Submitted to *Transportation Science*
manuscript (Please, provide the manuscript number!)

Hub Network Design Problem with Capacity, Congestion and Stochastic Demand Considerations

Vedat Bayram
1. Industrial Engineering Department, TED University, Ankara, Turkey, vedat.bayram@tedu.edu.tr,
2. Department of Analytics, Operations and Systems, Centre for Logistics and Heuristic Optimisation, Kent Business School, The University of Kent, Canterbury, Kent, United Kingdom, v.bayram@kent.ac.uk

Barış Yıldız
Industrial Engineering Department, Koç University, Istanbul, Turkey, byildiz@ku.edu.tr

M. Saleh Farham
Alberta School of Business, University of Alberta, Edmonton, Canada, farham@ualberta.ca

Our study introduces the hub network design problem with congestion, capacity, and stochastic demand considerations (HNDC), which generalizes the classical hub location problem in several directions. In particular, we extend state-of-the-art by integrating capacity acquisition decisions and congestion cost effect into the problem and allowing dynamic routing for origin-destination pairs. Connecting strategic and operational level decisions, HNDC jointly decides hub locations and capacity acquisitions by considering the expected routing and congestion costs. A path-based mixed-integer second-order cone programming (SOCP) formulation of the HNDC is proposed. We exploit SOCP duality results and propose an exact algorithm based on Benders decomposition and column generation to solve this challenging problem. We use a specific characterization of the capacity-feasible solutions to speed up the solution procedure and develop an efficient branch-and-cut algorithm to solve the master problem. We conduct extensive computational experiments to test the proposed approach's performance and derive managerial insights based on realistic problem instances adapted from the literature. In particular, we found that including hub congestion costs, accounting for the uncertainty in demand, and whether the underlying network is complete or incomplete have a significant impact on hub network design and the resulting performance of the system.

Key words: hub location problem; hub congestion; capacity building; multiple allocation; second order cone programming; Benders decomposition; column generation; branch-and-cut
History:

1. Introduction

Hubs play critical roles in many transportation and telecommunication networks with different names and functionalities. They are consolidation points in distribution networks to provide economies of scale (Alumur and Kara 2008). In multi-modal transportation, hubs are strategically located facilities where the travelers/commodities switch the mode of transportation (Contreras and O'Kelly 2019). In green logistics, they become the recharging/refueling stations extending the reach of alternative fuel vehicles (Yıldız, Karaşan, and Yaman 2018). Hubs are also at the center of many innovative transportation applications/concepts, such as physical internet (Montreuil 2011), crowdsourced delivery (Macrina et al. 2020, Kızıl and Yıldız 2022), and mobility as a service (Jittrapirom et al. 2017), which are poised to revolutionize the transport sector. The widening use of hub networks in various real-world applications urges several extensions in classical approaches for hub network design problems (HND). In this study, we focus on one such crucial extension and develop a new modeling approach to address the congestion issues (at hubs) that can impair the system performance and introduce significant costs.

Ignoring congestion in HND may lead to high overall costs and poor service quality (Alumur et al. 2018). For example, airlines estimate the broader economic cost of congestion in hub airports to be more than $18.5 billion per year, in addition to the more than one million metric tons of avoidable carbon dioxide emissions released while aircraft circle aimlessly waiting for an opening to land or run their engines on the tarmac before finally taking off (Forbes 2019). Despite the critical importance, a limited number of studies explicitly address the congestion issues in hub-and-spoke networks that are paradoxically designed to increase consolidation opportunities, which make them prone to negative externalities of congestion. Network management can take measures both at the strategic and operational levels to avoid congestion. At the strategic level, the hub capacities can be decided by taking the congestion costs and demand uncertainty into account, and at the operational level, dynamic routing and hub assignments can be considered to use available capacity in the most efficient way. Although these two approaches have been studied separately in the literature, to the best of our knowledge, this paper is the first one to propose a joint model that links the strategic level location and capacity decisions with the dynamic network management decisions (routing and hub assignment) to solve hub network design problem with congestion.

Considering the congestion costs and integrating the strategic level network design decisions with the tactical and operational level commodity transportation decisions is a challenging research area, which requires novel formulation approaches and efficient algorithms to solve them (Yildiz, Yaman, and Karasan 2021). Focusing on complete networks and formulating OD paths with only two hubs in between, as a common practice in the literature, prevents one from dealing with a range of realistic situations such as telecommunication networks with particular backbone structures (Klincewicz 1998) or inter-modal public transportation with several numbers of stops on an itinerary (Marín et al. 2002). Even when one assumes a complete network for inter-hub transfers (i.e., the triangular inequality holds for the transportation costs), including the congestion costs invalidates the use of formulations that assume OD paths with at most two hubs because an OD path with the minimum transportation cost is not necessarily the one that has the minimum total cost when the congestion is considered. Introducing an additional level of difficulty, the congestion costs have a nonlinear relation with the utilized hub capacities (Elhedhli and Wu 2010). Moreover, considering the demand uncertainties along with the dynamic hub assignments and OD routes requires novel two-stage stochastic programming formulations to extend HND as an already challenging design problem.

We introduce the hub network design problem with capacity, congestion, and stochastic demand considerations (HNDC) to cover a broad range of strategic to operational level decisions. HNDC aims to find the optimal network design that minimizes the total setup, capacity acquisition, congestion, and routing (transportation) costs. We assume a multiple-allocation setting where a demand point can send and receive flows through more than one hub. To reflect the congestion cost, a Kleinrock function is used that models each hub as an $M/M/1$ queue in the steady-state condition (Kleinrock 1964, Elhedhli and Wu 2010). The introduced problem has the following key characteristics: (i) Hubs are capacitated, i.e., the total flow through a hub location is restricted. The network management incurs a congestion cost that depends on how much of the available capacity is used at a hub. (ii) A general directed network structure is assumed. Therefore, the underlying transportation network can be complete or incomplete, and the arc costs do not have to satisfy the triangle inequality. (iii) An OD path can visit up to κ number of hubs, where κ is simply a problem parameter to limit the number of hub visits for operational requirements stemming from operator/user perspectives. (iv) The hub assignments and OD paths can be dynamically determined (after observing the demand) to make the best use of the available capacity in the network, considering the congestion and transportation costs together.

