Relay-Hub Network Design for Consolidation Planning Under Demand Variability

Onkar Kulkarni, Mathieu Dahan, and Benoit Montreuil
H. Milton Stewart School for Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30363
onkar.kulkarni@gatech.edu, mathieu.dahan@isye.gatech.edu, benoit.montreuil@isye.gatech.edu

Problem description: We study the problem of designing large-scale resilient relay logistics hub networks. We propose a model of *Capacitated Relay Network Design under Stochastic Demand and Consolidation-Based Routing* (CRND-SDCR), which aims to improve a network's efficiency and resilience against commodity demand variability through integrating tactical decisions. **Methodology:** We formulate CRND-SDCR as a two-stage stochastic optimization program where we locate relay logistics hubs and decide their capacities in the first stage and design a minimum-cost consolidation plan in the second stage. As an exact solution approach, we design a branch-and-cut algorithm with a nested Benders decomposition and integer L-shaped method. We decompose CRND-SDCR twice: (i) across the stochastic demand scenarios, and (ii) across each origin-destination pair within the scenario-dependent subproblems; and utilize Benders decomposition at each of these decomposition stages to add the associated Benders feedback cuts. We guarantee the exactness of our solution approach by adding integer L-shaped cuts, obtained by solving the second-stage subproblem exactly through Benders decomposition as well. **Results:** We apply our methodology to design large-scale resilient relay networks to be used for finished vehicle deliveries for a US-based car manufacturer partner. Our computational experiments demonstrate that our developed approach can obtain near-optimal solutions for practically relevant instances using sample average approximation. The resulting logistics networks showcase a significant improvement in capabilities to sustain commodity demand variability, in comparison with relay networks designed to fulfill average commodity demand. In particular, our networks lead to a $\sim 7\%$ decrease in average delivery costs as compared to networks designed under a deterministic demand setting. Moreover, we depict the importance of considering consolidation-based routing at the network design stage through benchmarking against literature-proposed relay networks that continuously approximate the routing operations. **Implications:** Our analysis provides decision-makers with recommendations regarding inducing network flexibility to hedge against commodity demand uncertainty.

Key words: Relay Network Design; Stochastic Demand; Consolidation-based Routing; Decomposition-based Branch-and-Cut; Nested Benders Decomposition; Integer L-Shaped Method; Sample Average Approximation

1. Introduction

1.1. Motivation

Freight transportation forms an integral part of the modern-day economy. It aids manufacturing processes through the timely availability of raw materials, facilitates global trade by delivering

finished goods to corresponding customers, and in turn fosters economic growth via creating job opportunities across geographies (Crainic 2000). These freight transportation services are provided through multiple modes, such as railways, trucks, container shipping liners, and airplanes. Among all these modes, truck transportation serves as the dominant land transportation mode and accounts for the largest fraction of revenue and freight movement. For instance, the trucking companies in the US transported 64.7% of total domestic tonnage in 2019, which corresponded to $940.8 billion in gross freight revenues (U.S. Bureau of Transportation Statistics 2023).

In order to transport such high freight volumes, trucking companies consolidate the loads between multiple locations to achieve better service offerings to customers, substantial transportation cost savings, and overall lesser emissions (Grove and O'Kelly 1986, Hall 1989). Although such trips provide economies of scale, they require the truck drivers to drive for more than 2,000 miles in a single journey thus keeping drivers away from their home for extended periods of time. These unsustainable working conditions for drivers force them to quit their jobs leading to a high driver turnover rate. This driver turnover rate has been consistently reported to be between 80% and 90% every year for the past decade (American Trucking Associations 2018) and has led the trucking industry to accrue substantially high unproductive costs estimated to be between $2.8 - $3 billion annually (Keller and Ozment 1999, Rodriguez et al. 2000). As a direct consequence, the trucking industry is facing a chronic issue of driver shortage as well. There was a reported shortage of 64,000 truck drivers in 2019 and it is expected to reach 160,000 by 2030 in the US (American Trucking Associations 2019). One potential remedy to the situation is to modify transportation networks and the operations conducted on them to better satisfy the needs of truck drivers. Relay transportation provides such an opportunity by making trucking a daily job.

Relay transportation consists of constructing relay facilities to permit the fulfillment of demand via short-haul transportation segments (Hunt 1998, Cabral et al. 2007, Montreuil 2011). These relay facilities or hubs act as sortation facilities where commodities are transferred between delivery vehicles, and serve as pit stops for drivers. In relay logistics, the delivery drivers can advance commodities for half of their daily driving limit from one relay to the next, and then return to the original relay—ideally with other commodities—before reaching their home by the end of the day (Ali et al. 2002, Hu, Askin, and Hu 2019).

All freight networks whether operated through long-haul or short-haul transportation, face a myriad of uncertainties in terms of travel time, commodity demand, vehicle breakdowns, etc., with commodity demand uncertainty causing the most undesirable consequences. These demand fluctuations occur across both space and time, due to which the trucking companies struggle to devise effective consolidation plans causing poor truck utilization in the delivery trips and ultimately high operational costs (Lium, Crainic, and Wallace 2009, Hewitt et al. 2019). Hence,

consideration of demand stochasticity becomes imperative, and even more so when there are more frequent commodity delivery trips, such as in the relay logistics paradigm. Furthermore, such consideration impacts the tactical plans for commodity transportation which in turn affects the strategic decisions of relay network design.

The literature on relay logistics network design primarily designs small-scale networks under deterministic commodity demand setting (Cabral et al. 2007, Yıldız, Karaşan, and Yaman 2018, Leitner et al. 2019), and the investigation that considers the stochastic counterpart assumes continuous flow routing decisions, which do not account for the important consolidation features in logistics planning (Hu, Askin, and Hu 2019). The current research addresses these gaps by answering the following research question: *How to design large-scale relay logistics hub networks under stochastic commodity demand and consolidation-based routing?*

1.2. Contributions

To address the research question, we introduce the problem of *Capacitated Relay Network Design under Stochastic Demand and Consolidation-Based Routing* (CRND-SDCR), which focuses on improving network efficiency and resilience against demand variability through integrating tactical planning decisions. We model commodity demand uncertainty through a finite set of stochastic scenarios and formulate CRND-SDCR as a two-stage stochastic program with mixed-integer recourse. The first stage comprises long-term decisions on where to locate relay hubs and decide their respective sizes, while the second stage designs a minimum-cost consolidation plan for a given stochastic commodity demand realization.

We develop a branch-and-cut algorithm with nested Benders decomposition and integer L-shaped methods (Algorithm 3). We employ Benders decomposition at two stages: Akin to classical stochastic programs, we solve for each stochastic scenario ω the dual $\mathcal{L}_\omega$ of the linear programming (LP) relaxation of the second-stage subproblem to add Benders feedback cuts to the first-stage master problem. However, as the LP relaxation of the second-stage problem entails solving a capacitated multi-commodity minimum-cost network flow problem, solving it and its dual by directly feeding it to an off-the-shelf optimization solver is not computationally efficient. Instead, we use Benders decomposition (again) to solve $\mathcal{L}_\omega$, by decomposing it on the basis of each O-D pair. Here, we leverage the underlying network-flow structure to generate Benders cuts in polynomial time using shortest-path routines. In order to enhance the computational efficiency of our solution approach even further, we add valid inequalities and strengthen the formulations by adding specific dummy variables. Finally, to guarantee the exactness of our solution approach, we add integer L-shaped cuts by solving the second-stage subproblem exactly through Benders decomposition as well.

We conduct an extensive case study pertaining to the design of large-scale relay networks to be used for finished vehicle deliveries for a US-based car manufacturer company that partnered with

our research team. We show that our solution approach can obtain near-optimal solutions for practically relevant instances using sample average approximation, and that our tailored implementation of decomposition-based branch-and-cut scales better in comparison to its classical counterpart with an increase in demand scenarios and network size. We assess the performance of the generated networks by determining minimum-cost consolidation plans for all available stochastic demand scenarios, and compare it with that of relay networks designed for an average parcel demand scenario. Our designed networks showcase significantly higher robustness against demand variability, as they can fulfill more demand by short-haul transportation and at a lower cost. Specifically, our networks are able to reduce the unfulfilled demand by a factor of 2 and lead to lower average delivery costs of around $\sim 7\%$ across instances. Finally, we benchmark the performance of our generated relay networks against literature-proposed relay networks designed by continuously approximating the tactical planning decisions and depict the importance of considering such tactical planning decisions at the network design stage.

The remainder of the article is organized as follows: Section 2 reviews the existing literature on Relay Network Design. We then formulate our two-stage stochastic optimization problem in Section 3. In Section 4 we present the developed solution approach to solve the problem exactly. Section 5 presents results from a case study to validate our model and solution approach when designing a relay hub network for finished vehicle delivery in the south-east of the US. Finally, Section 6 presents the concluding remarks and lays out the avenues for future research.

2. Related Work

The Hub Location Problem (HLP), one of the widely studied problems in location science, aims to select locations from a discrete set of candidates to build one or more hub facilities and serve commodity demand between a given set of origin-destination (O-D) pairs (Hall 1989). HLP seeks to optimize the trade-off between hub construction and transportation costs. Due to typically high hub construction costs and substantial transportation cost savings through consolidation, the resulting network designs contain very few hubs and rely heavily on long-haul delivery trips. Such long-haul delivery trips often last for multiple days and result in large away-from-home time for the delivery drivers, ultimately taking a toll on the drivers' mental and physical well-being (Sieber 2015, Nosowitz 2017). Such unsustainable working conditions for delivery drivers warrant the need to modify trucking operations. Recently, relay logistics has been proposed as an alternate logistics operational paradigm to render trucking a daily job.

Relay logistics involves building relay facilities and operating them to support commodity delivery for a given set of O-D pairs while respecting the drivers' driving limit (Kewcharoenwong, Li, and Üster 2023). Relay operations decompose shipments into short segments in a network of *relay*

hubs, each traveled by a separate driver. Thus, a driver moves commodities from one relay hub to the next, drops the commodities for another driver to pick up, and then travels back to the original relay hub (ideally, moving another set of commodities in the opposite direction) (Jacquillat, Schmid, and Wang 2022). In addition, relay logistics can provide other benefits as well such as: (i) improved truck utilization and, consequently, higher driver utilization, which leads to better compensation for drivers (drivers are primarily paid based on mileage and relay logistics facilitates a more continuous driving schedule as opposed to a long waiting time between each pickup and delivery in classical operations); (ii) reduction in delivery times as the commodities do not have to wait (due to driver rests) because they are relayed by multiple drivers; and (iii) the reduction in accidents, training costs, and insurance rates because of more experienced drivers with job continuity (Taylor, Whicker, and Usher 2001, Üster and Kewcharoenwong 2011).

While the advantages of adopting relay logistics are enticing, it also presents a distinct set of challenges. One of the major challenges the logistics service provider faces in this operational setting is increased overall transportation costs due to additional distance traveled and coordination between multiple commodities across space and time (Üster and Kewcharoenwong 2011). To tackle these challenges, appropriate relay hub networks have to be designed. This *relay network design* problem aims to select relay hub locations that minimize total relay hub construction and transportation costs to satisfy commodity demand while respecting the driver tour length constraints (Hunt 1998, Ali et al. 2002, Cabral et al. 2007, Üster and Maheshwari 2007, Ballot, Gobet, and Montreuil 2012).

The earliest efforts in this direction simply involved locating a minimum number of relay hubs on the shortest-path delivery routes and did not consider fixed-charge costs for the relay hub construction (Hunt 1998, Ali et al. 2002). Although results from these studies provided efficient delivery routes that respected the driver tour length constraints, they did not optimize the trade-off between relay hub construction costs and transportation costs. Such an issue was partially addressed by considering a fixed-charge network design type of formulation wherein relay hubs were established and deterministic commodity demand was routed between O-D pairs (Cabral et al. 2007, Üster and Maheshwari 2007, Kulturel-Konak and Konak 2008, Üster and Kewcharoenwong 2011, Konak 2012, Kewcharoenwong and Üster 2017, Yıldız, Karaşan, and Yaman 2018, Leitner et al. 2019, Kewcharoenwong, Li, and Üster 2023, Ziaeifar and Üster 2023). However, one limitation of these investigations, is that they designed small-scale relay networks for narrower geographical regions and did not consider relay-hub sizing nor commodity demand variability.

