivity and low diameter are desirable features for wireless communication networks and airline networks. Consider airline networks as an example. When an airport cannot function normally due to severe weather, it is extremely important that an airline carrier can provide alternative connecting flights with a limited number of stops to maintain a high customer satisfaction rate and thus market share. Therefore, our k-core approach can be applied in such domains where network elements are prone to failure due to various reasons like severe weather and targeted attack.

Wuellner et al. (2010) conducted an insightful analysis on the structures of the seven largest airline passenger carriers (namely, Southwest, American Airlines, Delta, United Airlines, Northwest, US Airways, and Continental). The authors collected public data from the US Department of Transportation and Bureau of Transportation Statistics and investigated the resilience of resulting networks after random or targeted vertex (airport) and/or edge (flight) deletion. They discovered that Southwest airlines network, a high minimum degree structure, was extremely resilient to both targeted removal of airports and random deletion of flights. Compared to others, it incurred minimum travel cost increase when airports/flights were deleted in the experiments conducted. The authors further conclude that although hub-and-spoke structure is more popular than point-to-point structure (or similar degree-regular structures such as k-core for large k values) in airline industry due to its economical advantages, the latter is a better option in applications where the requirement of resilience is emphasized.

1.3 Service System Capacity Planning Under Demand Uncertainty

Often, we seek a capacity decision which balances the cost invested in provisioning the system capacity against the benefit of satisfying the system performance requirements specified in service level agreement with clients. Such a decision is made based on information on parameters such as transaction arrival rate, service level required by the client, and fixed and variable cost rates. In most cases, some input parameters such as customer/transaction arrival rate are uncertain at design time and values used are estimates subject to errors, making it challenging for engineers to make a capacity decision.

In the domain of business service outsourcing, inadequate service capacity often incurs an immediate and direct financial consequence to the service provider. For instance, consider a third-party logistics provider who provides a warehousing and customer order fulfillment service to their clients. Throughout this dissertation, we use the term “client” to represent a business that contracts the service provider and the term “customer” to denote the individual consumers or orders arriving at the service facility contracted by the client. Suppose the client requires an incoming customer order for its goods to be shipped within 24 hours of order receipt on the average. At the end of each accounting cycle, the logistics provider has to report statistics on the customer order handling times for all orders received that cycle. In the case that order handling requirement is violated, the service provider may have to pay a financial penalty to its client. Other similar situations include customer service centers which can be walk-in facilities, or more commonly nowadays, call centers. In these applications, a common system performance measure is how long an incoming customer has to wait before being served by an agent. Typically, key performance measures of an operation and their target values are specified in the service level agreement (SLA) of an outsourcing relationship.

To be more specific, for an outcome of the uncertain parameter and a capacity decision, one can define a corresponding penalty value depending on the extent of violation of given requirements in the SLA. After all, ensuring performance requirements being satisfied in all possible outcomes is economically prohibitive and even physically impractical in many cases. By defining penalties, one can discourage the violation of system performance requirements by quantitatively limiting the value of a risk measure defined over the penalties across different outcomes. Different approaches to define and evaluate penalty functions have been discussed by Kosinski et al. (2008).

1.4 Organization

The remainder of this dissertation is structured as follows. Chapter 2 presents an extensive background for our work including an introduction to graph notations and definitions, the resilient network design problem, graph-theoretic properties of k-cores, the deterministic minimum spanning k-core problem, and different risk measures employed in engineering decision making along with their characteristics. In Chapter 3, we present a literature review of the generic network design problem, the survivable network design problem, the hop-constrained survivable network design problem, and available models for the service system design problem. Chapter 4 states the research objectives addressed in this dissertation.

Chapter 5 centers on obtaining network designs satisfying given connectivity and diameter requirements via k-cores under a conditional value-at-risk (CVaR) constraint. We specifically introduce two CVaR-constrained optimization models to obtain risk-averse solutions for the minimum spanning k-core problem under probabilistic edge failures. We present polyhedral reformulations of the convex piecewise linear loss functions used in these models that enable Benders-like decomposition approaches. A decomposition and branch-and-cut approach is then developed to solve the scenario-based approximation of the CVaR-constrained minimum spanning k-core problem for the aforementioned loss functions. The computational performance of the algorithm is studied in detail.

Chapter 6 focuses on a chance-constrained version of the spanning k-core problem under probabilistic edge failures. We first establish the intractability of this problem by proving that it is NP-hard. Subsequently, we conduct a polyhedral study investigating facet-inducing inequalities and then develop a strengthened formulation. The effectiveness of the strengthened formulation is demonstrated through a computational study.

In Chapter 7, we first present mathematical models for the capacity planning of single- and two-stage service systems that can be represented by queueing models when there is uncertainty in a key parameter such as the external arrival or demand rate. Analytical solutions are derived for single-stage systems modeled by the standard M/M/1 queue under the assumption that arrival rate is varied uniformly over a specified range. For two-stage service systems that pose bigger modeling and computational challenges, we provide scenario-based numerical methods (i.e., search-based and mathematical reformulation) to solve the capacity planning problem of a tandem configuration with two stages and a single external arrival process. The dissertation is concluded in Chapter 8 with a discussion of the results and future directions for research.

Some materials in Chapter 3 have appeared in (Ma et al., 2014) and most materials in Chapter 5 have been accepted for publication at the time of this writing (Ma et al., 2015). Materials in Chapter 6 have been submitted for journal publication at the same time. All the figures in this dissertation were generated using PowerPoint®, Matlab®, and Mathematica®.

CHAPTER 2

BACKGROUND

We first present a brief introduction to the relevant graph theory terminology in this chapter. Next, we review the concepts of the generic network design problem, the survivable network design problem, and the resilient network design problem. We illustrate the differences and highlight the relevance of the resilient network design problem. Also discussed are the theoretical properties of k-cores in terms of connectivity and diameter, and the formulation and tractability of the minimum spanning k-core problem in the deterministic setting. Finally, different risk measures that are commonly employed in operations research applications are explained.

2.1 Notations and Definitions

This subsection presents a concise introduction to most of the terminology used later in this document. Most of the terms used in this document are consistent with those described in well-known graph theory books (see Diestel, 1997; West, 2001), and are thus easy to understand. The few that may be better understood within the proper context will be introduced later.

Let $G = (V, E)$ denote a simple undirected graph with V being the vertex set and E being the set of edges. We say graph G is complete if it contains all possible edges. We denote the subgraph induced by $S \subseteq V$ as $G[S]$ and $G[S] = (S, E \cap (S \times S))$. In other words, graph $G[S]$ consists of vertex set S and all edges in E whose endpoints are both in S. A vertex set $S \subseteq V$ is a *clique* if $G[S]$ is complete.

If $(v, \hat{v}) \in E$, we say that edge $(v, \hat{v})$ is incident at vertex v (and at $\hat{v}$). Let $\gamma(v)$

represent the cut of vertex v, the set of edges incident at v. The cardinality of set $\gamma(v)$, denoted by $d(v)$, is called the degree of vertex v in G. The minimum degree of a vertex in G is denoted by $\delta(G)$; $\delta(G) = \min\{|\gamma(v)| \ : \ v \in V\}$. A graph is k-regular if $|\gamma(v)| = k$ for all $v \in V$. The density of graph G, denoted by $\rho(G)$ is the ratio of $|E|$ to the number of all possible edges (i.e. $\frac{|V|(|V|-1)}{2}$).

The *diameter* of graph G, denoted as $diam(G)$, is defined as the maximum over shortest path lengths between every pair of vertices in G. Note that the length of a path is defined as the number of edges on this path. By $\kappa(G)$ we denote the *vertex connectivity* of graph G which is defined as the minimum number of vertices whose deletion results in a disconnected or single-vertex graph. A graph is k-connected if deleting any k vertices in V does not result in a disconnected or single-vertex graph. Similarly, the *edge connectivity* denoted by $\lambda(G)$ is defined as the minimum number of edges whose removal disconnects the graph. A graph is k-edge-connected if the resulting graph stays connected after removing any k edges in E. It is worth noting that removing a vertex implies removing this vertex together with all its incident edges, while removing an edge refers to removing only this edge. In particular, for any non-trivial graph G, we have (Whitney, 1932):

$$\kappa(G) \leq \lambda(G) \leq \delta(G).$$

Therefore, a high vertex connectivity implies a high edge connectivity which in turn implies a large minimum degree, but not vice versa.

Definition 1 *A graph G is a k-*core *if the degree of every vertex in G is at least k.*

The k-core concept was introduced in social network analysis by Seidman (1983) to identify regions of the social network containing cohesive subgroups. Graphical instances of a clique and a 2-core with 5 vertices are presented in Figure 1.

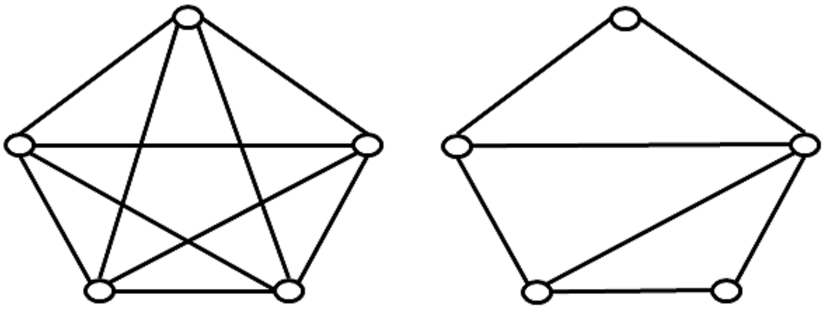

Figure 1: Examples of a clique (left) and a 2-core (right) as the minimum vertex degree is 2.

2.2 The Resilient Network Design Problem

The general *network design problem* (also termed as the network synthesis problem) can be stated as follows: Given a vertex set V and a set of candidate edges E that can be created, each at a nonnegative cost, the objective is to find an optimal set of edges $E^* \subseteq E$ such that a given set of design requirements are satisfied by $G^* = (V, E^*)$. Take the telecommunication network design problem as an example, the vertex set could represent switching centers; the set of edges may represent data transmission channels with different costs due to different transmission distances or geographical terrains; the requirements could be the maximum number of hops between any two switching centers or the minimum level of quality of transmission. In this context, the objective is to design a telecommunication network in the most cost efficient manner such that, for instance, it takes at most three hops to send a message packet from one switching center to another.

Given a network design solution, its reliability is crucial because the failure of a vertex (e.g. switching center) or an edge (e.g. a cable segment) may have a significant impact. This fact further motivates the study of the *survivable network design problem*. A survivable network design is essentially a network design with connectivity requirements so that there are redundant vertex-disjoint or edge-disjoint paths to guarantee reliable information/resources interchange under vertex/edge failures.

We formally define the survivable network design problem as follows. Given an undirected graph $G = (V, E)$, a cost vector $c \in \mathbb{R}_+^{|E|}$ on the edges, and a symmetric $|V| \times |V|$ matrix $R = [r_{ij}]$. Each entry r_{ij} in R represents the minimum number of vertex-disjoint or edge-disjoint paths between vertices i and j. We are seeking a design solution $G^* = (V, E^*)$ where $E^* \subseteq E$ such that total cost of edges in E^* is a minimum and there exist r_{ij} vertex-/edge- disjoint paths between every $i, j \in V$. Note that if r_{ij} is the minimum number of vertex-disjoint paths between vertex i and vertex j, we are seeking a design with vertex connectivity $\kappa(G^*) \geq \min\{r_{ij} \mid \forall i, j \in V \text{ and } i \neq j\}$; if r_{ij} represents the minimum number of edge-disjoint paths between vertex i and vertex j, we are accordingly seeking a design with edge connectivity $\lambda(G^*) \geq \min\{r_{ij} \mid \forall i, j \in V \text{ and } i \neq j\}$. Since $\kappa(G) \leq \lambda(G)$, imposing a vertex connectivity requirement yields a more robust design than imposing a edge connectivity requirement at the same level. Depending on the application, the value of r_{ij} can vary. For example, local telephone networks (Cardwell et al., 1989; Mahjoub, 1994; Grötschel et al., 1992), use a low connectivity requirement (i.e. $r_{ij} \in \{0, 1, 2\}$). In other cases a high connectivity requirement may be imposed (Grötschel et al., 1995), for instance, in a battlefield wireless communication network.

A *resilient network design problem* seeks a minimum cost design satisfying given connectivity requirements (like in survivable network design), and in addition, diameter requirements (different from survivable network design). In other words, a resilient design guarantees not only redundant vertex-disjoint/edge-disjoint paths (survivability) but also short distances. The latter requirement is important in applications like wireless communication networks and airline networks. While real costs present in, for example, traditional transportation networks or water distribution networks can be incorporated into cost vector c, the number of hops in a communication network or the number of legs in an airline network cannot be captured by c. Therefore, in

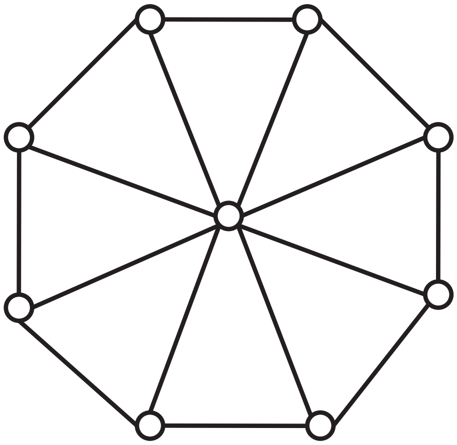

Figure 2: A 3-connected diameter-2 survivable design that does not preserve diameter-2 upon single vertex deletion.

addition to cost minimization and connectivity requirements, extra requirements on diameter are needed, which leads to the resilient network design problem. The difference between a survivable design and a resilient design is illustrated in Figure 2, which shows a 3-connected diameter-2 survivable design where between every pair of vertices there are three internally vertex disjoint paths. Consequently, upon single vertex deletion, at least two internally vertex disjoint paths exist between every pair of vertices. However, if the central vertex is deleted, the diameter increases to 4. Hence, the survivable design in Figure 2 is not a resilient design.

It is worth noting that other terminologies have been used in the literature to describe the network design problem with both connectivity and diameter constraints, such as *hop-constrained survivable network design problem* (Botton et al., 2013; Mahjoub et al., 2013) and *strongly attack-tolerant network design problem* (Veremyev and Boginski, 2012; Pastukhov et al., 2014). Throughout this dissertation, we use the term "resilient network design problem" because: (1) it suggests that more is expected from the design, beyond "survivability"; (2) it differentiates from models where the hop-constraints are required to hold only for specified pairs of nodes, not every pair of nodes.

2.3 Graph-Theoretic Properties of k-Cores

This section focuses on graph-theoretic properties of k-core in detail. A key point outlined in this section is that a k-core, if the k value is appropriately chosen, yields a resilient network which satisfies given connectivity and diameter requirements upon limited vertex/edge failures. The relationship between connectivity and diameters of k-cores when the k value varies has been studied by researchers. Next, we present theoretical results found in the literature establishing bounds on connectivity and diameter of k-cores.

Proposition 1 (Seidman (1983)) *Let G be a k-core on n vertices.*

1. *If $n \leq 2k - r + 2$ and $k \geq r$, then $\kappa(G) \geq r$.*
2. *If $k > \frac{n-2}{2}$ then $diam(G) \leq 2$.*

Proposition 2 (Seidman (1983)) *Let G be a k-core on n vertices with $\kappa(G) = r$ with $1 \leq r \leq k < n$, and $k \leq \frac{n-2}{2}$ then,*

$$diam(G) \leq \frac{3(n-2k-2)}{\beta} + \tau + 3,$$

where $\beta = \max\{k+1, 3r\}$ and

$$\tau = \begin{cases} 0, & \text{if } n-2k-2 \pmod{\beta} < r \\ 1, & \text{if } r \leq n-2k-2 \pmod{\beta} < 2r \\ 2, & \text{if } 2r \leq n-2k-2 \pmod{\beta}. \end{cases}$$

Proposition 3 (Pattillo et al. (2013)) *Let G be a connected k-core on n vertices, then*

$$diam(G) \leq \max_{\bar{\tau} \in \{0,1,2\}} \left\{ \left\lceil \frac{n}{k+1} \right\rceil, 3\left(\left\lfloor \frac{n-\bar{\tau}}{k+1} \right\rfloor - 1 \right) + \bar{\tau} \right\}.$$

This bound is sharp if $\kappa(G) = 1$.

These results can guide the choice of parameter k in order to design k-cores with known bounds on diameter and connectivity. Suppose we design a 10-vertex graph to be a 5-core, Proposition 1 guarantees that it will be 2-connected and have diameter at most 2. Alternately, if we ensure that the designed graph is connected, and control the minimum degree requirement k, Propositions 2 and 3 provide bounds on the resulting diameter. For example, if we design a k-core G with $n = 10, k = 2$ and $\kappa(G) = r = 1$, Proposition 2 implies that $diam(G) \leq 8$, whereas Proposition 3 offers a tighter upper bound of 7. However, for the same values of n and k, if design G has $\kappa(G) = r = 2$, Proposition 2 implies that $diam(G) \leq 5$, while the bound from Proposition 3 is unchanged. Given our focus on 2-hop resilient designs, the following corollary of Proposition 1 that we can derive is particularly useful.

Corollary 1 *For $r \geq 2$ and $k = \lceil \frac{n+r-2}{2} \rceil$, if $G = (V, E)$ is a k-core on n vertices then, $diam(G) \leq 2$ and $\kappa(G) \geq r$. Furthermore, for any $v \in V$,*

1. *$G - v$ is a $(k-1)$-core;*
2. *$\kappa(G - v) \geq r - 1$;*
3. *$diam(G - v) \leq 2$.*

By Corollary 1, if G is designed to be a k-core with $k = \lceil \frac{n+r-2}{2} \rceil$, the graph obtained by deleting any vertex from G is $(r-1)$-connected, and has diameter at most 2. In particular, when $r = 2$ and $k = \lceil \frac{n}{2} \rceil$, G is a 2-connected, diameter-2, k-core which upon single vertex (or edge) deletion still has diameter at most two.

Viewed for instance from an airline hub-network perspective, nearly half the hub airports have direct flights from every other hub, while the rest are reachable with one stop at another hub. More importantly, this continues to be the case if one of the hubs has been disabled with no flights in or out. Naturally, such questions are of a strategic nature compared to the more complicated tactical and operational problems.

However, the strategy employed in the topological design of such networks (airlines and others) critically affects the flexibility required to cope with the dynamic and uncertain operational needs.

Throughout this dissertation, we place considerable attention on k-core design that is at least 2-connected and diameter-2; in other words we let $r = 2$ and accordingly $k = \lceil \frac{n}{2} \rceil$ in most cases when it comes to numerical study or graphical illustration. The rationale is in applications like telecommunication systems, a low connectivity (i.e. 2-connected such that the system survives upon single switch center failures) is generally sufficient and often yields the best tradeoff between capital investment and system reliability (Cardwell et al., 1989). On the other hand, diameter-2 networks are also employed in applications like wireless communication where 2-hop communication helps limit possible information loss.

2.4 Deterministic Minimum Spanning k-Core Problem

Consider a graph $G = (V, E)$, not necessarily complete, where E represents the set of edges that can be created. Further suppose the cost c_e of creating an edge $e \in E$ and an appropriately chosen positive integer k are available. Consider the following optimization problem:

$$\textbf{(MSkCP)} \qquad \min_{x \in \{0,1\}^{|E|}} \Big\{ \sum_{e \in E} c_e x_e \mid \sum_{e \in \gamma(v)} x_e \geq k,\ \forall v \in V \Big\}. \tag{2.1}$$

A feasible solution to this formulation is an incidence vector of a subset of edges J, such that the graph (V, J) is a spanning subgraph of G with minimum degree at least k, i.e., (V, J) is a *spanning k-core* of G. Hence, (2.1) is a formulation of *the minimum spanning k-core problem* (MSkCP), which is to identify an $E^* \subseteq E$ that yields a spanning k-core of G such that the total cost of the edges in E^* is a minimum. The MSkCP can be solved efficiently using a polynomially bounded transformation to a generalized graph matching problem (Balasundaram, 2007).

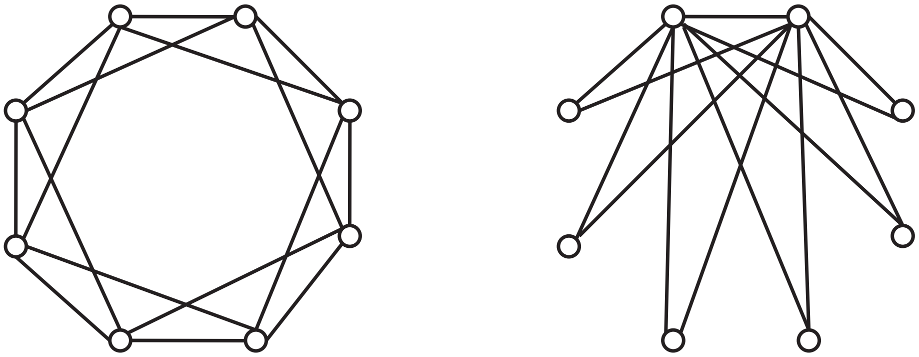

Figure 3: Both graphs are 2-connected and preserve the diameter-2 requirement under single vertex deletion. The design on the left is a 4-core while the design on the right is not.

The MSkCP uses a *sufficient* condition on the minimum degree to design a network that satisfies the second and third properties identified in Corollary 1. However, the minimum degree requirement is not *necessary* to achieve those results. It is possible to design a 2-connected network that preserves the diameter-2 requirement upon vertex deletion that is not a $\lceil \frac{n}{2} \rceil$-core. In Figure 3, the graph on the left is an 8-vertex 4-core that is 2-connected and has diameter-2 upon deleting any vertex as implied by Corollary 1. The graph on the right also satisfies those properties but it is not a 4-core. Note that this design is essentially a reinforced hub-and-spoke design and it is also provably the sparsest (in number of edges) such design (Veremyev and Boginski, 2012). It should be noted that *the r-robust k-club network design problem* studied by Pastukhov et al. (2014), and by Veremyev and Boginski (2012) directly controls diameter and connectivity requirements and can recognize such designs. However, the deterministic version of this problem is NP-hard, and presently there appears to be no work studying this general model in a stochastic setting.

The spanning k-core approach used in this dissertation employs a *sufficient* (but not necessary) condition on the minimum degree to design 2-hop resilient networks, and consequently overlooks designs such as the reinforced hub-and-spoke in Figure 3.

This is a drawback inherent to our approach as it typically produces denser and nearly degree regular network designs and is not suitable for applications where a (reinforced) hub-and-spoke design is desirable. However, it is better suited for 2-hop resilient *inter-hub* network design problems, since hub nodes are typically much smaller in number compared to the nodes in the overall network, and denser inter-connections between hubs are generally more desirable. In the remainder of this dissertation, we focus on MSkCP under uncertainty modeled as independent probabilistic edge failures.

2.5 Uncertainty Modeling and Risk Measures

In a generic design optimization problem, it is often the case that input parameters are uncertain. For example, in the resilient network design problem, the edge cost vector c may not be exactly known and only an estimation can be made which is subject to error, or even the vertices or edges may be subject to failures. Similarly in the service system design, usually customer/transaction arrival rate is uncertain. It usually takes significant monitoring/forecasting effort to get an adequate estimation of the uncertain input parameters.

A typical reactive approach to handle the uncertainty issue is to use the expected values or modes of the uncertain parameters to come up with a cost effective design and then conduct sensitivity analysis (Saltelli et al., 2000) or input parameter tolerance analysis (Leung et al., 2013; Barton et al., 2014). Based on the results of the analysis, some steps may be taken to rectify or enhance the design. However, this reactive approach may not be effective if the response steps are considerably expensive or, even worse, unavailable. In strategic design problems such as those we consider, for example, the capacity planning problem of a service system where a manager needs to make a one-time decision on what equipment to purchase and install, to re-order a new equipment may be cost-prohibitive even if that is suggested by the results of the sensitivity analysis.

Depending on how uncertainty is modeled, a design optimization problem could be approached using robust optimization models (Bertsimas et al., 2011) or stochastic optimization models (Shapiro et al., 2009), invariably using certain risk measures. A discussion of the impact of uncertainty in decision making and a discussion of different models can be found in (Sen and Higle, 1999).

In this dissertation, we focus on the case where one has to make a decision "today" based on the uncertain input parameters modeled as random variables. The random variables will only be realized "in the future" after the decision has been implemented, and it may only be revealed partially. So we focus on proactive approaches for the so-called "single-stage" decision models which do not explicitly model response or recourse actions.

Next, we present a brief introduction to the most commonly used risk measures of interest to us in the literature. Note that recent advances in risk modeling in optimization are reviewed in (Krokhmal et al., 2011; Rockafellar and Uryasev, 2013).

Let $(\Omega, \mathcal{F})$ be a measurable space equipped with probability measure $\mathbb{P}$ and $\mathcal{X} = (\Omega, \mathcal{F}, \mathbb{P})$ be a space of all $\mathcal{F}$-measurable functions $X : \Omega \mapsto \mathbb{R}$. The function X is called a random variable. Generally speaking, a risk measure is a functional $\mathcal{R}$ that "assigns" a number $\mathcal{R}(X)$ to the random variable X; i.e., it is a mapping $\mathcal{R} : \mathcal{X} \mapsto \mathbb{R}$ (Rockafellar and Royset, 2015). To avoid ambiguity, we assume in our context a smaller $\mathcal{R}(\cdot)$ value is better; in other words, $\mathcal{R}(X)$ is preferable to $\mathcal{R}(\hat{X})$ if $\mathcal{R}(X) < \mathcal{R}(\hat{X})$ for any $X, \hat{X} \in \mathcal{X}$. The commonly used risk measures are as follows.

Mean. The basic risk measure of expectation,

$$\mathcal{R}(X) := \mathbb{E}[X],$$

is widely used. It is generally classified as risk-neutral, as it is not sensitive to the existence of "heavy tail". There are many stochastic optimization models based on minimization or bounding of mean in the literature (see Birge and Louveaux, 1997).

Mean-variance. The mean-variance risk measure, also termed as safety margin (Rockafellar and Royset, 2015), can be expressed as:

$$\mathcal{R}(X) := \mathbb{E}[X] + aSD[X],$$

where a is a positive scalar and $SD[X]$ represents the standard deviation of X. Although this risk measure has incorporated variability, it is possible that the large variability on the high end (e.g., large penalty/risk values) remain undetected because of the compensation from the small variability on the low end (e.g., small penalty/risk values). Optimization models using mean-variance risk can be found, for example, in (Ahmed, 2006).

Worst case. Defined as,

$$\mathcal{R}(X) := \sup\{X\},$$

the worst case measure is generally overly conservative. By considering the worst case, the condition that any realization of X is acceptable is imposed. Related studies can be found, for example, in (Mulvey et al., 1995).

Failure probability. Let an event where X realizes as a non-positive number be viewed as a loss or failure. The failure probability is:

$$\mathcal{R}(X) := \mathbb{P}(X \leq 0).$$

This choice is often used in chance-constrained/probabilistic models (see Prékopa, 2003).

Value-at-risk. The risk measure α-value-at-risk (α-VaR) of X is the α-quantile of X (see Pflug, 2000; Krokhmal et al., 2005) defined as:

$$\mathcal{R}(X) := q_\alpha(X) = \inf\{\ell \mid \Psi(\ell) \geq \alpha\}$$

where $\Psi(\ell) = \mathbb{P}(X \leq \ell)$ is the cumulative distribution function of random variable X and $\alpha \in (0, 1)$. Further observe that α-VaR is related to failure probability as a risk measure in the sense that,

$$q_\alpha(X) \leq 0 \;\Leftrightarrow\; \mathbb{P}(X \leq 0) \geq \alpha. \tag{2.2}$$

Therefore, α-VaR and failure probability have similar statistical features and computational complexity (Ahmed and Shapiro, 2008). Although α-VaR has been widely used in the field of financial risk management, it is not without issues: 1) By definition, it does not take into account the realizations beyond the α-quantile point; 2) α-VaR is difficult to deal with inside an optimization model due to nonconvexity; 3) α-VaR is often discontinuous with respect to the parameter α. The last issue also indicates that a small perturbation of α may result in a big jump in the α-VaR value.

Conditional-value-at-risk. The risk measure α-conditional-value-at-risk (α-CVaR), which is also known as "superquantile" (Rockafellar and Royset, 2015) and "average value-at-risk" (Chun et al., 2012) is defined as the mean of the $(1-\alpha)$-tail distribution of $\Psi(\cdot)$, equivalently given by the minimization formula (Rockafellar and Uryasev, 2000):

$$\mathcal{R}(X) := \min_{\zeta}\{\zeta + \frac{1}{1-\alpha}\mathbb{E}[(X-\zeta)^+]\},$$

where $(\cdot)^+ = \max\{0, \cdot\}$.

The choice of risk measures is generally guided by the risk preference of the end-user and by specific problem characteristics (Bertsimas and Sim, 2004). This is of course in addition to computational characteristics of the chosen modeling approach. However, rigorous standards defining a good and useful risk measure are the properties of *coherency* and *regularity*. The concept of coherency was first proposed by Artzner et al. (1999). A risk measure $\mathcal{R}$ is coherent if the following axioms (1)-(4) are satisfied.

- **(Axiom 1)** Monotonicity: If $X, \hat{X} \in \mathcal{X}$ and $X \geq \hat{X}$ with probability 1, then $\mathcal{R}(X) \geq \mathcal{R}(\hat{X})$.

- **(Axiom 2)** Subadditivity: $\mathcal{R}(X + \hat{X}) \leq \mathcal{R}(X) + \mathcal{R}(\hat{X})$ for all $X, \hat{X} \in \mathcal{X}$.

- **(Axiom 3)** Positive homogeneity: $\mathcal{R}(tX) = t\mathcal{R}(X)$ for all $X \in \mathcal{X}$ and $t \in \mathbb{R}^+$.
- **(Axiom 4)** Translation invariance: $\mathcal{R}(t + X) = t + \mathcal{R}(X)$ for all $X \in \mathcal{X}$ and $t \in \mathbb{R}$.

Among the six aforementioned risk measures, mean, worst-case, and α-CVaR are coherent. Mean-variance is not monotonic, α-VaR is not subadditive, and failure probability is not positively homogenous.

The concept of regularity of risk measures is relatively new (Rockafellar and Uryasev, 2013). A risk measure is regular if the following axioms (5)-(8) are satisfied.