To model this challenging problem, we propose a novel path-based mixed-integer second-order cone programming (MISOCP) formulation and develop a Benders decomposition (BD; Benders 1962) approach for its solution. The Benders subproblem is a path-based second-order cone program (SOCP) where the optimal flow paths are determined using a column generation (CG) technique. Duality results of the SOCP are used to generate new columns and Benders cuts (Bayram and Yaman 2018). The master problem is reformulated to guarantee the generation of capacity-feasible solutions and solved through a branch-and-cut algorithm. We aim to find the optimal solution to the HNDC and answer the following research questions: (i) What is the benefit of considering capacity and congestion at the time of network design? (ii) What is the benefit of including demand uncertainty and taking advantage of the dynamic routing as a recourse action? (iii) How does the optimal network topology change when these extensions are considered? To this end, extensive computational results are conducted on a set of problem test instances adapted from the Turkish (TR) data set (Yaman, Kara, and Tansel 2007), and Australia Post (AP) data set (Ernst and Krishnamoorthy 1996) from the literature. The largest problem instance solved by our approach in the TR data set contains 81 nodes (40 of which are potential hub locations) and 200 nodes in the AP data set (20 of which are potential hub locations).

The rest of the paper is organized as follows. Section 2 reviews the related literature on HLPs with congestion consideration. In Section 3, we formulate the HLPCC as a mixed-integer nonlinear programming (MINLP) and transform it into an MISOCP formulation. We propose BD solution approach in Section 4 and provide a CG scheme for generating candidate OD paths in Section 5. Section 6 presents the computational study, and Section 7 concludes the paper with some final remarks.

2. Literature Review

Ever since the introduction of the first model by O'Kelly (1986a), a vast body of research has flourished on hub location problems (HLP) with contributions to formulations and/or solution methodologies proposed, which are too numerous to list here. For a comprehensive coverage of the HLP literature, we refer the reader to Alumur and Kara (2008), Campbell and O'Kelly (2012), Farahani et al. (2013), Contreras (2015), Contreras and O'Kelly (2019), Alumur et al. (2020). Here, we focus on recent studies involving a problem setting where congestion cost, capacity acquisition, and demand uncertainty considerations are included in the decision-making process and for which path-based formulations are proposed to deal with incomplete networks and to allow the use of more than two hubs in routing decisions.

The existing literature on hub location problems lacks a comprehensive approach to simultaneously account for both congestion effects and variability (fluctuations) in demand. Some studies address congestion effects but overlook demand variability, consider a single level for demand arrivals (e.g., the peak time or average demand arrivals) for the transportation demand between origin-destination pairs and use steady-state queuing theory results to model congestion (Guldmann and Shen 1997, Elhedhli and Hu 2005, Elhedhli and Wu 2010, De Camargo, de Miranda Jr, and Ferreira 2011, de Camargo and Miranda 2012, Ishfaq and Sox 2012, Kian and Kargar 2016, Azizi, Vidyarthi, and Chauhan 2018, Alumur et al. 2018, Najy and Diabat 2020, Bütün, Petrovic, and Muyldermans 2021). However, in practice, the passenger or goods transport demand varies significantly over the planning horizon, making it unrealistic to assume a single demand level for designing the network. On the other hand, the studies that address the changing demand levels neglect the congestion effect and model demand variability in a simplified manner to account for various scenarios (Contreras, Cordeau, and Laporte 2011b, Alumur, Nickel, and Saldanha-da Gama 2012, Meraklı and Yaman 2016, 2017, de Sá, Morabito, and de Camargo 2018, Taherkhani, Alumur, and Hosseini 2020). In our research, we adopt a more holistic approach by modeling the hub-and-spoke system as a network of $M/M/1$ queues, which accounts for congestion effects. Moreover, we incorporate the variability in the demand levels by considering two scenarios. The modeling approach for demand stochasticity considered in this paper follows the stream of research on two-stage stochastic programming in which the amount of demand is assumed to be a random variable that follows a Poisson process. Each scenario represents an independent network of $M/M/1$ queues with a different arrival rate for hub $h \in \mathcal{H}$ in scenario $s \in S$, denoted as $u_h(s)$. Specifically, scenario 1 reflects typical conditions of the hub-and-spoke system, whereas scenario 2 represents peak demand times, such as holidays (Christmas, Ramadan, etc.). Our study aims to explore the joint effect of considering and the tradeoff between the cost of acquiring capacity and incurring congestion costs due to increased demand. Neglecting such fluctuations in demand or considering only the peak time demand would result in underutilized capacity during the majority of the planning horizon. Conversely, neglecting peak demand times and focusing only on the average demand levels could lead to significant congestion costs and even system infeasibilities, as we demonstrate.

2.1. Capacity Acquisition and Congestion

O'Kelly (1986b) was the first to point out the negative effects of congestion due to traffic consolidation, which may result in schedule delays and increased transit times. Delays at an airport, for instance, can be due to airport activity exceeding capacity and/or queuing effects at the landing and take-off runways (Grove and O'Kelly 1986, Mayer and Sinai 2003). While economies of scale suggests consolidating flows, congestion costs require distributing flows in the network and/or building extra capacity. Therefore, the trade-off between economies of scale and congestion should be accounted for in HND.