Another important characteristic of relay networks designed in the literature is their hierarchical nature: The non-hub nodes such as commodity origins and destinations, commonly referred to as

spoke nodes, are allocated to specific hubs, forcing commodities to travel only along the associated legs (Cabral et al. 2007, Üster and Maheshwari 2007, Kulturel-Konak and Konak 2008, Üster and Kewcharoenwong 2011, Konak 2012, Kewcharoenwong and Üster 2017, Yıldız, Karaşan, and Yaman 2018, Hu, Askin, and Hu 2019, Leitner et al. 2019, Kewcharoenwong, Li, and Üster 2023, Ziaeifar and Üster 2023). However, such restriction constrains commodity flows and leads to an increase in traveled distances and potential commodity congestion at the hub nodes (Tu and Montreuil 2019, Montreuil et al. 2018). To overcome the aforementioned issues, the concept of Physical Internet recently emerged to design hyperconnected networks, which are multi-tier hub networks that interconnect open-access hub facilities at multiple planes (Montreuil 2011, Montreuil, Meller, and Ballot 2013). This hyperconnectivity provides better degrees of freedom for commodity movement while preserving cost savings achieved through consolidation opportunities (Kulkarni, Dahan, and Montreuil 2023). To the best of our knowledge, only (Kulkarni et al. 2021, Kulkarni, Dahan, and Montreuil 2022, 2024) designed such large-scale hyperconnected relay logistics networks. These studies, in particular, addressed the problem of relay network design from a logistics resilience perspective but again did not consider decisions related to hub sizing and demand variability.

To the best of our knowledge, only Hu, Askin, and Hu (2019) considered stochastic commodity demands in the relay network design problem with sizing considerations. However, one of the major drawbacks of Hu, Askin, and Hu (2019) and other investigations that considered fixed-charge relay network design lies in their restrictive modeling approach to transportation costs. The transportation costs are considered linear in terms of commodity flow. Such a representation does not favor commodity consolidation and is less indicative of trucking industry operations. We address this gap by accounting for fixed trucking costs in addition to linear commodity costs to represent real-life trucking operating costs well. Hence, in this work, we design large-scale capacitated relay hub networks under stochastic demand and consolidation-based routing.

3. Capacitated Relay Network Design under Stochastic Demand and Consolidation-Based Routing Modeling

In this section, we first present the *Capacitated Relay Network Design under Stochastic Demand and Consolidation-Based Routing* (CRND-SDCR) problem. We then formulate it as a two-stage stochastic optimization program with mixed-integer recourse.

3.1. Problem Definition

We consider a logistics service provider (trucking carrier) or a consortium of such providers interested in designing a large-scale logistics hub network for efficient relay transportation. This problem is motivated by existing unsustainable long-haul delivery trip schedules that affect drivers' mental and physical health (Sieber 2015, Nosowitz 2017). The *Capacitated Relay Network Design under*

Stochastic Demand and Consolidation-Based Routing (CRND-SDCR) problem seeks to locate logistics hubs and decide their respective capacities to minimize costs of hub construction and future transportation costs for satisfying stochastic commodity demand between each origin-destination (O-D) pair via consolidation-based trucking.

Formally, let $\mathcal{S}$ (respectively, $\mathcal{T}$) represent the discrete set of origin (respectively, destination) locations of all future commodities. Considering the inherent uncertainty in demand, we model demand variability through a finite set of stochastic demand scenarios denoted by Ω with each demand scenario $\omega \in \Omega$ having a probability of occurrence π_ω. Within each demand scenario $\omega \in \Omega$, and for each O-D pair $p \in \mathcal{P} \subseteq \mathcal{S} \times \mathcal{T}$, we denote by d_p^ω the associated commodity demand.

To serve such uncertain commodity demand, the service provider intends to open relay logistics hubs from a pre-selected set of discrete candidate locations $\mathcal{H}$ aimed to facilitate commodity deliveries from their origins to their corresponding destinations through short-haul segments. These logistics hubs act as sortation facilities where commodities are transferred between delivery vehicles depending on the respective destinations, and serve as pit stops for drivers. Specifically, to support commodity transfer between delivery vehicles, the logistics hubs require appropriate (un)loading capacities, usually measured in terms of total number of available vehicle docking lanes (or bays). Let $\mathcal{K}$ represent the pre-selected set of pragmatic capacity configurations, where each configuration $k \in \mathcal{K}$ corresponds to a specific number of vehicle docking lanes denoted by S_k. The service provider incurs a cost of C_i^k to open a logistics hub at the candidate location $i \in \mathcal{H}$ with a capacity configuration $k \in \mathcal{K}$. These hub opening costs include land acquisition costs and hub facility construction costs.

Building upon the principles of the Physical Internet, we permit each origin, destination, and potential hub location to be connected to multiple other locations to allow the design of a hyper-connected network. We represent as $\mathcal{A} \subseteq (\mathcal{S} \cup \mathcal{T} \cup \mathcal{H})^2$ the set of potential (directed) transportation legs, which satisfy the traveled distance or driving time regulations to ensure a daily return for all drivers to their respective homes. The distance of each leg $(i, j) \in \mathcal{A}$ is denoted by $\ell_{i,j}$.

In order to transport commodities on each leg $(i, j) \in \mathcal{A}$, the service provider incurs a fixed scheduling cost C_s for each required truck, and a variable fuel-related cost C_f per commodity and per unit of distance traveled. Such fixed plus linear cost structure realistically represents several real-world freight costs (Greening, Dahan, and Erera 2023). Then at the commodity origin location and at each relay hub, we suppose that commodities are sorted depending on their next transportation leg. Thus, the service provider faces a handling cost of C_h per unit commodity at every visited relay hub and their origin locations. After sorting, these commodities are consolidated in trucks. We denote by q the average cubic volume of a commodity, and by Q the total cubic volume capacity of each truck.

The goal of the CRND-SDCR problem is then to select a subset of hub locations $\mathcal{H}_o \subseteq \mathcal{H}$ and their respective sizes so as to minimize the total cost of hub opening and expected transportation costs across the stochastic commodity demand scenarios Ω.

3.2. Two-Stage Stochastic Programming Model

To model CRND-SDCR, we formulate a two-stage stochastic program with mixed-integer recourse. In the first stage, the design decisions of hub locations and their respective sizes are determined: For every hub $i \in \mathcal{H}$ and every capacity configuration $k \in \mathcal{K}$, we consider a binary variable y_i^k that takes a value of 1 if hub at location i with configuration k is opened, and 0 otherwise.

In the second stage, after a demand scenario ω is realized, the hub network is used to design a minimum-cost consolidation plan of all commodities. To this end for each scenario ω, we define continuous decision variables $f_{i,j}^{p,\omega} \in \mathbb{R}_{\geq 0}$ that denote the volume of commodity for O-D pair $p \in \mathcal{P}$ transported on leg $(i,j) \in \mathcal{A}$, and discrete variables $x_{i,j}^{\omega} \in \mathbb{Z}_{\geq 0}$ that represent the number of delivery trucks used to transport commodities on leg $(i,j) \in \mathcal{A}$. We then derive the following two-stage stochastic optimization problem:

$$\text{CRND-SDCR}: \quad \min_{y} \quad \sum_{i\in\mathcal{H}}\sum_{k\in\mathcal{K}} C_i^k \cdot y_i^k + \mathbb{E}_{\omega}[\mathbb{T}(y,\omega)] \tag{1a}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} y_i^k \leq 1, \quad \forall i \in \mathcal{H}, \tag{1b}$$

$$y_i^k \in \{0,1\}, \quad \forall i \in \mathcal{H}, \forall k \in \mathcal{K}. \tag{1c}$$

The objective (1a) minimizes the total cost of logistics hub opening and expected transportation costs across demand scenarios. We note that relay hub opening cost C_i^k are scaled to match the temporality of the long-term opening costs with the that of the short-term transportation costs. Constraints (1b) ensure at most one capacity configuration is selected at each hub location. Subsequently, the transportation costs for every $\omega \in \Omega$ are given by:

$$\mathbb{T}(y,\omega) = \min_{x,f} \quad \sum_{p\in\mathcal{P}}\sum_{(i,j)\in\mathcal{A}} (C_f \cdot \ell_{i,j} + C_h) f_{i,j}^{p,\omega} + \sum_{\{(i,j)\in\mathcal{A} | i<j\}} C_s \cdot x_{i,j}^{\omega} \tag{2a}$$

$$\text{s.t.} \quad \sum_{\substack{\{j\in\mathcal{H}\cup\{t\}\\ |\,(i,j)\in\mathcal{A}\}}} f_{i,j}^{p} - \sum_{\substack{\{j\in\mathcal{H}\cup\{s\}\\ |\,(j,i)\in\mathcal{A}\}}} f_{j,i}^{p} = \begin{cases} d_p^{\omega} & \text{if } i = s \\ 0 & \text{if } i \in \mathcal{H}\,, \\ -d_p^{\omega} & \text{if } i = t \end{cases} \quad \forall p = (s,t) \in \mathcal{P},\ \forall i \in \mathcal{H}\cup\{s,t\}, \tag{2b}$$

$$\sum_{p\in\mathcal{P}} q \cdot f_{i,j}^{p,\omega} \leq Q \cdot x_{i,j}^{\omega}, \qquad \forall (i,j) \in \mathcal{A}, \tag{2c}$$

$$\sum_{\{j\in\mathcal{S}\cup\mathcal{H}\cup\mathcal{T} | (i,j)\in\mathcal{A}\}} x_{i,j}^{\omega} \leq \sum_{k\in\mathcal{K}} S_k \cdot y_i^k, \qquad \forall i \in \mathcal{H}, \tag{2d}$$

$$x^{\omega}_{i,j} = x^{\omega}_{j,i}, \qquad \forall (i,j) \in \mathcal{A} | i < j, \tag{2e}$$

$$x^{\omega}_{i,j} \in \mathbb{Z}_{\geq 0}, \qquad \forall (i,j) \in \mathcal{A}, \tag{2f}$$

$$f^{p,\omega}_{i,j} \geq 0, \qquad \forall (i,j) \in \mathcal{A}, \forall p \in \mathcal{P}. \tag{2g}$$

The transportation costs are given by the objective (2a), obtained by minimizing the total cost of commodity handling, truck scheduling, and truck fuel consumption for commodity delivery. Constraints (2b) are flow conservation constraints to route commodities from their respective origins to their destinations. Constraints (2c) ensure that enough delivery vehicles are scheduled on each leg to feasibly transport the commodities planned to travel along that leg. Constraints (2d) ensure that the number of trucks visiting each hub respects the corresponding hub capacity. Finally, Constraints (2e) ensure that scheduled trucks on each transportation leg return to their origin locations, which in turn facilitates the single-day driver trips. We note that the generated consolidation plan will leverage the backhauling opportunities offered by the back-and-forth trips in the relay network to efficiently transport commodities that may travel in opposite directions.

The two-stage stochastic program is a challenging optimization formulation to solve exactly due to two major reasons. First, the number of decision variables and constraints grows with the number of demand scenarios ($|\Omega|$). Second, for each demand scenario $\omega \in \Omega$, the second-stage subproblem (2) aims to consolidate multiple commodities at minimum cost, which in itself is computationally a challenging problem. Embedding the minimum-cost consolidation planning problem (2) into a facility location, resulting in CRND-SDCR, complicates its solvability even further.