- **(Axiom 5)** Convexity: $\mathcal{R}(tX + (1-t)\hat{X}) \leq t\mathcal{R}(X) + (1-t)\mathcal{R}(\hat{X})$ for all $X, \hat{X} \in \mathcal{X}$ and $t \in [0, 1]$.
- **(Axiom 6)** Closedness: $\{X \mid \mathcal{R}(X) \leq t\}$ is a closed set for any constant t.
- **(Axiom 7)** Averseness: $\mathcal{R}(X) > \mathbb{E}(X)$ for nonconstant X.
- **(Axiom 8)** Constant equivalence: $\mathcal{R}(X) = t$ for constant random variable $X \equiv t$

We note that worst-case, α-CVaR, and mean-variance are regular risk measures. α-VaR and failure probability are irregular risk measures because the convexity axiom fails. Mean risk measure is not regular due to the averseness requirement. As observed by Rockafellar and Royset (2015), the coherency axioms and regularity axioms overlap while coherency emphasizes convexity and regularity emphasizes averseness.

In this dissertation, our choices of risk measure are: (1) α-CVaR, which is both coherent and regular (see Chapter 5); (2) failure probability, which is neither coherent nor regular (see Chapter 6); and (3) mean, which is risk neutral (see Chapter 7). The reasons for our choices of risk measures are discussed in the corresponding chapters respectively. Note that we call a stochastic model a CVaR-constrained/chance-constrained optimization model if α-CVaR/failure probability is incorporated in the constraints.

CHAPTER 3

LITERATURE REVIEW

This chapter presents a literature review of the generic network design problem, the survivable network design problem, and different optimization models for the service system design problem under uncertainty.

3.1 Generic Network Design

Network design problem is an important topic in combinatorial optimization and operations research due to its numerous applications in telecommunication networks (Monma and Shallcross, 1989; Cardwell et al., 1989; O'Kelly and Miller, 1994), transportation networks (Magnanti and Wong, 1984; Bell and Iida, 1997; Luathep et al., 2011), and water/electricity networks (Jeźowski, 2010; Hrasnica et al., 2005; Binato et al., 2001). Despite the diversity of practical models that have been developed to cope with domain specific problems, the underlying mathematical models fall under three categories (Minoux, 1989): models using minimum cost multi-commodity flows, models using tree-like networks, and models using non-simultaneous single-commodity or multi-commodity flows. An early review of solution methods for different network design models is provided in (Minoux, 1989). A more recent review is available in (Yang and Bell, 1998). Due to the abundant applications of network design problem, the more recent reviews of models and algorithms for network design problem are often restricted to one particular domain, such as water network design (Jeźowski, 2010) and freight transportation network design (Wieberneit, 2008).

3.2 Survivable Network Design

A survivable network design problem is essentially a network design problem with connectivity requirements so that there exist redundant vertex-disjoint or edge-disjoint paths between every distinct pair of nodes to guarantee reliable information/resources interchange upon network vertex or edge failures. Survivable network design problem captures many celebrated combinatorial optimization problems as special cases, such as the minimum spanning tree problem (Goemans, 2006), the minimum Steiner tree problem (Goemans and Bertsimas, 1993), the minimum Steiner forest problem (Agrawal et al., 1995), the minimum cost k-vertex connected spanning subgraph problem (Monma and Shallcross, 1989), and the minimum cost λ-edge connected spanning subgraph problem (Mahjoub, 1994).

It is not surprising that many survivable network design problems are NP-hard. Consider, for example, the survivable network design problem which seeks a two-edge connected spanning subgraph and all edges carry the same cost. This problem is NP-hard as it generalizes the *Hamiltonian cycle problem* which is NP-hard. Due to this fact, numerous studies have been conducted to develop approximation algorithms and heuristics for survivable network design problems during the past two decades. Note that an approximation algorithm is one that can find a solution whose value is within specified factor of the true optimum in polynomial time (Vazirani, 2001). Specifically for a minimization problem, an algorithm is called a *ρ-approximation algorithm*, $\rho > 0$, if the algorithm produces for any input a solution that is at most ρ**true optimum*. In particular, ρ is called *approximation ratio* or *approximation factor*.

A notable success in terms of approximation algorithms development is presented in (Jain, 2001) where a 2-approximation algorithm was proposed for the survivable network design problem with edge connectivity requirements. Meanwhile, the problem with vertex connectivity requirements was proven to be more difficult to approximate by Kortsarz et al. (2004). In (Lau et al., 2009), approximation algorithms for the survivable network design problem with edge connectivity and degree requirements are developed. This specific type of survivable network design problem seeks a minimum cost subgraph satisfying edge connectivity lower bounds and degree upper bounds which generalizes the *minimum bounded degree spanning tree problem* (Goemans, 2006).

There are relatively fewer studies conducted to develop heuristics for the survivable network design problem. The very first heuristic was presented in (Steiglitz et al., 1969) where the authors first produce an initial feasible solution via a randomized greedy algorithm and next improve the initial solution through a local search approach. A few other heuristic methods can be found in (Ko and Monma, 1989) and (Clarke and Anandalingam, 1995).

A comprehensive review of the literature prior to 1999 on the survivable network design problem can be found in (Soni et al., 1999). A more recent review centering on polyhedral approaches to survivable network design problem is presented by Kerivin and Mahjoub (2005). Other research attempts on case-specific variations of the basic survivable network design formulations can be found in (Magnanti and Raghavan, 2005; Tomaszewski et al., 2010; Song and Luedtke, 2013).

3.3 Service System Design Optimization Models [1]

In the business situation where a service provider performs an operation on incoming transactions from a client, the critical problem for the service provider is to balance the cost invested in provisioning the system capacity against the benefit of satisfying the system performance requirements specified in service level agreements (SLA) with clients. It is often the case that the service request rate from the client is uncertain

[1] Parts of this section are reprinted with permission from “Service system design under uncertainty” by J. Ma, Y. T. Leung, and M. Kamath, 2014. IIE Annual Conference Proceedings, pp. 3564-3573, Copyright [2014] by IIE.

at design time. This is similar to the resilient network design problem as both involve strategic decision made well in advance of system operations. In this section, we survey various modeling approaches available in the literature of service system design and related areas.

Mean Outcome Models. A natural way to address parameter uncertainty is to analyze the mean outcome across all possible realizations of the uncertain parameter. This is possible if we can estimate the probability distribution of the uncertain parameter, possibly using historical data or a surrogate (e.g., demand data for a service similar to a newly offered service). At the very least, the plausible range of the uncertain parameter can be guessed reasonably accurately and a uniform distribution can be assumed over that range. A uniform distribution implies that we have no information on the parameter besides its range, so that any value in the range may be realized. Given the distribution of the uncertain parameter, one can calculate the mean of the objective function, which may be the total cost of operation and can then find an optimal design. Similarly, the operational performance or quality constraints are also expressed in terms of mean quantities. With a given arrival rate, Mandelbaum and Zeltyn (2009) discuss some of these different types of constraints in the context of call centers. An example of mean-outcome models with parameter uncertainty is discussed in (Bassamboo et al., 2010).

Chance-constrained Models. Chance-constrained models arise from various applications in the field of reliability and risk management. A tutorial is provided by Ahmed and Shapiro (2008). In a service system design setting, a chance constraint requires that the system performance requirement specified in a service level agreement is satisfied with a client specified probability. Intuitively, 100% SLA satisfaction under any realization of the underlying random factors is physically impractical and/or economically prohibitive in most cases. As a compromise, imposing a chance constraint on the design becomes an appealing alternative, especially when the impact of

SLA violation is difficult to quantify and thus an explicit penalty/risk/loss function is not readily available, making other approaches like mean outcome, mean-variance, value-at-risk, and conditional-value-at-risk less attractive.

A few researchers have utilized a chance-constrained approach in their problems where the objective is to manage risk in service systems such as call centers and medical service systems. Gans and Zhou (2003) consider a queueing system in telephone call centers where there is a “H-type job” with service level constraints of the forms $\mathbb{E}[\text{delay}] \leq$ “user-specified threshold” and $\mathbb{P}\{\text{delay} \leq$ “user-specified threshold”$\} \geq$ “user-specified probability level”, and a “L-type job” without a service level constraint. In this scenario, they want to find the optimal routing policy such that the rate at which L-type jobs are processed is maximized. The second form of service level constraint is exactly a chance constraint.

Gurvich et al. (2010) study the call center staffing problem with an uncertain demand rate, multiple customer classes and agent types under a chance constraint imposed on the abandonment rate. Beraldi et al. (2004) study the problem of designing emergency medical services under uncertain service requests at the demand points. They model the service facility location and vehicle routing problem as a joint chance-constrained formulation. Since they have a discrete sample space, the problem is converted to a large-scale integer programming problem which is solved using a commercial optimizer. Additionally, Liu et al. (2001) develop a model for maximizing profits in general e-commerce companies where revenue is generated by satisfying quality requirements of service and a penalty is incurred otherwise. In their case, the SLA specifies that the probability of response time for a customer request being less than or equal to a threshold must be at least $(1 - \alpha)$.

Robust Optimization Models. One can formulate a general service system design problem under uncertainty following the robust optimization (RO) framework (Ben-Tal et al., 2009). The RO approach models uncertainty using uncertainty sets as opposed to random variables and employs a min-max objective. Soteriou and Chase (2000) utilize the RO approach to improve service quality by linking operational variables to service quality metrics. They develop a linear service quality optimization program where the coefficients (weights) relating the operational variables to the service quality metrics are uncertain. The model yields optimal guidelines in allocation of operational resources, which is validated by an application to a large healthcare facility. Other examples are contained in (Liang et al., 2009; Marques et al., 2012).

Mean-variance Models. Mean-variance models are similar to mean outcome models, except that the objective and/or the constraints include both the first and second moments of the random variables. For example, the objective may be the mean total cost plus a scalar times the standard deviation of the total cost, representing a utility function of the service provider. Including a variance term ensures, to some degree, that the cost of any one realization of the uncertain parameter is not likely to be very unusual. This is often critical as poor performance for one client may have a significant negative effect on a service provider's reputation. In some cases constraining the variance of the response time is equivalent to specifying a certain percentage of clients should have a certain maximum waiting time. Choi et al. (2008) and Choi and Chow (2008) use mean-variance analysis for a supply chain consisting of a manufacturer and a retailer. We note that mean-variance models originated in and are widely used in financial portfolio management (Markowitz, 1989).

Minimax Regret Models. Models aimed at minimizing the maximum regret (or simply minimax regret) have been employed in the facility location literature for decision making under uncertainty (Snyder, 2006), where the high level, aggregate customer demand is known fairly accurately but the day-to-day demand by location is uncertain. Regret is defined as "the sense of loss felt by a decision maker upon learning that an alternative action could have been preferable to the one actually selected (Mausser and Laguna, 1998)." One can measure regret as some kind of distance from optimality once uncertainty reveals itself. Although the minimax regret criterion is appealing to decision makers when the regret associated with each potential course of action is measurable, the regret-based mathematical formulations tend to be more challenging computationally than regular stochastic problems. To the best of our knowledge, application of minimax regret principles in service system design problems has been sparse if any. In other areas such as energy management, Dong et al. (2011) study a pertinent problem of power management systems planning under uncertainty and show that results from regret models are of great help in balancing the minimized economic loss and system failure risk.

Value-at-risk and Conditional-value-at-risk Models. Both value-at-risk and conditional-value-at-risk are gaining popularity as competitive risk measures against mean and combinations of mean and variance, especially in the fields of finance and portfolio optimization. Despite the popularity of VaR and CVaR in finance, very little research work has been done in applying VaR or CVaR in managing risks in service systems.

We note that some preliminary work has been done by Sodhi (2005) who studies the problem of managing demand risk in tactical supply chain planning for a global consumer electronics company. The company has been utilizing a deterministic replenishment and planning process in spite of considerable uncertainty in demand. To solve the problem, the author develops two models utilizing value-at-risk, employing concepts such as "demand-at-risk" (DaR) on the demand side and "inventory at risk" (IaR) on the supply side. Similarly, they utilize conditional value at risk and introduce cDaR and cIaR. The author concludes that the risk measures can guide the company in reallocating capacity amongst different products and thus are useful in managing demand and inventory risks. Another pertinent paper by Hassan et al. (2005) studies the topic of designing for flexibility in engineering systems via a case study of a satellite fleet design. Their purpose is to identify a satellite's optimal design via value-at-risk under random demand for satellite services such that profitability is maximized.

Among the aforementioned models, mean-variance models, minimax regret models, and value-at-risk/conditional-value-at-risk models are modeling approaches that have been employed in areas such as facility location and portfolio management but not yet in service system design domain. Each model employs a different risk measure relevant to the problem, reflecting different risk preferences on the part of the decision maker and the nature of the design problem.

CHAPTER 4

STATEMENT OF RESEARCH

The overall research goals in this dissertation are threefold.

1. We develop useful stochastic models for design optimization problems under uncertainty. The design optimization problems of our interest are specifically the network design problem with connectivity and diameter requirements, and service system design problem under uncertain demand rate. We employ conditional-value-at-risk, failure probability, and mean, as risk measures in our stochastic optimization formulations.

2. For the network design problem, we develop complexity results, theoretical assertions, and solution techniques in order to tackle the developed stochastic models for moderate- to large-scale instances.

3. For the service system design problem, we develop optimization models to aid a service provider's decision on system capacity. Both single-stage and multi-stage systems are considered in order to develop practical guidelines.

Specifically, we work towards the following objectives in order to fulfil the overall goals in this dissertation.

- **Objective 1.** Developing stochastic models to obtain risk-averse solutions for the resilient 2-hop network design problem under uncertainty via the notion of k-core.
  - **Task 1-1.** Developing polyhedral reformulation for the CVaR-constrained minimum spanning k-core problem under piecewise-linear loss functions.

Polyhedral reformulation for CVaR-constrained models with linear loss function have been studied in the literature but other loss functions have not been considered yet.

  - **Task 1-2.** Developing empirical bounds on the sample size needed in order to attain a reasonable sample-based approximation of CVaR.
  - **Task 1-3.** Designing decomposition algorithms for CVaR constrained programs with integral decision variables and compare against the existing algorithms.

- **Objective 2.** Developing chance-constrained models for the resilient 2-hop network design problem under uncertainty via the notion of k-core.
  - **Task 2-1.** Establishing the intractability of the chance-constrained minimum spanning k-core problem.
  - **Task 2-2.** Conducting polyhedral study on the chance-constrained minimum spanning k-core problem including identifying facet-inducing inequalities to strengthen the formulation to facilitate quicker solution times for moderate-sized instances.
  - **Task 2-3.** Empirically assessing the impact of strengthened formulation via computational studies.

- **Objective 3.** Developing stochastic models and solution techniques for capacity planning of service systems under uncertain demand rates utilizing mean as risk measure.
  - **Task 3-1.** Developing a stochastic model to decide the optimal capacity of a single-stage system represented by an M/M/1 queue where transaction arrival rate is uncertain and the mean of loss due to violation of a given

performance requirement specified in service level agreement needs to be acceptable.

- **Task 3-2.** Developing analytical solutions to the stochastic model in Task 3-1 under the assumption that arrival rates vary uniformly over a specified range.

- **Task 3-3.** Extending the work in Task 3-1 to a two-stage tandem line configuration where different servers may have different service rates.

- **Task 3-4.** Developing numerical methods including scenario-based searching and scenario-based reformulation to solve the stochastic models formulated in Task 3-3.

CHAPTER 5

CVaR-CONSTRAINED SPANNING k-CORE PROBLEM [1]

In this chapter, we study a CVaR-constrained optimization model that captures violation in the minimum degree requirement under independent probabilistic edge failures as a random loss function and limits the CVaR of this loss function. As we discuss next, this approach allows us to quantify the amount of violation in the degree constraints due to uncertainty, and limit this amount in a user-specified fraction of the worst-case scenarios. Consequently, this approach helps drive risk-averse decisions, which is suitable for this strategic topology design problem.

5.1 Model Formulation

Consider a random graph $\widetilde{G} = (V, \widetilde{E})$, where $\widetilde{E}$ denotes the random subset of edges and each edge exists with probability p_e. We associate with this random graph a (deterministic) support graph $G = (V, E)$ where $e \in E \iff p_e > 0$. Then, the sample space $\Omega = \{G^1, \ldots, G^N\}$ is a collection of spanning subgraphs of G. If each edge exists independently of the others, then $N = 2^{|E|}$. We denote the probability measure by $\mathbb{P} : 2^{\Omega} \longrightarrow [0, 1]$. The vector of indicator random variables $\boldsymbol{\xi} : \Omega \longrightarrow \{0, 1\}^{|E|}$ denotes the existence of edges with $\mathbb{P}\{\boldsymbol{\xi}_e = 1\} = p_e$. For each $s = 1, \ldots, N$, we refer to $G^s \in \Omega$, or equivalently, the realization of $\boldsymbol{\xi}$ denoted by ξ^s as

[1] Parts of this chapter are reprinted with permission from "The minimum spanning k-core problem with bounded CVaR under probabilistic edge failures" by J. Ma, F. M. Pajouh, B. Balasundaram, and V. Boginski. *INFORMS Journal on Computing.* Forthcoming. Copyright (2015), the Institute for Operations Research and the Management Sciences, 5521 Research Park Drive, Suite 200, Catonsville, Maryland 21228 USA.

scenario s and denote the probability of realization of scenario s by $\pi^s = \mathbb{P}\{\boldsymbol{\xi} = \xi^s\}$.

In this setting, the degree of vertex $v \in V$ in any design specified by the binary vector $x \in \{0,1\}^{|E|}$ is a random variable given by,

$$d(v) = \sum_{e \in \gamma(v)} x_e \boldsymbol{\xi}_e.$$

Recall that $\gamma(v)$ denotes the set of edges that are incident at v in support graph G. An edge contributes to the degree of its incident vertex only when it is included in the solution ($x_e = 1$) and it survives the random failure process (realization of $\boldsymbol{\xi}_e$ is 1). Since we require our design to have minimum degree k, vertices with degree less than k given a design x and scenario G^i, are in violation of the k-core requirement. The *degree deficiency* quantified as $[k - d(v)]^+ = \max\{k - d(v), 0\}$, is a measure of loss due to uncertainty at vertex v, leading to the following two loss functions:

$$\begin{aligned} L^1(x, \boldsymbol{\xi}) &= \sum_{v \in V} [k - \sum_{e \in \gamma(v)} x_e \boldsymbol{\xi}_e]^+ \qquad (5.1) \\ L^2(x, \boldsymbol{\xi}) &= \max_{v \in V} [k - \sum_{e \in \gamma(v)} x_e \boldsymbol{\xi}_e]^+ \qquad (5.2) \end{aligned}$$

Both loss functions quantify the violation of the minimum degree requirement due to edge failures; L^1 measures the *cumulative degree deficiency* and L^2 measures the *maximum degree deficiency.* Both loss functions are piecewise linear and convex over $x \in \mathbb{R}^{|E|}$ for every given scenario ξ^s.

Suppose $L(x, \boldsymbol{\xi})$ denotes one of the aforementioned loss functions. For a given $\alpha \in (0,1)$, let $\mathrm{VaR}_\alpha[L(x, \boldsymbol{\xi})]$ denote the α-quantile of $L(x, \boldsymbol{\xi})$, and $\mathrm{CVaR}_\alpha[L(x, \boldsymbol{\xi})]$ denote the α-conditional-value-at-risk of the loss functions. The CVaR-constrained MSkCP can be stated as,

$$\textbf{(CVaR-MSkCP)} \quad \min_{x \in \mathcal{P}} \Big\{ \sum_{e \in E} c_e x_e \mid \mathrm{CVaR}_\alpha[L(x, \boldsymbol{\xi})] \le C \Big\}, \qquad (5.3)$$

where $\mathcal{P} = \Big\{ x \in \{0,1\}^{|E|} \mid \sum_{e \in \gamma(v)} x_e \ge k, v \in V \Big\}$ and C is a user specified bound.

The goal of the above model is to identify a spanning k-core of minimum cost among all spanning k-cores of the support graph, such that the downside loss quantified by CVaR of the loss distribution for this solution is bounded. Note that we make a modeling choice to only consider x that correspond to spanning k-cores in the support graph, i.e., $x \in \mathcal{P}$. By doing so, we only discard solutions that are "infeasible" (not spanning k-cores) with probability one. This restriction could be removed and x could be any binary vector. However, from a modeling standpoint, we feel our approach makes it easier to explain to an end-user what the solutions will satisfy from a structural/graph-theoretic perspective. This choice also becomes crucial when tuning the bound C through empirical studies, to avoid choosing unreasonably loose bounds. Furthermore, the relaxation used in the decomposition approach we describe in Section 5.5 will be tighter in the presence of the support graph constraints, especially in the initial iterations.

CVaR is a coherent measure of downside risk (Artzner et al., 1999) that is more conservative than VaR (Rockafellar and Uryasev, 2002). As a result, for larger α values we restrict ourselves to risk-averse solutions to the problem. We may also employ this framework with low values of α, which will tend to drive less risk-averse or more risk-neutral decisions. Hence, we can vary the parameter α to continuously adjust our risk preference.

Bounding CVaR in an optimization model can be equivalently achieved (Rockafellar and Uryasev, 2002; Krokhmal et al., 2002) by bounding the function,

$$F_\alpha(x,\zeta) = \zeta + \frac{1}{1-\alpha}\mathbb{E}[(L(x,\boldsymbol{\xi}) - \zeta)^+] = \zeta + \frac{1}{1-\alpha}\sum_{s=1}^{N} \pi^s [L(x,\xi^s) - \zeta]^+ \tag{5.4}$$

where $\mathbb{E}[.]$ is the expectation operator and $\zeta \in \mathbb{R}$. As a function of $\zeta \in \mathbb{R}$, $F_\alpha(x,\zeta)$ is convex, and moreover if $L(x,\xi^s)$ is convex with respect to x, then $F_\alpha(x,\zeta)$ is jointly convex in $(x,\zeta) \in \mathbb{R}^{|E|} \times \mathbb{R}$ (Rockafellar and Uryasev, 2002). Since the loss functions defined in (5.1) and (5.2) are convex and piecewise linear for $x \in \mathbb{R}^{|E|}$, the aforementioned equivalence allows us to reformulate the optimization problem (5.3) without the explicit knowledge of the loss distribution $\Phi_x(t) = \mathbb{P}\{L(x, \boldsymbol{\xi}) \leq t\}$ for every $x \in \mathcal{P}$, or a closed-form expression for CVaR as a function of x. We refer to the following as the Rockafellar and Uryasev reformulation (RUR) of the CVaR-constrained MSkCP.

$$\textbf{(RUR-MSkCP)} \quad \min_{x \in \mathcal{P}, \zeta \in \mathbb{R}} \Big\{ \sum_{e \in E} c_e x_e \mid F_\alpha(x, \zeta) \leq C \Big\} \tag{5.5}$$

An obvious challenge working with RUR is the number of scenarios in Ω, since $N = 2^{|E|}$. One approach here is to approximate Ω through uniform random sampling to produce a set of equally likely samples $\mathcal{S}$, in which case one can substitute $F_\alpha(x, \zeta)$ with its scenario-based approximation:

$$\tilde{F}_\alpha(x, \zeta) = \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|} \sum_{s \in \mathcal{S}} [L(x, \xi^s) - \zeta]^+. \tag{5.6}$$

The CVaR-constrained MSkCP based on uniform random samples $\mathcal{S}$ is as follows.

$$\textbf{(RUR-MSkCP-}\mathcal{S}\textbf{)} \quad \min_{x \in \mathcal{P}, \zeta \in \mathbb{R}} \Big\{ \sum_{e \in E} c_e x_e \mid \tilde{F}_\alpha(x, \zeta) \leq C \Big\}. \tag{5.7}$$

RUR-MSkCP-$\mathcal{S}$ can now be reformulated into a large-scale mixed integer linear program (MILP) once the term $[L(x, \xi^s) - \zeta]^+$ and the loss function $L(x, \xi^s)$ have been linearized using auxiliary variables. If $\mathcal{S}$ is small enough this MILP could be solved directly using a general-purpose solver. Pertinently, the quality of the approximation $\tilde{F}_\alpha(x, \zeta)$ based on $\mathcal{S}$ would benefit from using a larger sample size. In this article, we adopt the scenario-based approach with uniform random sampling in our computational studies, but we utilize an alternate reformulation that is more amenable to the use of decomposition techniques.

5.2 Remarks on Sample Sizes

In this section, we show that for L^1 and L^2, $F_\alpha(x, \zeta)$ is empirically well approximated by $\tilde{F}_\alpha(x, \zeta)$ with a sample $\mathcal{S}$ sized polynomially in input graph size.

Given an $x \in \mathcal{P}$ defined in Equation (5.3), $\zeta \in \mathbb{R}$ and a loss function $L(x, \boldsymbol{\xi})$ with finite mean and variance, define the following.

$$\begin{aligned} \boldsymbol{\eta} &= [L(x, \boldsymbol{\xi}) - \zeta]^+ \\ \eta^s &= [L(x, \xi^s) - \zeta]^+, \ \forall s \in \mathcal{S} \end{aligned}$$

It follows immediately that $F_\alpha(x, \zeta) = \zeta + \frac{1}{(1-\alpha)}\mathbb{E}[\boldsymbol{\eta}]$ and $\bar{F}_\alpha(x, \zeta) = \zeta + \frac{1}{|\mathcal{S}|(1-\alpha)} \sum_{s \in \mathcal{S}} \eta^s$. Note that by definition, $\mathbb{E}[\boldsymbol{\eta}]$ is finite for a given point (x, ζ) because $L(x, \boldsymbol{\xi})$ is bounded by nk under the loss function L^1 and by k under the loss function L^2. We can view the samples $\eta^s, \forall s \in \mathcal{S}$ as a sequence of random variables, each having the same distribution as $\boldsymbol{\eta}$. Therefore, for any feasible point (x, ζ), we have $\mathbb{E}[\eta^s] = \mathbb{E}[\boldsymbol{\eta}]$. More importantly,

$$\mathbb{E}[\bar{F}_\alpha(x, \zeta)] = F_\alpha(x, \zeta).$$

That is, $\bar{F}_\alpha(x, \zeta)$ is an *unbiased* estimator of $F_\alpha(x, \zeta)$. In addition, by applying Strong Law of Large Numbers (SLLN), $\bar{F}_\alpha(x, \zeta)$ converges with probability one to $F_\alpha(x, \zeta)$ as $|\mathcal{S}| \to \infty$. Thus, we say that $\bar{F}_\alpha(x, \zeta)$ is a *consistent* estimator of $F_\alpha(x, \zeta)$. This consistency provides a certain assurance that as the sample size grows to infinity, the estimation error approaches zero in the limit. While this is important conceptually, insights on the magnitude of estimation error for a given finitely sized sample set $\mathcal{S}$ are desirable.

By the Central Limit Theorem (CLT), we have

$$\{\bar{F}_\alpha(x, \zeta) - F_\alpha(x, \zeta)\} \ \xrightarrow{d} \ N(0, \frac{\sigma^2(\boldsymbol{\eta})}{N(1-\alpha)^2}).$$

In other words, for large enough $|\mathcal{S}|$, $\bar{F}_\alpha(x, \zeta)$ approximately follows a normal distribution with mean $F_\alpha(x, \zeta)$ and variance $\frac{\sigma^2(\boldsymbol{\eta})}{N(1-\alpha)^2}$. In this case, the $100(1-\theta)$ percent interval for $F_\alpha(x, \zeta)$ is

$$[\bar{F}_\alpha(x, \zeta) - \frac{z_{\theta/2}\sigma(\boldsymbol{\eta})}{\sqrt{|\mathcal{S}|}(1-\alpha)}, \ \bar{F}_\alpha(x, \zeta) + \frac{z_{\theta/2}\sigma(\boldsymbol{\eta})}{\sqrt{|\mathcal{S}|}(1-\alpha)}]$$

where $z_{\theta/2} = \Phi^{-1}(1-\theta/2)$ and $\Phi^{-1}(\cdot)$ is the inverse of the cdf for the standard normal distribution. The estimation error, denoted by ϵ, (i.e. $\epsilon = \frac{z_{\theta/2}\sigma(\boldsymbol{\eta})}{\sqrt{|\mathcal{S}|}(1-\alpha)}$) is of order $O(|\mathcal{S}|^{-1/2})$. The constant here is proportional to the standard deviation $\sigma(\boldsymbol{\eta})$, for which an estimate will be derived later.

Consider the case where the values for the maximum estimation error ϵ and confidence level $(1-\theta)$ are given, to decide the required number of samples, we have

$$|\mathcal{S}| \cong \frac{z_{\theta/2}^2 \sigma^2(\boldsymbol{\eta})}{\epsilon^2(1-\alpha)^2}.$$

Specifically for the required sample sizes in regard to loss functions L^1 and L^2, we note the following. For any fixed $\epsilon > 0$ and $\theta \in (0,1)$, $\mathbb{P}\{|\bar{F} - F_\alpha(x,\zeta)| \leq \epsilon\} \geq 1-\theta$,

1. if $|\mathcal{S}|$ is $O(n^2k^2)$ for the cumulative degree deficiency loss function L^1 in Equation (5.1), and

2. if $|\mathcal{S}|$ is $O(k^2)$ for the maximum degree deficiency loss function L^2 in Equation (5.2).

To obtain these estimates for $|\mathcal{S}|$ under each loss function, we observe that $\sigma^2[\boldsymbol{\eta}]$ is $O(n^2k^2)$ when $L \equiv L^1$ and it is $O(k^2)$ when $L \equiv L^2$. Note that $L(x,\boldsymbol{\xi})$ has a finite support $\{0, 1, ..., \bar{m}\}$ where $\bar{m} = nk$ if $L \equiv L^1$ and $\bar{m} = k$ if $L \equiv L^2$. Therefore,

$$\begin{aligned}\sigma^2(\boldsymbol{\eta}) = \sigma^2([L(x,\boldsymbol{\xi}) - \zeta]^+) \leq \sigma^2[L(x,\boldsymbol{\xi}) - \zeta] = \sigma^2[L(x,\boldsymbol{\xi})] &\leq \mathbb{E}[(L(x,\boldsymbol{\xi}))^2] \\ &= \sum_{i=0}^{\bar{m}} \mathbb{P}\{L(x,\boldsymbol{\xi}) = i\} i^2 \leq \bar{m}^2.\end{aligned}$$

The asymptotic estimates of the required sample size may be loose estimates for practical use, but they give insights on the complexity of approximating the true $F_\alpha(x,\zeta)$ value. For a particular CVaR-constrained MSkC-$\mathcal{S}$ instance, it is advisable to conduct experiments in order to find out the appropriate sample size for a given precision and confidence level.