One way of lessening the negative effects of congestion is to impose explicit capacity limits on the total flow through hubs and/or arcs. Studies in the literature that adopted this approach include Campbell (1992), Aykin (1994), Ernst and Krishnamoorthy (1999), Ebery et al. (2000), Marianov and Serra (2003), Yaman and Carello (2005), Rodríguez-Martín and Salazar-González (2008), Correia, Nickel, and Saldanha-da Gama (2010), Contreras, Cordeau, and Laporte (2012), Tanash, Contreras, and Vidyarthi (2017), Meraklı and Yaman (2017), Bütün, Petrovic, and Muyldermans (2021), Taherkhani, Alumur, and Hosseini (2020). Except for Correia, Nickel, and Saldanha-da Gama (2010), Contreras, Cordeau, and Laporte (2012), these studies include capacity acquisition decisions at the time of network design. However, none of these studies employ a congestion cost function and consider uncertainty in demand. They work with complete network structures and do not allow the use of more than two hubs on a route.

Modeling congestion through capacity constraints, however, does not properly reflect the exponential behavior of the congestion effect (de Camargo et al. 2009). When the traffic flow through a hub approaches its capacity, the transit time, i.e., the cost incurred due to congestion, increases steeply. Therefore, nonlinear modeling of the congestion cost yields more realistic results. Several studies in the literature have explicitly considered congestion costs in the HLP (Guldmann and Shen 1997, Elhedhli and Hu 2005, Elhedhli and Wu 2010, De Camargo, de Miranda Jr, and Ferreira 2011, de Camargo and Miranda 2012, Ishfaq and Sox 2012, Kian and Kargar 2016, Azizi, Vidyarthi, and Chauhan 2018, Alumur et al. 2018, Najy and Diabat 2020, Bütün, Petrovic, and Muyldermans 2021). Two types of congestion cost functions are commonly used in this context. The first type, called power-law functions, have the form au^b, where a and b are constants and u indicates the flow through a hub. Several researchers have modeled the HLP with congestion as a queuing system. This led to the second type of nonlinear congestion cost functions, called Kleinrock average delay functions of the form $bu/(z-u)$, where z is the capacity level of the hub. Compared to the power cost function, Kleinrock functions capture the congestion effect more realistically as they consider the relative difference between hub flow and hub capacity rather than the hub flow alone (Elhedhli and Wu 2010).

The first study in the literature that explicitly considers congestion cost in the objective function is by Guldmann and Shen (1997). They employ piecewise linearization of a Kleinrock-type congestion cost function. Similarly, Elhedhli and Hu (2005), De Camargo, de Miranda Jr, and Ferreira (2011) and De Camargo, de Miranda Jr, and Ferreira (2011), Elhedhli and Wu (2010), Azizi, Vidyarthi, and Chauhan (2018), Bütün, Petrovic, and Muyldermans (2021) use a linearization (outer approximation) of a power law and Kleinrock type congestion cost functions, respectively. While Elhedhli and Hu (2005), Elhedhli and Wu (2010) propose a Lagrangian relaxation-based heuristic, De Camargo, de Miranda Jr, and Ferreira (2011) present a Benders decomposition-based approach and Bütün, Petrovic, and Muyldermans (2021) a tabu search heuristic to solve the problem. de Camargo and Miranda (2012) consider two types of design perspectives using a power law congestion cost function: the network owner and the network user. They employ a generalized Benders decomposition to solve the problem. Ishfaq and Sox (2012) model multi-modal hub operations as a queuing network by representing congestion in service time constraints. They employ a lower bounding procedure based on a partial linear relaxation of a subproblem for the original problem and a tabu search solution procedure to solve the problem. Kian and Kargar (2016) use power law congestion cost function and transform it to a SOCP formulation and solve their problem by using a commercial solver. Alumur et al. (2018) employ a modeling framework with a service time limit considering discretized congestion costs at hubs and use a commercial solver to solve their problem. Najy and Diabat (2020) consider an uncapacitated hub location problem where both flow-dependent economies of scale and congestion considerations are incorporated into the problem. They use a piecewise linear function to model congestion and a Benders decomposition-based solution methodology to solve the problem.

Steady-state queuing models with homogeneous (time-invariant) arrival processes may not be appropriate for certain systems such as airport operations because the demand for such systems varies considerably over time, with peaks during certain periods and little demand during others. To overcome the challenges posed by the fundamentally stochastic nature of such systems, simulation or approximation methods (Ignaccolo 2003, Hübl and Altendorfer 2015) or a step function indicating the expected demand during successive time periods (FAA 1976) is used, under the assumption that transient effects are of negligible importance. Other research in the literature utilizes dynamic queues (Odoni and Roth 1983, Pyrgiotis, Malone, and Odoni 2013, Schwarz, Selinka, and Stolletz 2016, Di Crescenzo et al. 2018) to address the in-homogeneous (time-variant) nature of such systems. However, in contrast to our problem setting, such studies assume that it is possible to switch from one arrival rate system to the other at any time step.

Most of the studies in the literature that explicitly consider nonlinear congestion costs in their formulation approximate the congestion cost function through linearization. Among these studies only Guldmann and Shen (1997), Elhedhli and Wu (2010), Alumur et al. (2018) consider capacity acquisition decisions and none consider stochasticity in demand. Only the modeling approach presented by Bütün, Petrovic, and Muyldermans (2021) allows more than two hubs on a path. None of these studies present a path-based formulation or a solution methodology that can be used for general networks.