4. Decomposition-Based Branch-and-Cut

Two-stage stochastic programs are typically solved using decomposition-based algorithms, including Benders decomposition (Benders 1962), integer L-shaped method (Laporte and Louveaux 1993), dual decomposition (Carøe and Schultz 1999), and progressive hedging (Rockafellar and Wets 1991). Our formulation (1)-(2) exhibits a nested block-angular structure, which enables us to derive an exact solution approach using the branch-and-cut algorithm, nested with Benders decomposition and the integer L-shaped method.

Classically, Benders decomposition and the integer L-shaped method reformulate a two-stage optimization problem to obtain a (relaxed) master problem and scenario-wise subproblems. The master problem comprises first-stage variables and an additional set of variables that estimate the second-stage objective function value, whereas the scenario-wise subproblems comprise second-stage variables. An iteration in such decomposition algorithms involves: (i) solving the master problem to fix the first-stage variables and in turn making the scenario-wise subproblems independently solvable, (ii) solving each independent scenario-wise subproblem with a fixed master

problem solution, (iii) deriving a cut (or multiple cuts) based on the subproblems' solutions and adding it (them) to the master problem, and (iv) repeating this process until a provably optimal solution to the original problem is found.

To this end, we first reformulate the model (Section 4.1) and then lay out the process of generating integer L-shaped cuts (Section 4.2), guaranteeing the exactness of our solution approach. To accelerate the typically slow convergence, we then describe the process of deriving Benders feedback cuts through nested decomposition (Section 4.3) and embed these cut generation schemes within a branch-and-bound process, resulting in a branch-and-cut algorithm (Section 4.4). Finally, we provide valid inequalities and additional computational enhancements employed to accelerate the convergence of the branch-and-cut algorithm (Section 4.5).

4.1. Model Reformulation

The two-stage stochastic program (1)-(2) exhibits a block angular structure: If we fix the y variables, then the second-stage subproblem for each $\omega \in \Omega$ can be independently solved. We leverage this structure to define an equivalent optimization problem of CRND-SDCR, master problem $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega)$. The master problem comprises the first-stage binary variables y, and a newly defined continuous decision variable $\theta \in \mathbb{R}$ that estimates the expected transportation costs across the demand scenarios. The master problem is then given by:

$$\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega): \min_{y,\theta} \quad \sum_{i \in \mathcal{H}} \sum_{k \in \mathcal{K}} C_i^k \cdot y_i^k + \theta \tag{3a}$$

$$\text{s.t.} \quad \sum_{k \in \mathcal{K}} y_i^k \leq 1, \qquad \forall i \in \mathcal{H}, \tag{3b}$$

$$\text{Integer L-shaped cuts } \mathcal{I}, \tag{3c}$$

$$\text{Benders feedback cuts } \mathcal{B}, \tag{3d}$$

$$y_i^k \in \{0,1\}, \qquad \forall i \in \mathcal{H}, \forall k \in \mathcal{K}. \tag{3e}$$

Due to the very large number of constraints (3c) and (3d), we solve $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega)$ using a branch-and-cut algorithm through lazy constraint callbacks in an off-the-shelf optimization solver. Such implementation entails dynamically adding cuts (3c), (3d) whenever the solver encounters a solution $(\hat{y}, \hat{\theta})$ that satisfies integrality requirements (i.e., $\hat{y}_i^k \in \{0,1\}, \forall i \in \mathcal{H}, \forall k \in \mathcal{K}$) but violates the corresponding integer L-shaped or Benders optimality cuts. The process is terminated when the search space is fully explored and the estimator variable θ correctly records the expected transportation costs. Next, we describe the process of generating integer L-shaped cuts.

4.2. Generating Integer L-Shaped Cuts

Given a feasible first-stage solution of the master problem $(\hat{y}, \hat{\theta})$, we solve the second-stage subproblem $\mathcal{D}_\omega(\hat{y})$ for each $\omega \in \Omega$, given by:

$$\mathcal{D}_\omega(\hat{y}): \min_{x,f} \quad \sum_{p\in\mathcal{P}} \sum_{(i,j)\in\mathcal{A}} (C_f \cdot \ell_{i,j} + C_h) f_{i,j}^{p,\omega} + \sum_{\{(i,j)\in\mathcal{A}|i<j\}} C_s \cdot x_{i,j}^{\omega} \tag{4a}$$

$$\text{s.t.} \quad \sum_{\substack{\{j\in\mathcal{H}\cup\{t\}\\ |\,(i,j)\in\mathcal{A}\}}} f_{i,j}^{p} - \sum_{\substack{\{j\in\mathcal{H}\cup\{s\}\\ |\,(j,i)\in\mathcal{A}\}}} f_{j,i}^{p} = \begin{cases} d_p^\omega & \text{if } i = s \\ 0 & \text{if } i \in \mathcal{H}\,, \\ -d_p^\omega & \text{if } i = t \end{cases} \quad \forall p = (s,t) \in \mathcal{P},\ \forall i \in \mathcal{H}\cup\{s,t\}, \tag{4b}$$

$$\sum_{p\in\mathcal{P}} q \cdot f_{i,j}^{p,\omega} \le Q \cdot x_{i,j}^{\omega}, \qquad \forall (i,j) \in \mathcal{A}, \tag{4c}$$

$$\sum_{\{j\in\mathcal{S}\cup\mathcal{H}\cup\mathcal{T}|(i,j)\in\mathcal{A}\}} x_{i,j}^{\omega} \le \sum_{k\in\mathcal{K}} S_k \cdot \hat{y}_i^k, \qquad \forall i \in \mathcal{H}, \tag{4d}$$

$$x_{i,j}^{\omega} = x_{j,i}^{\omega}, \qquad \forall (i,j) \in \mathcal{A}|i<j, \tag{4e}$$

$$x_{i,j}^{\omega} \in \mathbb{Z}_{\ge 0}, \qquad \forall (i,j) \in \mathcal{A}, \tag{4f}$$

$$f_{i,j}^{p,\omega} \ge 0, \qquad \forall (i,j) \in \mathcal{A}, \forall p \in \mathcal{P}. \tag{4g}$$

The second-stage subproblem $\mathcal{D}_\omega(\hat{y})$ devises for demand scenario ω a minimum-cost consolidation plan on the capacitated hub network given by the first stage decisions $\hat{y}$. However, this problem is challenging to solve. Instead of directly feeding $\mathcal{D}_\omega(\hat{y})$ to an optimization solver, we decompose it by first selecting the number of delivery trucks on each leg, and then determining the flow of commodities:

$$\mathcal{D}_\omega(\hat{y}): \min_{x} \quad \sum_{\{(i,j)\in\mathcal{A}|i<j\}} C_s \cdot x_{i,j}^{\omega} + \psi_\omega^*(x,\hat{y}) \tag{5a}$$

$$\text{s.t.} \quad (4d) - (4e),$$

$$x_{i,j}^{\omega} \in \mathbb{Z}_{\ge 0},\ \forall (i,j) \in \mathcal{A}. \tag{5b}$$

where the innermost subproblem and its dual are given by:

$$\begin{aligned} \psi_\omega^*(x,\hat{y}) &= \min_{f} \sum_{p\in\mathcal{P}} \sum_{(i,j)\in\mathcal{A}} (C_f \cdot \ell_{i,j} + C_h) f_{i,j}^{p,\omega} \\ &\quad \text{s.t.} \quad (4b) - (4c), \\ &\quad\quad f_{i,j}^{p,\omega} \ge 0, \qquad \forall (i,j) \in \mathcal{A}, \forall p \in \mathcal{P}. \\ &= \max_{\tau,\rho} \sum_{p\in\mathcal{P}} d_p^\omega (\tau_s^{p,\omega} - \tau_t^{p,\omega}) - \sum_{(i,j)\in\mathcal{A}} \frac{Q}{q} \cdot x_{i,j}^{\omega} \cdot \rho_{i,j}^{\omega} \\ &\quad \text{s.t.} \quad \tau_i^{p,\omega} - \tau_j^{p,\omega} - \rho_{i,j}^{\omega} \le C_f \cdot \ell_{i,j} + C_h, \qquad \forall (i,j) \in \mathcal{A}, \forall p \in \mathcal{P}, \\ &\quad\quad \rho_{i,j}^{\omega} \ge 0, \qquad \forall (i,j) \in \mathcal{A}. \end{aligned} \tag{6}$$

To circumvent feasibility issues in the primal innermost subproblem (6), we add dummy truck capacities with very high costs between each O-D pair that are sufficient to fulfill commodity demands even when no hub is opened. Then, $\mathcal{D}_\omega(\hat{y})$ is equivalent to the following master problem:

$$\mathcal{E}_\omega(\hat{y}, \mathcal{F}_\omega): \min_{x,\psi} \quad \sum_{\{(i,j)\in\mathcal{A}|i<j\}} C_s \cdot x^\omega_{i,j} + \psi_\omega \tag{7a}$$

$$\text{s.t.} \quad \sum_{\{j\in\mathcal{S}\cup\mathcal{H}\cup\mathcal{T}|(i,j)\in\mathcal{A}\}} x^\omega_{i,j} \le \sum_{k\in\mathcal{K}} S_k \cdot \hat{y}^k_i, \qquad \forall i \in \mathcal{H}, \tag{7b}$$

$$x^\omega_{i,j} = x^\omega_{j,i}, \qquad \forall (i,j) \in \mathcal{A}|i<j, \tag{7c}$$

$$\psi_\omega \ge \sum_{p\in\mathcal{P}} d^\omega_p(\tau^{s,\omega}_p - \tau^{t,\omega}_p) - \sum_{(i,j)\in\mathcal{A}} \frac{Q}{q} \cdot x^\omega_{i,j} \cdot \rho^\omega_{i,j}, \quad \forall (\tau,\rho) \in \mathcal{F}_\omega, \tag{7d}$$

$$x^\omega_{i,j} \in \mathbb{Z}_{\ge 0}, \qquad \forall (i,j) \in \mathcal{A}. \tag{7e}$$

where $\mathcal{F}_\omega$ contains all basic feasible solutions to the dual innermost subproblem (6). We solve $\mathcal{E}_\omega(\hat{y}, \mathcal{F}_\omega)$ itself with a branch-and-cut algorithm using constraint generation through lazy constraint callback implementation of a modern off-the-shelf optimization solver. The algorithm starts with the LP relaxation of the master problem $\mathcal{E}_\omega(\hat{y}, \mathcal{F}'_\omega)$ where $\mathcal{F}'_\omega = \emptyset$ at the root node of a branch-and-bound tree. During the process, whenever the solver encounters a feasible solution $(\hat{x}, \hat{\psi})$ that satisfies integrality constraints (7e), we determine if it violates any of the constraints (7d) in the original relaxed master problem $\mathcal{E}_\omega(\hat{y}, \mathcal{F}_\omega)$. To this end, we compute an optimal dual solution (τ^*, ρ^*) of the innermost subproblem (6). If the corresponding constraint (7d) is violated by $(\hat{x}, \hat{\psi})$, then (τ^*, ρ^*) is added to $\mathcal{F}'_\omega$ and the LP relaxation of $\mathcal{E}_\omega(\hat{y}, \mathcal{F}'_\omega)$ is solved again. The branching and pruning process is handled by the solver to explore the search space efficiently. This Branch-and-Benders decomposition algorithm terminates when the optimal solution to the relaxed master problem satisfies every constraint (7d) and the search space is completely explored. Our implementation of Branch-and-Benders decomposition for solving $\mathcal{D}_\omega(\hat{y})$ for each $\omega \in \Omega$ is detailed in Algorithm 1.