5.3 Reformulations of the CVaR Constraint Under Linear Losses

Following the pioneering work of Rockafellar and Uryasev (2000, 2002), several notable developments on computational aspects of bounding or minimizing CVaR have appeared in recent literature (Krokhmal et al., 2002; Künzi-Bay and Mayer, 2006; Wang, 2007; Fábián, 2008; Wang and Ahmed, 2008; Hong and Liu, 2009; Huang et al., 2010; Huang and Subramanian, 2012). In particular, Künzi-Bay and Mayer (2006) study CVaR minimization with linear loss functions in financial applications. They reformulate the CVaR minimization problem as a two-stage stochastic programming problem with recourse (Birge and Louveaux, 1997) and specialize the L-shaped method (Van Slyke and Wets, 1969) for their particular structure. A central result in their study is a polyhedral reformulation of the CVaR constraint under a linear loss function through exponentially many constraints. This result (stated in Theorem 1), as well as the algorithm "CVaRMin" developed by Künzi-Bay and Mayer (2006) are closely related to the work of Haneveld and van der Vlerk (2006) on "integrated chance-constrained optimization" that limits the expected constraint violation under uncertainty either individually or jointly.

Theorem 1 (Künzi-Bay and Mayer (2006)) *Define $\mathcal{Q}$ and $\mathcal{Q}'$ as follows.*

$$\begin{aligned} \mathcal{Q} &:= \Big\{(x,\zeta) \mid \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{s\in\mathcal{S}}[L(x,\ \xi^s)-\zeta]^+ \le C\Big\} \\ \mathcal{Q}' &:= \bigcap_{\mathcal{A}\subseteq\mathcal{S}}\Big\{(x,\zeta) \mid \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{s\in\mathcal{A}}[L(x,\ \xi^s)-\zeta] \le C\Big\} \end{aligned}$$

with the sum defined as zero for $\mathcal{A}=\emptyset$ in $\mathcal{Q}'$. Then, $\mathcal{Q}=\mathcal{Q}'$.

Note the absence of $[\cdots]^+$ in $\mathcal{Q}'$. The proof of Theorem 1 (Künzi-Bay and Mayer, 2006) implies that the result holds true for any loss function although it was introduced in the context of linear loss functions. In particular, the set $\mathcal{Q}'$ is a polyhedron when the loss function is linear and $\mathcal{S}$ is finite. We refer to the following as the

Künzi-Bay and Mayer reformulation (KBMR) of the CVaR-constrained MSkCP-$\mathcal{S}$.

$$\textbf{(KBMR-MSkCP-}\mathcal{S}\textbf{)} \qquad \min_{x\in\mathcal{P},\ \zeta\in\mathbb{R}} \left\{ \sum_{e\in E} c_e x_e \mid x \in \mathcal{Q}' \right\} \tag{5.8}$$

By Theorem 1, the constraint $\tilde{F}_\alpha(x,\zeta) \leq C$ in (5.7) can be replaced with exponentially many constraints as in (5.8). Although KBMR has $O(2^{|\mathcal{S}|})$ constraints compared to $O(|\mathcal{S}|)$ constraints that would result from linearizing RUR, the KBMR reformulation facilitates the adoption of a row-generation framework (Benders, 1962) as outlined next.

For simplicity, consider the problem in (5.9), and let us assume that the loss function $L(x,\ \xi^s)$ is linear in x and the set $\mathcal{P}'$ is a nonempty polytope in order to explain the approach presented in (Künzi-Bay and Mayer, 2006).

$$\textbf{(KBMR-R)} \qquad \min_{x\in\mathcal{P}',\ \zeta\in\mathbb{R}} \left\{ c^T x \mid \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|} \sum_{s\in\mathcal{A}^i} [L(x,\ \xi^s) - \zeta] \leq C,\ \forall i = 1,\ldots,t \right\} \tag{5.9}$$

The KBMR relaxation (5.9) only considers a subset (possibly empty) of the KBMR constraints. Suppose we solve KBMR-R and it is infeasible; then KBMR is infeasible and we terminate. Otherwise, let the optimal solution found be (x^*,ζ^*). Construct $\mathcal{A}^{t+1} = \{s \in \mathcal{S} \mid L(x^*,\xi^s) - \zeta^* > 0\}$. If the KBMR constraint for $\mathcal{A}^{t+1}$ is satisfied, then no violated constraint exists; (x^*,ζ^*) is optimal to KBMR. Otherwise, add the violated constraint corresponding to $\mathcal{A}^{t+1}$ to KBMR-R and repeat until an optimum is found or KBMR-R becomes infeasible.

The aforementioned sequential cutting plane method (SCPM) was suggested by Künzi-Bay and Mayer (2006) for CVaR minimization and they also noted that their approach extends to CVaR-constrained optimization in a straightforward manner; see also (Fábián, 2008; Subramanian and Huang, 2009). This SCPM is correct and finitely convergent, since $\mathcal{S}$ is finite and no subset of $\mathcal{S}$ will be repeated.

We argue that the favorable computational performance of the SCPM witnessed in empirical studies (Künzi-Bay and Mayer, 2006; Fábián, 2008; Subramanian and

Huang, 2009) is due at least in part to the following characteristics:

1. The set $\mathcal{A}^{t+1}$ corresponds to the most violated KBMR constraint (if one exists).

2. The x variables were continuous and the loss function was linear.

The second characteristic is crucial since KBMR-R solved in each major iteration was a large-scale linear program in (Künzi-Bay and Mayer, 2006). If x is binary, SCPM requires solving an MILP in each major iteration, which can be challenging. Furthermore, if $L(x, \xi^s)$ was piecewise linear and convex, as is the case with our loss functions (5.1) and (5.2), then SCPM would be adding a piecewise linear cut in each major iteration. Hence, we would require additional linearizing variables to handle the piecewise linear loss function in order to solve KBMR-R using an MILP solver. The upshot of this discussion is that while the SCP developed by Künzi-Bay and Mayer (2006) was effective for the problem they considered, direct application/extension of their ideas has clear drawbacks in our setting where x is binary and $L(x, \xi^s)$ is piecewise linear and convex.

In the next section, we present our reformulation ideas that can be viewed as a nested reformulation in the same vein as KBMR that allows us to handle the piecewise linear loss functions more effectively, rather than use auxiliary linearizing variables. In Section 5.5, we develop decomposition branch-and-cut algorithms based on this reformulation, which preserves the desirable features of the approach developed by Künzi-Bay and Mayer (2006), but is better suited to handle the binary variables.

5.4 Reformulations of the CVaR Constraint Under Piecewise Linear Losses

In this section, we develop polyhedral reformulations of the CVaR constraint under both cumulative and maximum degree deficiency loss functions. Note that our results in this section and the decomposition algorithm discussed in Section 5.5 will continue to apply for the more general and exact formulation (5.6) in which the approximation $\tilde{F}_\alpha(x,\zeta)$ in (5.7) is replaced by the exact function $F_\alpha(x,\zeta)$. We discuss our results in this form as it is more relevant in practice given that the actual number of scenarios is exponentially large.

In order to develop an effective approach to solve the CVaR-constrained MSkCP-$\mathcal{S}$, we propose two key computational ideas. First, we extend the polyhedral reformulation ideas of Künzi-Bay and Mayer (2006) to our convex piecewise linear loss functions. Specifically, we introduce an equivalent reformulation for our loss functions (5.1) and (5.2) that uses linear constraints (but more than $2^{|\mathcal{S}|}$ are used). This reformulation also allows us to find the most violated cutting plane by considering every sample in $\mathcal{S}$, similar to the original ideas of Künzi-Bay and Mayer (2006). The binary restriction on x discourages the use of a sequential cutting plane method in favor of a branch-and-cut (BC) algorithm. In Section 5.5, we integrate the reformulation and decomposition ideas into the BC algorithm, leading to an integer programming-based approach for the CVaR-constrained MSkCP-$\mathcal{S}$, similar to those developed recently for chance-constrained optimization (Luedtke et al., 2010; Shen et al., 2010).

Theorem 2 *Let $L^1(x,\boldsymbol{\xi}) = \sum\limits_{v\in V}[k - \sum\limits_{e\in\gamma(v)} x_e\boldsymbol{\xi}_e]^+$ denote the cumulative degree deficiency loss function. Define $\mathcal{Q}_1$ and $\mathcal{T}_1$ as the points (x,ζ) satisfying the CVaR constraint with loss L^1 and the KBMR reformulation of the CVaR constraint, respectively.*

$$\begin{aligned}
\mathcal{Q}_1 &:= \Big\{(x,\zeta) \mid \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{s\in\mathcal{S}}[L^1(x,\ \xi^s)-\zeta]^+ \le C\Big\}\\
\mathcal{T}_1 &:= \bigcap_{\mathcal{A}\subseteq\mathcal{S}}\Big\{(x,\zeta) \mid \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{s\in\mathcal{A}}[L^1(x,\ \xi^s)-\zeta] \le C\Big\}
\end{aligned}$$

Define $\mathcal{R}_1$ as the set of points (x,ζ) satisfying the following constraints, for each $\mathcal{A} = \{s_1,\ldots,s_p\} \subseteq \mathcal{S}$, and for each p-tuple $(V_1,\ldots,V_p)$ such that $V_i \subseteq V,\ \forall i =$

$1, \ldots, p$:

$$\zeta + \frac{1}{(1-\alpha)|\mathcal{S}|} \sum_{i=1}^{p} \Big[\Big\{ \sum_{v \in V_i} (k - \sum_{e \in \gamma(v)} x_e \xi_e^{s_i}) \Big\} - \zeta \Big] \leq C \tag{5.10}$$

The sums are defined as zero for $\mathcal{A} = \emptyset$ $(p = 0)$ or $V_i = \emptyset, i = 1, \ldots, p$. Then, $\mathcal{Q}_1 = \mathcal{T}_1 = \mathcal{R}_1$.

Proof of Theorem 2. The claim $\mathcal{Q}_1 = \mathcal{T}_1$ follows from Theorem 1. By definition, the set $\mathcal{T}_1$ is described by the following constraints, for each $\mathcal{A} = \{s_1, \ldots, s_p\} \subseteq \mathcal{S}$,

$$\zeta + \frac{1}{(1-\alpha)|\mathcal{S}|} \sum_{i=1}^{p} \Big[\Big\{ \sum_{v \in V} (k - \sum_{e \in \gamma(v)} x_e \xi_e^{s_i})^+ \Big\} - \zeta \Big] \leq C.$$

On the other hand, for arbitrary real numbers $a_1, a_2, \ldots, a_{|V|}$, the equation

$$\sum_{i \in V} a_i^+ = \max_{\hat{V} \subseteq V} \sum_{i \in \hat{V}} a_i \tag{5.11}$$

holds with the maximum achieved at $\hat{V} = \{i \mid a_i > 0\}$ and the sum defined as zero when $\hat{V} = \emptyset$. Hence,

$$\sum_{v \in V} (k - \sum_{e \in \gamma(v)} x_e \xi_e^{s_i})^+ = \max_{\hat{V} \subseteq V} \sum_{v \in \hat{V}} (k - \sum_{e \in \gamma(v)} x_e \xi_e^{s_i}).$$

Therefore, $\mathcal{T}_1$ is equivalently described by the following constraints for each $\mathcal{A} = \{s_1, \ldots, s_p\} \subseteq \mathcal{S}$,

$$\zeta + \frac{1}{(1-\alpha)|\mathcal{S}|} \sum_{i=1}^{p} \Big[\Big\{ \max_{\hat{V} \subseteq V} \sum_{v \in \hat{V}} (k - \sum_{e \in \gamma(v)} x_e \xi_e^{s_i}) \Big\} - \zeta \Big] \leq C.$$

This description of $\mathcal{T}_1$ is then equivalent to the description of $\mathcal{R}_1$ using inequalities (5.10). Hence, $\mathcal{Q}_1 = \mathcal{T}_1 = \mathcal{R}_1$. ■

Theorem 3 *Let $L^2(x, \boldsymbol{\xi}) = \max_{v \in V} [k - \sum_{e \in \gamma(v)} x_e \boldsymbol{\xi}_e]^+$ denote the maximum degree deficiency loss function. Define $\mathcal{Q}_2$ and $\mathcal{T}_2$ as the points (x, ζ) satisfying the CVaR constraint with loss L^2 and the KBMR reformulation of the CVaR constraint, respec-*

tively.

$$
\begin{aligned}
\mathcal{Q}_2 &:= \Big\{(x,\zeta) \mid \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{s\in\mathcal{S}}[L^2(x,\ \xi^s)-\zeta]^+ \le C\Big\} \\
\mathcal{T}_2 &:= \bigcap_{\mathcal{A}\subseteq\mathcal{S}}\Big\{(x,\zeta) \mid \zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{s\in\mathcal{A}}[L^2(x,\ \xi^s)-\zeta] \le C\Big\}
\end{aligned}
$$

As before, the sum is defined as zero when $\mathcal{A}=\emptyset$. Define $\mathcal{R}_2$ as the set of points (x,ζ) satisfying the following constraints for each $\mathcal{A}=\{s_1,\dots,s_p\}\subseteq\mathcal{S}$, for each p-tuple $(v_1,\dots,v_p)$ such that $v_i\in V,\ \forall i=1,\dots,p$, and for each $\mathcal{B}\subseteq\mathcal{A}$:

$$
\zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\Big[\Big\{\sum_{s_i\in\mathcal{B}}(k-\sum_{e\in\gamma(v_i)}x_e\xi_e^{s_i})\Big\}-p\zeta\Big] \le C \tag{5.12}
$$

The sum is defined as zero when $\mathcal{B}=\emptyset$. Then, $\mathcal{Q}_2=\mathcal{T}_2=\mathcal{R}_2$.

Proof of Theorem 3. As before, the claim $\mathcal{Q}_2=\mathcal{T}_2$ follows from Theorem 1. By definition, $\mathcal{T}_2$ is described by the following constraints for each $\mathcal{A}=\{s_1,\dots,s_p\}\subseteq\mathcal{S}$,

$$
\zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{i=1}^{p}\Big[\Big\{\max_{v_i\in V}(k-\sum_{e\in\gamma(v_i)}x_e\xi_e^{s_i})^+\Big\}-\zeta\Big] \le C
$$

which is equivalent to the following inequality.

$$
\zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\Big[\Big\{\sum_{i=1}^{p}\big(\max_{v_i\in V}(k-\sum_{e\in\gamma(v_i)}x_e\xi_e^{s_i})\big)^+\Big\}-p\zeta\Big] \le C
$$

Using the identity (5.11) again, we see that $\mathcal{T}_2$ can be equivalently described by

$$
\zeta + \frac{1}{(1-\alpha)|\mathcal{S}|}\Big[\Big\{\max_{\mathcal{B}\subseteq\mathcal{A}}\sum_{s_i\in\mathcal{B}}\max_{v_i\in V}(k-\sum_{e\in\gamma(v_i)}x_e\xi_e^{s_i})\Big\}-p\zeta\Big] \le C
$$

for each $\mathcal{A}=\{s_1,\dots,s_p\}\subseteq\mathcal{S}$. It now follows that the description of $\mathcal{T}_2$ is equivalent to that of $\mathcal{R}_2$ using inequalities (5.12). Hence, $\mathcal{Q}_2=\mathcal{T}_2=\mathcal{R}_2$. ■

Note that in Theorems 2 and 3, the sets $\mathcal{R}_1$ and $\mathcal{R}_2$ are polyhedral. Furthermore, no additional variables are used in the description.

5.5 Decomposition and Branch-and-Cut with CVaR Constraints

In this section, we present a decomposition and branch-and-cut (DBC) algorithm to solve the CVaR-constrained MSkCP-$\mathcal{S}$ with the cumulative degree-deficiency loss function reformulated as in Theorem 2. The approach naturally extends to the maximum degree-deficiency loss function. In the DBC algorithm, we denote by MRP($\mathcal{D}, \mathcal{N}_0, \mathcal{N}_1$) the master linear programming relaxation (5.13) that is solved at every node of the DBC tree to obtain lower bounds. Here, $\mathcal{D}$ is the set of globally valid type (5.10) constraints that have been generated when the DBC node associated with MRP($\mathcal{D}, \mathcal{N}_0, \mathcal{N}_1$) is processed. The sets $\mathcal{N}_0$ and $\mathcal{N}_1$ denote the variables that have been fixed at 0 and 1, respectively, as a result of variable dichotomy branching on the binary variables x at the DBC node being processed.

$$\begin{aligned}
\text{MRP}(\mathcal{D}, \mathcal{N}_0, \mathcal{N}_1): \quad & \min \sum_{e\in E} c_e x_e && (5.13\text{a})\\
\text{subject to} \quad & \sum_{e\in\gamma(v)} x_e \geq k, \ \forall v \in V && (5.13\text{b})\\
& (x, \zeta) \text{ satisfies the constraints in } \mathcal{D} && (5.13\text{c})\\
& x_e = 0, \quad \forall e \in \mathcal{N}_0 && (5.13\text{d})\\
& x_e = 1, \quad \forall e \in \mathcal{N}_1 && (5.13\text{e})\\
& 0 \leq \ x_e \leq 1, \quad \forall e \in E \setminus (\mathcal{N}_0 \cup \mathcal{N}_1) && (5.13\text{f})
\end{aligned}$$

Given a point $(x^*, \zeta^*) \in [0,1]^{|E|} \times \mathbb{R}$ that satisfies inequality (5.13b), we can find the most violated inequality (5.10), or conclude that none exists by solving the following separation problem.

$$lhs(x^*, \zeta^*) = \max\Big\{\zeta^* + \frac{1}{(1-\alpha)|\mathcal{S}|}\sum_{i=1}^{p}\Big[\Big\{\sum_{v\in V_i}(k - \sum_{e\in\gamma(v)} x_e^* \xi_e^{s_i})\Big\} - \zeta^*\Big]\Big\}, \qquad (5.14)$$

where the maximum is over each $\mathcal{A} = \{s_1, \ldots, s_p\} \subseteq \mathcal{S}$, and for each p-tuple $(V_1, \ldots, V_p)$ such that $V_i \subseteq V, \ \forall i = 1, \ldots, p$. An optimal solution to (5.14), $\mathcal{A}^*, (V_1^*, \ldots, V_p^*)$, can be constructed as follows. Let $\mathcal{A}^* = \{s \in \mathcal{S} \mid L^1(x^*, \ \xi^s) - \zeta^* > 0\}$. If

$\mathcal{A}^* = \emptyset$, then $lhs(x^*, \zeta^*) = \zeta^*$. Furthermore, if $\zeta^* > C$, the cutting plane detected is $\zeta \le C$. Suppose $\mathcal{A}^* \neq \emptyset$, and assume $\mathcal{A}^* = \{s_1, \ldots, s_p\}$. For each $s_i \in \mathcal{A}^*$, define $V_i^* = \{v \in V \mid k - \sum\limits_{e \in \gamma(v)} x_e^* \xi_e^{s_i} > 0\}$. If $lhs(x^*, \zeta^*) > C$, then the most violated inequality of type (5.10) is associated with $\mathcal{A}^*, (V_1^*, \ldots, V_p^*)$, which can then be added to $\mathcal{D}$ to cut-off (x^*, ζ^*).

The DBC algorithm is presented in Algorithm 1 and the separation procedure is described in Algorithm 2. We utilize a warm-up procedure (Algorithm 3) to initialize the set $\mathcal{D}$ of globally valid inequalities. This procedure (step 2 in Algorithm 1) is optional and one can start with $\mathcal{D} = \emptyset$. Pertinently in step 18, the separation procedure is not invoked when the optimal solution x^* at any node of the DBC tree is fractional; it is only invoked when x^* is integral to check its feasibility for the original problem. Alternately, the separation procedure could also be invoked in step 18. There is naturally a trade-off between finding cuts before branching begins versus during the branching process. Adding a large number of cuts during the warm-up procedure leads to a large problem being solved at each DBC node, but it also has the potential to produce tighter bounds early and thereby reduce the size of the search tree. We will explore this trade-off in our numerical experiments presented next.

5.6 Computational Experience

The objectives of this computational study are twofold. First, we assess the performance enhancements attributable to the reformulation ideas introduced in Section 5.4 and the DBC algorithmic framework discussed in Section 5.5. Second, we assess the trade-off between adding more cuts of type (5.10) at the root node, potentially resulting in a smaller tree that solves a larger system at each node, versus a potentially larger tree with a smaller system being solved at each node of the search tree if the cuts are only generated during the branching process. To this end, we test two versions of the DBC Algorithm 1 with warm-up (denoted by DBC-WU) and without warm-

Algorithm 1 Decomposition and Branch-and-Cut

```
Require: $G = (V, E), c_e \forall e \in E, \mathcal{S}, k, C, \alpha$
$\mathcal{D} \leftarrow \emptyset$, $ACTIVE \leftarrow \emptyset$, $ub \leftarrow +\infty$
Warm-Up($\mathcal{D}$)                                        ▷ optional
$ACTIVE \leftarrow \{\text{MRP}^0(\mathcal{D}, \emptyset, \emptyset)\}$
while $ACTIVE \neq \emptyset$ do
    Select and delete MRP$^\ell$ from $ACTIVE$
    repeat
        Solve MRP$^\ell(\mathcal{D}, \mathcal{N}_0^\ell, \mathcal{N}_1^\ell)$
        if MRP$^\ell(\mathcal{D}, \mathcal{N}_0^\ell, \mathcal{N}_1^\ell)$ is infeasible then
            $CUTFOUND \leftarrow false$, $lb^\ell \leftarrow +\infty$
        else
            $(x^*, \zeta^*)$ be the optimal solution to MRP$^\ell(\mathcal{D}, \mathcal{N}_0^\ell, \mathcal{N}_1^\ell)$ found and $lb^\ell$ the optimal cost
            if $x^* \in \{0, 1\}^{|E|}$ and $lb^\ell < ub$ then
                $CUTFOUND \leftarrow$ SepCuts($x^*, \zeta^*, \mathcal{D}$)
                if $CUTFOUND = false$ then
                    $ub \leftarrow lb$, incumbent $\leftarrow x^*$          ▷ incumbent update
                end if
            else
                $CUTFOUND \leftarrow false$          ▷ optional: $CUTFOUND \leftarrow$ SepCuts($x^*, \zeta^*, \mathcal{D}$)
            end if
        end if
    until $CUTFOUND = false$ or $lb^\ell \geq ub$
    if $lb^\ell < ub$ then
        Branch on some $x_e^* \in (0,1)$
        Generate and add child nodes with appropriate node-ID to $ACTIVE$
    end if
end while
```

Algorithm 2 SepCuts$(x^*, \zeta^*, \mathcal{D})$

```
$FLAG \leftarrow false$
$i \leftarrow 1$
for each $s \in \mathcal{S}$ do
    if $L^1(x^*, \xi^s) - \zeta^* > 0$ then
        $s_i \leftarrow s$, $\mathcal{A}^* \leftarrow \mathcal{A}^* \cup \{s_i\}$, $i \leftarrow i+1$
    end if
end for
for each $s_i \in \mathcal{A}^*$ do
    $V_i^* = \{v \in V \mid k - \sum_{e \in \gamma(v)} x_e^* \xi_e^s > 0\}$
end for
if $lhs(x^*, \zeta^*) > C$ then                                  ▷ defined in (5.14)
    Add inequality of type (5.10) associated with $\mathcal{A}^*, (V_1^*, \ldots, V_{|\mathcal{A}^*|}^*)$ to $\mathcal{D}$
    $FLAG \leftarrow true$
end if
return $FLAG$
```

Algorithm 3 Warm-Up($\mathcal{D}$)

```
repeat
    $FLAG \leftarrow false$
    Solve MRP($\mathcal{D}, \emptyset, \emptyset$)
    if MRP($\mathcal{D}, \emptyset, \emptyset$) is optimal then
        $FLAG \leftarrow$ SepCuts($x^*, \zeta^*, \mathcal{D}$)
    end if
until $FLAG = false$
```

up (denoted by DBC-NoWU), alongside the following approaches from literature for single-stage CVaR constrained problems. For the purposes of this computational study we limit our attention to just the cumulative degree deficiency loss function (5.1).

1. The first approach is to solve RUR-MSkCP-$\mathcal{S}$ (5.7) directly as a large-scale MILP after linearizing the piecewise linear terms. This serves as a baseline to assess the impact of both reformulation and DBC on overall performance.

2. The second approach is to use the SCPM discussed in Section 5.3. In our case, we repeatedly solve (5.9) with only a subset of the reformulation constraints (5.10) in each iteration. Hence, an MILP is solved in each iteration, and if a violated constraint from (5.10) is detected, it is added and the process is repeated. This lends the SCPM the benefits of the reformulation ideas developed by us for the piecewise linear loss function (5.1). Comparing the two versions of DBC Algorithm 1 this SCPM helps assess the impact of using a branch-and-cut framework, as opposed to a sequential approach, as they both employ the same reformulation.

It should be noted that if the SCPM terminates by reaching the user-specified time limit, then we do not have a feasible solution to the CVaR-constrained MSkCP-

$\mathcal{S}$. By contrast, DBC-NoWU and DBC-WU could return a feasible solution even if the procedure is terminated by reaching the time limit as long as at least one incumbent update occurred prior to termination. This is naturally an important practical consideration while using either approach. It is also preferable to place a time limit on the warm-up procedure called in DBC-W to directly control the maximum amount of time it can take.

5.6.1 Test Instances and Settings

We conduct computational experiments on instances with a complete support graph with $|V| = 10, 50, 100$. According to Corollary 1, we study the case where $r = 2$ and $k = \lceil \frac{n}{2} \rceil$ to obtain 2-hop resilient designs. In each instance, edge failure probabilities are chosen randomly and uniformly from the interval $[0.00, 0.25]$. We chose a uniform distribution so that the edge probabilities in any given instance possess high variance for the given range, which from our past experience results in a reasonably difficult test-bed of instances. The range itself is chosen to ensure that all edges are more likely to exist so that the test instances are meaningful. Similarly, the edge costs are chosen from the interval $[1, \frac{n^2}{2}]$. Note that the range of edge costs increases as the support graph size grows. This is again done to ensure high variance among edge costs as instances of different sizes are considered.

We set $\alpha = 0.9$ in all our numerical studies. Note that if the bound C in the CVaR constraint is too small then the problem becomes infeasible. By contrast, the constraint is redundant if it is too large and the optimal solution would be the same as that of the corresponding deterministic MSkCP. We can however, choose C on the basis of the loss function being used, especially since it has a finite support for each x. For example, all realizations of the cumulative degree deficiency loss function (5.1) are contained in the interval $[0, kn]$. Based on our preliminary experiments we found smaller values of C to result in more challenging instances. In our experiments, C is chosen to be approximately 10%-20% of the maximum loss.

We randomly generate equally likely samples according to the edge failure probabilities for each instance. The number of scenarios varied in our experiments from $|\mathcal{S}| = 500$ to 10,000. For a given number of vertices and number of scenarios, we generate and test 30 replications (or 30 different sets of scenarios). We also impose a 1-hour time limit for each replication, for each instance. We report statistics based on the replications that solved to optimality under the time limit; 20 replications, typically more, are solved to optimality in our experiments.

The entire 1-hour time limit is available to the branch-and-cut in DBC-NoWU. However, in DBC-WU up to 90% of the total wall-clock computing time is allocated to initializing $\mathcal{D}$. This ensures a clear contrast between DBC-NoWU and DBC-WU algorithms. Note that if the warm-up procedure takes less time than maximum allotted, the remaining time under the 1-hour time limit is available to the branch-and-cut.

All algorithms are implemented in C++. All experiments are conducted on a 64-bit Linux system with eight Intel Xeon E5620 2.40GHz processors and 96GB RAM. Gurobi® Optimizer v5.0.1 is used as the MILP solver. All implementations inherited the default settings for branching, node selection, general purpose cutting planes, preprocessing and heuristics. The Gurobi® parameter *GRB_IntParam_Threads*, number of threads used by the parallel MILP solver, is set to its default value 0, which means the thread count is equal to the number of logical cores in the machine, which is eight in our case.

Cut addition in the SCPM is implemented using *addConstr* to avoid the extra effort of rebuilding the MILP model in each iteration. The separation procedure in DBC-NoWU and DBC-WU algorithms is implemented using the Gurobi® Callback feature with "lazy constraints". The valid inequalities of type (5.10) generated in the warm-up phase are added directly to the model using *addConstr* while those generated while branching are added to the "lazy constraint pool" of the MILP solver which are included in the node relaxation only when they are violated.

5.6.2 Numerical Results

Table 1 presents the results of the deterministic MSkCP solved directly by Gurobi® MILP solver under default settings. All instances are solved to optimality within 0.05 seconds. Table 1 reports $|V|, |E|, k$, total cost of all edges given by $\sum_{e \in E} c_e$, the optimal cost, and number of edges in the optimal solution found E^*.

Table 1: Results from solving the deterministic MSkCP on our test-bed.

$\lvert V\rvert$	$\lvert E\rvert$	k	Total cost	Optimal cost	$\lvert E^*\rvert$
10	45	5	1226	431	25
50	1225	25	753228	206620	625
100	4950	50	12560300	3273150	2500

The wall-clock running time statistics over all replications that solved to optimality under the time limit on 10- and 50-vertex test instances are reported in Tables 2 and 3. The first observation is that the approach directly solving RUR-MSkCP-$\mathcal{S}$ is competitive only for the smallest values of $|V|$ and $|\mathcal{S}|$ we tested. In general, it is consistently outperformed by the other three approaches that employ our reformulation and decomposition ideas from Section 5.4 in some form, highlighting the computational benefits of our approach.

On our test instances, with the exception of one set of parameter values ($|V| = 50, |\mathcal{S}| = 10000$), DBC-WU is consistently the fastest based on the average running times. It is also significantly faster than the next fastest algorithm taking into account the difference in average running times and the range of running time measurements (Max-Min) observed for each algorithm, and for each setting of $|V|, |\mathcal{S}|$. In that exceptional case, DBC-WU is the second fastest, behind by about 15 seconds on average. But even so, DBC-WU results in a much narrower range of running times.

Table 2: Running time (secs) results on 10-vertex instances with $k = 5$, $\alpha = 0.9$, and $C = 10$. Statistics are over 30 replications.

		Number of samples					
Algorithm	Measure	500	1000	2500	5000	7500	10000
RUR-MSkCP-$\mathcal{S}$	Min	1.23	4.28	15.84	50.00	88.22	179.37
	Max	3.10	11.21	47.93	169.14	423.33	662.58
	Avg	2.15	7.29	29.27	91.86	209.70	347.52
SCPM	Min	0.22	1.09	3.28	5.36	8.52	10.44
	Max	64.95	35.06	83.61	31.18	34.93	39.53
	Avg	7.74	10.02	13.89	10.82	16.76	18.39
DBC-NoWU	Min	0.38	1.55	4.35	9.06	11.98	13.43
	Max	4.78	8.45	14.28	27.41	46.50	51.34
	Avg	2.11	3.66	8.91	15.80	25.25	29.07
DBC-WU	Min	0.25	0.77	2.19	4.63	7.20	9.95
	Max	2.01	2.37	6.28	8.17	14.33	17.99
	Avg	0.80	1.36	3.38	6.01	9.95	12.68

Tables 4 and 5 present a comparison of DBC-NoWU and DBC-WU on 10- and 50-vertex graphs focusing on average tree size across 30 replications and the average of total number of calls to the separation procedure across 30 replications. For the DBC-WU approach, the number of separation calls includes those made in the warm-up phase. DBC-WU typically results in a much smaller tree and more interestingly, fewer total separation calls compared to DBC-NoWU, reinforcing the merits of the warm-up procedure.