2.2. Path-based Formulations allowing Multiple Hub Visits and a Generalized Network Structure

A common restriction in HLP literature is the use of at most two hubs on an OD path. However, if the arc costs do not follow the triangle inequality, or in case of an incomplete network, the optimal solution might incorporate more than two hubs on some routes. Assuming a complete network with triangular inequality and formulating OD paths with at most two hubs in between, as a common practice in the literature, results in disregarding a range of realistic problem instances such as telecommunication networks with special backbone structure (Klincewicz 1998) or inter-modal public transportation with several numbers of stops on an itinerary (Marín et al. 2002), parcel delivery systems (van Essen 2009), and express shipment networks (Meuffels 2015, Pérez, Lange, and Tancrez 2018). Since the majority of the current mathematical models rely on arc variables, allowing more than two hubs on a route makes the problem large and intractable to solve (van Essen 2009).

The number of studies that propose a path-based formulation in the literature (Contreras, Cordeau, and Laporte 2012, Rothenbächer, Drexl, and Irnich 2016, Tanash, Contreras, and Vidyarthi 2017, de Sá, Morabito, and de Camargo 2018, Brimberg et al. 2019, Taherkhani, Alumur, and Hosseini 2020) are limited. Except for the study by Contreras, Cordeau, and Laporte (2012), these studies use a general network structure and allow more than two hubs on an OD path. Although they do not present a path-based formulation, there exist another group of studies that do not rely on a complete network structure (Nickel, Schöbel, and Sonneborn 2001, Campbell, Ernst, and Krishnamoorthy 2005b,a, Yaman 2008, Alumur, Kara, and Karasan 2009, Contreras, Fernández, and Marín 2010, O'Kelly et al. 2015, Alibeyg, Contreras, and Fernández 2016, Rothenbächer, Drexl, and Irnich 2016, de Camargo et al. 2017, Tanash, Contreras, and Vidyarthi 2017, Brimberg et al. 2019), or which allow more than two hubs on a path (O'Kelly et al. 2015, de Camargo et al. 2017, Bütün, Petrovic, and Muyldermans 2021). Except for Bütün, Petrovic, and Muyldermans (2021) and de Sá, Morabito, and de Camargo (2018), Taherkhani, Alumur, and Hosseini (2020), these studies, however, do not take into consideration congestion costs and the uncertainty in demand, respectively, and they generally do not include capacity acquisition decisions.

2.3. Uncertainty in Demand

HLP incorporates strategic location and capacity acquisition decisions as well as operational routing decisions. The information regarding the OD demands is not fully known at the design stage and is revealed after the strategic decisions are made. There is a trade-off between building an expensive hub network with idle capacities and the adverse effect of an unexpected demand on system service times due to congestion. This requires taking into consideration the uncertainty in demand.

There do not exist many studies in the HLP literature that consider uncertainty in demand. Some examples include studies by Contreras, Cordeau, and Laporte (2011b), Alumur, Nickel, and Saldanha-da Gama (2012), Meraklı and Yaman (2016, 2017), de Sá, Morabito, and de Camargo (2018), Taherkhani, Alumur, and Hosseini (2020). Contreras, Cordeau, and Laporte (2011b) propose a two-stage stochastic optimization formulation for uncapacitated, multiple allocation HLP and a solution methodology based on Benders decomposition and sample average approximation to solve the problem. Meraklı and Yaman (2016, 2017) consider robust optimization formulations under polyhedral and hose demand uncertainty and employ a Benders decomposition-based solution methodology. Taherkhani, Alumur, and Hosseini (2020) model the profit-maximizing capacitated hub location problem with multiple demand classes and propose a two-stage stochastic optimization approach integrating Benders decomposition and sample average approximation through a Monte-Carlo simulation.

In summary, these studies do not consider congestion costs and capacity acquisition decisions. They all assume a complete network structure with triangular inequality, and allow at most two hubs on an OD path, and except for Taherkhani, Alumur, and Hosseini (2020) they are not path-based. To the best of our knowledge, there do not exist any studies in HLP literature that consider congestion costs, capacity acquisition decisions and uncertainty in demand, simultaneously.

Therefore, considering the current gap in the literature, we highlight our contribution as follows. (i) This work relaxes several assumptions commonly used in hub location literature. We introduce a new hub location problem that generalizes the classical multi-allocation hub location problem in the literature, as we allow more than two intermediate points (hubs) on a path, consider a general network structure, i.e., complete and incomplete, and cope with hub location, capacity acquisition, congestion, and routing decisions, simultaneously under demand uncertainty. (ii) A scenario-based two-stage stochastic programming approach and a path-based MISOCP formulation of the problem are presented for the first time in the literature. (iii) We develop an efficient exact solution algorithm based on Benders decomposition and column generation approaches. As the second stage of the problem requires solving SOCP subproblems, the duality results for the SOCP are used to generate new columns and Benders cuts. (iv) For both complete and incomplete networks, the solution generated by the master problem may be infeasible for the subproblem. To guarantee that the master problem generates capacity-feasible solutions for the scenario subproblems, we reformulate it to characterize capacity-feasible solutions and solve it through a branch-and-cut algorithm. (v) We carry out an extensive computational study on a set of real-world problem test instances to analyze the trade-off between network design cost and traffic congestion, as well as the effect of different problem parameters on the final solution and the total cost. The results show that our proposed algorithm is able to solve problem instances with reasonable sizes using off-the-shelf solvers.

3. Problem Definition and Formulation

The HNDC links strategic level hub location and capacity decisions with operational level hub assignment and routing decisions to achieve an optimal hub network design, i.e., number, location, and capacity of hubs, such that the sum of location, capacity acquisition, congestion, and transportation cost is minimized. The problem is defined on a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is the set of nodes and $\mathcal{A}$ is the set of arcs in the network. The unit transportation cost on an arc $(i,j) \in \mathcal{A}$ is denoted by c_{ij}. Graph $\mathcal{G}$ can be complete or incomplete. Let $\mathcal{H} \subseteq \mathcal{N}$ be the set of potential hub locations. We say a graph $\mathcal{G}$ is κ-connected, for some $\kappa \geq 2$, if for any simple walk $w = \{h_1, \ldots, h_n\}$ in the hub sub-graph $\mathcal{G}(H)$ with $n > \kappa$, there exist an alternative walk $\bar{w}$, between h_1 and h_n, that uses a proper subset of the nodes visited by w.