Let $S^*_\omega(\hat{y})$ be the optimal objective of $\mathcal{D}_\omega(\hat{y})$ obtained through Algorithm 1 for each $\omega \in \Omega$. If the corresponding integer L-shaped cut (3c) is violated by the current feasible first-stage solution $(\hat{y}, \hat{\theta})$, we add it to $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega)$ as follows:

$$\theta \ge \Big(\sum_{\omega\in\Omega} \pi_\omega \cdot S^*_\omega(\hat{y})\Big)\Big(1 - \sum_{\{(i,k)\in\mathcal{H}\times\mathcal{K}\,|\,\hat{y}^k_i=1\}} (1-y^k_i) \quad - \sum_{\{(i,k)\in\mathcal{H}\times\mathcal{K}\,|\,\hat{y}^k_i=0\}} y^k_i\Big). \tag{8}$$

4.3. Generating Benders Optimality Cuts

Despite guaranteeing termination to an exact solution of CRND-SDCR, solely adding integer L-shaped cuts leads to a slow overall convergence. Hence, in order to accelerate the procedure's

Algorithm 1: Branch-and-Benders Decomposition for Solving $\mathcal{D}_\omega(\hat{y})$ (BD($\mathcal{D}_\omega(\hat{y})$))

Input : Stochastic demand scenario $\omega \in \Omega$ and integer feasible first-stage solution $\hat{y} \in \{0,1\}^{|\mathcal{H}|\times|\mathcal{K}|}$

Output: Optimal objective function cost of $\mathcal{D}_\omega(\hat{y})$ given by $S^*_\omega(\hat{y})$

Initialize: List of branch-and-bound tree nodes $\mathcal{N} \leftarrow \{\text{root node}\}$, $S^*_\omega(\hat{y}) \leftarrow +\infty$, $(x^*, \psi^*) \leftarrow \emptyset$, $\mathcal{F}'_\omega \leftarrow \emptyset$;
while $\mathcal{N} \neq \emptyset$ **do**
  Choose a node $i \in \mathcal{N}$ and solve the LP relaxation of $\mathcal{E}_\omega(\hat{y}, \mathcal{F}'_\omega)$ at node i: $(\hat{x}, \hat{\psi}) \leftarrow$ optimal solution, $\hat{S}_\omega(\hat{y}) \leftarrow$ optimal value ;
  if $\hat{S}_\omega(\hat{y}) < S^*_\omega(\hat{y})$ **then**
    if $\hat{x}_{i,j} \in \mathbb{Z}_{\geq 0}, \forall (i,j) \in \mathcal{A}$ **then**
      if $\hat{\psi} < \sum_{p\in\mathcal{P}} d^\omega_p(\tau^{p,\omega}_s - \tau^{p,\omega}_t) - \sum_{(i,j)\in\mathcal{A}} \frac{Q}{q} \cdot \hat{x}^\omega_{i,j} \cdot \rho^\omega_{i,j}$ **then**
        $\mathcal{F}'_\omega \leftarrow \mathcal{F}'_\omega \cup \{\tau, \rho\}$;
      if $\hat{\psi} = \sum_{p\in\mathcal{P}} d^\omega_p(\tau^{p,\omega}_s - \tau^{p,\omega}_t) - \sum_{(i,j)\in\mathcal{A}} \frac{Q}{q} \cdot \hat{x}^\omega_{i,j} \cdot \rho^\omega_{i,j}$ **then**
        $\mathcal{N} \leftarrow \mathcal{N} \setminus \{i\}$;
        $(x^*, \psi^*) \leftarrow (\hat{x}, \hat{\psi})$, $S^*_\omega(\hat{y}) \leftarrow \hat{S}_\omega(\hat{y})$;
    else
      Branch on a variable $\hat{x}_{i,j} \notin \mathbb{Z}_{\geq 0}$ to create two nodes i_1 and i_2 with additional constraints of $x_{i,j} \leq \lfloor \hat{x}_{i,j} \rfloor$ and $x_{i,j} \geq \lceil \hat{x}_{i,j} \rceil$ in LP relaxations of $\mathcal{E}_\omega(\hat{y}, \mathcal{F}'_\omega)$ at i_1 and i_2 respectively ;
      $\mathcal{N} \leftarrow (\mathcal{N} \setminus \{i\}) \cup \{i_1, i_2\}$;
  else
    $\mathcal{N} \leftarrow \mathcal{N} \setminus \{i\}$;
return $S^*_\omega(\hat{y})$

convergence, we add Benders optimality cuts to $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega)$ as well. To this end, at any feasible first-stage solution $(\hat{y}, \hat{\theta})$, we solve the dual of the LP relaxation of $\mathcal{D}_\omega(\hat{y})$, given by:

$$\mathcal{L}_\omega(\hat{y}): \max_{\alpha,\beta,\gamma,\lambda} \quad \sum_{p\in\mathcal{P}} d^\omega_p(\alpha^{s,\omega}_p - \alpha^{t,\omega}_p) - \sum_{i\in\mathcal{H}}\sum_{k\in\mathcal{K}} S_k \cdot \hat{y}^k_i \cdot \gamma^\omega_i \tag{9a}$$

$$\text{s.t.} \quad \alpha^{i,\omega}_p - \alpha^{j,\omega}_p - \beta^\omega_{i,j} \leq C_f \cdot \ell_{i,j} + C_h, \qquad \forall (i,j) \in \mathcal{A}, \forall p \in \mathcal{P}, \tag{9b}$$

$$\frac{Q}{q} \cdot \beta^\omega_{i,j} - \gamma^\omega_i \cdot \mathbb{1}_{\{i\in\mathcal{H}\}} = \begin{cases} C_s - \lambda^\omega_{i,j} & \text{if } i < j \\ \lambda^\omega_{j,i} & \text{if } i > j \end{cases}, \quad \forall (i,j) \in \mathcal{A}, \tag{9c}$$

$$\beta^\omega_{i,j} \geq 0, \qquad \forall (i,j) \in \mathcal{A}, \tag{9d}$$

$$\gamma^\omega_i \geq 0, \qquad \forall i \in \mathcal{H}, \tag{9e}$$

$$\lambda^\omega_{i,j} \geq 0, \qquad \forall (i,j) \in \mathcal{A} | i < j. \tag{9f}$$

We observe that $\mathcal{L}_\omega(\hat{y})$ is the dual of a capacitated multi-commodity minimum-cost network flow problem that requires substantial computational effort and time to be solved even by a modern optimization solver due to the large number of commodities. Thus, we decompose the problem by first selecting the dual variables (β,γ,λ), and then determining the optimal dual variables α for each O-D pair independently. The resulting Benders decomposition is formulated as follows:

$$\mathcal{L}_\omega(\hat{y}):\ \max_{\beta,\gamma,\lambda}\ \sum_{p\in\mathcal{P}} d_p^\omega\cdot\eta_p^{\omega^*}(\beta,\gamma,\lambda,\hat{y})-\sum_{i\in\mathcal{H}}\sum_{k\in\mathcal{K}} S_k\cdot\hat{y}_i^k\cdot\gamma_i^\omega \tag{10a}$$
$$\text{s.t.}\quad (9c)-(9f)$$

where for every O-D pair $p\in\mathcal{P}$, the innermost subproblem and its dual are given by

$$\eta_p^{\omega^*}(\beta,\gamma,\lambda,\hat{y})=\max_\alpha\ \alpha_p^{s,\omega}-\alpha_p^{t,\omega}\quad=\min_u\ \sum_{(i,j)\in\mathcal{A}}(C_f\cdot\ell_{i,j}+C_h+\beta_{i,j}^\omega)\cdot u_{i,j}^{p,\omega}$$
$$\text{s.t.}\quad (9b),\qquad\qquad \text{s.t.}\ \sum_{\{j\in\mathcal{H}\cup\{t\}\,|\,(i,j)\in\mathcal{A}\}} u_{i,j}^{p,\omega}-\sum_{\{j\in\mathcal{H}\cup\{s\}\,|\,(j,i)\in\mathcal{A}\}} u_{j,i}^{p,\omega}=\begin{cases}1 & \text{if } i=s\\ 0 & \text{if } i\in\mathcal{H}\\ -1 & \text{if } i=t\end{cases}, \tag{11}$$
$$u_{i,j}^{p,\omega}\geq 0,\qquad \forall(i,j)\in\mathcal{A}.$$

With this, $\mathcal{L}_\omega(\hat{y})$ is equivalent to the following problem:

$$\mathcal{J}_\omega(\hat{y},\mathcal{C}_\omega):\quad \max_{\beta,\gamma,\lambda,\eta}\ \sum_{p\in\mathcal{P}} d_p^\omega\cdot\eta_p^\omega-\sum_{i\in\mathcal{H}}\sum_{k\in\mathcal{K}} S_k\cdot\hat{y}_i^k\cdot\gamma_i^\omega \tag{12a}$$
$$\text{s.t.}\quad \frac{Q}{q}\cdot\beta_{i,j}^\omega-\gamma_i^\omega\cdot\mathbb{1}_{\{i\in\mathcal{H}\}}=\begin{cases}C_s-\lambda_{i,j}^\omega & \text{if } i<j\\ \lambda_{j,i}^\omega & \text{if } i>j\end{cases},\qquad \forall(i,j)\in\mathcal{A}, \tag{12b}$$
$$\eta_p^\omega\leq\sum_{(i,j)\in\mathcal{A}}(C_f\cdot\ell_{i,j}+C_h)u_{i,j}^{p,\omega}+\sum_{(i,j)\in\mathcal{A}} u_{i,j}^{p,\omega}\cdot\beta_{i,j}^\omega,\quad \forall u^{p,\omega}\in\mathcal{C}_{p,\omega},\forall p\in\mathcal{P}, \tag{12c}$$
$$\beta_{i,j}^\omega\geq 0,\qquad \forall(i,j)\in\mathcal{A}, \tag{12d}$$
$$\gamma_i^\omega\geq 0,\qquad \forall i\in\mathcal{H}, \tag{12e}$$
$$\lambda_{i,j}^\omega\geq 0,\qquad \forall(i,j)\in\mathcal{A}|i<j. \tag{12f}$$

Here, $\mathcal{C}_{p,\omega}$ contains all basic feasible solutions to the dual innermost subproblem (11). Due to the very large number of Benders optimality cuts (12c), we solve it using constraint generation. We note that due to the presence of only continuous variables in $\mathcal{J}_\omega(\hat{y},\mathcal{C}_\omega)$, lazy constraint callback implementation does not provide any additional benefit as compared to the classical implementation of Benders decomposition. Hence, we opt for the classical implementation where we start with solving a relaxed version of $\mathcal{J}_\omega(\hat{y},\mathcal{C}'_\omega)$ with $\mathcal{C}'_{p,\omega}=\emptyset$ for every $p\in\mathcal{P}$; let $(\beta^\omega,\gamma^\omega,\lambda^\omega,\eta^\omega)$ be its optimal solution. We then determine if this solution violates any of the constraints (12c) by computing for each O-D pair $p\in\mathcal{P}$ an optimal dual solution $u^{p,\omega}$ of the innermost subproblem (11).

If the corresponding constraint (12c) is violated by $(\beta^\omega, \gamma^\omega, \lambda^\omega, \eta^\omega)$, then $u^{p,\omega}$ is added to $\mathcal{C}'_{p,\omega}$ and the relaxed master problem is solved again. Benders decomposition terminates when the optimal solution to the relaxed master problem satisfies every constraint (12c).

At each iteration of the algorithm, we must solve the dual innermost subproblem (11) for each O-D pair $p \in \mathcal{P}$. A close inspection reveals that (11) can be cast as a shortest path problem between O-D pair p. We construct a new graph $\mathcal{G}^p_\omega$ that is identical to $\mathcal{G}$ and set the length of each edge (i,j) to $C_f \cdot \ell_{i,j} + C_h + \beta^\omega_{i,j}$. Then, the dual innermost subproblem (11) can be solved by determining a shortest path μ^ω_p between O-D pair p in the graph $\mathcal{G}^p_\omega$, which can be computed using Dijkstra's algorithm. Thus, the corresponding Benders optimality cut (12c) can be computed in polynomial time. Our detailed implementation of Benders decomposition tailored for solving $\mathcal{L}_\omega(\hat{y})$ is described in Algorithm 2. We note that we use the multi-cut version of Benders decomposition, i.e., we add a Benders optimality cut for each O-D pair instead of the single-cut option due to its better convergence properties (Birge and Louveaux 1988).