Table 3: Running time (secs) results on 50-vertex instances with $k = 25$, $\alpha = 0.9$, and $C = 180$. Statistics over 30 replications are reported unless indicated otherwise.

		Number of samples					
Algorithm	Measure	500	1000	2500	5000	7500	10000
RUR-MSkCP-$\mathcal{S}$	Min	19.54	48.17	142.19	287.06[a]	1004.49[b]	1632.97[c]
	Max	203.78	423.01	1349.63	1781.79[a]	3491.30[b]	3417.73[c]
	Avg	49.34	108.75	441.63	1135.10[a]	2615.27[b]	2347.14[c]
SCPM	Min	3.94	6.86	18.14	34.57	63.96	86.78
	Max	443.46	612.37	1183.87	370.24	547.56	324.86
	Avg	53.41	75.35	174.88	95.99	170.30	171.58
DBC-NoWU	Min	3.17	18.41	22.81	45.35	78.54	129.10
	Max	94.63	322.76	689.63	314.37	747.96	605.00
	Avg	24.39	65.38	159.70	139.58	273.87	298.74
DBC-WU	Min	5.09	8.45	21.55	46.14	81.98	101.99
	Max	38.03	63.01	250.34	168.84	282.14	339.17
	Avg	12.19	22.24	70.59	90.73	149.40	186.59

[a] Statistics reported over 29 replications that solved to optimality within the time limit; 1 replication terminated with a feasible solution. [b] Statistics reported over 26 replications that solved to optimality within the time limit; 4 replications terminated with a feasible solution. [c] Statistics reported over 5 replications that solved to optimality within the time limit; 9 replications terminated with a feasible solution; remaining 16 replications failed at the root node.

Clearly, our experiments focused more on scaling with respect to $|\mathcal{S}|$ rather than $|V|$ as the inter-hub network design problem is typically for the design of medium-scale networks. However, it is also interesting to note the behavior of different algorithms

Table 4: Tree size and number of cuts on 10-vertex instances with $k = 5$, $\alpha = 0.9$, and $C = 10$.

	DBC-WU		DBC-NoWU	
# of scenarios	# of BC nodes	# of Sep-Cut calls	# of BC nodes	# of Sep-Cut calls
500	1196	122	5454	389
1000	1449	142	5912	418
2500	1675	177	6233	490
5000	1122	173	6011	460
7500	1346	197	6250	502
10000	1312	190	6039	435

as $|\mathcal{S}|$ increases. Despite the fact that the 30 replications are uniformly random sets of samples, a wide range of running times is observed even when $|\mathcal{S}| = 10000$. However, the range narrows as $|\mathcal{S}|$ increases in all algorithms. Pertinently, the range of running time measurements (Max-Min) is significantly smaller for DBC-WU compared to all the other algorithms in general. As expected, we observe in Table 6 that the optimal objective values fall in a much tighter interval as the number of scenarios increases.

Our experiments demonstrate that in general, DBC-WU is faster and more consistent compared to the other approaches for solving the CVaR-constrained MSkCP-$\mathcal{S}$. The dominance is more pronounced when we consider 100-vertex instances that are relatively more challenging. Table 7 presents summary results for each $|\mathcal{S}|$; we report the number of replications out of 30 that are solved to optimality within the time limit, the number for which only a feasible solution is found, and the rest where the algorithm failed to return a feasible solution. DBC-WU solved the most number of

Table 5: Tree size and number of cuts on 50-vertex instances with $k = 25$, $\alpha = 0.9$, and $C = 180$.

	DBC-WU		DBC-NoWU	
# of scenarios	# of BC nodes	# of Sep-Cut calls	# of BC nodes	# of Sep-Cut calls
500	3712	116	13376	235
1000	3761	118	17751	353
2500	4013	169	19557	381
5000	1724	110	13619	162
7500	1509	123	14513	219
10000	958	116	11602	177

Table 6: Optimal costs of 10-, 50-vertex instances (range and standard deviation over 30 replications reported).

		Number of samples					
$\|V\|$	Measure	500	1000	2500	5000	7500	10000
10	Min	479	489	498	496	501	501
	Max	508	511	508	508	505	505
	Std. Dev.	6.91	4.34	1.74	2.12	0.63	0.52
50	Min	206718	206981	207101	207119	207153	207175
	Max	207752	207616	207506	207434	207375	207352
	Std. Dev.	238.45	177.34	108.92	74.45	48.60	42.70

replications to optimality by a large margin, and it is able to find a feasible solution in all other cases. By contrast, directly solving the RUR-MSkCP-$\mathcal{S}$ fails to find an optimal solution, often stuck at the root node, especially when the scenario size becomes large. Note that the SCPM fails to guarantee feasibility whenever it terminates reaching the time limit, which is frequently observed. The performance of DBC-NoWU is also inferior to DBC-WU as shown in Table 7, although it does terminate finding a feasible solution in all replications for each value of $|\mathcal{S}|$, and optimal solutions in some replications when $|\mathcal{S}|$ is small.

Table 7: Results on a 100-vertex graph with $k = 50$, $\alpha = 0.9$, and $C = 650$.

		Number of samples					
Algorithm	Termination count	500	1000	2500	5000	7500	10000
RUR-MSkCP-$\mathcal{S}$	# optimal	25	23	14	6	0	0
	# feasible	5	7	16	21	24	0
	# failed	0	0	0	3	6	30
SCPM	# optimal	2	4	14	4	6	1
	# failed	28	26	16	26	24	29
DBC-NoWU	# optimal	6	5	0	0	0	0
	# feasible	24	25	30	30	30	30
DBC-WU	# optimal	21	26	27	22	19	21
	# feasible	9	4	3	8	11	9

On the trade-off between cut addition at the root versus elsewhere in the search tree, our experiments with this test-bed seem to overwhelmingly favor adding the KBMR-type cuts we have introduced, generated by the separation Algorithm 2, at the root-node before branching begins. However, we suspect that this behavior may not be observed if we solved an NP-hard combinatorial optimization problem, such as the maximum clique problem or the traveling salesman problem under a similar CVaR-constrained framework. Hence, it would be interesting to empirically compare the behavior of say, the shortest path problem and the maximum clique problem, in a similar CVaR-constrained framework to see whether emphasizing cut generation at the root node is more favorable in the former compared to the latter.

CHAPTER 6

CHANCE-CONSTRAINED SPANNING k-CORE PROBLEM

A variety of modeling frameworks may be used to formulate the minimum spanning k-core problem under uncertainty with respect to edge failures. With uncertainty characterized by probabilistic edge failures, the chance-constrained optimization framework is another appropriate choice of risk measure. We view this work as complementary to the CVaR-constrained model as discussed in Chapter 5. We formally introduce the optimization problem of interest next.

6.1 Problem Formulation

Similar to Chapter 5, we consider a random graph $\widetilde{G} = (V, \widetilde{E})$ in which the edges are subject to independent, probabilistic failures. The indicator random variable $\boldsymbol{\xi}_e$ denotes the existence of an edge. In this setting, we use $G = (V, E)$ to denote the support graph of $\widetilde{G}$ where $e \in E \iff \mathbb{P}(\boldsymbol{\xi}_e = 1) > 0$. Accordingly, the degree of vertex $v \in V$ in any design specified by the binary vector $x \in \{0,1\}^{|E|}$ is a random variable given by, $\sum\limits_{e \in \gamma(v)} x_e \boldsymbol{\xi}_e$, where $\gamma(v) = \{(u,v) \mid (u,v) \in E\}$. The *chance-constrained minimum spanning k-core problem* (CCkCP) can be formulated as follows for a user-specified parameter $\epsilon \in [0,1]$.

$$\min_{x \in \{0,1\}^{|E|}} \Big\{ \sum_{e \in E} c_e x_e \mid \mathbb{P}\Big(\bigwedge_{v \in V} \Big(\sum_{e \in \gamma(v)} x_e \boldsymbol{\xi}_e \geq k \Big)\Big) \geq 1 - \epsilon \Big\} \tag{6.1}$$

The random vector $\boldsymbol{\xi}$ has a finite discrete support $\{0,1\}^{|E|}$ which is exponentially large under independent edge failures. Hence, we may approximate the problem by generating a set of samples/scenarios denoted by $\mathcal{S} = \{\xi^1, \xi^2, \ldots, \xi^N\}$. Recall from Chapter 5 that we denote the graph associated with each scenario $\xi^s \in \mathcal{S}$ by $G^s = (V, E^s)$ where $E^s = \{e \in E \mid \xi^s_e = 1\}$. The CCkCP can now be rewritten as follows:

$$\min_{x\in\{0,1\}^{|E|}, I\subseteq\mathcal{S}} \Big\{ \sum_{e\in E} c_e x_e \mid \sum_{e\in\gamma(v)} x_e \xi^s_e \geq k, \forall v \in V, \xi^s \in I; \sum_{\xi^s\in I} \mathbb{P}(\boldsymbol{\xi} = \xi^s) \geq 1 - \epsilon \Big\}. \quad (6.2)$$

Suppose $\mathcal{S}_\emptyset = \{\xi^s \in \mathcal{S} \mid \delta(G^s) < k\}$ and $\mathcal{S}_\emptyset$ is nonempty. In other words, $\xi^s \in \mathcal{S}_\emptyset$ if and only if $G^s = (V, E^s)$ does not contain a spanning k-core. If $\sum_{\xi^s\in\mathcal{S}_\emptyset} \mathbb{P}(\boldsymbol{\xi} = \xi^s) > \epsilon$, then problem (6.2) is infeasible. If $\sum_{\xi^s\in\mathcal{S}_\emptyset} \mathbb{P}(\boldsymbol{\xi} = \xi^s) \leq \epsilon$, the scenarios in $\mathcal{S}_\emptyset$ can simply be eliminated from consideration without affecting the feasible solutions to problem (6.2). Therefore, we make the following simplifying assumption about the sample set for the remainder of this chapter.

Assumption 1 *For any $\xi^s \in \mathcal{S}$, we assume that the associated graph $G^s = (V, E^s)$ is a k-core.*

Now consider $\mathcal{S}_\epsilon = \{s \mid \mathbb{P}(\boldsymbol{\xi} = \xi^s) > \epsilon\}$. For every scenario $\xi^s \in \mathcal{S}_\epsilon$, inequalities $\sum_{e\in\gamma(v)} x_e \xi^s_e \geq k, \ \forall v \in V$ have to hold valid at any feasible solution x to problem (6.2), otherwise the chance constraint will be violated. Based on this observation, the CCkCP can be rewritten as:

$$\begin{aligned}
&\min \sum_{e\in E} c_e x_e && (6.3a)\\
s.t. \quad \sum_{e\in\gamma(v)} x_e \xi^s_e \ &\geq \ k, \quad \forall v \in V, \xi^s \in I \subseteq \mathcal{S} \setminus \mathcal{S}_\epsilon && (6.3b)\\
\sum_{\xi^s\in I} \mathbb{P}(\boldsymbol{\xi} = \xi^s) \ &\geq \ 1 - \epsilon - \sum_{\xi^s\in\mathcal{S}_\epsilon} \mathbb{P}(\boldsymbol{\xi} = \xi^s) && (6.3c)\\
x \ &\in \ \mathcal{P} && (6.3d)
\end{aligned}$$

where $\mathcal{P} = \big\{x \in \{0,1\}^{|E|} \mid \sum_{e\in\gamma(v)} x_e \xi^s_e \geq k, \ \forall v \in V, \xi^s \in \mathcal{S}_\epsilon\big\}$, which is of the same form as (6.2). The upshot of this observation is that in the integer programming reformulation of problem (6.2), binary variables corresponding to scenarios in $\mathcal{S}_\epsilon$ can be fixed to 1 to simplify the integer program. So we make the next assumption to focus on the case where such implicit equations (variables fixed to one) are eliminated from consideration, simplifying the polyhedral analysis that follows.

Assumption 2 *We assume* $\mathbb{P}(\boldsymbol{\xi} = \xi^s) \leq \epsilon$ *for every* $\xi^s \in \mathcal{S}$ *and that* $N \geq 2$*.*

Note that the application of our theoretical assertions and solution techniques in the remainder of this chapter does not suffer if either Assumption 1 or Assumption 2 is violated, because a scan of $\mathcal{S}$ can identify $\mathcal{S}_\epsilon$ and $\mathcal{S}_\emptyset$ by checking $\mathbb{P}(\boldsymbol{\xi} = \xi^s)$ and $\delta(G^s)$ respectively.

Also note that parameter $k \in \{0, 1, \ldots, n-1\}$ where $n = |V|$. When $k = 0$, it is obvious that $x = (0, ..., 0)$ is an optimal solution and the problem becomes trivial. When $k = |V|-1$, the solution $x = (1, ..., 1)$ is optimal if the sum of probability of the scenarios in which G^s is complete exceeds $1-\epsilon$. Otherwise, the problem is infeasible.

6.2 Computational Complexity

We next establish the intractability of the CCkCP. Let $q = \lceil N(1-\epsilon) \rceil$. The decision version of the CCkCP is as follows:

(D-CCkCP) Given a support graph $G = (V, E)$, edge cost vector $c \in \mathbb{R}^{|E|}$, a set of scenarios $\mathcal{S} = \{\xi^1, \xi^2, \ldots, \xi^N\}$, fixed nonnegative integers $k < |V| - 1$, $q \leq N$, and B, is there a binary vector x and a subset $I \subseteq \mathcal{S}$ with $|I| \geq q$ such that x is a k-core in G^s for every $\xi^s \in I$ and $\sum\limits_{e \in E} c_e x_e \leq B$?

We show that D-CCkCP is NP-complete by reduction from the NP-complete problem DPCLP in (Luedtke et al., 2010), which is a linear program with joint chance constraints where only the right-hand side is random with a finite support. The decision version of the DPCLP problem is as follows:

(DPCLP) Given binary integers η^s_ℓ, $s \in \mathcal{S}' = \{1, 2, \ldots, N\}$, $\ell = 1, 2, \ldots, M$, and integers $K \leq N$ and J, is there an $L \subseteq \mathcal{S}'$ such that $|L| \geq K$ and $\sum\limits_{\ell=1}^{M} \max\limits_{s \in L}\{\eta^s_\ell\} \leq J$?

Theorem 4 *D-CCkCP is NP-Complete, even in the special case in which* $\mathbb{P}(\boldsymbol{\xi} = \xi^s) = \frac{1}{N}$, $\forall s \in \mathcal{S}$.

Proof of Theorem 4. Consider an instance of the DPCLP problem, we show how to construct an instance of D-CCkCP, $< G, c, \mathcal{S}, k, q, B >$ in polynomial time such that the answer to D-CCkCP is "yes" $\iff$ the answer to DPCLP is "yes". Let $k = M$, $B = J$, and $q = K$. Create node sets $V^i = \{v_1^i, v_2^i, \ldots, v_M^i\}, i = 1, 2, 3, 4$ and $V = \bigcup_{i=1}^{4} V^i$. Note that $|V| = 4M$. Create edge set $E = E_1 \cup E_2 \cup E_3$ where,

1. $E_1 = \bigcup_{i=1}^{4} \bigcup_{a=1}^{M-1} \bigcup_{b=a+1}^{M} \{(v_a^i, v_b^i)\} \cup \bigcup_{a=1}^{M} \{(v_a^3, v_a^4)\}$,
2. $E_2 = \bigcup_{a=1}^{M} \{(v_a^1, v_a^2)\}$, and
3. $E_3 = \bigcup_{a=1}^{M} \{(v_a^1, v_a^4), (v_a^2, v_a^3)\}$.

Thus $|E| = 2M(M + 1)$. Note that in $G = (V, E)$, $d(v) = k + 1, \forall v \in V$. Figure 4 illustrates a constructed graph $G = (V, E)$ for $M = 3$. Let $c_e = 1$ if $e = (v_a^1, v_a^4), a = 1, \ldots, M$ and $c_e = 0$ otherwise. We construct the sample set $\mathcal{S}$ as follows. Suppose edges in E_2 are subject to failures and edge sets E_1 and E_3 are deterministic. Let u_i^M, in this proof and beyond, denote a unit vector of dimension M with component i being one and $\mathbf{1}^M = \sum_{i=1}^{M} u_i^M$. Subsequently, let $\xi^s = \mathbf{1}^M - \eta^s$ where $\xi^s,\ \eta^s \in \{0, 1\}^M$, be the incidence vector of edges in E_2 under scenario s. That is, for $i = 1, \ldots, M$, if $\xi_i^s = 1$, then edge (v_i^1, v_i^2) is present in scenario s; if $\xi_i^s = 0$, edge (v_i^1, v_i^2) fails in scenario s. Therefore, the scenario graphs associated with sample set $\mathcal{S}$ for D-CCkCP consists of $G^s = (V, E_1 \cup E_3 \cup E_2^s)$ and $|\mathcal{S}| = N$. This completes the reduction. Note that instance $< G, c, \mathcal{S}, k, q, B >$ can be constructed in polynomial time.

We now show that if the answer to a DPCLP instance $< \eta_\ell^s, K, J >$ is "yes", then the answer to $< G, c, \mathcal{S}, k, q, B >$ is "yes" as well. Suppose $\exists L \subset \mathcal{S}'$ such that $|L| \geq K$ and $\sum_{\ell=1}^{M} \max_{s \in L} \{\eta_\ell^s\} \leq J$. We let $I = \{\xi^s \in \mathcal{S} \mid s \in L\}$. Obviously, $|I| = |L| \geq K = q$.

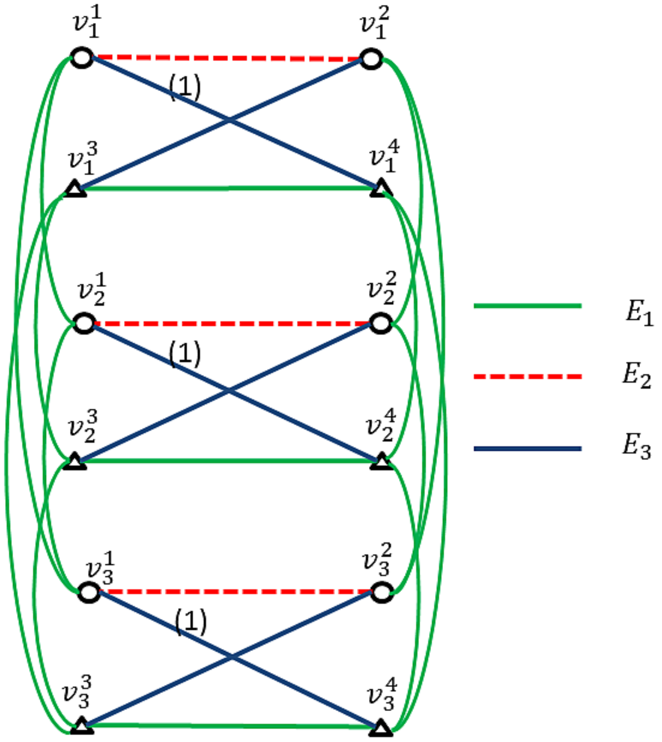

Figure 4: Illustration of D-CCkCP instance G when $M = 3$.

Next, we assign the values of decision vector x and show that x is a k-core in G^s for every $\xi^s \in I$ and $\sum_{e\in E} c_e x_e = \sum_{a=1}^{M} x_{(v_a^1, v_a^4)} \leq B = J$.

1. Set $x_e = 1$, $\forall e \in E_1$, which ensures that $d(v) \geq k$, for all $v \in V^3, V^4$.

2. If $\max_{s\in L}\{\eta_i^s\} = 0$ where $i = 1, \ldots, M$, let $x_{(v_i^1, v_i^2)} = 1$ and $x_{(v_i^1, v_i^4)} = x_{(v_i^2, v_i^3)} = 0$. Here $\max_{s\in L}\{\eta_i^s\} = 0 \Rightarrow \max_{\xi^s\in I}\{1 - \xi_i^s\} = 0 \Rightarrow \min_{\xi^s\in I}\{\xi_i^s\} = 1$. Namely, the edge (v_i^1, v_i^2) is present in all $\xi^s \in I$. Therefore, the assignment of $x_{(v_i^1, v_i^2)} = 1$ and $x_{(v_i^1, v_i^4)} = x_{(v_i^2, v_i^3)} = 0$ ensures that the degree of nodes v_i^1 and v_i^2 is exactly k in this case. Besides, the amount of contribution to the total cost is zero.

3. If $\max_{s\in L}\{\eta_i^s\} = 1$ where $i = 1, \ldots, M$, let $x_{(v_i^1, v_i^2)} = 0$ and $x_{(v_i^1, v_i^4)} = x_{(v_i^2, v_i^3)} = 1$. In this case, $\max_{s\in L}\{\eta_i^s\} = 1 \Rightarrow \max_{\xi^s\in I}\{1 - \xi_i^s\} = 1 \Rightarrow \min_{\xi^s\in I}\{\xi_i^s\} = 0$. That is, edge (v_i^1, v_i^2) fails in some scenario $\xi^s \in I$. The assignment here ensures that for any $\xi^s \in I$, the possible degree loss of nodes v_i^1 and v_i^2 due to the failure of (v_i^1, v_i^2)

is avoided by setting $x_{(v_i^1,v_i^2)} = 0$ and $x_{(v_i^1,v_i^4)} = x_{(v_i^2,v_i^3)} = 1$. Again the degree of nodes v_i^1 and v_i^2 is exactly k.

The above assignment guarantees that x is a k-core in G^s for every $\xi^s \in I$. Furthermore, $\sum\limits_{e\in E} c_e x_e = \sum\limits_{e\in E_1} c_e x_e + \sum\limits_{e\in E_2\cup E_3} c_e x_e = 0 + \sum\limits_{\ell=1}^{M} \max\limits_{s\in L}\{\eta_\ell^s\} \leq J = B$. Hence, the answer to D-CCkCP is "yes".

Next we establish that if the answer to $< G, c, \mathcal{S}, k, q, B >$ is "yes", then $\exists L \subseteq \mathcal{S}'$ such that $|L| \geq K$ and $\sum\limits_{\ell=1}^{M} \max\limits_{s\in L}\{\eta_\ell^s\} \leq J$. Suppose $\exists I \subseteq \mathcal{S}$ with $|I| \geq q$ such that x is a k-core in G^s for every $\xi^s \in I$ and $\sum\limits_{i=1}^{M} x_{(v_i^1,v_i^4)} \leq B = J$. Let $L \subseteq \mathcal{S}'$ be the set of indices corresponding to $I \subseteq \mathcal{S}$, i.e., $L = \{s \in \mathcal{S}' \mid \xi^s \in I\}$. Obviously, $|L| = |I| \geq q = K$. Since x is a k-core in G^s for every $\xi^s \in I$, for $i = 1, ..., M$:

1. If $\min\limits_{\xi^s\in I}\{\xi_i^s\} = 0$, in order that the degree of nodes v_i^1 and v_i^2 is at least k for $\xi^s \in I$ where edge (v_i^1, v_i^2) fails, it must hold that $x_{(v_i^1,v_i^4)} = x_{(v_i^2,v_i^3)} = 1$. Thus, $\max\limits_{s\in L}\{\eta_i^s\} = \max\limits_{\xi^s\in I}\{1 - \xi_i^s\} = 1$.

2. If $\min_{\xi^s\in I}\{\xi_i^s\} = 1$, then $\max\limits_{s\in L}\{\eta_i^s\} = 0$.

Therefore, $\sum\limits_{\ell=1}^{M} \max\limits_{s\in L}\{\eta_\ell^s\} \leq \sum\limits_{i=1}^{M} x_{(v_i^1,v_i^4)} = \sum\limits_{e\in E} c_e x_e \leq B = J$. The answer to DPCLP is "yes" as well. ■

6.3 Deterministic Minimum Spanning k-Core Polytope

To the best of our knowledge, there are no prior studies in the published literature exploring the MSkCP polyhedron directly. This is expected as the problem was introduced only recently (Balasundaram, 2007; Ma and Balasundaram, 2013) and most of the MSkCP polyhedral study results introduced later can be "translated" from the polyhedral results for the generalized b-matching problem. However, the translated results are not so intuitive to understand and it often requires one to be familiar with many terms specifically created in the context of the matching problem (such as near-perfect-match, hypo-matchable, odd polygon, etc). Therefore, we next present a basic observation regarding the full-dimensionality of the k-core polytope that provides a useful intuition for our results concerning the chance-constrained version.

Given a MSkCP defined on graph $G(V, E)$, suppose $\mathcal{P}$ denotes the set of all feasible solutions:

$$\mathcal{P} := \Big\{ x \in \{0,1\}^{|E|} \mid \sum_{e \in \gamma(v)} x_e \geq k, v \in V \Big\}.$$

Proposition 4 *The convex hull conv($\mathcal{P}$) is full-dimensional if and only if $\delta(G) \geq k+1$.*

Proof of Proposition 4. (*Sufficiency*) Suppose $\delta(G) \geq k+1$, it is obvious that the point $x = \mathbf{1}^m$ and points $\mathbf{1}^m - u_e^m$ for all $e \in E$ form $m+1$ feasible and affinely independent points in *conv*($\mathcal{P}$). Accordingly *conv*($\mathcal{P}$) is full-dimensional.

(*Necessity*) Now suppose $\delta(G) = k$ with $|\gamma(\hat{v})| = k$, then the inequality $x_e \leq 1$, $\forall e \in \gamma(\hat{v})$ must hold as equalities for any point in *conv*($\mathcal{P}$). The dimension of *conv*($\mathcal{P}$) is accordingly at most $m-k$, not full-dimensional. It is straightforward that if $\delta(G) < k$ then the MSkCP is infeasible, and $\mathcal{P} = \emptyset$. ■

The sufficient and necessary condition for *conv*($\mathcal{P}$) to be full dimensional is analogous to the condition that vector b has only positive components for the general maximum weighted b-matching polytope to be full-dimensional.

6.4 Chance-Constrained Spanning k-Core Polytope

A deterministic equivalent reformulation DEF of (6.2) by introducing a logical variable z_s for each $\xi^s \in \mathcal{S}$ is as follows.

$$
\begin{aligned}
\textbf{(DEF)} \qquad & \min \sum_{e \in E} c_e x_e && \text{(6.4a)}\\
\text{s.t.} \qquad & \sum_{s \mid \xi^s \in \mathcal{S}} z_s \mathbb{P}(\boldsymbol{\xi} = \xi^s) \geq 1 - \epsilon && \text{(6.4b)}\\
& \sum_{e \in \gamma(v)} x_e \xi_e^s \geq z_s k, \forall v \in V, \xi^s \in \mathcal{S} && \text{(6.4c)}\\
& x \in \{0,1\}^m, \quad z \in \{0,1\}^N && \text{(6.4d)}
\end{aligned}
$$

We next study the feasible solutions of formulation (6.4), the corresponding convex hull, and the corresponding linear relaxation denoted by F, $conv(F)$, and F_{LP}, respectively. Then, $F \subseteq conv(F) \subseteq F_{LP}$ where

$$F = \{(x,z) \in \{0,1\}^m \times \{0,1\}^N \mid (x,z) \text{ satisfies constraints (6.4b) and (6.4c)}\},$$

$$F_{LP} = \{(x,z) \in [0,1]^m \times [0,1]^N \mid (x,z) \text{ satisfies constraints (6.4b) and (6.4c)}\}.$$

By Assumption 1, $F \neq \emptyset$ because the point $(x,z) = (\mathbf{1}^m, \mathbf{1}^N) \in F$. Additionally based on Assumption 2, points $(\mathbf{1}^m, \mathbf{1}^N - u_s^N)$ for all s such that $\xi^s \in \mathcal{S}$ are also feasible to formulation (6.4). Hence, there are at least $N+1$ affinely independent points in $conv(F)$.

To establish the dimension of polytope $conv(F)$, we first introduce some additional notations. For each $e \in E$, we let $\mathcal{D}_e$ be the set of edge incidence samples in $\mathcal{S}$ where in the associated graph G^s, the absence of merely edge e will result in k-core "deficiency"; namely, $\mathcal{D}_e = \{\xi^s \in \mathcal{S} \mid \delta(G^s - e) = k - 1\}$. Also let $E_D = \{e \in E \mid \mathbb{P}(\mathcal{D}_e) \leq \epsilon\}$.

Proposition 5 *The dimension of $conv(F)$ is $N + |E_D|$.*

Proof of Proposition 5. Suppose $E \setminus E_D \neq \emptyset$ and $E_D \neq \emptyset$. For any $e \in E \setminus E_D$, by definition $\mathbb{P}(\mathcal{D}_e) > \epsilon$. Recall that $\mathcal{D}_e$ is the set of scenarios which are k-core structure deficient if $x_e = 0$. Hence, to satisfy the probability constraint (6.4b), the value of x_e in any feasible solution must be 1. So for $\forall(x,z) \in conv(F)$, $x_e \leq 1$ holds at equality for any $e \in E \setminus E_D$. Therefore, the dimension of $conv(F)$ is at most $m+N-|E\setminus E_D| = N+|E_D|$. On the other hand, we can find that for any $e \in E_D$, the point $(x,z) = (\mathbf{1}^m - u_e^m, \sum_{s|\xi^s \in \mathcal{S}\setminus\mathcal{D}_e} u_s^N) \in F$. Also recall that points $(x,z) = (\mathbf{1}^m, \mathbf{1}^N)$ and $(x,z) = (\mathbf{1}^m, \mathbf{1}^N - u_s^N)$ for every s such that$\xi^s \in \mathcal{S}$ are also feasible points in F. Thus, we have obtained $|E_D| + 1 + N$ affinely independent points in F. So the dimension of $conv(F)$ is at least $N + |E_D|$. To conclude, the dimension is exactly $N + |E_D|$.

Now we consider the case where $E = E_D$. Then, it follows that points $(x,z) = (\mathbf{1}^m - u_e^m, \sum_{s|\xi^s \in \mathcal{S}\setminus\mathcal{D}_e} u_s^N) \in F$, for all $e \in E$ and are affinely independent points. Together with points $(x,z) = (\mathbf{1}^m, \mathbf{1}^N)$ and $(x,z) = (\mathbf{1}^m, \mathbf{1}^N - u_s^N), \forall s$ such that $\xi^s \in \mathcal{S}$, we have $m + N + 1$ such points. Therefore the dimension of $conv(F)$ is $N + |E_D|$.