Let $\mathcal{K}$ be the set of commodities corresponding to origin-destination pairs in $\mathcal{N}$. The origin and destination of a commodity $k \in \mathcal{K}$ are given as o_k and d_k, respectively. Without loss of generality, we assume that none of the candidate locations for the hubs is a source or a destination node for any demand. We define $\alpha \in [0,1]$ as the discount factor for the inter-hub transportation and denote the unit cost of transportation on an arc $a \in \mathcal{A}$ with c_a, which includes the discount for the arcs between hubs $A_{\mathcal{H}} \subset \mathcal{A}$.

The traffic demand for commodity k is denoted by w_k and has to be routed through at least one opened hub. A hub can be opened in different sizes (capacity levels) listed in the set L. The fixed and variable cost of opening a size $\ell \in L$ hub at a location $h \in \mathcal{H}$ is given by f_h^ℓ. The capacity associated with a hub size ℓ is denoted by q_ℓ and the maximum possible capacity for a hub (the capacity of the largest size hub) is shown as Q.

Let $\mathcal{P}_k$ be the set of all alternative feasible paths for commodity k. A feasible path $p \in \mathcal{P}_k$ is required to visit at least one and at most κ number of hubs, and defined as an ordered set of nodes $\{o_k = n_p^0, \cdots, n_p^{m_p}, n_p^{m_p+1} = d_k\} \subseteq \mathcal{N}$, such that:

- $\{n_p^1, \cdots, n_p^{m_p}\} \subseteq \mathcal{H}$, i.e., all intermediate nodes of a path are hubs and no direct transshipment of commodities are allowed.
- $\{(o_k, n_p^1), (n_p^m, d_k)\} \cup \{(n_p^i, n_p^{i+1}) : i = 1, \ldots, m_p - 1\} \subseteq \mathcal{A}$, i.e., the connectivity limitations are respected.

Let A_p indicate the set of arcs in $\mathcal{A}$ that are visited by a path p. We define the unit transportation cost on a path as $c_p = \sum_{a \in A_p} c_a$.

In a two-stage stochastic setting, the first stage of HNDC is about strategic hub network design decisions, i.e., hub location and capacity decisions. The binary variable y_h^ℓ indicates whether a hub of size $\ell \in L$ is opened at $h \in \mathcal{H}$ or not. Given the hub location and capacity decisions from the first stage and the realization of the origin-destination demands, the second stage is about operational level routing decisions. We define S as the set of possible demand scenarios and associate a probability $\varphi(s)$ to each scenario $s \in S$. To model variability in demand, we assume the flow rate from origin o_k to destination d_k is an independent random variable that follows a Poisson process with mean $w_k(s)$ for commodity $k \in \mathcal{K}$ in scenario $s \in S$. Moreover, due to the superposition property of Poisson processes, the aggregate traffic flow rate through hub $h \in \mathcal{H}$ in scenario $s \in S$ is also a random variable that follows a Poisson process with mean $u_h(s) = \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k : h \in p} w_k(s) v_p(s)$. We assume that transit times at each hub $h \in \mathcal{H}$ are independent and identically distributed random variables that follow an exponential distribution and are denoted by $\sum_{\ell \in L} q^\ell y_h^\ell$. The transit time at a hub $h \in \mathcal{H}$ denotes the maximum flow rate that can be processed at the hub during a given time period and is the sum of waiting and service times. It is a measure of the hub's capacity and represents the amount of flow that the hub is able to process in the specified time period. We further assume that total flow through a hub cannot exceed its capacity, i.e., $u_h(s) \leq \sum_{\ell \in L} q^\ell y_h^\ell$, for hub $h \in \mathcal{H}$ in scenario $s \in S$. Under these assumptions, the hub-and-spoke system can be modeled as a network of $M/M/1$ queues where each hub $h \in \mathcal{H}$ behaves as a single server with an exponential service rate $\sum_{\ell \in L} q^\ell y_h^\ell$ with a first-come, first-served arrival process with rate $u_h(s)$ for each scenario $s \in S$. Then, for a unit flow, mean transit time through a hub under steady-state conditions is modelled as: $\frac{1}{\sum_{\ell \in L} q^\ell y_h^\ell - u_h(s)}$ (Little 1961). The fraction of flow from an origin to a destination on path p is denoted by the continuous variable $v_p(s)$, $s \in S$. To capture the effects of congestion, we adopt the cost function as in Elhedhli and Wu (2010), inspired by $M/M/1$ queues, which has been widely used in previous studies to model congestion at hub locations. However, our solution framework allows us to use any congestion function such as power functions, as long as it is convex or can be transformed into a convex function by taking advantage of the binary variables as we do for the queuing function studied in this paper. For a fixed scenario $s \in S$, the congestion cost function for hub h is given by

$$b_h \frac{u_h}{\sum_{\ell \in L} q^\ell y_h^\ell - u_h + \epsilon}, \quad h \in \mathcal{H}, \tag{1}$$

where b_h is a scaling factor used to calculate the congestion cost of hub h and ϵ is an arbitrarily small positive number to avoid the cases with zero divided by itself. Below, we formulate the HNDC as a path-based two-stage stochastic MINLP.