Algorithm 2: Benders Decomposition for solving $\mathcal{L}_\omega(\hat{y})$ (BD($\mathcal{L}_\omega(\hat{y})$))

Input : Stochastic demand scenario $\omega \in \Omega$, integer feasible first-stage solution $\hat{y} \in \{0,1\}^{|\mathcal{H}|\times|\mathcal{K}|}$, optimality gap $\epsilon \geq 0$

Output: Optimal objective function value of $\mathcal{L}_\omega(\hat{y})$ given by $L^*_\omega(\hat{y})$ and optimal solution of $\mathcal{J}_\omega(\hat{y}, \mathcal{C}_\omega)$ given by $(\beta^\omega, \gamma^\omega, \lambda^\omega, \eta^\omega)$

Initialize: $LB \leftarrow -\infty$, $UB \leftarrow +\infty$, $\mathcal{C}'_{p,\omega} \leftarrow \emptyset \ \forall p \in \mathcal{P}$;
while $UB - LB > \epsilon$ **do**
  Solve $\mathcal{J}_\omega(\hat{y}, \mathcal{C}'_{p,\omega}) : (\beta^\omega, \gamma^\omega, \lambda^\omega, \eta^\omega) \leftarrow$ optimal solution, $UB \leftarrow$ optimal value ;
  for every $p \in \mathcal{P}$ **do**
    Create an auxiliary graph: $\mathcal{G}^p_\omega \leftarrow \mathcal{G}$;
    for each edge (i,j) in $\mathcal{A}$ **do**
      Set the length of edge (i,j) in $\mathcal{G}^p_\omega$ to $C_f \cdot \ell_{i,j} + C_h + \beta^\omega_{i,j}$;
    $\mu_{p,\omega} \leftarrow$ shortest path between O-D pair p in $\mathcal{G}^p_\omega$ using Dijkstra's algorithm , $\ell_{\mu_{p,\omega}} \leftarrow$ shortest path length in $\mathcal{G}^p_\omega$;
    $u^{p,\omega}_{i,j} \leftarrow \mathbb{1}_{\{(i,j)\in\mu_{p,\omega}\}}$ for every $(i,j) \in \mathcal{A}$;
    if $\eta^\omega_p > \ell_{\mu_{p,\omega}}$ **then**
      $\mathcal{C}'_{p,\omega} \leftarrow \mathcal{C}'_{p,\omega} \cup \{u^{p,\omega}\}$;
  $LB \leftarrow \max\{LB, \sum_{p\in\mathcal{P}} d^\omega_p \cdot \ell_{\mu_{p,\omega}} - \sum_{i\in\mathcal{H}} \sum_{k\in\mathcal{K}} S_k \cdot \hat{y}^k_i \cdot \gamma^\omega_i\}$, $L^*_\omega(\hat{y}) \leftarrow LB$;
return $(L^*_\omega(\hat{y}), (\beta^\omega, \gamma^\omega, \lambda^\omega, \eta^\omega))$

We can now complete the generation of Benders optimality cuts for the original master problem $\mathcal{M}(\mathcal{I},\mathcal{B},\Omega)$. Once Algorithm 2 computes the optimal solutions $(\beta^\omega,\gamma^\omega,\lambda^\omega,\eta^\omega)$ of $\mathcal{L}_\omega(\hat{y})$ for every $\omega\in\Omega$, we then add the following cut, if violated by the current feasible first-stage solution $(\hat{y},\hat{\theta})$:

$$\theta \geq \sum_{\omega\in\Omega} \pi_\omega \cdot \left(\sum_{p\in\mathcal{P}} d_p^\omega \cdot \eta_p^\omega - \sum_{i\in\mathcal{H}} \sum_{k\in\mathcal{K}} S_k \cdot y_i^k \cdot \gamma_i^\omega \right). \tag{13}$$

4.4. Overall Branch-and-Cut Algorithm

In order to solve (1)-(2), we initially solve the LP relaxation of $\mathcal{M}(\mathcal{I}',\mathcal{B}',\Omega)$ with $\mathcal{I}'_\omega=\emptyset$ and $\mathcal{B}'_\omega=\emptyset$ for every $\omega\in\Omega$ at the root node of a branch-and-bound tree. During the solution process, whenever the solver encounters a feasible first-stage solution $(\hat{y},\hat{\theta})$ that satisfies integrality constraints at a node, we determine if it violates any Benders optimality cuts (13) and integer L-shaped cuts (8). For the Benders optimality cuts, we solve the dual of the LP relaxation of $\mathcal{D}_\omega(\hat{y})$, given by $\mathcal{L}_\omega(\hat{y})$ for every $\omega\in\Omega$ using Benders decomposition via Algorithm 2. The expected optimal objective function value of $\mathcal{L}_\omega(\hat{y})$, given by $\sum_{\omega\in\Omega}\pi_\omega\cdot L^*_\omega(\hat{y})$, provides a lower bound on θ, and we add the corresponding Benders optimality cut (13) if $\hat{\theta}$ violates that lower bound. For the integer L-shaped cuts, we solve $\mathcal{D}_\omega(\hat{y})$ for every $\omega\in\Omega$ through branch-and-Benders decomposition using Algorithm 1. The expected optimal objective function value of $\mathcal{D}_\omega(\hat{y})$, given by $\sum_{\omega\in\Omega}\pi_\omega\cdot S^*_\omega(\hat{y})$, provides the exact value of θ at the current first-stage solution $\hat{y}$. Thus, we add the integer L-shaped cut (8) if $\hat{\theta}$ does not record this value correctly.

In the branch-and-bound process, we update the incumbent solution (y^*,z^*) whenever we encounter an integer feasible solution $(\hat{y},\hat{\theta})$, the solution records the subproblem objective function value correctly, and its objective function value is strictly better than that of the previous incumbent. Due to the lazy constraint callback implementation, the optimization solver handles the branch-and-bound process including branching and pruning. At termination, we return the incumbent as the optimal solution for (1)-(2). The detailed algorithm is described in Algorithm 3 and the overview is presented in Figure 1.

4.5. Computational Enhancements

To further improve the computational performance of the decomposition-based branch-and-cut algorithm presented above, we next develop strategies to strengthen our formulation and warm start the algorithm.

4.5.1. Strengthening $\mathcal{M}(\mathcal{I},\mathcal{B},\Omega)$. Typically, decomposition-based branch-and-cut algorithms suffer from ineffective initial iterations due to the generation of low-quality solutions (Rahmaniani et al. 2018). One of the strategies that we adopt to tackle this issue is to add valid inequalities to $\mathcal{M}(\mathcal{I},\mathcal{B},\Omega)$. In particular, all commodities originating at a location $s\in\mathcal{S}$ must travel to hubs that are adjacent in $\mathcal{G}$ to s. Thus, a valid inequality is ensures that there is sufficient

Algorithm 3: Decomposition-Based Branch-and-Cut for Solving (1)-(2)

Input : Graph $\mathcal{G}=(\mathcal{S}\cup\mathcal{H}\cup\mathcal{T},\mathcal{A})$, stochastic demand scenarios $\omega\in\Omega$, vector of commodity demand $(d_p^\omega)_{p\in\mathcal{P},\omega\in\Omega}$, vector of occurrence probability $(\pi_\omega)_{\omega\in\Omega}$, Cost parameters C_s, C_h, C_f, and $(C_i^k)_{(i\in\mathcal{H}.k\in\mathcal{K})}$

Output: Optimal subset of hubs to open and their respective sizes y^* and optimal objective function cost of (1) z^*

Initialize: List of branch-and-bound tree nodes $\mathcal{N}'\leftarrow\{\text{root node}\}$, $z^*\leftarrow+\infty$, $y^*\leftarrow\emptyset$, $\mathcal{I}'_\omega\leftarrow\emptyset$, $\mathcal{B}'_\omega\leftarrow\emptyset$, $\forall\omega\in\Omega$;
while $\mathcal{N}\neq\emptyset$ **do**
  Choose a node $i\in\mathcal{N}'$ and solve the LP relaxation of $\mathcal{M}(\mathcal{I}',\mathcal{B}',\Omega)$ at node i: $(\hat{y},\hat{\theta})\leftarrow$ optimal solution, $\hat{z}\leftarrow$ optimal value ;
  if $\hat{z}<z^*$ **then**
    if $\hat{y}\in\{0,1\}^{|\mathcal{H}|\times|\mathcal{K}|}$ **then**
      for every $\omega\in\Omega$ **do**
        $S^*_\omega(\hat{y})\leftarrow \text{BD}(\mathcal{D}_\omega(\hat{y}))$ using Algorithm 1 ;
        $(L^*_\omega(\hat{y}),(\beta^\omega,\gamma^\omega,\lambda^\omega,\eta^\omega))\leftarrow \text{BD}(\mathcal{L}_\omega(\hat{y}))$ using Algorithm 2 ;
      if $\hat{\theta}<\sum_{\omega\in\Omega}\pi_\omega\cdot S^*_\omega(\hat{y})$ **then**
        Add Integer L-Shaped Cut (8): $\mathcal{I}'_\omega\leftarrow\mathcal{I}'_\omega\cup\{S^*_\omega(\hat{y})\}$;
      if $\hat{\theta}<\sum_{\omega\in\Omega}\pi_\omega\cdot L^*_\omega(\hat{y})$ **then**
        Add Benders Optiamlity Cut (13): $\mathcal{B}'_\omega\leftarrow\mathcal{B}'_\omega\cup\{(\beta^\omega,\gamma^\omega,\lambda^\omega,\eta^\omega)\}$;
      if $\hat{\theta}=\sum_{\omega\in\Omega}\pi_\omega\cdot S^*_\omega(\hat{y})$ **then**
        $\mathcal{N}'\leftarrow\mathcal{N}'\setminus\{i\}$;
        $y^*\leftarrow\hat{y}$, $z^*\leftarrow\hat{z}$;
    else
      Branch on a variable $\hat{y}_i^k\notin\{0,1\}$ to create two nodes i_1 and i_2 with additional constraints of $y_i^k=0$ and $y_i^k=1$ in LP relaxations of $\mathcal{M}(\mathcal{I}',\mathcal{B}',\Omega)$ at i_1 and i_2, respectively ;
      $\mathcal{N}'\leftarrow(\mathcal{N}'\setminus\{i\})\cup\{i_1,i_2\}$;
  else
    $\mathcal{N}'\leftarrow\mathcal{N}'\setminus\{i\}$;
return (y^*,z^*)

hub capacity adjacent to s to handle the largest commodity demand originating at that location. Analogously, another valid inequality ensures that there is enough hub capacity adjacent to any destination location $t\in\mathcal{T}$ to handle the largest commodity demand terminating at that location. The valid inequalities are given by:

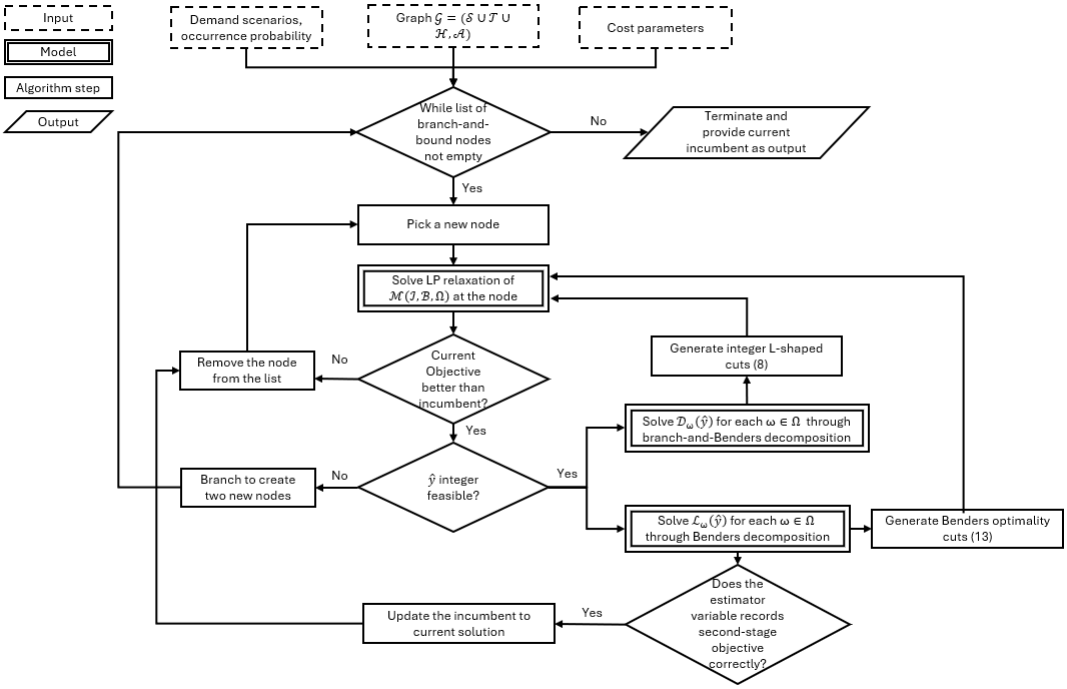

Figure 1 **Overview of the decomposition-based branch-and-cut algorithm.**

$$\sum_{\{i \in \mathcal{H} \,|\, (s,i) \in \mathcal{A}\}} \sum_{k \in \mathcal{K}} S_k \cdot y_i^k \;\geq\; \frac{q}{Q} \cdot \max_{\omega \in \Omega} \left\{ \sum_{\substack{\{t \in \mathcal{T} \\ |\, p=(s,t) \in \mathcal{P}\}}} d_p^\omega \right\}, \qquad \forall s \in \mathcal{S} \tag{14}$$

$$\sum_{\{i \in \mathcal{H} \,|\, (i,t) \in \mathcal{A}\}} \sum_{k \in \mathcal{K}} S_k \cdot y_i^k \;\geq\; \frac{q}{Q} \cdot \max_{\omega \in \Omega} \left\{ \sum_{\substack{\{s \in \mathcal{S} \\ |\, p=(s,t) \in \mathcal{P}\}}} d_p^\omega \right\}, \qquad \forall t \in \mathcal{T} \tag{15}$$

Another strategy we adopt is to include information from the scenario subproblems (2) in the master problem $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega)$. However, instead of including explicit information as in Crainic et al. (2021), we only include implicit information about the origin and destination locations from the subproblems. The underlying idea is to route a very small amount of commodity flow between the O-D pairs in order to obtain a connected relay network even in the initial iterations of the decomposition-based branch-and-cut algorithm.

To this end, we create a new graph $\mathcal{G}'$, which consists of extending the original graph $\mathcal{G}$ by adding a "super" sink node t' and a set of transportation legs $\mathcal{A}'$ between each $t \in \mathcal{T}$ and t'. Then, we define continuous dummy flow variables $w_{i,j}^{s,t'}$ for each origin $s \in \mathcal{S}$ and each transportation leg $(i,j) \in \mathcal{A}^\dagger = \mathcal{A} \cup \mathcal{A}'$. We also define $\mathcal{T}_s$ as the set of destination locations each origin serves (i.e., $\mathcal{T}_s = \{t \in \mathcal{T} \mid \exists \omega \in \Omega, p = (s,t) \in \mathcal{P} : d_p^\omega > 0\}$). Then, given a small number $\epsilon > 0$, the updated master problem $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega)$ becomes:

$$\min_{y,\theta} \quad \sum_{i \in \mathcal{H}} \sum_{k \in \mathcal{K}} C_i^k \cdot y_i^k + \theta \tag{16a}$$

s.t. (3*b*) – (3*d*), (14) – (15),

$$\sum_{\substack{\{j\in\mathcal{H}\cup\mathcal{T}_s\cup\{t'\}\\ |\,(i,j)\in\mathcal{A}^\dagger\}}} w_{i,j}^{s,t'} - \sum_{\substack{\{j\in\mathcal{H}\cup\mathcal{T}_s\cup\{s\}\\ |\,(j,i)\in\mathcal{A}^\dagger\}}} w_{j,i}^{s,t'} = \begin{cases} |\mathcal{T}_s|\cdot\epsilon & \text{if } i=s \\ 0 & \text{if } i\in\mathcal{H}\cup\mathcal{T}_s \\ -|\mathcal{T}_s|\cdot\epsilon & \text{if } i=t' \end{cases}, \quad \forall s\in\mathcal{S},\ \forall i\in\mathcal{H}\cup\mathcal{T}_s\cup\{s,t'\}, \tag{16b}$$

$$\sum_{s\in\mathcal{S}}\sum_{\substack{\{j\in\mathcal{H}\cup\mathcal{T}_s\cup\{t'\}\\ |\,(i,j)\in\mathcal{A}^\dagger\}}} w_{i,j}^{s,t'} \le \sum_{k\in\mathcal{K}} S_k\cdot y_i^k, \qquad \forall i\in\mathcal{H}, \tag{16c}$$

$$w_{t,t'}^{s,t'} \le \epsilon, \qquad \forall (t,t')\in\mathcal{A}' : t\in\mathcal{T}_s, \forall s\in\mathcal{S}, \tag{16d}$$

$$y_i^k\in\{0,1\}, \qquad \forall i\in\mathcal{H}, \forall k\in\mathcal{K}, \tag{16e}$$

$$w_{i,j}^{s,t'} \ge 0, \qquad \forall (i,j)\in\mathcal{A}^\dagger, \forall s\in\mathcal{S}. \tag{16f}$$

Constraints (16b)-(16d) ensure a feasible path through the relay hub network between each possible O-D pair with non-zero commodity demand in at least one demand scenario. Such constraints remove solutions $(\hat{y},\hat{\theta})$ in which the relay network is not connected. We also note that such a specific dummy flow variable definition limits the number of dummy variables to add to the master problem.

4.5.2. Acceleration strategies for Algorithm 3. In order to increase the convergence rate of Algorithm 3, we employ two acceleration strategies. First, we initialize the algorithm by adding a pre-determined number of Benders cuts (13), which we compute by solving the LP relaxation of $\mathcal{M}(\mathcal{I},\mathcal{B},\Omega)$ iteratively. Such an initialization yields a lesser number of branches in the branch-and-bound tree and helps reduce the overall computational burden of Algorithm 3 (Chen and Luedtke 2022).

Another acceleration strategy that we employ is to add a no-good type of cut whenever the algorithm encounters a characteristic first-stage solution $(\hat{y},\hat{\theta})$ that has positive amount of commodity flow on the dummy arcs. To ensure that the subproblem was $\mathcal{D}_\omega(\hat{y})$ always feasible, we previously added dummy arcs between each O-D pair at a high premium unit price. Thus, whenever the algorithm obtains a solution for each some commodity flow traverses these dummy arcs, it actually represents an infeasible solution in practice. As a result, we add the following no-good cut to discard the current first-stage solution:

$$\sum_{\{(i,k)\in\mathcal{H}\times\mathcal{K}\,|\,\hat{y}_i^k=1\}}(1-y_i^k) \quad + \quad \sum_{\{(i,k)\in\mathcal{H}\times\mathcal{K}\,|\,\hat{y}_i^k=0\}} y_i^k \quad \ge 1. \tag{17}$$

5. Computational Study

This section presents the results from the computational experiments conducted using data instances from a US-based car manufacturer company for its finished vehicle deliveries to (i) validate the CRND-SDCR model, (ii) test the scalability of the solution approaches developed in

Section 4, and (iii) analyze the efficiency of the designed networks under commodity demand variability. All algorithms were implemented in Python v.3.8 and all optimization problems were solved using Gurobi v.10.0.3 on an AMD EPYC processor (with IBPB) 2.50 GHz (1 core), with 255 GB assigned RAM and Windows Server 2012 R2 standards, 64-bit operating system, and x64-based processor.

5.1. Data Instances

Using the data of one of the US-based car manufacturer companies that partnered with our research team, we create 5 representative problem instances of increasing size and complexity. These instances differ in the number of O-D pairs, hub candidate locations, and transportation legs with each instance representing end-to-end logistics operations of delivering finished vehicles spanning over a broad geographical region of the south-east US. In every instance, the finished vehicle or commodity demand originates at one of the company's production facilities, railheads, or seaports ($\mathcal{S}$) and is destined for one of the dealerships of the company ($\mathcal{T}$). As the company intends to implement relay transportation for finished vehicle delivery operations, it identified a set $\mathcal{H}$ of candidate locations to open relay hubs, situated at the intersection of major highways and in major cities. These relay hubs will serve as facilities where commodities are unloaded from incoming trucks, sorted, and then loaded into the outgoing trucks. Figure 2 shows the locations of hub candidates, origins, and destinations for Instance 5.

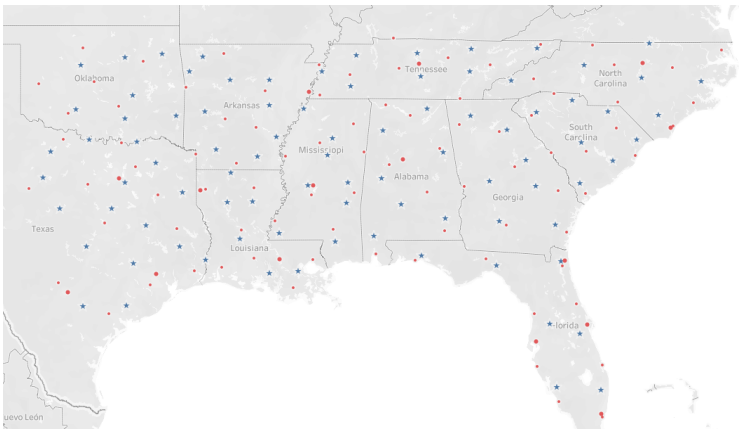

Figure 2 **Instance 5 with hub candidate locations (blue asterisks), and origins and/or destinations (red circles)**

To determine the set $\mathcal{A}$ of feasible transportation legs between facilities, we first compute the distance $\ell_{i,j}$ between every pair of locations $(i,j) \in (\mathcal{S} \cup \mathcal{T} \cup \mathcal{H})^2$ using the Haversine formula for the estimated traveled distance and using an average driving speed of 45 miles/hour. Since the Federal Motor Carrier Safety Administration imposes an 11-hour daily driving limit for truck drivers, we then only retain the transportation legs for which the distance does not exceed $45 \times 5.5 = 250$ miles.

This ensures that commodities travel towards their respective destinations while drivers return home daily. The characteristics for all data instances used in this study are described in Table 1.

Table 1 Data instance characteristics

Data instance	# Origins $\lvert\mathcal{S}\rvert$	# Destinations $\lvert\mathcal{T}\rvert$	# O-D pairs $\lvert\mathcal{P}\rvert$	# Candidates $\lvert\mathcal{H}\rvert$	# Edges $\lvert\mathcal{A}\rvert$
1	4	15	27	31	617
2	8	26	94	50	1,207
3	8	42	134	65	1,487
4	10	47	158	76	1,890
5	15	87	460	88	3,626

To model demand stochasticity, we create a set of 272 stochastic demand scenarios (Ω) based on the annual finished vehicle delivery demand recorded by our industry partner. Each stochastic scenario $\omega \in \Omega$ represents a duration of 96 hours (4 days) of logistics operations of our partner and provides the details of the upcoming finished vehicle demand d_p^ω that each O-D pair $p \in \mathcal{P}$ will face in the next 96 hours. Figure 3 represents the demand variability that each O-D pair $p \in \mathcal{P}$ faces in Instance 1. It can be observed that the demand d_p^ω varies dramatically across the scenarios and hence requires a relay hub to be designed with the necessary capacity to withstand it. We remark that designing relay networks for such instances is significantly complex. These instances, coupled with the stochastic demand scenarios, serve as one of the largest instances for relay network design. Next, we test our developed solution methodology and showcase their computational performance in solving the CRND-SDCR problem for these representative data instances.