For the case where $E_D = \emptyset$, $x = \mathbf{1}^m$ for any $(x,z) \in F$. Hence, by Assumption 2, $(x,z) = (\mathbf{1}^m, \mathbf{1}^N)$ and $(x,z) = (\mathbf{1}^m, \mathbf{1}^N - u_s^N), \forall s$ such that $\xi^s \in \mathcal{S}$ are $N + 1$ affinely independent points in F. It follows that the dimension of $conv(F)$ is $N + |E_D|$. ■

Proposition 5 implies that convex hull $conv(F)$ is full-dimensional if and only if for any $e \in E$, $\mathbb{P}(\mathcal{D}_e) \leq \epsilon$, i.e., $E_D = E$. Let $\mathcal{S}_{k+1} = \{\xi^s \in \mathcal{S} \mid \delta(G^s) \geq k+1\}$. A useful corollary we can derive from Proposition 5 is the following.

Corollary 2 *If $\mathbb{P}(\mathcal{S}_{k+1}) \geq 1 - \epsilon$, $conv(F)$ is full-dimensional.*

Proof of Corollary 2. Given $\mathbb{P}(\mathcal{S}_{k+1}) \geq 1-\epsilon$, it immediately follows that for all $e \in E$, $\mathbb{P}(\mathcal{D}_e) \leq \epsilon$ because we have $\delta(G^s - e) \geq k$ for every s that satisfies $\delta(G^s) \geq k + 1$. Therefore, $E_D = E$ and $conv(F)$ is full-dimensional. ■

Compared with the sufficient and necessary condition $E_D = E$ for $conv(F)$ to be full-dimensional, the sufficient condition in Corollary 2 is more straightforward because one just needs to check the minimum vertex degree associated with each sample followed by verifying whether cumulative probability is greater than $1-\epsilon$. Additionally, let $F_s = \{x \in \{0,1\}^m \mid \sum\limits_{e\in\gamma(v)} x_e\xi_e^s \geq k, \forall v \in V\}$. That is, F_s denotes the set of feasible solutions to the deterministic minimum spanning k-core problem associated with scenario s. It is worth noting that $conv(F_s)$ is full-dimensional as well for any s such that $\delta(G^s) \geq k+1$. Therefore, we have the following.

Corollary 3 *If* $\sum\limits_{\xi^s\in\mathcal{S}} \mathbb{P}(\xi^s \mid conv(F_s) \textit{ is full-dimensional}) \geq 1-\epsilon$, *$conv(F)$ is full-dimensional.*

Proof of Corollary 3. Corollary 3 follows from Proposition 5 and Corollary 2. ■

Needless to say, if $conv(F_s)$ is full-dimensional for every $\xi^s \in \mathcal{S}$, $conv(F)$ is full-dimensional. The condition in Corollary 2 or Corollary 3 is not necessary, which is illustrated by the counter-example shown in Figure 5. Suppose $\epsilon = 0.5$ and $k = 1$. In the example, the cumulative probability of s satisfying $\delta(G^s) \geq k+1 = 2$ is 0.25 which is less than $1-\epsilon = 0.5$, violating the sufficient condition stated in Corollary 2 and Corollary 3. But the corresponding CCkCP polytope is still full-dimensional because $(x,z)^T = (1\ 1\ 1,\ 1\ 1\ 1\ 1)$, $(1\ 1\ 1,\ 0\ 1\ 1\ 1)$, $(1\ 1\ 1,\ 1\ 0\ 1\ 1)$, $(1\ 1\ 1,\ 1\ 1\ 0\ 1)$, $(1\ 1\ 1,\ 1\ 1\ 1\ 0)$, $(0\ 1\ 1,\ 1\ 0\ 1\ 0)$, $(1\ 0\ 1,\ 1\ 1\ 0\ 0)$, and $(1\ 1\ 0,\ 1\ 0\ 0\ 1)$ are $1+N+m$ affinely independent points in F. Note that for $e = 1,\ 2,\ 3$, $\mathbb{P}(\mathcal{S}\setminus\mathcal{D}_e) = 0.5 \leq 1-\epsilon = 0.5$, meaning the sufficient and necessary condition $E_D = E$ is satisfied.

In the remainder of this section, we investigate facet-inducing conditions for variable bounds of x and z, for the probability constraint, and for the degree constraints. Note that Propositions 6-11 are valid even if $conv(F)$ is not full-dimensional.

6.4.1 Facets From Variables Bounds

Proposition 6 *Given an edge $e \in E$, the inequality $x_e \geq 0$ induces a facet of $conv(F)$ if and only if*

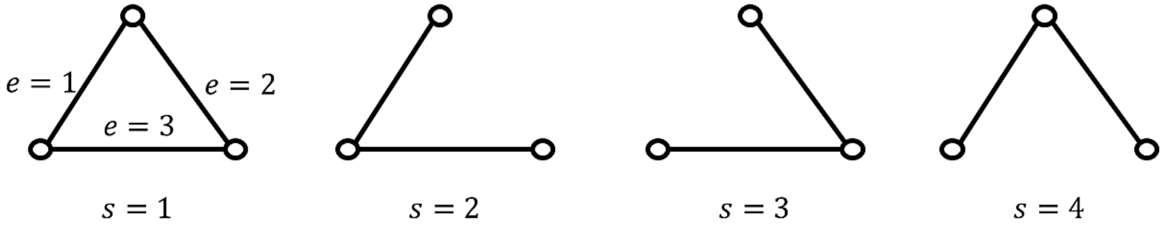

Figure 5: A CCkCP instance with $n = 3$, $m = 3$, $N = 4$, $k = 1$, and $\mathbb{P}(\boldsymbol{\xi} = \xi^s) = 0.25$, which corresponds to a full-dimensional $conv(F)$ even though the sufficient condition of Corollary 2 is violated.

(i) $\mathcal{D}_e = \emptyset$*;*

(ii) $\mathbb{P}(\mathcal{D}_{e,a}) \leq \epsilon$, $\forall a \in E_D \setminus \{e\}$ *where* $\mathcal{D}_{e,a} = \{\xi^s \in \mathcal{S} \mid \delta(G^s - e - a) < k\}$*.*

Proof of Proposition 6. Let $F' = \{(x, z) \in conv(F) \mid x_e = 0\}$. We first prove the necessity of conditions (*i*) and (*ii*). Suppose condition (*i*) is violated; namely $\mathcal{D}_e \neq \emptyset$. Then for any $(x, z) \in F'$, $z_s = 0$ for any $s \in \mathcal{D}_e$. Therefore, $dim(F') \leq |E_D| + N - 1 - |\mathcal{D}_e| < |E_D| + N - 1$, indicating that F' is not a facet of $conv(F)$. Now suppose condition (*ii*) is violated. That is, $\exists a \in E_D \setminus \{e\}$ such that $\mathbb{P}(\mathcal{D}_{e,a}) > \epsilon$. In this case, for any $(x, z) \in F'$, $x_a = 1$. Therefore, F' is not a facet of $conv(F)$ because its dimension is strictly less than $|E_D| + N - 1$.

Now we show that the two conditions are sufficient. By Assumption 1, the feasible point $(x, z) = (\mathbf{1}^m, \mathbf{1}^N)$ satisfies $x_e \geq 0$ with strict inequality. Hence, $x_e \geq 0$ is not an implicit equation. Suppose $\mathcal{D}_e = \emptyset$, it follows from Assumption 2 that $(x, z) = (\mathbf{1}^m - u_e^m, \mathbf{1}^N)$ and $(x, z) = (\mathbf{1}^m - u_e^m, \mathbf{1}^N - u_\ell^N)$, for any $\ell \in \mathcal{S}$ are feasible points in $conv(F)$ that satisfies $x_e \geq 0$ at equality. Note that $\mathcal{D}_e = \emptyset$ immediately implies that $e \in E_D$. Additionally, suppose $\mathbb{P}(\mathcal{D}_{e,a}) \leq \epsilon$, $\forall a \in E_D \setminus \{e\}$, then the point $(x, z) = (\mathbf{1}^m - u_e^m - u_a^m, \sum_{\ell \in \mathcal{S} \setminus \mathcal{D}_{e,a}} u_\ell^N)$ is feasible where $x_e \geq 0$ holds at equality. Thus, we have obtained $1 + N + |E_D| - 1 = N + |E_D|$ feasible points satisfying $x_e \geq 0$ with equality and it is easy to verify that these points are affinely independent. ■

Proposition 7 *Given an edge $e \in E$, the inequality $x_e \leq 1$ induces a facet of conv(F) if and only if $\mathbb{P}(\mathcal{D}_e) \leq \epsilon$.*

Proof of Proposition 7. (*Sufficiency*) Given $\mathbb{P}(\mathcal{D}_e) \leq \epsilon$, inequality $x_e \leq 1$ is valid for *conv*(F) and holds with equality at feasible points $(\mathbf{1}^m - u_a^m, \sum\limits_{s|\xi^s \in \mathcal{S} \setminus \mathcal{D}_a} u_s^N)$, $\forall a \in E_D \setminus \{e\}$, $(\mathbf{1}^m, \mathbf{1}^N)$, $(\mathbf{1}^m, \mathbf{1}^N - u_s^N), \forall s$ such that $\xi^s \in \mathcal{S}$, which consist of $(|E_D|-1)+1+N = |E_D| + N$ affinely independent points. Also, the feasible point $(\mathbf{1}^m - u_e^m, \sum\limits_{s|\xi^s \in \mathcal{S} \setminus \mathcal{D}_e} u_s^N)$ satisfies $x_e < 1$, indicating that $x_e \leq 1$ is not an implicit equation. Therefore, $x_e \leq 1$ defines a facet of *conv*(F) and the condition is sufficient.

(*Necessity*) Suppose $\mathbb{P}(\mathcal{D}_e) > \epsilon$, then $x_e \leq 1$ must hold as equality for any $(x, z) \in$ *conv*(F). Further, $F' = \{(x, z) \in conv(F) \mid x_e = 1\} = conv(F)$, i.e., F' is not a proper face of *conv*(F). Hence, $x_e \leq 1$ is not facet-defining. ■

Proposition 8 *The inequality $z_s \geq 0$ induces a facet of conv(F) if and only if*

(i) for any $\ell \in \mathcal{S} \setminus \{s\}$, $\mathbb{P}(\boldsymbol{\xi} = \xi^s) + \mathbb{P}(\boldsymbol{\xi} = \xi^\ell) \leq \epsilon$;

(ii) for any $e \in E_D$, either $s \in \mathcal{D}_e$, or $s \notin \mathcal{D}_e$ but $\mathbb{P}(\mathcal{S} \setminus \mathcal{D}_e) - \mathbb{P}(\boldsymbol{\xi} = \xi^s) \geq 1 - \epsilon$.

Proof of Proposition 8. Let $F' = \{(x, z) \in conv(F) \mid z_s = 0\}$. We first show the necessity of both conditions. Suppose condition (*i*) is violated, meaning $\exists \ell \in \mathcal{S} \setminus \{s\}$ such that $\mathbb{P}(\boldsymbol{\xi} = \xi^s) + \mathbb{P}(\boldsymbol{\xi} = \xi^\ell) > \epsilon$. It follows that for any $(x, z) \in F'$, the equation $z_\ell = 1$ holds valid. Hence, the dimension of F' is at most $|E_D| + N - 2$. Now suppose condition (*ii*) is violated. That is, $\exists e \in E_D$ such that $s \notin \mathcal{D}_e$ and $\mathbb{P}(\mathcal{S} \setminus \mathcal{D}_e) - \mathbb{P}(\boldsymbol{\xi} = \xi^s) < 1 - \epsilon$. Then for this edge e, any point $(x, z) \in F'$ satisfies $x_e = 1$, which implies that the dimension of F' is at most $|E_D| + N - 2$.

To show sufficiency, we first observe that feasible point $(\mathbf{1}^m, \mathbf{1}^N)$ satisfies inequality $z_s \geq 0$ with strictly inequality, which is therefore not an implicit equation. We next construct $|E_D| + N$ affinely independent points in F' under the two conditions. Due to condition (*i*), points $(\mathbf{1}^m, \mathbf{1}^N - u_s^N)$ and $(\mathbf{1}^m, \mathbf{1}^N - u_s^N - u_\ell^N)$, $\forall \ell \in \mathcal{S} \setminus \{s\}$ are feasible points at which $z_s = 0$ holds. Additionally, $(\mathbf{1}^m - u_e^m, \sum_{\ell \in \mathcal{S} \setminus \mathcal{D}_e} u_\ell^N)$ belongs to F' for $e \in E_D$ such that $s \in \mathcal{D}_e$, and $(\mathbf{1}^m - u_e^m, \sum_{\ell \in \mathcal{S} \setminus \mathcal{D}_e} u_\ell^N - u_s^N)$ belongs to F' for $e \in E_D$ such that $s \notin \mathcal{D}_e$ and $\mathbb{P}(\mathcal{S} \setminus \mathcal{D}_e) - \mathbb{P}(\boldsymbol{\xi} = \xi^s) \geq 1 - \epsilon$. So we have $1 + (N-1) + |E_D|$ feasible points in F' and it is easy to verify that these points are affinely independent. ■

Proposition 9 *The inequality $z_s \leq 1$ induces a facet of conv(F), if and only if either (i) $\delta(G^s) \geq k+1$ or (ii) $\delta(G^s) = k$ with $\gamma(\hat{v}) \cap E_D = \emptyset$ for every $\hat{v}$ of degree k in G^s.*

Proof of Proposition 9. Let $F' = \{(x,z) \in conv(F) \mid z_s = 1\}$. We first show that condition (i) is sufficient. If $\delta(G^s) \geq k+1$, we know that $s \notin \mathcal{D}_e$, for all $e \in E$. The inequality $z_s \leq 1$ holds with equality at feasible points $(\mathbf{1}^m, \mathbf{1}^N)$ and $(\mathbf{1}^m, \mathbf{1}^N - u_\ell^N)$, $\forall \ell \in \mathcal{S} \setminus \{s\}$. Additionally given $E_D \neq \emptyset$ (the case where $E_D = \emptyset$ is trivial), feasible points $(\mathbf{1}^m - u_e^m, \sum_{\ell \in \mathcal{S} \setminus \mathcal{D}_e} u_\ell^N)$, $\forall e \in E_D$ satisfy $z_s \leq 1$ with equality. Therefore, we have obtained $1 + (N-1) + |E_D|$ affinely independent points satisfying $z_s \leq 1$ with equality. Also, at the feasible point $(\mathbf{1}^m, \mathbf{1}^N - u_s^N)$, $z_s \leq 1$ holds as a strict inequality, which indicates $z_s \leq 1$ is not an implicit equation. Therefore, the inequality $z_s \leq 1$ induces a facet.

Next, we show condition (ii) is sufficient. Suppose $\delta(G^s) = k$ and $\hat{v}$ is a node of degree k in the graph associated with scenario s, that is, $\sum_{e \in \gamma(\hat{v})} \xi_e^s = k$. Given $\gamma(\hat{v}) \cap E_D = \emptyset$, the inequality $x_e \leq 1$ for any $e \in \gamma(\hat{v})$ must hold as equality for any $(x,z) \in conv(F)$. Therefore, points $(\mathbf{1}^m, \mathbf{1}^N)$, $(\mathbf{1}^m, \mathbf{1}^N - u_\ell^N)$, $\forall \ell \in \mathcal{S} \setminus \{s\}$, and $(\mathbf{1}^m - u_e^m, \sum_{\ell \in \mathcal{S} \setminus \mathcal{D}_e} u_\ell^N)$, $\forall e \in E_D$ are still $1 + (N-1) + |E_D|$ affinely independent points in F', which indicates that the dimension of F' is $N + |E_D| - 1$. Hence, $z_s \leq 1$ induces a facet.

Now suppose $\delta(G^s) = k$ and $\gamma(\hat{v}) \cap E_D \neq \emptyset$ for some $\hat{v}$ that satisfies $\sum_{e \in \gamma(\hat{v})} \xi_e^s = k$. It follows that for any $e \in \gamma(\hat{v})$, $x_e = 1$ for all $(x,z) \in F'$. Therefore, the dimension of F' is at most $|E_D| + N - 1 - |\gamma(\hat{v}) \cap E_D| < |E_D| + N - 1$, implying that F' is not a facet. ■

As condition (ii) in Proposition 9 is less intuitive, we illustrate this case with the example in Figure 5. Now let us suppose $\epsilon = 0.25$. According to Proposition 5, the dimension of the associated CCkCP polytope is $N + |E_D| = 4$ where $|E_D| = 0$. Take scenario $s = 2$ as an example. Obviously, $\delta(G^s) = 1 = k$ and $\gamma(\hat{v}) \cap E_D = \emptyset$ where $\hat{v}$ is a node of degree 1 in G^2. We can observe that inequality $z_2 \leq 1$ is facet-defining because $(1\ 1\ 1,\ 1\ 1\ 1\ 1)^T$, $(1\ 1\ 1,\ 0\ 1\ 1\ 1)^T$, $(1\ 1\ 1,\ 1\ 1\ 0\ 1)^T$, $(1\ 1\ 1,\ 1\ 1\ 1\ 0)^T$ are 4 feasible and affinely independent points satisfying inequality $z_2 \leq 1$ as equality.

Consider the special case where $conv(F)$ is full-dimensional. A corollary we can derive from Proposition 9 is as follows.

Corollary 4 *Given that $conv(F)$ is full-dimensional, the inequality $z_s \leq 1$ induces a facet if and only if $\delta(G^s) \geq k + 1$.*

Proof of Corollary 4. The sufficiency follows from Proposition 9. By Proposition 5, $conv(F)$ being full-dimensional implies $E_D = E$, which further indicates $\gamma(\hat{v}) \cap E_D \neq \emptyset$ for any $\hat{v}$ with degree k in G^s. Hence, $\delta(G^s) \geq k + 1$ is also a necessary condition. ■

6.4.2 CCkCP Probability Inequality

Proposition 10 *The inequality (6.4b) induces a facet of $conv(F)$ in the special case $\mathbb{P}(\boldsymbol{\xi} = \xi^s) = \frac{1}{N}$.*

Proof of Proposition 10. In this special case, we notice that constraint (6.4b) can be rewritten as the following.

$$\sum_{s|\xi^s \in \mathcal{S}} z_s \geq q \tag{6.5}$$

where $q = \lceil N(1 - \epsilon) \rceil$. By Assumption 2, $\frac{1}{N} \leq \epsilon \Rightarrow q < N$. The feasible point $(x, z) = (\mathbf{1}^m, \mathbf{1}^N)$ satisfies (6.5) with strict inequality. Hence, inequality (6.5) is not an implicit equation. To complete the proof, we next show that there are $|E_D| + N$ feasible and affinely independent points at which (6.5) holds as an equation. For each edge $e \in E_D$ (the case where $E_D = \emptyset$ is trivial), $(x, z) = (\mathbf{1} - u^m_e, \sum_{\ell \in \mathcal{S} \setminus \mathcal{D}_e} u^N_\ell)$ is a feasible point. Notice that $|\mathcal{S} \setminus \mathcal{D}_e| \geq q$ because E_D is the set of edges satisfying $\mathbb{P}(\mathcal{D}_e) \leq \epsilon$ and thus $|\mathcal{D}_e| \leq \lfloor N\epsilon \rfloor$. Now let $\bar{\mathcal{D}}_e$ be an arbitrary subset of $\mathcal{D}_e$ with size q; i.e., $\bar{\mathcal{D}}_e \subseteq \mathcal{D}_e$ and $|\bar{\mathcal{D}}_e| = q$. Then, $(x, z) = (\mathbf{1} - u^m_e, \sum_{\ell \in \bar{\mathcal{D}}_e} u^N_\ell)$ is feasible in F and satisfies (6.5) at equality. In this manner, we can construct $|E_D|$ such points. Now consider the polytope $F' = \{z \in [0, 1]^N \mid \sum_{s|\xi^s \in \mathcal{S}} z_s = q\}$. Indeed, $F' \neq \emptyset$ and $dim(F') = N - 1$. Hence, we can find N affinely independent integral points in F' and we label these N points as $z^1, ..., z^N$. It is easy to verify that $(x, z) = (\mathbf{1}^m, z^i),\ i = 1, ..., N$ are N feasible and affinely independent points in *conv*(F). Further these N points together with $(x, z) = (\mathbf{1} - u^m_e, \sum_{\ell \in \mathcal{S} \setminus \mathcal{D}_e} u^N_\ell), \forall e \in E_D$ yield $N + |E_D|$ feasible and affinely independent points in *conv*(F) that satisfy (6.5) as equality. Hence, (6.5) is facet-inducing in this case. ■

6.4.3 CCkCP-Degree Inequalities

Suppose $\mathbb{P}(\mathcal{S}_{k+1}) \geq 1 - \epsilon$, that is, *conv*$(F)$ is full-dimensional according to Corollary 2. An intuitive question is, given $\xi^s \in \mathcal{S}_{k+1}$, does the degree inequality $\sum_{e \in \gamma(v)} \xi^s_e x_e \geq z_s k$ induce a facet of *conv*(F) for some $v \in V$? Let us take an arbitrary node $\hat{v} \in V$ and first consider the simple case where $\sum_{e \in \gamma(\hat{v})} \xi^s_e = k + 1$. In the following, we check the number of feasible and affinely independent points at which the CCkCP-degree inequality for scenario s, vertex $\hat{v}$ holds as an equality.

Let $(x, z) = (\mathbf{1}^m - u^m_e, \sum_{\ell \in \mathcal{S}_{k+1}} u^N_\ell)$, for an $e \in \gamma(\hat{v})$ with $\xi^s_e = 1$. For all $\ell \in \mathcal{S}_{k+1}$, $x = \mathbf{1}^m - u^m_e$ is a k-core in G^ℓ because if $\xi^\ell_e = 1$, the degree of $\hat{v}$ in G^ℓ under solution x is equal to k and if $\xi^\ell_e = 0$, the degree is $k + 1$. Accordingly, the degree of vertex $\hat{v}$ in G^s under solution x is k. Meanwhile, $z_s = 1$ in the solution (x, z) defined above. Therefore, (x, z) is feasible and satisfies the CCkCP-degree constraint with equality. Note that in this way we can construct $\sum_{e \in \gamma(\hat{v})} \xi^s_e$ points.

Let $(x, z) = (\mathbf{1} - u^m_e - u^m_a, \sum_{\ell \in \mathcal{S}_{k+1}} u^N_\ell)$. Here e is an edge in $\gamma(\hat{v})$ with smallest index satisfying $\xi_e^s = 1$, and $a \in E \setminus \gamma(\hat{v})$. That is, a is an arbitrary edge outside the cut of vertex $\hat{v}$. Similarly, for any $\ell \in \mathcal{S}_{k+1}$, $x = \mathbf{1}^m - u_e^m - u_a^m$ is a k-core in G^ℓ because $\delta(G^\ell) = k+1$. In this way we can construct $m - |\gamma(\hat{v})|$ feasible points.

By this direct construction, we can obtain at least $m - |\gamma(\hat{v})| + \sum\limits_{e \in \gamma(\hat{v})} \xi_e^s$ feasible and affinely independent points at which the CCkCP-degree inequality for scenario s, vertex $\hat{v}$ holds valid as an equality.

Lifting the CCkCP-degree constraints. Due to the fact described above, it appears that we can find some feasible points in F satisfying CCkCP-degree constraint at equality, but not enough of them. It suggests that the hyperplane induced by CCkCP-degree constraint supports the convex hull *conv*(F) but is not facet-inducing. With this understanding, we believe strengthening CCkCP-degree constraints by lifting will be helpful.

Given an arbitrary $\hat{s} \in \mathcal{S}$, an arbitrary node $\hat{v} \in V$, let $F^1 = F \bigcap \{(x,z) \mid z_s = 1\}$ and $F^0 = F \bigcap \{(x,z) \mid z_s = 0\}$. Indeed, $\sum\limits_{e \in \gamma(\hat{v})} \xi_e^{\hat{s}} x_e \geq k$ is valid for F^1 (not necessarily facet-inducing). Recall that according to Assumption 2, $\mathbb{P}(\boldsymbol{\xi} = \xi^s) \leq \epsilon$ and $|\mathcal{S}| > 1$. Hence $F^0 \neq \emptyset$. Based on lifting theorem (Nemhauser and Wolsey, 1999), the following is valid for F.

$$\beta z_{\hat{s}} + \sum_{e \in \gamma(\hat{v})} \xi_e^{\hat{s}} x_e \geq k + \beta \text{ where } \beta \leq \min\{\sum_{e \in \gamma(\hat{v})} \xi_e^{\hat{s}} x_e \mid (x,z) \in F^0\} - k$$

To get the bound for β, one needs to minimize $\sum\limits_{e \in \gamma(\hat{v})} \xi_e^{\hat{s}} x_e$ over exactly the same CCkCP polytope except for the only difference that corresponding sample set is now $\mathcal{S} \setminus \hat{s}$ instead of $\mathcal{S}$, which is almost as difficult as to solve CCkCP itself. But we notice that,

$$\min\{\sum_{e \in \gamma(\hat{v})} \xi_e^{\hat{s}} x_e : (x,z) \in F^0\} \geq \max\{0, \sum_{e \in \gamma(\hat{v})} \xi_e^{\hat{s}} - (|\gamma(\hat{v})| - k)\}.$$

The above inequality holds valid due to the naive observation that the maximum number of elements in $\gamma(\hat{v})$ which can take zero-value is $|\gamma(\hat{v})| - k$. Since the set of incident edges at node $\hat{v}$ in any scenario is a subset of $\gamma(\hat{v})$, the number of elements in the cut set of node $\hat{v}$ in $\hat{s}$ that can take zero-value is no more than $|\gamma(\hat{v})| - k$.

Let $\beta = \max\{0, \sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} - (|\gamma(\hat{v})| - k)\} - k$, then the lifted degree constraint is:

$$\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} x_e \geq k + (1 - z_{\hat{s}})(\max\{0, \sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} - (|\gamma(\hat{v})| - k)\} - k). \tag{6.6}$$

When $\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} \leq (|\gamma(\hat{v})| - k)$, the inequality (6.6) is equivalent to CCkCP-degree constraint. Whenever $\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} > (|\gamma(\hat{v})| - k)$, the inequality (6.6) is reduced to $\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} x_e \geq k + (1 - z_{\hat{s}})(\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} - (|\gamma(\hat{v})|)$, which dominates the original CCkCP-degree constraint. A special case is when $\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} = |\gamma(\hat{v})|$, in other words no edges in the cut set of node $\hat{v}$ fail in scenario $\hat{s}$, then the coefficient of $z_{\hat{s}}$ becomes zero, and the inequality reduces to the degree inequality for deterministic spanning k-core. Based on the lifted CCkCP-degree constraints, we obtain the following strengthened CCkCP formulation.

$$\begin{aligned}
\textbf{(DEFS)} \quad \min \quad & \sum_{e\in E} c_e x_e && (6.7a)\\
\text{s.t.} \quad & \sum_{e\in\gamma(v)} \xi_e^s x_e \geq k + (1 - z_s)(\max\{0, \sum_{e\in\gamma(v)} \xi_e^s - (|\gamma(v)| - k)\} - k), \\
& \qquad\qquad\qquad\qquad \forall v \in V,\ \xi^s \in \mathcal{S} && (6.7b)\\
& \sum_{s|\xi^s\in\mathcal{S}} z_s \mathbb{P}(\boldsymbol{\xi} = \xi^s) \geq 1 - \epsilon && (6.7c)\\
& x \in \{0,1\}^m,\ z \in \{0,1\}^N && (6.7d)
\end{aligned}$$

Note that $\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} - (|\gamma(\hat{v})| - k) > 0$ is often the case when edge failure probabilities are low and k is not too small. Also from implementation perspective, the sign of $\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} - (|\gamma(\hat{v})| - k)$ can be decided before building MIP model, making program (6.7) a better alternative to the original CCkCP program (6.4).

Proposition 11 *Given that the support graph of $\tilde{G} = (V, \tilde{E})$ where $|V| = n > 3$ is complete and $k = n - 2$, the lifted degree constraint (6.6) induces a facet of conv(F) if $\sum_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} = k$ and $\sum_{e\in\gamma(v)} \xi_e^{\hat{s}} > k$, $\forall v \in V$ and $v \neq \hat{v}$.*

Proof of Proposition 11. As the support graph is complete, $|\gamma(\hat{v})| = n-1$ and $|\gamma(\hat{v})| - k = 1$. Hence, $\sum\limits_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} - (|\gamma(\hat{v})| - k) = k-1 > 0$. The constraint (6.6) is now $\sum\limits_{e\in\gamma(\hat{v})} \xi_e^{\hat{s}} x_e \geq k-(1-z_{\hat{s}})$. The feasible point $(x,z) = (\mathbf{1}^m, \mathbf{1}^N - u_{\hat{s}}^N)$ satisfies (6.6) under given conditions with strict inequality, indicating that (6.6) is not an implicit equation. We next show that there are $|E_D| + N$ feasible and affinely independent points where constraint (6.6) holds as an equation. These points are $(\mathbf{1}^m, \mathbf{1}^N)$, $(\mathbf{1}^m, \mathbf{1}^N - u_\ell^N)$ for every $\ell \in \mathcal{S} \setminus \{\hat{s}\}$, and $(\mathbf{1}^m - u_e^m, \sum\limits_{\ell\in\mathcal{S}\setminus\mathcal{D}_e} u_\ell^N)$ for every $e \in E_D$. We elaborate on the feasibility of the last $|E_D|$ points next. If $e \notin \gamma(\hat{v})$, or if $e \in \gamma(\hat{v})$ and $\xi_e^{\hat{s}} = 0$, $\hat{s} \notin \mathcal{D}_e$ and $z_{\hat{s}} = 1$. Then, the left-hand side of (6.6) equals the right-hand side, which is k. If $e \in \gamma(\hat{v})$, $\hat{s} \in \mathcal{D}_e$ and $z_{\hat{s}} = 0$, the left-hand-side of (6.6) equals the right-hand-side, which is $k-1$. ■

6.4.4 Other Valid Inequalities

Let $b_v^s = \sum\limits_{e\in\gamma(v)} \xi_e^s - k$. According to Assumption 1, b_v^s is nonnegative integer for any $\xi^s \in \mathcal{S}, v \in V$. We further do the variable substitution $x_e = 1 - y_e$, $\forall e \in E$, $y_e \in \{0,1\}$, then it follows that

$$\begin{aligned}
\mathcal{P}_s \quad & := \{y \in \{0,1\}^m \mid \sum_{e\in\gamma(v)} (1-y_e)\xi_e^s \geq k, \forall v \in V\} \\
& := \{y \in \{0,1\}^m \mid \sum_{e\in\gamma(v)} y_e \xi_e^s \leq \sum_{e\in\gamma(v)} \xi_e^s - k, \forall v \in V\} \\
& := \{y \in \{0,1\}^m \mid \sum_{e\in\gamma(v)} y_e \xi_e^s \leq b_v^s, \forall v \in V\}
\end{aligned}$$

Note that if $b_v^s = 0$, the corresponding degree constraint implies $\xi_e^s y_e = 0$ for all $e \in \gamma(v)$, thus $\mathcal{P}_s$ is not full-dimensional. Now let us assume $\mathcal{P}_s$ is full-dimensional, i.e. b_v^s, $\forall v \in V$ is positive. Then $\mathcal{P}_s$ is the general 1-capacitated b-matching feasible solution set whose convex hull, as shown in (Edmonds, 1965) is given by

$$\begin{aligned}\sum_{e\in\gamma(v)} y_e\xi_e^s &\leq b_v^s,\ v\in V && \text{(6.8a)}\\ \sum_{e\in E(W)}\xi_e^s y_e + \sum_{e\in F}\xi_e^s y_e &\leq \frac{1}{2}(\sum_{v\in W} b_v^s + \sum_{e\in F}\xi_e^s - 1), &&\\ & W\subset V,\ F\subset\gamma(W) \text{ with } \sum_{v\in W} b_v^s + \sum_{e\in F}\xi_e^s \text{ odd} && \text{(6.8b)}\\ & \mathbf{0}^m \leq y \leq \mathbf{1}^m && \text{(6.8c)}\end{aligned}$$

where $E(W)\subseteq E$ is defined as the set of edges with both end nodes in W and $\gamma(W)$ is the set of edges with only one end in W. The convex hull of $\mathcal{P}_s$ is characterized by degree constraints (6.8a), blossom constraints (6.8b), and bounds (6.8c). Specifically the blossom constraints grow exponentially with support graph size n, i.e., for $b_v^s = 1,\ \forall v\in V,\ \xi^s\in\mathcal{S}$, the number of odd sets is $O(2^{n-1})$. Pulleyblank (1973) pointed out that the inequalities set of (6.8) is not minimal and studied the unique minimal subset of (6.8) defining $conv(\mathcal{P}_s)$, given that $conv(\mathcal{P}_s)$ is full-dimensional.