$$\begin{aligned}
\text{Minimize} \quad & \sum_{h \in \mathcal{H}} \sum_{\ell \in L} f_h^\ell y_h^\ell + \sum_{s \in S} \varphi(s) \left(\sum_{h \in \mathcal{H}} b_h \frac{u_h(s)}{\sum_{\ell \in L} q^\ell y_h^\ell - u_h(s) + \epsilon} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k} c_p v_p(s) \right) && && (2)\\
\text{subject to:} \quad & \sum_{p \in \mathcal{P}_k} v_p(s) = 1 && \forall k \in \mathcal{K}, s \in S, && (3)\\
& \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k : h \in p} w_k(s) v_p(s) = u_h(s) && \forall h \in \mathcal{H}, s \in S, && (4)\\
& u_h(s) \le \sum_{\ell \in L} q^\ell y_h^\ell && \forall h \in \mathcal{H}, s \in S, && (5)\\
& \sum_{\ell \in L} y_h^\ell \le 1 && \forall h \in \mathcal{H}, && (6)\\
& y_h^\ell \in \{0,1\} && \forall h \in \mathcal{H}, \ell \in L, && (7)\\
& v_p(s) \ge 0 && \forall k \in \mathcal{K}, p \in \mathcal{P}_k, s \in S, && (8)\\
& u_h(s) \ge 0 && \forall h \in \mathcal{H}, s \in S. && (9)
\end{aligned}$$

Objective function (2) minimizes the total cost of opening new hubs with a certain capacity, and expected congestion and transportation costs. Constraint (3) ensures that the demand for every commodity $k \in \mathcal{K}$ in each scenario $s \in S$ is satisfied. Constraint (4) computes the total flow transiting through each hub in each scenario $s \in S$. Constraint (5) limits the flow at each hub to its allocated capacity in each scenario $s \in S$. Constraint (6) restricts the selection of only one capacity level for every hub. Constraints (7)–(8) define variable domains.

There are two important issues with using (1) in a mathematical model. First, it results in a non-convex model, and second, the use of arbitrarily small values for ϵ may cause numerical issues when one tries to solve it in practice. To address these issues, we make the following modifications to the model.

- We define $u_h^\ell(s)$ as the total flow through a hub at location $h \in \mathcal{H}$ with capacity $\ell \in L$.
- We replace the objective (2) with the following convex function.

$$\text{Minimize} \sum_{h\in\mathcal{H}}\sum_{\ell\in L} f_h^\ell y_h^\ell + \sum_{s\in S}\varphi(s)\left(\sum_{h\in\mathcal{H}}\sum_{\ell\in L} b_h \frac{u_h^\ell(s)}{q^\ell - u_h^\ell(s)} + \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k} c_p v_p(s)\right) \tag{10}$$

- We update the constraints (4),(5) and (9) with (11),(12) and (13), which we define as follows.

$$\sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k: h\in p} w_k(s)v_p(s) = \sum_{\ell\in L} u_h^\ell(s) \quad \forall h\in\mathcal{H}, s\in S, \tag{11}$$

$$u_h^\ell(s) \le q^\ell y_h^\ell \quad \forall h\in\mathcal{H}, \ell\in L, s\in S, \tag{12}$$

$$u_h^\ell(s) \ge 0 \quad \forall h\in\mathcal{H}, s\in S. \tag{13}$$

With the stated updates, the resulting problem is a convex problem. However, the objective function (10) is still nonlinear, due to the congestion term. Next, we discuss the SOCP transformation we propose to obtain the MISOCP formulation for the problem with a linear objective function and SOCP constraints.

3.1. A Mixed-Integer Second-Order Cone Programming Formulation

Due to the advances in theoretical findings and development of efficient interior point/barrier methods (Nesterov, Nemirovskii, and Ye 1994, Potra and Wright 2000), SOCP techniques have been applied to solve a wide range of optimization problems (see, for example, Atamtürk, Berenguer, and Shen 2012, Şen, Atamtürk, and Kaminsky 2018, Bayram and Yaman 2018). For further information on the SOCP and its complexity, we refer the reader to Lobo et al. (1998), Ben-Tal and Nemirovski (2001) and Alizadeh and Goldfarb (2003). Here, we reformulate the HNDC (2)–(8) as a MISOCP where the nonlinearity is transferred from the objective function to the constraint set in the form of second order quadratic constraints. To achieve this, we define an auxiliary variable $r_h^\ell(s)$, for each $h\in\mathcal{H}$, $s\in S$, and $\ell\in L$ as follows.

$$r_h^\ell(s) \ge 0 \quad \forall h\in\mathcal{H}, s\in S, \ell\in L, \tag{14}$$

$$r_h^\ell(s) \ge \frac{u_h^\ell(s)}{q^\ell - u_h^\ell(s)} \quad \forall h\in\mathcal{H}, s\in S, \ell\in L. \tag{15}$$

We transform inequality (15) into a second-order cone constraint by multiplying both sides of it by q^ℓ and adding $(u_h^\ell(s))^2$ to both sides, which yields:

$$(u_h^\ell(s))^2 \le (q^\ell r_h^\ell(s) - u_h^\ell(s))(q^\ell - u_h^\ell(s)) \quad \forall h\in\mathcal{H}, s\in S, \ell\in L. \tag{16}$$

Constraint (16) is a hyperbolic inequality of the form $\zeta^2 \le \xi_1\xi_2$ where $\zeta, \xi_1, \xi_2 \ge 0$. The constraint $\zeta^2 \le \xi_1\xi_2$ can be transformed into the quadratic form $\|(2\zeta, \xi_1 - \xi_2)\| \le \xi_1 + \xi_2$, where $\|\cdot\|$ is the Euclidean norm (see Lobo et al. 1998, Alizadeh and Goldfarb 2003). Hence, we can represent constraint (16) as the following second-order cone constraint (Günlük and Linderoth 2008, Salimian 2013):

$$\left\|\left(2u_h^\ell(s), q^\ell r_h^\ell(s) - q^\ell\right)\right\| \le q^\ell r_h^\ell(s) + q^\ell - 2u_h^\ell(s) \quad \forall h \in \mathcal{H}, s \in S, \ell \in L. \tag{17}$$