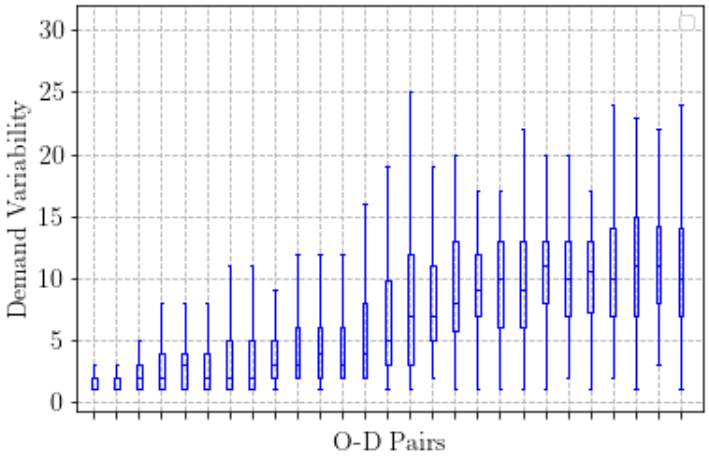

Figure 3 Demand variability for each O-D pair across the stochastic demand scenarios in Instance 1

5.2. Solution Methodology Performance Assessment

We run the solution approach developed in Section 4 to solve CRND-SDCR for each data instance. However, due to the large number of stochastic demand scenarios $\omega \in \Omega$, solving CRND-SDCR to optimality for these instances is computationally challenging. Hence, we utilize sample average approximation (SAA) to solve CRND-SDCR and provide a near-optimal solution with provable statistical guarantees and moderate computational time (Kleywegt, Shapiro, and Homem-de Mello 2002). To this end, we create a set $\mathcal{R}$ of independent samples of stochastic scenarios. Each sample $r \in \mathcal{R}$ consists of N demand scenarios $\Omega_r = \{\omega_1, \ldots, \omega_N\}$ drawn randomly from the set of available demand scenarios Ω with a probability distribution of π^r_ω. Then, instead of solving $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega)$, we solve $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega_r)$ for each sample $r \in \mathcal{R}$. Let $\hat{z}_r$ and $\hat{y}_r$ be the optimal objective function value and optimal solution to $\mathcal{M}(\mathcal{I}, \mathcal{B}, \Omega_r)$, respectively. Then, the provably near-optimal solution y^* to CRND-SDCR is given by the solution $\hat{y}_r$ that provides the lowest expected consolidation cost across *all* stochastic scenarios, $\mathbb{E}_\omega[\mathbb{T}(\hat{y}_r, \omega)]$. For the purpose of this case study, we created a set of 50 independent samples ($|\mathcal{R}|$) and within each sample we picked 30 random demand scenarios i.e., $N = 30$ for every instance.

In order to benchmark our developed solution approach, we compare it with a classical decomposition-based branch-and-cut algorithm wherein at every integer feasible first-stage solution $\hat{y}$, $\mathcal{D}_\omega(\hat{y})$ and $\mathcal{L}_\omega(\hat{y})$ are solved for each $\omega \in \Omega$ directly using the off-the-shelf optimization solver. For our proposed decomposition-based branch-and-cut algorithm, we run the computational experiments with and without including the computational enhancements described in Section 4.5. Table 2 compares the resulting average optimality gaps for the instances after 72-hour time limit.

Table 2 Average optimality gaps after a 72-hour time limit.

Data instance	Classical decomposition-based branch-and-cut (%)	Proposed decomposition-based branch-and-cut (%)	Proposed decomposition-based branch-and-cut + computational enhancements (%)
1	32.74	11.28	3.69
2	47.91	17.83	6.25
3	60.02	23.19	10.93
4	69.77	30.55	18.44
5	81.32	34.18	24.89

We observe that despite the large-scale nature of these instances, our solution approach is capable of solving CRND-SDCR to near-optimal solutions for most of the instances. As expected, the optimality gaps increase with the size and complexity of the data instance. This trend can greatly be attributed to the fact that these decomposition-based branch-and-cut algorithms rely on repeatedly solving $\mathcal{D}_\omega(\hat{y})$ and $\mathcal{L}_\omega(\hat{y})$ at every integer-feasible first-stage solution $\hat{y}$, which requires more time with increase in network complexity (in terms of $|\mathcal{P}|$ and $|\mathcal{A}|$).

More importantly, we observe that our proposed decomposition-based branch-and-cut algorithm outperforms its classical counterpart for all the data instances. In particular, our proposed approach performs at least twice as better as compared to the classical decomposition-based branch-and-cut algorithm. The underlying reason for this occurrence can be traced to the fact that our proposed algorithm leverages the structure of the subproblems $\mathcal{D}_\omega(\hat{y})$ and $\mathcal{L}_\omega(\hat{y})$ in a better manner as opposed to the off-the-shelf solver and hence, the process of solving these subproblems is accelerated in our proposed approach, which ultimately results in more cuts being added and better optimality gaps are achieved at the termination of the algorithm.

A closer inspection of Table 2 also reveals that the computational enhancements employed in addition to our proposed decomposition-based branch-and-cut algorithm reach better optimality gaps. Especially these enhancements improve the performance by reducing the optimality gaps by a factor of 1/3. The reason for this occurrence is that decomposition-based branch-and-cut algorithms suffer from ineffective initial iterations due to the generation of low-quality solutions. These computational enhancement strategies help bypass such ineffective initial iterations by generating better solutions from the get-go.

Overall, Table 1 shows the effectiveness of our tailored solution method in solving CRND-SDCR and designing relay networks for large-scale logistics operations. Next, we analyze the designed networks and quantify their capabilities to sustain commodity variability.

5.3. Value of Incorporating Demand Uncertainty

To evaluate the importance of considering demand uncertainty in designing relay networks, we compare the performance of the relay networks obtained by solving CRND-SDCR with that of relay networks obtained by considering deterministic average commodity demand. In this deterministic commodity demand setting, the demand for each O-D pair $p \in \mathcal{P}$ is then given by $\lceil \sum_{\omega \in \Omega} \pi_\omega d_p^\omega \rceil$. To evaluate the performance of these designed relay networks, we optimally solve (2) to determine a minimum-cost consolidation plan that transports the commodities from their respective origins to their respective destinations for each of the 272 stochastic demand scenarios. We note that to ensure feasibility while solving $\mathcal{D}_\omega(y^*)$, we included dummy long-haul transportation arcs between each O-D pair with very high transportation costs that are only used when the commodity demand cannot be transported through the relay network. Hence, we evaluate 3 key performance indicators for each network: (i) the cost of constructing the relay hub network, (ii) the percentage amount of demand that cannot be fulfilled through the relay network, and (iii) the average delivery cost of the commodities that are transported through the relay network. Table 3 compares these evaluation metrics.

Tables 3 showcases that with the increase in data instance size, the average commodity delivery costs decrease, providing compelling evidence that larger relay networks are able to provide better

Table 3 Network performance comparison against demand variability, averaged across demand scenarios

Commodity demand setting	Data instance	Hub construction costs ($)	Unfulfilled demand (%)	Average delivery costs ($/car)
Deterministic commodity demand (average scenario)	1	10,164.66	32.86	1,726.34
	2	49,156.07	27.39	1,447.28
	3	66,235.61	20.55	1,389.05
	4	82,967.34	17.83	1,266.17
	5	91,459.57	13.41	1,135.88
Stochastic commodity demand (multiple scenarios)	1	13,875.68	18.96	1,645.96
	2	56,012.05	15.77	1,387.52
	3	78,620.58	11.54	1,231.30
	4	96,121.26	9.25	1,218.95
	5	110,309.95	7.92	1,049.56

consolidation opportunities to transport commodities and achieving the necessary economies of scale. Moreover, the comparison results show that for every data instance, the relay networks designed under a stochastic commodity demand setting outperform the relay networks designed under the deterministic average commodity demand setting by fulfilling a larger proportion of commodity demand through relay transportation and doing so at lower average delivery costs. Consideration of commodity demand uncertainty induces the required flexibility in relay networks in terms of relay hub capacities that help devise consolidation plans that are more economical and able to transport higher amounts of commodities through relay transportation.

Overall, we observe that although relay networks designed under the deterministic demand setting have lower hub construction costs, they are not robust against demand variability and provide substandard transportation operations outcomes. Such issues are overcome by considering demand stochasticity while designing relay networks. Next, we analyze the importance of considering consolidation-based routing while designing relay networks.

5.4. Value of Considering Consolidation-based Routing

To evaluate the importance of considering consolidation-based routing in designing relay networks, we compare the performance of relay networks designed from CRND-SDCR with that of networks that approximate routing decisions. Specifically, we also design networks by solving (1)-(2), but with relaxed integrality constraints (2f). This approach assumes a continuous flow routing of commodities without consolidation considerations, and can be carried out using a simplified branch-and-cut algorithm that only involves adding Benders optimality cuts (13) at every integer-feasible first-stage solution encountered in the search tree.

We design both these types of networks for 3 different types of demand settings where the amount of commodity demand that has to be transported is respectively low, medium, and high. To this end, we define a scaling factor ζ that scales the commodity demand d_p^ω for each $p \in \mathcal{P}$ and $\omega \in \Omega$ in these 3 different demand settings. For the low-demand scenario, we use $\zeta = 0.1$, meaning that

the commodity demand here $d_p^{\omega,\prime} = 0.1 \times d_p^{\omega}$ for each $p \in \mathcal{P}$ and $\omega \in \Omega$. Similarly, we use $\zeta = 0.5$ and $\zeta = 1$ for medium and high settings, respectively.

Table 4 Network performance comparison under various types of demand settings, averaged across demand scenarios

Demand setting	Data instance	Approximated routing			Consolidation-based routing		
		Hub construction costs ($)	Unfulfilled demand (%)	Average delivery costs ($/car)	Hub construction costs ($)	Unfulfilled demand (%)	Average delivery costs ($/car)
Low ($\zeta = 0.1$)	1	10,784.14	15.7	1,896.43	8,046.23	12.94	1,719.69
	2	47,628.11	13.28	1,553.90	50,681.93	11.56	1,439.54
	3	72,056.09	10.07	1,423.81	69,869.36	9.57	1,307.66
	4	92,269.39	7.96	1,366.19	92,079.26	7.22	1,283.17
	5	101,652.55	6.02	1,245.39	102,016.88	5.18	1,148.61
Medium ($\zeta = 0.5$)	1	11,028.95	17.32	1,791.44	11,264.08	15.29	1,698.35
	2	57,153.26	16.19	1,499.02	54,291.02	13.07	1,408.64
	3	76,269.16	11.48	1,379.23	75,953.57	10.33	1,290.75
	4	94,008.66	9.13	1,303.77	95,647.29	8.97	1,244.19
	5	110,369.26	7.86	1,181.57	108,268.15	6.01	1,127.93
High ($\zeta = 1$)	1	12,059.98	19.56	1,658.04	13,875.68	18.76	1,645.96
	2	59,061.33	16.44	1,403.08	56,012.05	15.77	1,387.52
	3	77,128.69	12.83	1,247.39	78,620.58	11.54	1,231.30
	4	95,388.42	10.04	1,228.77	96,121.26	9.25	1,218.95
	5	114,035.78	8.71	1,055.94	110,309.95	7.92	1,049.56

To compare the performance of these two types of networks, we employ the same evaluation metric used in Section 5.3. These results are portrayed in Table 4. For both types of network designs, we observe that with an increase in commodity demand, i.e., moving from a low to a high-demand setting, the average delivery costs decrease. The networks leverage economies of scale by providing better consolidation opportunities. As expected, we also observe that when the commodity demand variability decreases, i.e., when moving from a high to a low-demand setting, a larger proportion of commodity demand is transported through relay transportation.