Proposition 12 *The following inequalities are valid inequalities for* $conv(F)$

CCkCP-blossom:

$$\sum_{e\in E(W)\cup F}\xi_e^s x_e \geq z_s\{\sum_{e\in E(W)\cup F}\xi_e^s - \frac{1}{2}(\sum_{v\in W} b_v^s + \sum_{e\in F}\xi_e^s - 1)\},$$

$$W\subset V,\ F\subset\gamma(W)\ \textit{with}\ \sum_{v\in W} b_v^s + \sum_{e\in F}\xi_e^s\ \textit{odd.}$$

Proof of Proposition 12. These inequalities are valid for $conv(F)$ as they are equivalent to the corresponding b-matching blossom inequalities for the scenario solution. In addition, for these inequalities to be valid, b_v^s, $\forall v\in V$, $\xi^s\in\mathcal{S}$ does not have to be positive. ■

6.5 Computational Experience

The goal of our computational study in this section is to evaluate the merits of the strengthened formulation (6.7) as opposed to the direct deterministic equivalent reformulation (6.4).

We conduct computational experiments on instances with $|V| = 10, 30, 50, 100$. Two levels of minimum vertex degree in support graph were considered for $|V| = 30, 50$, and 100; namely $\delta(G) = 29$ and 20 for $|V| = 30$, $\delta(G) = 49$ and 35 for $|V| = 50$, and $\delta(G) = 99$ and 65 for $|V| = 100$. In other words, we consider a complete support graph and a less dense support graph for each $|V|$ considered. According to Corollary 1, we study the case where $r = 2$ and $k = \lceil \frac{n}{2} \rceil$ to design 2-connected diameter-2 networks that preserve their diameter upon vertex deletion. The edge failure probabilities and edge costs are set in the same manner as described in Section 5.6.1.

We randomly generate equally likely samples according to the edge failure probabilities for each instance. The number of scenarios is varied in our experiments from $|\mathcal{S}| = 100$ to 5,000. For a given number of vertices and number of scenarios, we generate and test 5 replications (or 5 different sets of scenarios). We also impose a 1-hour time limit for each replication, for each instance. We report either average solution time based on the replications that solved to optimality or average optimality gap based on the replications in which a feasible solution is returned under the time limit.

All experiments are conducted on a 64-bit Linux system with eight Intel® Xeon® E5620 2.40GHz processors and 96GB RAM. Gurobi® Optimizer v5.5 is used as the MILP solver. Both DEF and DEFS are implemented in C++. All implementations inherited the default settings for branching, node selection, general purpose cutting planes, preprocessing and heuristics. The Gurobi® parameter *GRB_IntParam_Threads*, number of threads used by the parallel MILP solver, is set to its default value 0, which means the thread count is equal to the number of logical cores in the machine, which is eight in our case.

Tables 8, 9, 10, and 11 present computational results comparing the average solution times or optimality gaps of DEF against DEFS. Two key observations we can make from the tables are as follows: (1) DEF performs consistently poorer than DEFS as the average solution time/optimality gap from the latter is much smaller. For example, the 50-vertex graph with $\delta(G) = 49$ as shown in Table 10, when the number of samples is equal to 5000, the optimality gap of DEF is 80.0% while that of DEFS is 23.3%. (2) Additionally, one can also observe that the more dense the support graph is, the more pronounced the advantage of DEFS over DEF becomes. Take the instance of the 30-vertex graph with 5000 scenarios in Table 9 as an example. When $\rho(G) = 1$ (recall that $\rho(G)$ denote the edge density of graph G), the average optimality gap is reduced by 19.3% using DEFS than using DEF. On the other hand, when $\rho(G) = 0.896$, the reduction is 13.9%.

Table 8: Results on a 10-vertex complete graph with $k = 5$, $\epsilon = 0.2$, and time limit = 1 hour. Average over 5 replications is reported.

	DEF		DEFS	
$\lvert\mathcal{S}\rvert$	Time/Gap	#BC Nodes	Time/Gap	#BC Nodes
100	556.37 sec	1059370	12.99 sec	9677
250	4.05%	1368782	184.91 sec	69116
500	7.21%	304885	878.11 sec	139760
750	8.71%	83171	2027.88[a] sec	179497[a]
1000	10.74%	43597	1.92%	112041
2500	17.47%	24000	5.69%	41345
5000	21.56%	1911	8.81%	1661

[a] 4 out of 5 instances were solved to optimality within 1-hour time limit and this average solution time was calculated based on these 4 instances. The 5th instance was only solved to feasibility with a gap of 1.69%.

Table 9: Results on 30–vertex graphs with different edge densities $\rho(G)$ and graph degrees $\delta(G)$. Parameters $k = 15$, $\epsilon = 0.2$, and time limit = 1 hour. Average solution time/optimality gap over 5 replications is reported.

	$\rho(G) = 1$, $\delta(G) = 29$		$\rho(G) = 0.896$, $\delta(G) = 20$	
$\|\mathcal{S}\|$	DEF	DEFS	DEF	DEFS
100	12.3%	5.2%	1.9%	587.67 sec
250	20.2%	9.8%	11.3%	5.9%
500	24.5%	11.7%	13.2%	7.5%
750	25.9%	12.4%	15.3%	8.0%
1000	26.5%	12.8%	15.5%	8.2%
2500	31.1%	16.8%	19.3%	11.4%
5000	40.0%	20.7%	26.5%	12.6%

Table 10: Results on 50-vertex graphs with different edge densities $\rho(G)$ and graph degrees $\delta(G)$. Parameters $k = 25$, $\epsilon = 0.2$, and time limit = 1 hour. Average solution time/optimality gap over 5 replications is reported.

	$\rho(G) = 1$, $\delta(G) = 49$		$\rho(G) = 0.805$, $\delta(G) = 35$	
$\|\mathcal{S}\|$	DEF	DEFS	DEF	DEFS
100	14.0%	5.8%	12.4%	4.7%
250	23.3%	10.6%	21.2%	9.4%
500	26.1%	13.2%	23.1%	11.2%
750	29.6%	13.8%	24.5%	11.9%
1000	35.7%	14.9%	26.7%	12.5%
2500	62.0%	17.7%	50.8%	14.7%
5000	80.0%	23.3%	27.6%	20.9%

Table 11: Results on 100-vertex graphs with different edge densities $\rho(G)$ and graph degrees $\delta(G)$. Parameters $k = 50$, $\epsilon = 0.2$, and time limit = 1 hour. Average solution time/optimality gap over 5 replications is reported.

	$\rho(G) = 1$, $\delta(G) = 99$		$\rho(G) = 0.758$, $\delta(G) = 65$	
$\lvert\mathcal{S}\rvert$	DEF	DEFS	DEF	DEFS
100	35.9%	8.9%	20.5%	5.3%
250	44.2%	11.6%	27.2%	6.5%
500	55.3%	15.2%	36.4%	9.1%
750	55.9%	17.1%	40.3%	9.9%
1000	56.6%	18.3%	41.7%	11.6%
2500	-	-	-	-
5000	-	-	-	-

-: Failed to solve the LP relaxation within 1-hour time limit at root nodes.

CHAPTER 7

SERVICE SYSTEM CAPACITY PLANNING UNDER DEMAND UNCERTAINTY

Various risk measures have been developed to capture potential loss depending on problem specific features and service provider's risk preferences, as discussed in Section 3.3. In this chapter, one of the most commonly used risk measures, mean outcome, is adopted due to its ease of interpretation and popularity in applications. Additionally, among model parameters, we assume only the demand rate is uncertain throughout this chapter. We chose the commonly used system performance measure–the average time a customer/transaction stays in the system for our study. In Section 7.1, we present our study on the capacity planning problem of a facility abstracted by a single-stage service system modeled by an M/M/1 queue. In Section 7.2, we extend our study by investigating the capacity planning problem in a two-stage service system.

7.1 Capacity Planning in a Single-Stage Service System

To begin with, a mathematical model is developed for the capacity planning problem in a single-stage service system modeled by an M/M/1 queue. We analytically derive optimal service rates for the model under the assumption that arrival rates vary uniformly over a specified range. As managing information uncertainty is of significant interest to practitioners, we next illustrate the concepts of *cost of uncertainty in arrival rate* and *minimum level of information quality* via numerical experiments.

7.1.1 Problem Formulation and Analytical Solutions

We assume that the overall design process is as follows: the range for an uncertain demand rate is given, while the costs and other performance requirements are known with certainty; the decision maker determines a service rate; the facility is then operated with a realized demand rate according to an M/M/1 model; and the business performance is finally measured for the realized model. To proceed, we first introduce the notations used.

$\boldsymbol{\lambda}$	transaction arrival rate (modeled as a random variable)
$[a, b]$	range over which $\boldsymbol{\lambda}$ is assumed to be uniformly distributed
μ	transaction service rate (transactions per unit time)
T	average time a transaction spends in the system
$\hat{T}$	upper bound of T promised in SLA (unit time)
θ	upper bound of server utilization, $\theta \in (0, 1)$
$\hat{\lambda}$	a realization of $\boldsymbol{\lambda}$
c	cost per unit increase in transaction service rate

Note that by assuming a uniform distribution for the transaction arrival rate over the specified range, we are implying that we have no information besides its rage and it is equally likely for the rate to take any value in the range.

All costs are measured in terms of accounting time unit (e.g., a month). Arrival rate and service rate are measured in terms of operational time unit (e.g., an hour). The design decision is service capacity μ which is the decision variable in our optimization problem. Recall that in our problem setting the performance measure is average time in system T, which is an often used performance metric in practice. Specifically, T has to satisfy the following constraint.

$$T = \frac{1}{\mu - \hat{\lambda}} \leq \hat{T} \tag{7.1}$$

Depending on the realizations of transaction arrival rate, $\hat{\lambda}$, SLA may be violated for a chosen μ. Whenever constraint (7.1) is violated for a given combination $(\hat{\lambda}, \mu)$, we assume that a penalty is imposed on the service provider, according to his/her agreement with the client firm. The definition of penalty functions depends on specific business settings. For example, in the case of a high performance computing cluster, the penalty is defined as the rate for compensating the user for failure to meet the deadline (Chee Shin and Buyya, 2005). In e-commerce, a fixed charge could be used as penalty whenever the response time to a customer request exceeds a prescribed threshold (Liu et al., 2001). A brief review of typical penalty functions is presented in (Kosinski et al., 2008).

Following one of the basic penalty types discussed in (Kosinski et al., 2008), we define a stepwise penalty function based on the SLA requirement and the utilization requirement. Since T is a function of the uncertain parameter $\boldsymbol{\lambda}$ and the decision variable μ, the penalty imposed is a function of $\boldsymbol{\lambda}$ and μ as well. We denote it by $f(\boldsymbol{\lambda}, \mu)$. For a given arrival rate realization $\hat{\lambda}$ and a given design decision μ, the penalty function is defined as follows.

$$f(\hat{\lambda}, \mu) = \begin{cases} 0, \ \text{if SLA is satisfied and utilization is no more than } \theta; \\ H_1, \ \text{if SLA is violated but utilization is no more than } \theta; \\ H_2, \ \text{if utilization is greater than } \theta. \end{cases} \tag{7.2}$$

By the definition of penalty function in (7.2), when system performance measure T satisfies SLA constraint (7.1), the penalty function value is zero. When T is greater than $\hat{T}$ but utilization requirement is satisfied, a penalty of amount H_1 will be imposed on the service provider. If for some realizations of $\boldsymbol{\lambda}$, the utilization requirement is violated for a given design decision, a penalty of H_2 will be charged. From a modeling point of view, SLA requirement should be stricter than utilization requirement. Hence, the following condition should hold through an appropriate selection of the θ value during model development phase.

Assumption 3 *For any $\hat{\lambda} \in [a, b]$, there does not exist a feasible capacity decision such that SLA requirement is satisfied but utilization requirement is violated.*

It should also be noted that $H_1 << H_2$, i.e., the penalty due to utilization requirement violation far exceeds that due to SLA violation. We consider the M/M/1 model for the single-stage service system in this subsection. By Assumption 3, $\nexists \mu \in \mathbb{R}$ such that $\frac{1}{\mu - \hat{\lambda}} \leq \hat{T}$ and $\frac{\hat{\lambda}}{\mu} > \theta$ for any $\hat{\lambda}$ in the range of $[a, b]$. It further implies that $\hat{\lambda} + 1/\hat{T} \geq \hat{\lambda}/\theta$ for any $\hat{\lambda}$. Hence, $\theta \geq \frac{b\hat{T}}{1+b\hat{T}}$, i.e., a minimum value should be imposed on θ following Assumption 3. Intuitively, when such a minimum value approaches 1, i.e., the value of θ approaches 1 such that utilization requirement approximately reduces to system stability requirement, SLA requirement is certainly stronger than utilization requirement (system stability requirement).

Suppose the input parameters are $\hat{T} = 0.2$ and the uncertain parameter $\boldsymbol{\lambda} \sim U[0, 25]$, where $U[a, b]$ denotes the uniform distribution over the interval $[a, b]$. By Assumption 3, $\theta \geq 0.2 \times 25/(1 + 0.2 \times 25) = 0.83$. We let $\theta = 0.85$. For this specific numerical example, a graphical representation of the penalty function (7.2) is presented in Figure 6. The horizontal axis denotes realizations of the uncertain parameter $\boldsymbol{\lambda}$. The vertical axis denotes different decisions of system capacity. The solid line outlines the area where utilization requirement is satisfied (above the solid line) and the area where utilization requirement is violated (below the solid line). The dashed line outlines the region where SLA is satisfied (above) and the region where SLA is violated (below). The dotted line outlines the region where Assumption 3 holds valid (left) and the region where the assumption is violated (right). Values of penalty for each region are also labeled in Figure 6.

The objective of the service system design problem is to identify the optimal capacity μ such that the overall service cost and expected penalty is minimized.

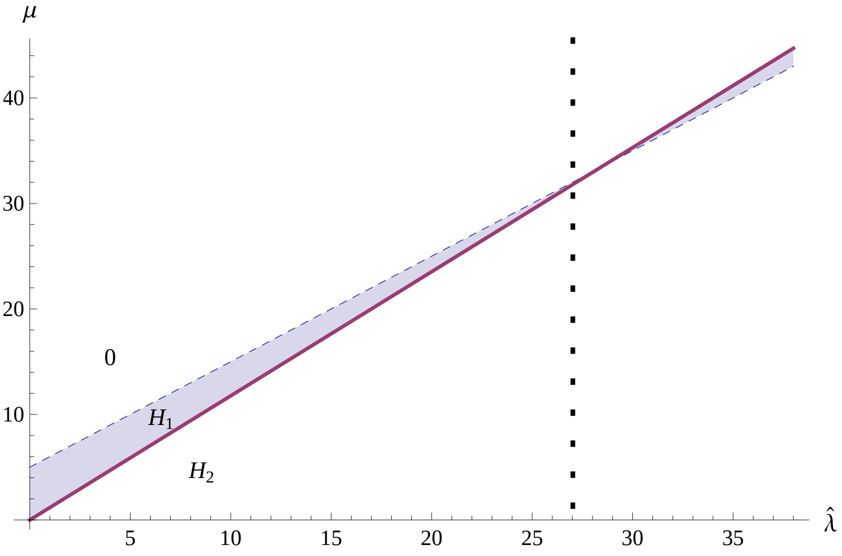

Figure 6: An illustration of the penalty function.

Therefore, the mathematical model can be described as follows in general.

$$\min_{\mu \in \mathbb{R}_+} \quad c\mu + \mathbb{E}[f(\boldsymbol{\lambda}, \mu)] \tag{7.3}$$

Without loss of generality, we assume $c = 1$ because one can always normalize other cost parameters (i.e., H_1 and H_2) based on service cost rate c. Optimal solutions for model (7.3) for an appropriately chosen θ value are presented in the following proposition.

Proposition 13 *Given the design optimization problem modeled by formulation (7.3), the optimal capacity is achieved at one of the following points:* 0, a/θ, b/θ, $a + 1/\hat{T}$, *and* $b + 1/\hat{T}$.

Proof of Proposition 13. Since $\boldsymbol{\lambda}$ is uniformly distributed in the interval [a,b] and

$$\mathbb{E}[f(\boldsymbol{\lambda}, \mu)] = H_1 \mathbb{P}\{1/(\mu - \boldsymbol{\lambda}) > \hat{T} \text{ and } \boldsymbol{\lambda} < \theta\mu\} + H_2 \mathbb{P}\{\boldsymbol{\lambda} \geq \theta\mu\},$$

it follows that $\mathbb{E}[f(\boldsymbol{\lambda}, \mu)]$ can be expressed as a piecewise linear function w.r.t. μ as in Equations (7.4) and (7.5) respectively depending on the value of θ.

Case 1: $\frac{b\hat{T}}{1+b\hat{T}} \leq \theta < \min\{1, \frac{b\hat{T}}{1+a\hat{T}}\}$

$$\mathbb{E}[f(\boldsymbol{\lambda}, \mu)] = \begin{cases} H_2, \quad 0 \leq \mu < \frac{a}{\theta}; \\ \frac{b-\theta\mu}{b-a} H_2 + \frac{\theta\mu - a}{b-a} H_1, \quad \frac{a}{\theta} \leq \mu < a + \frac{1}{\hat{T}}; \\ \frac{b-\theta\mu}{b-a} H_2 + \frac{\theta\mu - (\mu - 1/\hat{T})}{b-a} H_1, \quad a + \frac{1}{\hat{T}} \leq \mu < \frac{b}{\theta}; \\ \frac{b-(\mu-1/\hat{T})}{b-a} H_1, \quad \frac{b}{\theta} \leq \mu < b + \frac{1}{\hat{T}}; \\ 0, \quad b + \frac{1}{\hat{T}} \leq \mu. \end{cases} \tag{7.4}$$

Case 2: $\min\{1, \frac{b\hat{T}}{1+a\hat{T}}\} \leq \theta < 1$

$$\mathbb{E}[f(\boldsymbol{\lambda}, \mu)] = \begin{cases} H_2, \quad 0 \leq \mu < \frac{a}{\theta}; \\ \frac{b-\theta\mu}{b-a} H_2 + \frac{\theta\mu - a}{b-a} H_1, \quad \frac{a}{\theta} \leq \mu < \frac{b}{\theta}; \\ H_1, \quad \frac{b}{\theta} \leq \mu < a + \frac{1}{\hat{T}}; \\ \frac{b-(\mu-1/\hat{T})}{b-a} H_1, \quad a + \frac{1}{\hat{T}} \leq \mu < b + \frac{1}{\hat{T}}; \\ 0, \quad b + \frac{1}{\hat{T}} \leq \mu. \end{cases} \tag{7.5}$$

Accordingly, a closed-form expression for the objective function in formulation (7.3) can be obtained as well by adding the service cost term $c\mu$ to the above equations, which is thus piecewise linear. Depending on the values of parameters a, b, $\hat{T}$, H_1, and H_2, the optimal solution of model (7.3) will be one of the five break points: 0, a/θ, b/θ, $a + 1/\hat{T}$, and $b + 1/\hat{T}$. This completes the proof. ■

According to Proposition 13, one can simply plug in the five candidates into model (7.3) and obtain the optimal solution in a straightforward manner. It is worth noting that when the optimum is achieved at $\mu = 0$, the business solution is equivalent to "do nothing". In addition, complete look-up tables for the optimal solutions and optimal objective values for the service system design problem formulated as (7.3) can be analytically derived. We present the results in Table 12 and Table 13 for Case

1 and Case 2 respectively. Furthermore, let

$$\beta = min\{1, \frac{b - (b/\theta - 1/\hat{T})}{b - a}\}.$$

We can combine Table 12 and Table 13 and provide a consolidated Table 14 using auxiliary parameter β, which serves as a clearly-defined quantitative guideline for practitioners to effectively determine the optimal solutions using given parameter.

Table 12: Optimal solutions for formulation (7.3) when $\frac{b\hat{T}}{1+b\hat{T}} \le \theta < \min\{1, \frac{b\hat{T}}{1+a\hat{T}}\}$.

Parameter conditions	Optimal solution	Optimal objective
$H_2 \le \frac{b-(b/\theta-1/\hat{T})}{b-a} H_1 + b/\theta$ and $H_2 \le b + 1/\hat{T}$	0	H_2
$H_2 \ge \frac{b-(b/\theta-1/\hat{T})}{b-a} H_1 + b/\theta$ and $b + 1/\hat{T} \ge \frac{b-(b/\theta-1/\hat{T})}{b-a} H_1 + b/\theta$	b/θ	$b/\theta + \frac{b-(b/\theta-1/\hat{T})}{b-a} H_1$
$H_2 \ge b+1/\hat{T}$ and $\frac{b-(b/\theta-1/\hat{T})}{b-a} H_1 + b/\theta \ge b + 1/\hat{T}$	$b + 1/\hat{T}$	$b + 1/\hat{T}$

Table 13: Optimal solutions for formulation (7.3) when $\min\{1, \frac{b\hat{T}}{1+a\hat{T}}\} \le \theta < 1$.

Parameter conditions	Optimal solution	Optimal objective
$H_2 \le H_1 + b/\theta$ and $H_2 \le b + 1/\hat{T}$	0	H_2
$H_2 \ge H_1 + b/\theta$ and $b + 1/\hat{T} \ge H_1 + b/\theta$	b/θ	$H_1 + b/\theta$
$H_2 \ge b + 1/\hat{T}$ and $H_1 + b/\theta \ge b + 1/\hat{T}$	$b + 1/\hat{T}$	$b + 1/\hat{T}$

In the next two subsections, we use the model developed and its optimal solutions to illustrate two concepts to help manage information uncertainty when determining the capacity of a service system. First, we use a baseline system with no uncertainty in the input parameters, i.e., every parameter of the system is known with a single value.

Table 14: Optimal solutions for formulation (7.3).

Parameter conditions	Optimal solution	Optimal objective
$H_2 \leq \beta H_1 + b/\theta$ and $H_2 \leq b + 1/\hat{T}$	0	H_2
$H_2 \geq \beta H_1 + b/\theta$ and $b + 1/\hat{T} \geq \beta H_1 + b/\theta$	b/θ	$\beta H_1 + b/\theta$
$H_2 \geq b + 1/\hat{T}$ and $\beta H_1 + b/\theta \geq b + 1/\hat{T}$	$b + 1/\hat{T}$	$b + 1/\hat{T}$

We can optimize the baseline system according to a given cost objective. Similarly we analyze the system with some uncertainty in the input parameters. The difference in the optimal costs of the baseline and the uncertain system is the cost of uncertainty. Uncertainty in the input parameters represents the quality of information we have at design time. Often the quality of information can be improved, for example, by investing in data collection efforts; the issue then is how much investment should be made. Our analysis below provides some insights through a concept of minimum level of information quality.

7.1.2 Cost of Uncertainty in Arrival Rate

We analyze the impact on the optimal objective function values when the range of the arrival rate $\boldsymbol{\lambda}$ varies. First, take a numerical example where $\hat{T} = 0.2$. Suppose the cost parameters are $H_1 = 7$ and $H_2 = 150$, i.e., H_1 is seven times and H_2 is 150 times the marginal service cost. We vary a in the range of $[0, 25]$, b in the range of $[1, 50]$, imposing the condition that $b \geq a + 1$. In other words, we consider a minimum interval length of 1 and a maximum interval length of 50. By Assumption 3, $\theta \geq \frac{b\hat{T}}{1+b\hat{T}}, \forall b \in [1, 50]$. Hence, $\theta \geq \frac{50\hat{T}}{1+50\hat{T}} = 0.91$. Graphs of the optimal cost versus a and b are shown in Figure 7 for $\theta = 0.95$ (left) and $\theta = 0.99$ (right) respectively.

Three straightforward observations from Figure 7 are as follows: (1) For a fixed a the optimal cost increases as the value of b increases because of higher average service

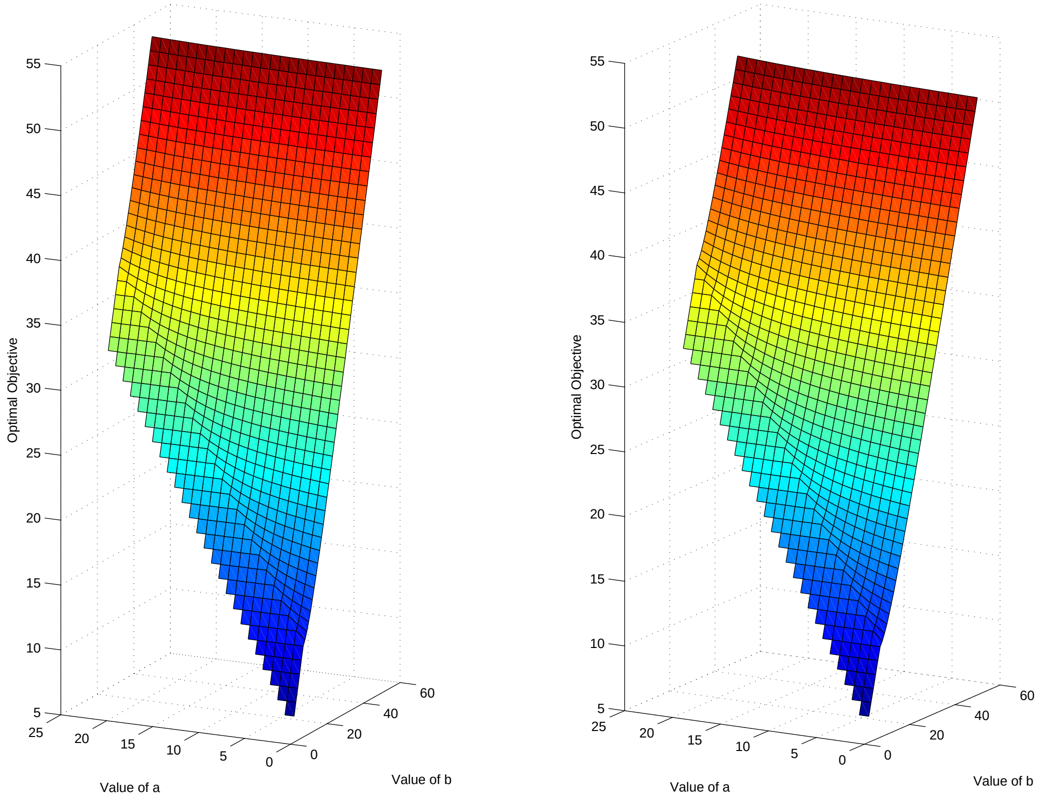

Figure 7: A graphical presentation of optimal design costs for different ranges of arrival rate when $\hat{T} = 0.2$, $c = 1$, $H_1 = 7$, $H_2 = 150$, and $\theta = 0.95$ (left) or $\theta = 0.99$ (right).

load and a larger degree of uncertainty in the arrival rate. (2) For a fixed b, as the value of a increases, the optimal service cost increases as well due to a higher average service load despite a lower degree of uncertainty in the arrival rate. This shows that uncertainty is a second order effect when compared to the first order effect of server utilization. However, the rate of increase is much lower than that in case (1) because of the offset from lower degree of uncertainty. (3) For the same a and b, the objective value is larger when $\theta = 0.95$ as opposed to $\theta = 0.99$. In other words, the more averse a service provider is towards high utilization rates, the more it costs to design the system.

For a more detailed examination, the optimal objective function values of selected cases of $[a, b]$ when $\theta = 0.95$ and 0.99 (a slice of the graphs in Figure 7) are shown in Tables 15 and 16 respectively. All cases there have the same mean arrival rate (midpoint of the range) but have different variability levels. The bottom row shows the case of a fixed arrival rate, with zero uncertainty. The column "cost of uncertainty in arrival rate" represents the difference in objective value of the row from the bottom row of zero uncertainty. As can be seen, for instance, in Table 15, when the length of the interval $[a, b]$ decreases, the optimal solution μ^* and the corresponding objective function value are decreasing, though relatively slowly. When the range length drops from 50 to 2, a decrease of 96%, the cost saving is 22.96, or 43%. The increase in cost savings observed with decreasing uncertainty in the arrival rate is very encouraging as higher service system capacities could involve a high capital investment, although improving information quality could be quite challenging in practice.