Using the above transformations, we can formulate the HLPCC as the following path-based MISOCP. Objective function (18) is the reformulation of (2), and (19) adds the required constraints.

$$\text{HNDC: Minimize} \quad \sum_{h\in\mathcal{H}}\sum_{\ell\in L} f_h^\ell y_h^\ell + \sum_{s\in S}\varphi(s)\left(\sum_{h\in\mathcal{H}}\sum_{\ell\in L} b_h r_h^\ell(s) + \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k} c_p v_p(s)\right) \tag{18}$$

$$\text{subject to: } (3), (7), (8), (11)\text{–}(14), (17). \tag{19}$$

Note that our methodology does not depend on this specific congestion cost function. Any function with the desired, exponentially increasing congestion cost, such as power functions and the Bureau of Public Roads (BPR) function (TAM 1964) could be employed (see online Appendix A for the required SOCP reformulation steps).

In the next section, we propose an exact solution approach based on Benders decomposition that decomposes the problem into two, an MIP and an SOCP, both of which we solve with efficient solution methodologies.

4. A Benders Decomposition Approach

We employ a Benders decomposition (BD) approach to decompose HNDC into smaller problems, a master problem (MP) and a subproblem for each scenario. MP is obtained by projecting out the second-stage decision variables and contains a large number of constraints called Benders cuts (BC). The solution of MP provides a lower bound on the optimal value of HNDC, as not all BCs exist (Geoffrion 1972). An iterative solution procedure is pursued, in which MP is solved and the information from the MP regarding temporarily fixed hub locations and capacities is passed to subproblems and then dual information from subproblems is obtained to generate BCs, until all of them are satisfied at a relaxed MP solution.

Fixing hub location and capacity decisions from MP, results in SOCP subproblems, one for each $s \in S$. We define the subproblem in Section 4.1, and discuss MP, identify BCs, and summarize the proposed BD for the HNDC in Section 4.2.

4.1. The Subproblem

In our problem, the binary design variables y are the complicating variables and are handled in the MP. Therefore, we can project out continuous variables u and v in the SP and determine them based on the given values of y found in the MP. The SP determines the traffic at each hub and the route(s) for each commodity such that the sum of congestion and routing costs is minimized. For a given scenario $s \in S$, the primal conic subproblem (PCSP) is presented by (20)–(29).

$$
\begin{aligned}
\text{PCSP}(\overline{y},s)\text{: Minimize} \quad & \sum_{h\in\mathcal{H}}\sum_{\ell\in L} b_h r_h^\ell(s) + \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k} c_p v_p(s) && && (20)\\
\text{subject to:} \quad & \sum_{p\in\mathcal{P}_k} v_p(s) = 1 && \forall k\in\mathcal{K}, && (21)\\
& \sum_{k\in\mathcal{K}}\sum_{p\in\mathcal{P}_k:h\in p} w_k(s)v_p(s) = \sum_{\ell\in L} u_h^\ell(s) && \forall h\in\mathcal{H}, && (22)\\
& u_h^\ell(s) \le q^\ell \overline{y}_h^\ell && \forall h\in\mathcal{H}, \ell\in L && (23)\\
& 2u_h^\ell(s) - t_h^\ell(s) = 0 && \forall h\in\mathcal{H}, \ell\in L && (24)\\
& q^\ell r_h^\ell(s) - t_h^{\ell-}(s) = q^\ell && \forall h\in\mathcal{H}, \ell\in L, && (25)\\
& q^\ell r_h^\ell(s) - 2u_h^\ell(s) - t_h^{\ell+}(s) = -q^\ell && \forall h\in\mathcal{H}, \ell\in L, && (26)\\
& (t_h^\ell(s))^2 + (t_h^{\ell-}(s))^2 \le (t_h^{\ell+}(s))^2 && \forall h\in\mathcal{H}, \ell\in L, && (27)\\
& u_h^\ell(s), r_h^\ell(s), t_h^{\ell+}(s) \ge 0 && \forall h\in\mathcal{H}, \ell\in L, && (28)\\
& v_p(s) \ge 0 && \forall k\in\mathcal{K}, p\in\mathcal{P}_k, && (29)
\end{aligned}
$$

where $\overline{y}_h^\ell$ are the fixed values for the design variables. Constraints (24)–(27) are used to represent the cone constraint (17).

The duality results of the SOCP in Benders decomposition has been applied in the literature (see Bayram and Yaman 2018). Using a similar methodology, we formulate the dual subproblem as follows. Let $\gamma_k(s), \delta_h(s), \eta_h^\ell(s), \lambda_h^\ell(s), \mu_h^\ell(s)$, and $\nu_h^\ell(s)$ be the dual variables associated with constraints (21)–(26), respectively, for a given scenario subproblem $s\in S$. Then, we can formulate the dual conic subproblem (DCSP) as follows.