We also observe that the networks designed by considering consolidation-based routing outperform the networks that do not consider it: they are able to fulfill more demand through relay transportation and are able to do so at lower average delivery costs. Interestingly, the difference in average delivery costs is more prominent for low and medium-demand settings, while the costs are comparable for high-demand settings. The reason can be attributed to the fact that when the demand is low to moderate, effective consolidation provides substantial transportation cost-savings, and considering such routing decisions while designing the relay networks helps achieve these cost-savings in operations, which can be seen from the Table 4. Overall, the comparable average delivery costs for high-demand settings provide evidence that approximating the routing decisions while designing relay networks is still acceptable in this case whereas, for the low to moderate-demand cases, it provides sub-standard outcomes and leads to non-economical transportation costs.

6. Conclusion

In this article, we studied the problem of designing large-scale resilient relay logistics hub networks with commodity demand uncertainty. We introduced a model that focuses on improving efficiency and resilience against demand variability through integrating tactical planning decisions. This model, for *Capacitated Relay Network Design under Stochastic Demand and Consolidation-Based Routing* (CRND-SDCR), consists of locating logistics hubs and deciding their respective capacities under uncertain demand, and routing stochastic commodity demand between each origin-destination pair through consolidation-based trucking. We formulated this problem as a two-stage stochastic optimization program where we located and capacitized the hubs in the first stage and designed a minimum-cost consolidation plan for the realized demand scenarios in the second stage.

We leveraged the structure of the problem and devised a branch-and-cut algorithm with nested Benders decomposition and integer L-shaped methods to solve CRND-SDCR exactly. We employed a nested decomposition scheme where we decomposed CRND-SDCR twice, once across the stochastic demand scenarios and the second across each origin-destination pair within the scenario subproblems, permitting us to efficiently add the associated Benders feedback cuts. We guaranteed the exactness of our solution approach by adding integer L-shaped cuts, which are computed by solving the second-stage subproblem exactly through Benders decomposition as well.

We then conducted computational experiments to design large-scale resilient relay logistics networks using data from a large US-based car manufacturer. We found that our developed algorithm can obtain near-optimal solutions in a reasonable time with sample average approximation. Furthermore, our tailored implementation of the decomposition-based branch-and-cut converges faster and scales better with the instance size, in comparison with its classical implementation. To validate our designed relay networks, we computed their performance by determining minimum-cost consolidation plans for various demand realizations. For comparison purposes, we also created baseline networks, designed solely to satisfy the average commodity demand. Overall, we observed that our designed relay networks induced flexibility in the network and significantly outperformed the baseline networks as they fulfill more demand through relay transportation and at a lower cost. Finally, we compared our designed networks with the literature-proposed relay networks that continuously approximate consolidation-based routing operations. We observed that for demand realizations with low to medium commodity demand, our networks outperform the literature-proposed networks, and for demand realizations with high commodity demand, our networks provided comparable performances.

While this work has focused on designing networks that can handle commodity demand variability, it does not factor in service level requirements. A natural extension, in this case, is to consider the temporal aspect of commodity demand stochasticity, which will require modeling the

tactical planning sub-problem more comprehensively and tailoring solution algorithms to facilitate the coordination of transportation services across space and time. Another extension is to design networks to tackle and sustain logistics disruptions such as hub failures, traffic congestion, etc. This will require evaluating the impact of such disruptions on various network configurations in order to guide the design of appropriate logistics networks. All this will result in complex multi-stage optimization problems, which will require new heuristics and approximation algorithms to provide practically relevant solutions. Finally, it will be worthwhile to study how the models and solution techniques proposed in this work can be extended to design multi-layered hyperconnected logistics networks for faster and resilient commodity delivery.

Acknowledgments

References

Ali TH, Radhakrishnan S, Pulat S, Gaddipati NC, 2002 *Relay network design in freight transportation systems. Transportation Research Part E: Logistics and Transportation Review* 38(6):405–422.

American Trucking Associations, 2018 *Turnover rate at large truckload carriers rises in first quarter*. Technical report, http://www.trucking.org/article/Turnover-Rate-at-Large-Truckload-Carriers-Rises-in-First-Quarter.

American Trucking Associations, 2019 *Truck driver shortage analysis*. Technical report, https://www.trucking.org/sites/default/files/2020-01/ATAs%20Driver%20Shortage%20Report%202019%20with%20cover.pdf.

Ballot E, Gobet O, Montreuil B, 2012 *Physical Internet Enabled Open Hub Network Design for Distributed Networked Operations*, 279–292 (Berlin, Heidelberg: Springer Berlin Heidelberg).

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

Birge JR, Louveaux FV, 1988 *A multicut algorithm for two-stage stochastic linear programs. European Journal of Operational Research* 34(3):384–392.

Cabral EA, Erkut E, Laporte G, Patterson RA, 2007 *The network design problem with relays. European Journal of Operational Research* 180(2):834–844.

Carøe CC, Schultz R, 1999 *Dual decomposition in stochastic integer programming. Operations Research Letters* 24(1):37–45.

Chen R, Luedtke J, 2022 *On generating lagrangian cuts for two-stage stochastic integer programs. INFORMS Journal on Computing* 34(4):2332–2349.

Crainic TG, 2000 *Service network design in freight transportation. European Journal of Operational Research* 122(2):272–288.

Crainic TG, Hewitt M, Maggioni F, Rei W, 2021 *Partial benders decomposition: General methodology and application to stochastic network design. Transportation Science* 55(2):414–435.

Greening LM, Dahan M, Erera AL, 2023 *Lead-time-constrained middle-mile consolidation network design with fixed origins and destinations. Transportation Research Part B: Methodological* 174:102782.

Grove PG, O'Kelly ME, 1986 *Hub networks and simulated schedule delay. Papers in Regional Science* 59(1):103–119.

Hall RW, 1989 *Configuration of an overnight package air network. Transportation Research Part A: General* 23(2):139–149.

Hewitt M, Crainic TG, Nowak M, Rei W, 2019 *Scheduled service network design with resource acquisition and management under uncertainty. Transportation Research Part B: Methodological* 128:324–343.

Hu Z, Askin RG, Hu G, 2019 *Hub relay network design for daily driver routes. International Journal of Production Research* 57(19):6130–6145.

Hunt GW, 1998 *Transportation Relay Network Design*. Ph.D. thesis, Georgia Institute of Technology.

Jacquillat A, Schmid A, Wang K, 2022 *Relay logistics: A multi-variable generation approach. SSRN Electronic Journal* URL http://dx.doi.org/10.2139/ssrn.4241031.

Keller SB, Ozment J, 1999 *Managing driver retention: effects of the dispatcher. Journal of Business Logistics* 20(2):97.

Kewcharoenwong P, Li Q, Üster H, 2023 *Lagrangean relaxation algorithms for fixed-charge capacitated relay network design. Omega* 121:102926.

Kewcharoenwong P, Üster H, 2017 *Relay network design with capacity and link-imbalance considerations: A lagrangean decomposition algorithm and analysis. Transportation Science* 51(4):1177–1195.

Kleywegt AJ, Shapiro A, Homem-de Mello T, 2002 *The sample average approximation method for stochastic discrete optimization. SIAM Journal on Optimization* 12(2):479–502.

Konak A, 2012 *Network design problem with relays: A genetic algorithm with a path-based crossover and a set covering formulation. European Journal of Operational Research* 218(3):829–837.

Kulkarni O, Cohen YM, Dahan M, Montreuil B, 2021 *Resilient hyperconnected logistics hub network design. 8th International Physical Internet Conference.*

Kulkarni O, Dahan M, Montreuil B, 2022 *Resilient hyperconnected parcel delivery network design under disruption risks. International Journal of Production Economics* 251:108499.

Kulkarni O, Dahan M, Montreuil B, 2023 *Resilience assessment of hyperconnected parcel logistic networks under worst-case disruptions. 9th International Physical Internet Conference.*

Kulkarni O, Dahan M, Montreuil B, 2024 *Resilient relay logistics network design: A k-shortest path approach. Optimization Online* https://optimization-online.org/?p=24297.

Kulturel-Konak S, Konak A, 2008 *A local search hybrid genetic algorithm approach to the network design problem with relay stations.* Raghavan S, Golden B, Wasil E, eds., *Telecommunications Modeling, Policy, and Technology*, 311–324 (Boston, MA: Springer US).

Laporte G, Louveaux FV, 1993 *The integer l-shaped method for stochastic integer programs with complete recourse. Operations Research Letters* 13(3):133–142.

Leitner M, Ljubić I, Riedler M, Ruthmair M, 2019 *Exact approaches for network design problems with relays. INFORMS Journal on Computing* 31(1):171–192.

Lium AG, Crainic TG, Wallace SW, 2009 *A study of demand stochasticity in service network design. Transportation Science* 43(2):144–157.

Montreuil B, 2011 *Toward a physical internet : meeting the global logistics sustainability grand challenge. Logistics Research* 3:71–87.

Montreuil B, Buckley SM, Faugère L, Khir R, Derhami S, 2018 *Urban parcel logistics hub and network design: The impact of modularity and hyperconnectivity. Progress in Material Handling Research.*

Montreuil B, Meller RD, Ballot E, 2013 *Physical Internet Foundations*, 151–166 (Berlin, Heidelberg: Springer Berlin Heidelberg).

Nosowitz D, 2017 *The long white line: The mental and physical effects of long-haul trucking.* Pacific Standard, http://tinyurl.com/ypsp3vvf.

Rahmaniani R, Crainic TG, Gendreau M, Rei W, 2018 *Accelerating the benders decomposition method: Application to stochastic network design problems. SIAM Journal on Optimization* 28(1):875–903.

Rockafellar RT, Wets RJB, 1991 *Scenarios and policy aggregation in optimization under uncertainty. Mathematics of Operations Research* 16(1):119–147.

Rodriguez J, Kosir M, Lantz B, Griffin G, Glatt J, 2000 *The costs of truckload driver turnover.* Technical report.

Sieber K, 2015 *Long-haul truck driver health survey results.* NIOSH Science Blog, CDC, http://tinyurl.com/ypsp3vvf.

Taylor G, Whicker GL, Usher JS, 2001 *Multi-zone dispatching in truckload trucking. Transportation Research Part E: Logistics and Transportation Review* 37(5):375–390.

Tu D, Montreuil B, 2019 *Hyper-connected megacity logistics: Multi-tier territory clustering and multi-plane meshed hub network design. 6th International Physical Internet Conference*, 159–167.

US Bureau of Transportation Statistics, 2023 *Transportation statistics annual report.* Technical report, U.S. Bureau of Labor Statistics, https://www.bls.gov/ooh/transportation-and-material-moving/heavy-and-tractor-trailer-truck-drivers.htm.

Üster H, Kewcharoenwong P, 2011 *Strategic design and analysis of a relay network in truckload transportation. Transportation Science* 45(4):505–523.

Yıldız B, Karaşan OE, Yaman H, 2018 *Branch-and-price approaches for the network design problem with relays. Computers & Operations Research* 92:155–169.

Ziaeifar A, Üster H, 2023 *Relay network design with direct shipment and multi-relay assignment. Annals of Operations Research* .

Üster H, Maheshwari N, 2007 *Strategic network design for multi-zone truckload shipments. IIE Transactions* 39(2):177–189.