7.1.3 Minimum Level of Information Quality

Continuing the above line of thought, when will the level of uncertainty in arrival rate (as represented by the range $[a, b]$) be too high to handle? Conceptually, when the level of uncertainty becomes very high, we will be better off by just paying the

Table 15: A numerical presentation of optimal design costs for different ranges of arrival rate when $\hat{T} = 0.2$, $c = 1$, $H_1 = 7$, $H_2 = 150$, and $\theta = 0.95$.

a	b	Range length	Optimal solution (μ^*)	Optimal objective	Cost of uncertainty in arrival rate
0	50	50	52.63	52.96	22.96
4	46	42	48.42	48.85	18.85
8	42	34	44.21	44.78	14.78
12	38	26	40.00	40.81	10.81
16	34	18	35.79	37.04	7.04
20	30	10	31.58	33.97	3.97
24	26	2	31.00	31.00	1.00
25	25	0	30.00	30.00	0.00

Table 16: A numerical presentation of optimal design costs for different ranges of arrival rate when $\hat{T} = 0.2$, $c = 1$, $H_1 = 7$, $H_2 = 150$, and $\theta = 0.99$.

a	b	Range length	Optimal solution (μ^*)	Optimal objective	Cost of uncertainty in arrival rate
0	50	50	50.51	51.13	21.13
4	46	42	46.46	47.22	17.22
8	42	34	42.42	43.37	13.37
12	38	26	38.38	39.63	9.63
16	34	18	34.34	36.15	6.15
20	30	10	30.30	33.59	3.59
24	26	2	31.00	31.00	1.00
25	25	0	30.00	30.00	0.00

penalty of not satisfying the SLA in every time period rather than operating the service system from a service provider's perspective. But this is not a realistic option as no customers will be served. Such a minimum level of information quality can be calculated precisely from our model, an example of which is as follows.

Consider the single-stage model (7.3). It follows from Table 14 that when $H_2 \leq b/\theta + \beta H_1$ and $H_2 \leq b + 1/\hat{T}$ where $\beta = \min\{1, \frac{b-(b/\theta-1/\hat{T})}{b-a}\}$, the optimal solution is $\mu^* = 0$. That is, the model tells us not to operate the service system and instead pay the penalty of not satisfying the SLA in every time period. This represents the boundary condition of $[a, b]$ under which the optimal decision is "do nothing". Strictly speaking, one should have another penalty scale H_3 that accounts for the lost opportunity of conducting business. However, when the utilization requirement is violated, some transactions may have extremely long waiting times, so one may argue that this is equivalent to not being in business. In any case, the concept illustrated in this section is applicable even with an H_3 penalty.

Again take the numerical example where $\hat{T} = 0.2$, $H_1 = 7$, and $H_2 = 150$, we plot the region of a and b such that $\mu^* = 0$ when $\theta = 0.99$ in Figure 8.

From Figure 8, for a given a, when b exceeds a certain threshold, the optimal decision is "do nothing". Meanwhile, for a given b, when a is smaller than a certain value, it is optimal to do nothing. The shape of the region is complex due, in part, to the compounding effect of utilization and uncertainty when a and/or b varies. If the design requirements happen to fall in this region of "do nothing," it means that the information given is subject to high uncertainty; either we attempt to get better information or we consider rejecting the business proposal. The boundary of this region defines the minimum level of information quality (for the arrival rate in this case) required for a sustainable business operation.

It is worth noting that for the same numerical example described in this subsection, when the value θ is set to be 0.95 instead of 0.99, the set of combinations of a and b

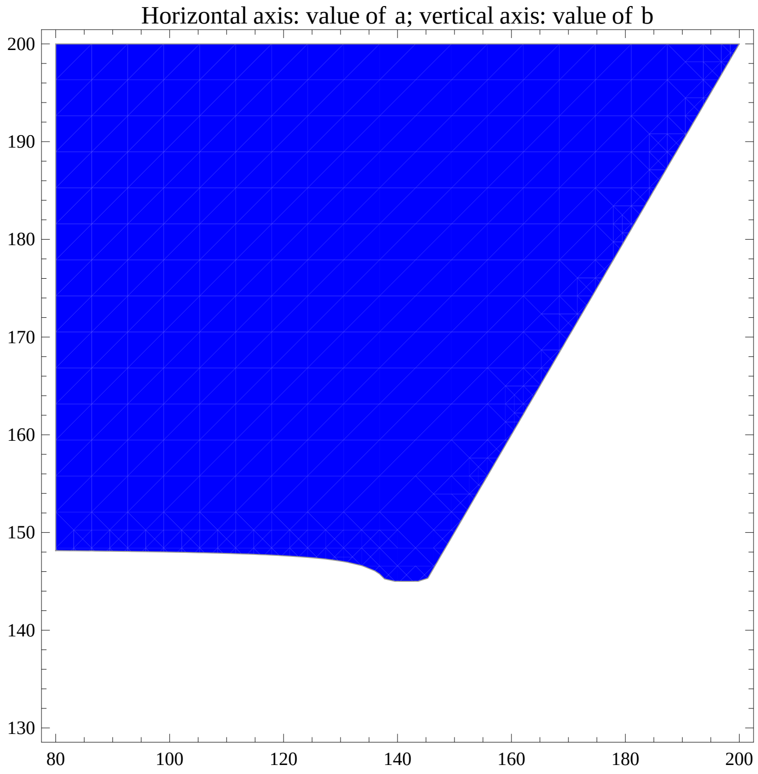

Figure 8: Shaded region defines a and b such that $\mu^* = 0$ when $\hat{T} = 0.2$, $c = 1$, $H_1 = 7$, $H_2 = 150$, and $\theta = 0.99$.

such that our solution is “do nothing” (i.e., $\theta \geq b\hat{T}/(1+b\hat{T})$ and $\mu^* = 0$) is empty. It indicates that when the service provider wants to impose a stricter control on the system utilization, it becomes less likely that the business solution is “do nothing” when all cost parameters remain the same.

Although this analysis is dependent on the structure of the penalty function $f(\boldsymbol{\lambda}, \mu)$, the concept of minimum level of information quality as defined this way is generally applicable and will be useful regardless of the structure of the penalty function.

7.2 Capacity Planning in a Two-Stage Service System

Multi-stage service systems that can be modeled as queuing networks, pose bigger challenges for researchers/practitioners in determining the server capacity at each stage. Due to their widespread applications in practice, we present our study for a simple multi-stage system. We consider the special case of a tandem configuration with two stages and a single external arrival process in this section.

7.2.1 Problem Formulation and Analytical Solutions

The tandem line system with two single-server nodes is depicted in Figure 9. To describe the model formulation, we first introduce some additional notations. Let μ_1 and μ_2 denote the transaction service rate at server 1 and 2 respectively; c_1 and c_2 denote the cost per unit increase in service rate at server 1 and 2 respectively.

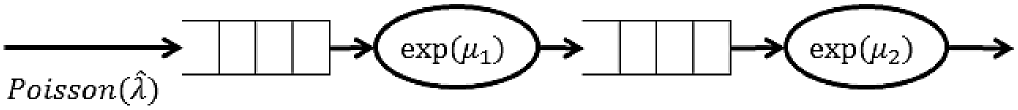

Figure 9: A two-stage tandem line system.

Accordingly, the SLA requirement and penalty function are

$$T = \frac{1}{\mu_1 - \hat{\lambda}} + \frac{1}{\mu_2 - \hat{\lambda}} \leq \hat{T} \tag{7.6}$$

and

$$f(\hat{\lambda}, \mu_1, \mu_2) = \begin{cases} 0, \text{ if SLA is satisfied and } \hat{\lambda}/\mu_1 \leq \theta \text{ and } \hat{\lambda}/\mu_2 \leq \theta; \\ H_1, \text{ if SLA is violated and } \hat{\lambda}/\mu_1 \leq \theta \text{ and } \hat{\lambda}/\mu_2 \leq \theta; \\ H_2, \text{ if } \hat{\lambda}/\mu_1 > \theta \text{ and/or } \hat{\lambda}/\mu_2 > \theta. \end{cases} \tag{7.7}$$

For this two-stage tandem line system, it follows from Assumption 3 that $\nexists(\mu_1, \mu_2) \in \mathbb{R}^2_+$ such that $1/(\mu_1 - \hat{\lambda}) + 1/(\mu_2 - \hat{\lambda}) \leq \hat{T}$ and $\hat{\lambda}/\mu_1 > \theta$ and/or $\hat{\lambda}/\mu_2 > \theta$ given any $\hat{\lambda}$ in the range of $[a, b]$. Similar to the case of single-stage system, we can derive that the value of θ should be chosen such that $\theta \geq (b\hat{T})/(1 + b\hat{T})$. To identify the optimal service rate such that the summation of penalty and service cost is minimized on the average, we need to solve the following optimization model:

$$\min_{(\mu_1,\mu_2)\in\mathbb{R}^2_+} c_1\mu_1 + c_2\mu_2 + \mathbb{E}[f(\boldsymbol{\lambda}, \mu_1, \mu_2)]. \tag{7.8}$$

Lemma 1 *Given a tandem line system consisting of two single server queues with exponential service time, an external Poisson arrival rate $\hat{\lambda}$, and total capacity M ($M = \mu_1 + \mu_2$) where $M > 2\hat{\lambda}$, in order to minimize the average time a transaction spends in the system, the optimal capacity allocation is $\mu_1^* = \mu_2^* = \frac{M}{2}$.*

Proof of Lemma 1. In order to satisfy stability conditions at both servers, a feasible solution (μ_1, μ_2) must satisfy the conditions that $\mu_1 > \hat{\lambda}$ and $\mu_2 > \hat{\lambda}$. The average time a transaction spends in the system is

$$T = \frac{1}{(\mu_1 - \hat{\lambda})} + \frac{1}{(\mu_2 - \hat{\lambda})} = \frac{1}{(\mu_1 - \hat{\lambda})} + \frac{1}{(M - \mu_1 - \hat{\lambda})}$$

Take the derivative of the above function in the domain $(\hat{\lambda}, M - \hat{\lambda})$. We have $T'(\mu_1) < 0$ when $\hat{\lambda} < \mu_1 < M/2$; $T'(\mu_1) = 0$ when $\mu_1 = M/2$; $T'(\mu_1) > 0$ when $M/2 < \mu_1 < M - \hat{\lambda}$. Therefore, the minimum T is achieved at $\mu_1 = M/2 = \mu_2$. ■

An interesting question raised by Lemma 1 is whether the optimal service rate at server 1 (μ_1^*) is equal to that at server 2 (μ_2^*) for formulation (7.8). Since arrival rate $\hat{\lambda}$ is the same for server 1 and server 2 in a tandem line system, in the special case where $\mu_1 = \mu_2$ is added as an extra constraint in formulation (7.8), the problem reduces to the single M/M/1 capacity planning under uncertainty with SLA upper bound modified as $\hat{T}/2$. As a result, the optimal solutions will follow look-up Table 14 presented in Section 7.1. For the general case where costs for the two servers are balanced, we establish the following proposition.

Proposition 14 *Given that $c_1 = c_2$, the optimal objective function value of formulation (7.8) can be achieved at $\mu_1 = \mu_2$.*

Proof of Proposition 14. Given a feasible solution to formulation (7.8) denoted by (μ_1, μ_2). Without loss of generality, we assume $\mu_1 > \mu_2 > 0$. Let another feasible solution be $(\frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2})$. It immediately follows that the service costs for these two solutions are equal because $c_1 = c_2$. To compare the mean penalty, we examine the penalty values at these two feasible solutions in the following three cases for an arbitrary arrival realization $\hat{\lambda}$.

Case 1 $\hat{\lambda} < \theta\mu_2$**.** We have $\hat{\lambda}/\mu_2 < \theta$ and $\hat{\lambda}/\mu_1 < \theta$. It follows from Lemma 1 that $T(\mu_1, \mu_2) > T(\frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2})$. By definition of penalty function, $f(\hat{\lambda}, \mu_1, \mu_2) \geq f(\hat{\lambda}, \frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2})$.

Case 2 $\theta\mu_2 \leq \hat{\lambda} < \theta\frac{\mu_1+\mu_2}{2}$**.** We have $\hat{\lambda}/\mu_2 \geq \theta$ and $\hat{\lambda}/(\frac{\mu_1+\mu_2}{2}) < \theta$. At solution (μ_1, μ_2), the penalty $f(\hat{\lambda}, \mu_1, \mu_2) = H_2$ by definition. On the other hand, $f(\hat{\lambda}, \frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2}) = H_1$ or 0. It immediately follows that $f(\hat{\lambda}, \mu_1, \mu_2) > f(\hat{\lambda}, \frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2})$.

Case 3 $\hat{\lambda} \geq \theta\frac{\mu_1+\mu_2}{2}$**.** We have $\hat{\lambda}/\mu_2 > \theta$ and $\hat{\lambda}/(\frac{\mu_1+\mu_2}{2}) \geq \theta$. It is obvious that utilization requirement is violated at least at one of the servers for both feasible solutions (μ_1, μ_2) and $(\frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2})$. Therefore, the penalty value $f(\hat{\lambda}, \mu_1, \mu_2) = f(\hat{\lambda}, \frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2}) = H_2$.

Overall, the inequality $f(\hat{\lambda}, \mu_1, \mu_2) \geq f(\hat{\lambda}, \frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2})$ holds for an arbitrary realization $\hat{\lambda}$. As a result, the objective function value for formulation (7.8) at solution (μ_1, μ_2) is consistently greater than or equal to objective function value at solution $(\frac{\mu_1+\mu_2}{2}, \frac{\mu_1+\mu_2}{2})$. ■

It is noteworthy that the optimal objective function value of formulation (7.8) may also be achieved at other solutions where $\mu_1 \neq \mu_2$. However, the benefit of Proposition 14 is the following implication: In order to solve formulation (7.8) and find an optimal solution where $\mu_1 = \mu_2$, one can decompose the two-stage problem into two identical single-stage problems where input parameters all remain the same except for $\hat{T}$, H_1, and H_2 which are now updated as half of their original values. Such a decomposition approach enables a direct application of analytical solution look-up Table 14 for the single-stage problem (7.3) to solve the two-stage problem here.

Now consider the generic imbalanced-cost case where $c_1 \neq c_2$. Without loss of generality, let $c_1 = c_2 + c'$ where $c_2 > 0$ and $c' > 0$.

Proposition 15 *Given that $c_1 > c_2$, the inequality $\mu_1 \leq \mu_2$ holds valid in every optimal solution for formulation (7.8).*

Proof of Proposition 15. Suppose $\mu_1 > \mu_2$, for a two-stage tandem line system with two M/M/1 queues, swapping μ_1 and μ_2 will not change the average time a transaction/customer stays in the system. Therefore, the average penalty will remain unchanged while the service cost will decrease after swap. In other words, the objective value at solution (μ_1, μ_2) is greater than that at solution (μ_2, μ_1). ■

7.2.2 Scenario-Based Grid Search

Intuitively, when the service rate increases, the penalty cost may decrease while the service cost increases. Therefore, when $c_1 > c_2$, whether $\mu_1 = \mu_2$ or $\mu_1 < \mu_2$ is an optimal solution depends on the trade-off between the decrease in penalty cost and the increase in service cost. In this subsection, we conduct numerical experiments following a scenario-based grid search approach to understand the trade-off.

Consider a numerical example where $\boldsymbol{\lambda} \sim U[20, 30]$, $H_2 = 150$, and $\hat{T} = 0.24$. Accordingly, $\theta \geq 0.878$. Hence, suppose $\theta = 0.95$. In the following, we adopt a sampling based 2-dimensional grid search to solve formulation (7.8) in order to observe the trade-off. We search in the range of [1, 100] at an increment of 0.1 for both μ_1 and μ_2. We randomly generate 1,000 scenarios following a uniform distribution $U[20, 30]$ and assume each scenario is equally likely to realize (i.e., with probability of 0.001). Numerical results are presented in Tables 17 and 18. From the two tables, we can see that $\mu_1 < \mu_2$ is an optimal solution in the cases where the penalty decrease outweighs the service cost increase, i.e., when penalty H_1 is large (e.g., $H_1 = 30$ in Table 17 instead of 7 in Table 18) or when service cost rate c_2 is small (e.g., $c_2 \leq 0.25$ in Table 18).

However, analytically deriving the optimal solutions when $c_1 > c_2$ is challenging. This is partially because in our business problem, SLA requirement is imposed on the overall system performance measure (i.e., T) instead of individual server performance measure. Hence, the decomposition idea, which has been shown to be applicable when $c_1 = c_2$, is less applicable in general cases. Next, we present scenario-based reformulations to solve the problem in a general setting.

7.2.3 Scenario-Based Reformulations

We start with a scenario-based reformulation for the single-stage problem formulation (7.3). Given a set of samples of the arrival rate $\mathcal{S} = \{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_N\}$ where each

Table 17: Numerical results for the two-stage tandem line system with $\boldsymbol{\lambda} \sim U[20, 30]$, $\theta = 0.95$, $\hat{T} = 0.24$, $H_1 = 30$, and $H_2 = 150$.

c_1	c_2	Optimal solution(μ_1^*)	Optimal solution(μ_2^*)	Optimal objective
1	0.05	35.3	49.4	37.77
1	0.1	35.5	47.1	40.21
1	0.15	35.8	44.4	42.52
1	0.2	36	43.6	44.72
1	0.25	36.1	42.2	46.86
1	0.3	36.3	41.5	48.96
1	0.35	36.5	40.9	51.025
1	0.4	36.5	40.9	53.07
1	0.45	36.7	40.4	55.09
1	0.5	36.7	40.4	57.11
1	0.55	37.2	39.4	59.08
1	0.6	37.2	39.4	61.05
1	0.65	37.2	39.4	63.02
1	0.7	37.6	38.8	64.97
1	0.75	37.6	38.8	66.91
1	0.8	37.6	38.8	68.85
1	0.85	37.6	38.8	70.79
1	0.9	38	38.3	72.71
1	0.95	38	38.3	74.625

Table 18: Numerical results for the two-stage tandem line system with $\boldsymbol{\lambda} \sim U[20, 30]$, $\theta = 0.95$, $\hat{T} = 0.24$, $H_1 = 7$, and $H_2 = 150$.

c_1	c_2	Optimal solution(μ_1^*)	Optimal solution(μ_2^*)	Optimal objective
1	0.05	31.6	45.6	36.358
1	0.1	31.6	40.8	38.473
1	0.15	31.6	38.2	40.438
1	0.2	31.6	38.2	42.348
1	0.25	31.6	32.7	44.129
1	0.3	31.6	31.6	45.714
1	0.35	31.6	31.6	47.294
1	0.4	31.6	31.6	48.874
1	0.45	31.6	31.6	50.454
1	0.5	31.6	31.6	52.034
1	0.55	31.6	31.6	53.614
1	0.6	31.6	31.6	55.194
1	0.65	31.6	31.6	56.774
1	0.7	31.6	31.6	58.354
1	0.75	31.6	31.6	59.934
1	0.8	31.6	31.6	61.514
1	0.85	31.6	31.6	63.094
1	0.9	31.6	31.6	64.674
1	0.95	31.6	31.6	66.254

sample $\hat{\lambda}_s$ has a probability of occurrence of π_s, formulation (7.3) can be rewritten as,

$$\begin{aligned}
\min \quad & c\mu + \sum_{s\in\mathcal{S}} \pi_s(H_1 y_s + H_2 z_s) && \text{(7.9a)}\\
\text{s.t.} \quad & 1 - \hat{T}\mu + \hat{T}\hat{\lambda}_s \leq M_s(y_s + z_s), \forall s \in \mathcal{S} && \text{(7.9b)}\\
& \hat{\lambda}_s - \mu\theta \leq \hat{M}_s z_s, \forall s \in \mathcal{S} && \text{(7.9c)}\\
& y_s, z_s \in \{0, 1\}, \forall s \in \mathcal{S} && \text{(7.9d)}\\
& \mu \geq 0 && \text{(7.9e)}
\end{aligned}$$

where M_s and $\hat{M}_s$ are sufficiently large values ("big-M"). As shown above, we introduce extra binary decision variables y_s and z_s for each scenario $s \in \mathcal{S}$. When $z_s = 1$, constraints (7.9c) and (7.9b) become redundant. When $z_s = 0$, constraint (7.9c) imposes utilization requirement to be satisfied. Between y_s and z_s, the minimization process always pushes z_s to be zero first as $H_2 >> H_1$. When y_s is further pushed to be zero, constraint (7.9c) becomes equivalent to the SLA requirement. In this setting, we can let $M_s = 1 + \hat{T}\hat{\lambda}_s, \forall s \in \mathcal{S}$ and $\hat{M}_s = \hat{\lambda}_s, \forall s \in \mathcal{S}$ for the sake of convenience. In practice, there often is a maximum service rate a service provider can achieve due to physical/economical limitation. Therefore, one can include an upper bound for the decision variable μ in formulation (7.9) when necessary.

In addition to y_s and z_s, let us introduce two more binary decision variables $\hat{z}_s$ and $\bar{z}_s$ for each $s \in \mathcal{S}$. We can similarly present a scenario based reformulation for the two-stage service system design model (7.8) as follows.

$$\begin{aligned}
\min \quad & c_1\mu_1 + c_2\mu_2 + \sum_{s\in\mathcal{S}} \pi_s(H_1 y_s + H_2 z_s) && \text{(7.10a)}\\
\text{s.t.} \quad & (\mu_1 - \hat{\lambda}_s) + (\mu_2 - \hat{\lambda}_s) + \hat{T}[(\mu_1 + \mu_2)\hat{\lambda}_s - \hat{\lambda}_s^2] \\
& \qquad -\hat{T}\mu_1\mu_2 \leq M_s(y_s + z_s), \forall s \in \mathcal{S} && \text{(7.10b)}\\
& \hat{\lambda}_s - \mu_1\theta \leq \hat{M}_s \hat{z}_s, \forall s \in \mathcal{S} && \text{(7.10c)}
\end{aligned}$$

$$\hat{\lambda}_s - \mu_2\theta \leq \bar{M}_s\bar{z}_s, \forall s \in \mathcal{S} \tag{7.10d}$$

$$z_s \geq \hat{z}_s, z_s \geq \bar{z}_s, \forall s \in \mathcal{S} \tag{7.10e}$$

$$y_s, z_s, \hat{z}_s, \bar{z}_s \in \{0, 1\}, \forall s \in \mathcal{S} \tag{7.10f}$$

$$\mu_1, \mu_2 \geq 0 \tag{7.10g}$$

Again parameters M_s, $\hat{M}_s$, $\bar{M}_s$ are sufficiently large values. Constraints (7.10c)-(7.10e) together imply that $z_s = 0$ when both $\hat{z}_s = 0$ and $\bar{z}_s = 0$, while $z_s = 1$ when at least one of the variables $\hat{z}_s$ and $\bar{z}_s$ is equal to 1. In other words, penalty H_2 is avoided when the utilization requirement is satisfied at both server 1 and server 2, while penalty H_2 is incurred when the utilization requirement is violated at one or more servers. Constraint (7.10b) implies penalty of H_1 is avoided for a given solution (μ_1, μ_2) and scenario s when both SLA requirement and utilization requirement are satisfied; penalty of H_1 is incurred when SLA requirement is satisfied but utilization requirement is violated at either/both of the two servers. Like the single-stage case, upper bounds on μ_1 and μ_2 can be included in the model if required by practical considerations.

However, the formulation above has a bilinear term $\mu_1\mu_2$ in Constraint (7.10b). Variables μ_1 and μ_2 are both continuous variables. We adopt a piecewise linear approximation approach for functions of two variables called the triangle method (see Vielma et al. (2010); D'Ambrosio et al. (2010)) to linearize the bilinear term $\mu_1\mu_2$.

The piecewise linear approximation for a one-variable nonlinear function can be obtained by introducing an adequate number of sampling coordinates and then approximate each interval with linear functions. The triangle method for a bilinear function is essentially an extension of the technique for the one-dimensional case to two dimensions.

Let $g(\mu_1, \mu_2) = \mu_1\mu_2$. Consider p sampling coordinates $\mu_{11}, \ldots, \mu_{1p}$ on the μ_1 axis and q sampling coordinates $\mu_{21}, \ldots, \mu_{2q}$ on the μ_2 axis. Consider the rectangle corresponding to interval $[\mu_{1i}, \mu_{1,i+1}]$ and $[\mu_{2j}, \mu_{2,j+1}]$ where $i = 1, \dots, p, j = 1, \dots, q$. We create binary decision variables h^u_{ij}, h^ℓ_{ij} for the upper and lower triangle of each rectangle and continuous decision variables $\beta_{ij} \in [0, 1]$ for every breakpoint/vertex of the rectangle. Then the piecewise linear approximation for g, denoted by $\hat{g}$ can be characterized by

$$\hat{g} = \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij} g(\mu_{1i}, \mu_{2j}),$$

together with the following constraints

$$\sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij} = 1$$

$$\sum_{i=1}^{p-1} \sum_{j=1}^{q-1} (h^u_{ij} + h^\ell_{ij}) = 1$$

$$\beta_{ij} \le h^u_{ij} + h^\ell_{ij} + h^u_{i-1,j-1} + h^\ell_{i-1,j-1} + h^\ell_{i-1,j} + h^u_{i,j-1} \forall i = 1, \dots, p, j = 1, \dots, q$$

where $\mu_1 = \sum\limits_{i=1}^{p} \sum\limits_{j=1}^{q} \beta_{ij}\mu_{1i}$, $\mu_2 = \sum\limits_{i=1}^{p} \sum\limits_{j=1}^{q} \beta_{ij}\mu_{2j}$, and the dummy values $h^\ell_{0,j} = h^u_{0,j} = h^\ell_{i,0} = h^u_{i,0} = h^\ell_{p,j} = h^u_{p,j} = h^\ell_{i,q} = h^u_{i,q} = 0$. The interested reader is referred to D'Ambrosio et al. (2010) for modeling details.

Thus, we obtain the piecewise linear approximation based on the triangle method for formulation (7.10) as follows.

$$\min \quad c_1 \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\mu_{1i} + c_2 \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\mu_{2j} + \sum_{s \in \mathcal{S}} \pi_s (H_1 y_s + H_2 z_s) \tag{7.12a}$$

$$\text{s.t.} \quad (\sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\mu_{1i} + \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\mu_{2j})(1 + \hat{T}\hat{\lambda}_s) - 2\hat{\lambda}_s - \hat{T}\hat{\lambda}_s^2$$

$$- \hat{T} \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\mu_{1i}\mu_{2j} \le M_s(y_s + z_s), \forall s \in \mathcal{S} \tag{7.12b}$$

$$\hat{\lambda}_s - \theta \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\mu_{1i} \le \hat{M}_s \hat{z}_s, \forall s \in \mathcal{S} \tag{7.12c}$$

$$\hat{\lambda}_s - \theta \sum_{i=1}^{p} \sum_{j=1}^{q} \beta_{ij}\mu_{2j} \le \bar{M}_s \bar{z}_s, \forall s \in \mathcal{S} \tag{7.12d}$$

$$z_s \ge \hat{z}_s, z_s \ge \bar{z}_s, \forall s \in \mathcal{S} \tag{7.12e}$$

$$y_s, z_s, \hat{z}_s, \bar{z}_s \in \{0, 1\}, \forall s \in \mathcal{S} \tag{7.12f}$$

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\beta_{ij} = 1 \tag{7.12g}$$

$$\sum_{i=1}^{p-1}\sum_{j=1}^{q-1}(h_{ij}^{u} + h_{ij}^{\ell}) = 1 \tag{7.12h}$$

$$\beta_{ij} \leq h_{ij}^{u} + h_{ij}^{\ell} + h_{i-1,j-1}^{u} + h_{i-1,j-1}^{\ell} + h_{i-1,j}^{\ell} + h_{i,j-1}^{u} \quad \forall i = 1,\ldots,p, j = 1,\ldots,q \tag{7.12i}$$

$$\beta_{ij} \in [0,1], \forall i = 1,\ldots,p, j = 1,\ldots,q \tag{7.12j}$$

$$h_{ij}^{u}, h_{ij}^{\ell} \in \{0,1\}, \forall i = 1,\ldots,p-1, j = 1,\ldots,q-1 \tag{7.12k}$$

7.2.4 Computational Experience for Scenario-Based Reformulations

In this subsection, we present our computational experiments for the scenario-based reformulations. We first present the computational settings of our experiments, followed by some numerical results from the scenario-based reformulation of the single-stage service system design problem. Our objective is to draw insights on the solution quality when the sample size varies by comparing our computational results with the analytical results. Last but not the least, we present computational results from the scenario-based reformulation of the two-stage service system design problem.

We consider solving formulation (7.9) for different ranges of arrival rate. For each given range, we randomly generate equally likely samples of three different sizes, i.e., 500, 1000, and 2000. For every instance (i.e., a given arrival rate range and a given sample size), we generate and test 5 replications (or 5 different sets of sample pool). We also impose a 1-hour time limit for each replication, for each instance. All experiments are conducted on a 64-bit Linux system with eight Intel Xeon E5620 2.40GHz processors and 96GB RAM. Gurobi® Optimizer v6.0 is used as the MILP solver. All implementations inherited the default settings for branching, node selection, general purpose cutting planes, preprocessing and heuristics. We report statistics (average optimal solutions and average optimal objective values) based on the replications that solved to optimality under the time limit.

The results for formulation (7.9) are presented in Table 19. Our observations from the table can be summarized as follows.

1. Comparison with the analytical optimal solutions reported in the last column (i.e., "Analytical solutions"), the scenario-based optimization model provides optimal solution with good accuracy. For instance, for the widest range we tested, [0, 50], the optimal objective is 52.81 when sample size is only 500 while the analytical optimal solution is 52.96.

2. For a given instance, as sample size increases, the reported optimal objective value approaches the analytical optimal objective value. However, the difference is not significant with the maximum absolute difference of optimal objective value in percentage being 0.28%.

Due to the second observation above, we choose $|\mathcal{S}| = 500$ in our experiments for the two-stage reformulation (7.10). We take the same numerical example as in Subsection 7.2.2 where $\boldsymbol{\lambda} \sim U[20, 30]$, $H_2 = 150$, $\hat{T} = 0.24$, and $\theta = 0.95$. Similar to the grid search approach, we consider 1000 sampling coordinates across the range of [1, 100] along μ_1 and μ_2 axis respectively. We found that for the same input parameters values (H_1, c_1, and c_2) as presented in Table 18, Gurobi optimizer invariably ran into the problem of "out of memory." This indicates that the scenario-based reformulation approach is computationally more expensive compared to the scenario-based grid search approach, though it may yield better solutions. Additionally, better computational performance of the scenario-based approach, especially for the two-stage problem, may be achieved by manually tuning the Gurobi® optimizer.

Table 19: Optimal design solutions and costs for the scenario-based single-stage service design formulation under different number of samples and different ranges of arrival rate when $\hat{T} = 0.2$, $H_1 = 7$, $H_2 = 150$, and $\theta = 0.95$.

				Number of samples			
a	b	Range Length	Measure	500	1000	2000	Analytical solutions
0	50	50	optimal rate	52.55	52.56	52.61	52.63
			optimal cost	52.81	52.95	52.93	52.96
4	46	42	optimal rate	48.28	48.37	48.40	48.42
			optimal cost	48.82	48.82	48.82	48.85
8	42	34	optimal rate	44.14	44.18	44.19	44.21
			optimal cost	44.73	44.73	44.77	44.78
12	38	26	optimal rate	39.96	39.99	39.99	40.00
			optimal cost	40.76	40.76	40.85	40.81
16	34	18	optimal rate	35.76	35.77	35.78	35.79
			optimal cost	37.13	37.03	36.99	37.04
20	30	10	optimal rate	31.61	31.61	31.59*	31.58
			optimal cost	33.92	34.07	34.00*	33.97
24	26	2	optimal rate	31.00	31.00	31.00	31.00
			optimal cost	31.00	31.00	31.00	31.00

*: Only feasible solution is returned within time limit for each of the 5 replications of this instance. All 5 replications of other instances reported in this table are solved to optimality within less than 5 minutes.

CHAPTER 8

SUMMARY AND FUTURE WORK

In this final chapter, we provide a summary of the research carried out in this dissertation effort, research contributions made, and some directions for future research.

8.1 Summary of Research

Throughout this dissertation, we focus on developing modeling methodologies and solution techniques for system design problems under uncertainty. Considerable emphasis has been placed on developing useful stochastic models using different risk measures and effective methods through analytical and computational study for (1) network design problem with connectivity and diameter requirements under probabilistic edge failures, and (2) capacity planning problem in a service system under uncertain demand rate.

To begin with, we study a combinatorial optimization problem called the minimum spanning k-core problem, which can be used to design networks that maintain their (low) diameter upon deletion of a vertex or an edge. With the deterministic version known to be polynomially solvable, we study the problem specifically under probabilistic edge failures.

In Chapter 5, a CVaR-constrained model is formulated in the stochastic setting using convex piecewise linear loss functions based on cumulative and maximum constraint violation. Polyhedral reformulations of the CVaR constraint for the aforementioned loss functions are then introduced, which allows us to extend a recent and successful decomposition approach to CVaR optimization with linear loss functions (Künzi-Bay and Mayer, 2006). Based on the reformulations, we develop a decomposition and branch-and-cut algorithm and evaluate two versions of this approach against two existing approaches from literature. Through numerical experiments, we find that the decomposition and branch-and-cut approach, when emphasis is placed on the initialization with cutting planes from the reformulation, offers significant computational advantages compared to the other approaches.

In Chapter 6, we formulate the chance-constrained spanning k-core problem to obtain resilient designs. We establish the intractability of the formulated problem by showing that it is NP-hard even in the special case where each scenario is equally likely to happen. We conduct a polyhedral study on the CCSkCP polytope and develop a strengthened formulation via lifting. Our numerical study shows that the strengthened formulation is computationally advantageous.

Chapter 7 presents our study on the capacity planning problem in a service system represented by queueing models under uncertain demand rate. First, a stochastic model is formulated for the single-stage system represented by M/M/1 to minimize the summation of service cost and mean penalty due to violation of prescribed system requirements. Analytical optimal solutions are derived under the assumption that demand rate is uniformly distributed in a given interval. We conduct a numerical study to illustrate the concepts of "cost of uncertainty in demand rate" and "minimum level of information quality" which are of interest to practitioners in particular. Subsequently, we investigate the capacity planning problem of a two-stage service system modeled as a tandem line system with two single server nodes. A similar stochastic model is formulated. While the optimal capacity can be obtained analytically under the condition that the cost rate of the two servers are equal, to determine the optimal capacity when this condition is violated appears to be more challenging. We develop two scenario-based approaches, i.e., grid search and mathematical reformulation, to solve this model.

As summarized above, we have looked into different stochastic models and adopted conditional-value-at-risk, chance constraint (i.e., failure probability), and mean as risk measures in Chapters 5, 6, and 7 respectively. In the following, we first provide a high-level summary on model selection for system design under uncertainty, followed by a discussion about the computational characteristics of different stochastic models.

Modeling. A system design where input parameters are subject to uncertainty can be obtained by solving various stochastic models involving different risk measures. Some of the commonly used risk measures are mean, mean-variance, worst case, failure probability, value-at-risk and conditional-value-at-risk. Different risk measures demonstrate different statistical features and different levels of computational challenges. The choice of a risk measure (and a corresponding stochastic model) depends mostly on problem-specific features and users' risk preferences. Figure 10 provides a view of which modeling approach should be considered under different circumstances.

The reactive approach, where one replaces an uncertain parameter with a nominal value, may be also informative and be considered as a basic step to be used alongside any of the other proactive approaches. The different proactive approaches will yield different solutions which may not be entirely intuitive. It is therefore of some value to see all the different solutions, especially when the business impact of the design problem is significant. A practitioner has to also balance the effort in building a possibly very sophisticated model versus finding information to narrow down the uncertainty in input parameters. Narrowing the range of the uncertain parameter, coupled with more straightforward reactive approach may make a better strategy in practice when computational resources and expertise are limited.

Computational Characteristics. In addition to problem-specific features and users' risk preferences, computational challenges should also be considered when choosing a modeling approach. For fundamental models, e.g., M/M/1 with expected penalty as risk measure, analytical solution could be derived under the assumption

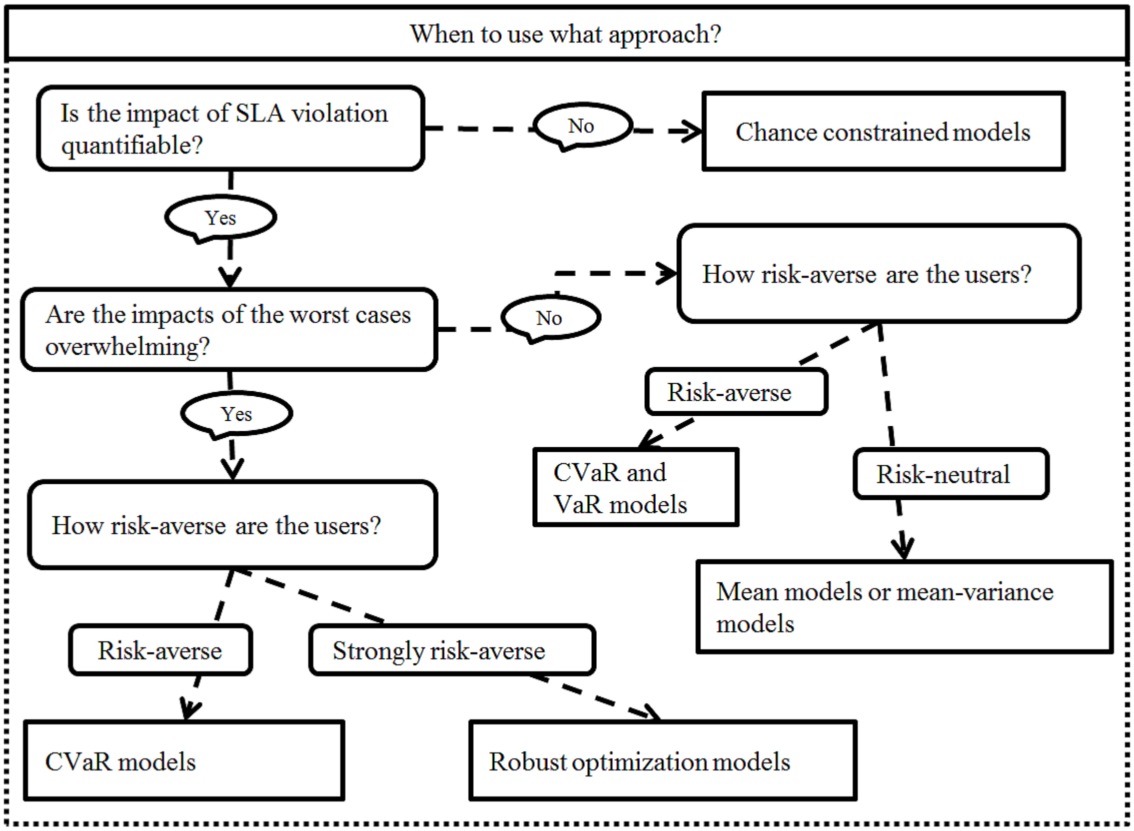
When to use what approach?
Is the impact of SLA violation quantifiable?
No
Chance constrained models
Yes
Are the impacts of the worst cases overwhelming?
No
How risk-averse are the users?
Risk-averse
CVaR and VaR models
Risk-neutral
Mean models or mean-variance models
Yes
How risk-averse are the users?
Risk-averse
Strongly risk-averse
CVaR models
Robust optimization models

Figure 10: Choice of modeling approaches.

that arrival rate is uniformly distributed in a given range. However, for more general models, deriving analytical solutions becomes challenging, which is why our focus was to develop computational techniques for CVaR- and chance-constrained models in Chapters 5 and 6 respectively. Between CVaR-constrained model and chance-constrained model, the former is relatively easier to cope with computationally either in the continuous case with linear penalty function (Künzi-Bay and Mayer, 2006) or the discrete case with convex piecewise linear penalty function (Ma et al., 2015) while the latter is more challenging as showed in Chapter 6.

8.2 Research Contributions

Firstly, we investigate *the minimum spanning k-core problem* in a probabilistic setting. We exploit the graph-theoretic properties of this model to introduce a new approach to resilient inter-hub network design that preserves connectivity and diameter under limited edge failures. We first study a conditional-value-at-risk constrained optimization model to obtain risk-averse solutions for the minimum spanning k-core problem under probabilistic edge failures. We investigate if a polynomial number of scenarios are sufficient to approximate CVaR of the convex piecewise linear loss functions we aim to employ in our formulation. Polyhedral reformulation of the CVaR constraint for piecewise linear loss functions is investigated in this dissertation. A decomposition and branch-and-cut approach is designed to solve the scenario-based approximation of the CVaR-constrained minimum spanning k-core problem.

The second stochastic optimization problem we study is the chance-constrained minimum spanning k-core problem. The complexity of this problem is established. In addition, a polyhedral study is conducted, which ultimately leads to effective solution techniques, contributing to the state-of-art of both chance-constrained programming and resilient network design.

In the capacity planning problem of a single-stage or two-stage system, it is critical for a service provider to control the risk of violating system performance requirements specified in service level agreement while minimizing the system configuration and operational cost. While one can model such systems using queuing models, to determine the optimal capacity is challenging for researchers/practitioners due to the nonlinearity of system performance measures and due to the uncertainty in demand rate. Our study in this regard enhances the literature of service system design by developing analytical solutions and computational techniques.

8.3 Future Work

Chapter 5 serves as a first study of the CVaR-constrained minimum spanning k-core problem focused only on using cutting planes based on the reformulation of the CVaR constraint. It would likely be computationally beneficial, especially for large-scale networks, if the facet-inducing inequalities of the deterministic version are also employed in the branch-and-cut, particularly during the early stages of branching. While the complexity of the deterministic version is settled, the complexity of the CVaR constrained version is still open. From a modeling perspective, it would be interesting to study the r-robust 2-club based designs (Veremyev and Boginski, 2012) that directly capture the requirements of 2-hop resilient network design.

Following the research in Chapter 6, a significant future research direction is the use of blossom inequalities in solving the CCkCP. Given a rational vector $(x^*, z^*) \in [0,1]^m \times \{0,1\}^N$ outside the CCkCP polytope, to identify a CCkCP-blossom inequality that cuts off (x^*, z^*) or prove that no such inequality exists, can be converted to a blossom inequality separation problem for general capacitated b-matching polyhedra (Padberg and Rao, 1982; Letchford et al., 2008).

Separating a blossom inequality, as proved by Padberg and Rao (1982), is equivalent to solving a minimum odd cut-set problem on a specially constructed graph, for which they developed a polynomial algorithm based a minor modification of Gomory-

Hu algorithm for the minimum cut-set problem. A minimum cut-set problem on a finite undirected graph $G = (V, E)$ is to find a partition of V into two nonempty subsets V_1 and V_2 such that the weight of the induced cut is minimum. Suppose V has been partitioned into two classes of nodes called odd and even and suppose a node subset is labeled odd if it contains odd number of odd nodes (assuming $|V|$ is even), a minimum odd cut-set problem is to find a partition of V into two nonempty odd subsets V_1 and V_2 such that the the weight of the induced cut is minimal.

The algorithm Padberg and Rao (1982) proposed can run in $O(n^2 m \log(n^2/m))$ for uncapacitated b-matching and in $O(nm^2 \log(n^2/m))$ for capacitated b-matching. Recently Letchford et al. (2008) developed a new version based again on the Gomory-Hu algorithm, which can run in $O(n^2 m \log(n^2/m))$ for the capacitated case. Note that Gomory-Hu algorithm is readily available in graph library LEMON[1].

Chapter 7 on the single-stage and two-stage service system design serves as a first step in studying the capacity planning problem in a multi-stage service system modeled either as an m-stage tandem queueing system or a queueing network. As scenario-based grid search and scenario-based reformulation each have their own limitations, developing heuristic approaches to obtain near optimal solutions for the capacity planning problem in a multi-stage system is an interesting direction for future research. A potential solution approach along this line is to appropriately "apportion" the overall SLA requirement to the individual stages, solve the capacity planning problem of a single-stage system separately for each stage, and then "aggregate" the solutions from different stages to construct a solution for the multistage system. Although this solution approach may only provide approximate solutions, our exploratory numerical experiments indicate that this approach holds promise for yielding good practical solutions[2].

[1] LEMON: Library for Efficient Modeling and Optimization in Networks. http://lemon.cs.elte.hu/

[2] Personal communication with Dr. Ying Tat Leung

BIBLIOGRAPHY

A. Agrawal, P. Klein, and R. Ravi. When trees collide: An approximation algorithm for the generalized Steiner problem on networks. *SIAM Journal on Computing*, 24 (3):440–456, 1995.

S. Ahmed. Convexity and decomposition of mean-risk stochastic programs. *Mathematical Programming*, 106:433–446, 2006.

S. Ahmed and A. Shapiro. Solving chance-constrained stochastic programs via sampling and integer programming. In Z. L. Chen and S. Raghavan, editors, *Tutorials in Operations Research*, volume 10, pages 261–270. INFORMS, 2008.

M. Altaf-Ul-Amine, K. Nishikata, T. Korna, T. Miyasato, Y. Shinbo, M. Arifuzzaman, C. Wada, M. Maeda, T. Oshima, H. Mori, and S. Kanaya. Prediction of protein functions based on k-cores of protein-protein interaction networks and amino acid sequences. *Genome Informatics*, 14:498–499, 2003.

P. Artzner, F. Delbaen, J-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

B. Balasundaram. *Graph Theoretic Generalizations Of Clique: Optimization and Extensions*. PhD thesis, Texas A&M University, College Station, Texas, USA, 2007.

R. R. Barton, B. L. Nelson, and W. Xie. Quantifying input uncertainty via simulation confidence intervals. *INFORMS Journal on Computing*, 26(1):74–87, 2014.

A. Bassamboo, R. S. Randhawa, and A. Zeevi. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56(10):1668–1686, 2010.

M. G. H. Bell and Y. Iida. *Transportation network analysis*. 1997.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4:238–252, 1962.

P. Beraldi, M. E. Bruni, and D. Conforti. Designing robust emergency medical service via stochastic programming. *European Journal of Operational Research*, 158(1):183–193, 2004.

D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

S. Binato, M. V. F. Pereira, and S. Granville. A new benders decomposition approach to solve power transmission network design problems. *Power Systems, IEEE Transactions on*, 16(2):235–240, 2001.

J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer, New York, NY, 1997.

Q. Botton, B. Fortz, L. Gouveia, and M. Poss. Benders decomposition for the hop-constrained survivable network design problem. *INFORMS Journal on Computing*, 25(1):13–26, 2013.

R. H. Cardwell, C. Monma, and T.-H. Wu. Computer-aided design procedures for survivable fiber optic networks. *Selected Areas in Communications, IEEE Journal on*, 7(8):1188–1197, 1989.

Y. Chee Shin and R. Buyya. Service level agreement based allocation of cluster resources: Handling penalty to enhance utility. In *IEEE International Conference on Cluster Computing Cluster Computing*, pages 1–10, 2005.

T.-M. Choi and P.-S. Chow. Mean-variance analysis of quick response program. *International Journal of Production Economics*, 114(2):456–475, 2008.

T.-M. Choi, D. Li, and H. Yan. Mean-variance analysis of a single supplier and retailer supply chain under a returns policy. *European Journal of Operational Research*, 184(1):356–376, 2008.

S. Y. Chun, A. Shapiro, and S. Uryasev. Conditional value-at-risk and average value-at-risk: Estimation and asymptotics. *Operations Research*, 60(4):739–756, 2012.

L. W. Clarke and G. Anandalingam. A bootstrap heuristic for designing minimum cost survivable networks. *Computers & operations research*, 22(9):921–934, 1995.

C. D'Ambrosio, A. Lodi, and S. Martello. Piecewise linear approximation of functions of two variables in milp models. *Operations Research Letters*, 38(1):39 – 46, 2010.

R. Diestel. *Graph Theory*. Springer-Verlag, Berlin, 1997.

C. Dong, G. H. Huang, Y. P. Cai, and Y. Xu. An interval-parameter minimax regret programming approach for power management systems planning under uncertainty. *Applied Energy*, 88(8):2835–2845, 2011.

J. Edmonds. Maximum matching and a polyhedron with 0,1-vertices. *Journal of Research of the National Bureau of Standards - B*, 69B:125–130, 1965.

C. I. Fábián. Handling CVaR objectives and constraints in two-stage stochastic models. *European Journal of Operational Research*, 191(3):888 – 911, 2008.

N. Gans and Y. P. Zhou. A call-routing problem with service-level constraints. *Operations Research*, 51(2):255–271, 2003.

M. Goemans. Minimum bounded degree spanning trees. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 273–282. IEEE, 2006.

M. Goemans and D. Bertsimas. Survivable networks, linear programming relaxations and the parsimonious property. *Mathematical Programming*, 60(1-3):145–166, 1993.

M. Grötschel, C. L. Monma, and M. Stoer. Computational results with a cutting plane algorithm for designing communication networks with low-connectivity constraints. *Operations Research*, 40(2):309–330, 1992.

M. Grötschel, C. L. Monma, and M. Stoer. Polyhedral and computational investigations for designing communication networks with high survivability requirements. *Operations Research*, 43(6):1012–1024, 1995.

I. Gurvich, J. Luedtke, and T. Tezcan. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56 (7):1093–1115, 2010.

W. Haneveld and M. van der Vlerk. Integrated chance constraints: Reduced forms and an algorithm. *Computational Management Science*, 3:245–269, 2006.

R. Hassan, R. Neufville, and D. McKinnon. Value-at-risk analysis for real options in complex engineered systems. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3697–3704, 2005.

L. J. Hong and G. Liu. Simulating sensitivities of conditional value at risk. *Management Science*, 55:281–293, 2009.

H. Hrasnica, A. Haidine, and R. Lehnert. *Broadband powerline communications: network design*. John Wiley & Sons, 2005.

P. Huang and D. Subramanian. Iterative estimation maximization for stochastic linear programs with conditional value-at-risk constraints. *Computational Management Science*, 9(4):441–458, 2012.

P. Huang, D. Subramanian, and J. Xu. An importance sampling method for portfolio CVaR estimation with gaussian copula models. In *Simulation Conference (WSC), Proceedings of the 2010 Winter*, pages 2790–2800, 2010.

K. Jain. A factor 2 approximation algorithm for the generalized Steiner network problem. *Combinatorica*, 21:39–60, 2001.

J. Jeźowski. Review of water network design methods with literature annotations. *Industrial & Engineering Chemistry Research*, 49(10):4475–4516, 2010.

H. Kerivin and A. R. Mahjoub. Design of survivable networks: A survey. *Networks*, 46(1):1–21, 2005.

C. W. Ko and C. L. Monma. Heuristic methods for designing highly survivable communication networks. In *Technical report*. Bell Communications Research Chichester, 1989.

G. Kortsarz, R. Krauthgamer, and J. R. Lee. Hardness of approximation for vertex-connectivity network design problems. *SIAM Journal on Computing*, 33(3):704–720, 2004.

J. Kosinski, D. Radziszowski, K. Zielinski, S. Zielinski, G. Przybylski, and P. Niedziela. Definition and evaluation of penalty functions in sla management framework. In *Fourth International Conference on Networking and Services*, pages 176–181, 2008.

P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4:43–68, 2002.

P. Krokhmal, S. Uryasev, and G. Zrazhevsky. Numerical comparison of conditional value-at-risk and conditional drawdown-at-risk approaches: application to hedge funds. In *Applications of stochastic programming*, volume 5 of *MPS/SIAM Ser. Optim.*, pages 609–631. SIAM, Philadelphia, PA, 2005.

P. Krokhmal, M. Zabarankin, and S. Uryasev. Modeling and optimization of risk. *Surveys in Operations Research and Management Science*, 16(2):49 – 66, 2011.

A. Künzi-Bay and J. Mayer. Computational aspects of minimizing conditional value-at-risk. *Computational Management Science*, 3:3–27, 2006.

L. C. Lau, J. Naor, Mohammad R. S., and M. Singh. Survivable network design with degree or order constraints. *SIAM Journal on Computing*, 39(3):1062–1087, 2009.

A. N. Letchford, G. Reinelt, and D. O. Theis. Odd minimum cut sets and *b*-matchings revisited. *SIAM Journal on Discrete Mathematics*, 22(4):1480–1487, 2008.

Y. T. Leung, M. Kamath, and J. Ma. Parameter tolerance in queueing models. Technical Report RJ10512, IBM Research Report, 2013.

Q. Liang, X. Wu, and H. C. Lau. Optimizing service systems based on application-level QoS. *IEEE Transactions on Services Computing*, 2(2):108–121, 2009.

Z. Liu, M. S. Squillante, and J. L. Wolf. On maximizing service-level-agreement profits, 2001.

P. Luathep, A. Sumalee, W. Lam, Z.-C. Li, and H. Lo. Global optimization method for mixed transportation network design problem: a mixed-integer linear programming approach. *Transportation Research Part B: Methodological*, 45(5):808–827, 2011.

J. Luedtke, S. Ahmed, and G. Nemhauser. An integer programming approach for linear programs with probabilistic constraints. *Mathematical Programming*, 122 (2):247–272, 2010.

J. Ma and B. Balasundaram. Solving chance-constrained spanning k-core problem via decomposition and integer programming. *IIE Annual Conference Proceedings*, pages 2774–2783, 2013.

J. Ma, Y. T. Leung, and M. Kamath. Service system design under uncertainty. *IIE Annual Conference Proceedings*, pages 3564–3573, 2014.

J. Ma, F. Mahdavi Pajouh, B. Balasundaram, and V. Boginski. The minimum spanning k-core problem with bounded CVaR under probabilistic edge failures. *INFORMS Journal on Computing*, 2015. Forthcoming.

T. L. Magnanti and S. Raghavan. Strong formulations for network design problems with connectivity requirements. *Networks*, 45(2):61–79, 2005.

T. L. Magnanti and R. T. Wong. Network design and transportation planning: Models and algorithms. *Transportation Science*, 18(1):1–55, 1984.

A. R. Mahjoub. Two-edge connected spanning subgraphs and polyhedra. *Mathematical Programming*, 64:199–208, 1994.

A. R. Mahjoub, L. Simonetti, and E. Uchoa. Hop-level flow formulation for the survivable network design with hop constraints problem. *Networks*, 61(2):171–179, 2013.

A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

H. M. Markowitz. Mean-variance analysis in portfolio choice and capital markets. *The Journal of Finance*, 44(2):531–535, 1989.

J. Marques, M. C. Cunha, J. Sousa, and D. Savic. Robust optimization methodologies for water supply systems design. *Drinking Water Engineering Science*, 5(1):31–37, 2012.

H. E. Mausser and M. Laguna. A new mixed integer formulation for the maximum regret problem. *International Transactions in Operational Research*, 5(5):389–403, 1998.

M. Minoux. Network synthesis and optimum network design problems: Models, solution methods, and applications. *Networks*, 19:313–360, 1989.

C. L. Monma and D. F. Shallcross. Methods for designing communications networks with certain two-connected survivability constraints. *Operations Research*, 37(4): 531–541, 1989.

J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios. Robust optimization of large-scale systems. *Operations Research*, 43(2):264–281, 1995.

G. L. Nemhauser and L. A. Wolsey. *Integer and Combinatorial Optimization*. Wiley, New York, 1999.

M. E. O'Kelly and H. J. Miller. The hub network design problem: A review and synthesis. *Journal of Transport Geography*, 2:31–40, 1994.

M. W. Padberg and M. R. Rao. Odd minimum cut-sets and *b*-matchings. *Mathematics of Operations Research*, 7(1):67–80, 1982.

G. Pastukhov, A. Veremyev, V. Boginski, and E. L. Pasiliao. Optimal design and augmentation of strongly attack-tolerant two-hop clusters in directed networks. *Journal of Combinatorial Optimization*, 27(3):462–486, 2014.

J. Pattillo, N. Youssef, and S. Butenko. On clique relaxation models in network analysis. *European Journal of Operational Research*, 226(1):9 – 18, 2013.

G. C. Pflug. Some remarks on the value-at-risk and the conditional value-at-risk. *Nonconvex optimization and its applications*, 49:272–281, 2000.

A. Prékopa. Probabilistic programming. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 267–351. Elsevier, 2003.

W. R. Pulleyblank. *Faces of matching polyhedra*. PhD thesis, University of Waterloo, Canada, 1973.

R. T. Rockafellar and J. Royset. Engineering decisions under risk averseness. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 1(2):04015003, 2015.

R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000.

R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

R. T. Rockafellar and S. Uryasev. The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, 18(1):33–53, 2013.

A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity analysis*, volume 134. Wiley New York, 2000.

S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.

S. Sen and J. L. Higle. An introductory tutorial on stochastic linear programming models. *Interfaces*, 29(2):33–61, 1999.

A. Shapiro, D. Dentcheva, and A. Ruszczynski, editors. *Lectures on stochastic programming: Modeling and theory*. Society for Industrial and Applied Mathematics (SIAM): MPS/SIAM Series on Optimization, Philadelphia, PA, 2009.

S. Shen, J. C. Smith, and S. Ahmed. Expectation and chance-constrained models and algorithms for insuring critical paths. *Management Science*, 56(10):1794–1814, 2010.

L. V. Snyder. Facility location under uncertainty: A review. *IIE Transactions*, 38(7): 547–564, 2006.

M. S. Sodhi. Managing demand risk in tactical supply chain planning for a global consumer electronics company. *Production and Operations Management*, 14(1): 69–79, 2005.

Y. Song and J. R. Luedtke. Branch-and-cut approaches for chance-constrained formulations of reliable network design problems. *Mathematical Programming Computation*, 5(4):397–432, 2013.

S. Soni, R. Gupta, and H. Pirkul. Survivable network design: The state of the art. *Information Systems Frontiers*, 1(3):303–315, 1999.

A. C. Soteriou and R. B. Chase. A robust optimization approach for improving service quality. *Manufacturing and Service Operations Management*, 2(3):264–286, 2000.

K. Steiglitz, P. Weiner, and D. Kleitman. The design of minimum-cost survivable networks. *IEEE Transactions on Circuit Theory*, 16(4):455–460, 1969.

D. Subramanian and P. Huang. An efficient decomposition algorithm for static, stochastic, linear and mixed-integer linear programs with conditional-value-at-risk constraints. Technical Report RC24752, IBM Research Report, 2009.

A. Tomaszewski, Michał Pióro, and M. Żotkiewicz. On the complexity of resilient network design. *Networks*, 55(2):108–118, 2010.

R. Van Slyke and R. Wets. L-shaped linear programs with applications to control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17:638–663, 1969.

V. V. Vazirani. *Approximation Algorithms*. Springer-Verlag, Berlin, 2001.

A. Veremyev and V. Boginski. Robustness and strong attack tolerance of low-diameter networks. In A. Sorokin, R. Murphey, M. T. Thai, and P. M. Pardalos, editors, *Dynamics of Information Systems: Mathematical Foundations*, volume 20 of *Springer Proceedings in Mathematics & Statistics*, pages 137–156. Springer New York, 2012.

J. P. Vielma, S. Ahmed, and G. Nemhauser. Mixed-integer models for nonseparable piecewise-linear optimization: Unifying framework and extensions. *Operations Research*, 58(2):303–315, 2010.

W. Wang. *Sample Average Approximation of Risk-Averse Stochastic Programs*. PhD thesis, Georgia Institute of Technology, 2007.

W. Wang and S. Ahmed. Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36(5):515 – 519, 2008.

D. West. *Introduction to Graph Theory*. Prentice-Hall, Upper Saddle River, NJ, 2001.

H. Whitney. Congruent graphs and the connectivity of graphs. *American Journal of Mathematics*, 54(1):150–168, 1932.

N. Wieberneit. Service network design for freight transportation: a review. *OR Spectrum*, 30(1):77–112, 2008.

S. Wuchty and E. Almaas. Peeling the yeast protein network. *Proteomics*, 5(2): 444–449, 2005.

D. R. Wuellner, S. Roy, and R. M. D'Souza. Resilience and rewiring of the passenger airline networks in the united states. *Physical Review E*, 82(5):056101, 2010.

H. Yang and M. G. H. Bell. Models and algorithms for road network design: a review and some new developments. *Transport Reviews*, 18(3):257–278, 1998.

VITA

Juan Ma

Candidate for the Degree of

Doctor of Philosophy

Dissertation: PROACTIVE APPROACHES FOR SYSTEM DESIGN UNDER UNCERTAINTY APPLIED TO NETWORK SYNTHESIS AND CAPACITY PLANNING

Major Field: Industrial Engineering and Management

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma, United States in December, 2015.

Completed the requirements for the Bachelor of Engineering in Logistics Systems Engineering at Huazhong University of Science and Technology, Wuhan, Hubei, China in July, 2010.

Experience:

Juan Ma's research interests are in the areas of combinatorial optimization, robust/resilient design of networks under uncertainty, service system design under uncertainty, and data mining. She has worked as a graduate research and teaching assistant in the School of Industrial Engineering and Management at Oklahoma State University from 2010 to 2015. She has also interned at Air Force Research Laboratory-Mathematical Modeling and Optimization Institute in Summers 2013 and 2014.