$$
\begin{aligned}
\text{DCSP}(\overline{y},s)\text{: Maximize} \quad & \sum_{k\in\mathcal{K}} \gamma_k(s) - \sum_{h\in\mathcal{H}}\sum_{\ell\in L} q^\ell \left(\eta_h^\ell(s)\overline{y}_h^\ell - \mu_h^\ell(s) + \nu_h^\ell(s)\right) && && (30)\\
\text{subject to:} \quad & \gamma_k(s) + \sum_{h\in\mathcal{H}:h\in p} w_k(s)\delta_h(s) \le c_p && \forall k\in\mathcal{K}, p\in\mathcal{P}_k, && (31)\\
& -\delta_h(s) - \eta_h^\ell(s) + 2\lambda_h^\ell(s) - 2\nu_h^\ell(s) \le 0 && \forall h\in\mathcal{H}, \ell\in L, && (32)\\
& q^\ell \mu_h^\ell(s) + q^\ell \nu_h^\ell(s) \le b_h && \forall h\in\mathcal{H}, \ell\in L, && (33)\\
& (\lambda_h^\ell(s))^2 + (\mu_h^\ell(s))^2 \le (\nu_h^\ell(s))^2 && \forall h\in\mathcal{H}, \ell\in L, && (34)\\
& \eta_h^\ell(s), \nu_h^\ell(s) \ge 0 && \forall h\in\mathcal{H}, \ell\in L. && (35)
\end{aligned}
$$

The DCSP is also a SOCP problem. Note that for any feasible PCSP$(\overline{y},s)$, there exists a strictly feasible point. Since this problem is also bounded, the corresponding DCSP$(\overline{y},s)$ is also feasible and bounded by the strong duality theorem (Ben-Tal and Nemirovski 2001). In other words, PCSP and DCSP attain the same optimal solutions over given values of $\overline{y}$. Since design variables y have bounded feasible domains, BD generates finitely many cuts and terminates in a finite number of steps (Geoffrion 1972).

4.2. The Master Problem

The master problem of the HNDC includes location and capacity allocation decisions and a surrogate decision variable Ω to represent the congestion and transportation costs of the subproblem. The MP of the HNDC is given below.

$$\text{MP: Min.} \sum_{h\in\mathcal{H}}\sum_{\ell\in L} f_h^\ell y_h^\ell + \sum_{s\in S}\Omega(s) \tag{36}$$

$$\text{s.t.:}\quad \Omega(s) \geq \varphi(s)\left[\sum_{k\in\mathcal{K}}\overline{\gamma}_k^j(s) - \sum_{h\in\mathcal{H}}\sum_{\ell\in L} q^\ell\left(\overline{\eta}_h^{\ell j}(s)y_h^\ell - \overline{\mu}_h^{\ell j}(s) + \overline{\nu}_h^{\ell j}(s)\right)\right] \quad \forall s\in S, j\in J, \tag{37}$$

$$\sum_{\ell\in L} y_h^\ell \leq 1 \qquad \forall h\in\mathcal{H}, \tag{38}$$

$$y_h^\ell \in \{0,1\} \qquad \forall h\in\mathcal{H}, \ell\in L \tag{39}$$

$$\Omega(s) \geq 0 \qquad \forall s\in S, \tag{40}$$

where $\overline{\gamma}_k^j(s), \overline{\eta}_h^{\ell j}(s), \overline{\mu}_h^{\ell j}(s)$ and $\overline{\nu}_h^{\ell j}(s)$ are the values of γ, η, μ and ν variables obtained from scenario subproblem $s\in S$ and J is the set of optimal multiplier vectors. The objective function (36) minimizes the cost of opening hubs with a certain capacity and sum of the surrogate variable values. Constraint (37) represents optimality cuts added to the MP. Constraints (39)–(40) satisfy variable domains.

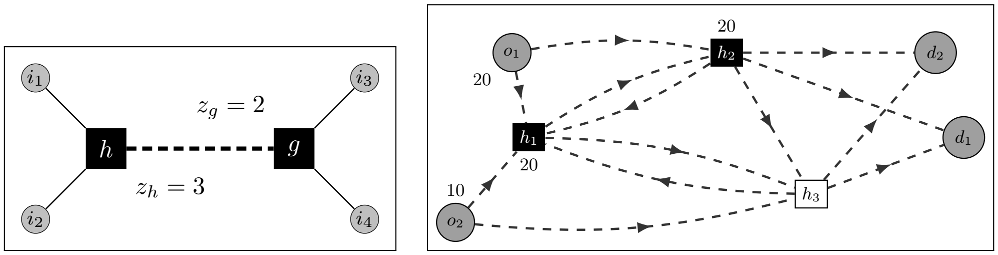
4.3. Solving the Master Problem

In each iteration of BD, the MP (36)–(40) is solved to decide on the strategic hub location and capacity acquisition variables. Next, based on these decisions, the SP is solved to determine flow routes. However, for a given MP solution $\overline{y}$, one or more scenario subproblems may turn out to be infeasible, if the allocated capacity in the MP is not enough to route all the demand. One way to address this issue is to look for Benders feasibility cuts first and check for optimality cuts when the subproblem if feasible. However, in BD, feasibility cuts are undesirable as they are usually much weaker compared to the optimality cuts and the need for adding a large number of them increases the solution time (Rahmaniani et al. 2017). One can avoid the burden of generating feasibility cuts by adding a set of valid inequalities that exclude the infeasible solution (see, for example, Contreras, Cordeau, and Laporte 2011a, de Sá, de Camargo, and de Miranda 2013). In the next section, we explain the approach we take to address such infeasible solutions.

4.3.1. Complete Networks. When the network is complete, i.e., all nodes (hub or non-hub) have access to all other nodes, Constraint (41) can generally guarantee a sufficient amount of capacity for the aggregate traffic demand.

$$\sum_{h\in\mathcal{H}}\sum_{\ell\in L} y_h^\ell q^\ell \geq \sum_{k\in\mathcal{K}} w_k(s), s\in S. \tag{41}$$

However, this constraint does not guarantee feasibility. See Figure 1a for such an example. Suppose there are 5 commodities, i_1–h, i_2–h, i_3–g, i_4–g, h–g, each with a demand equal to 1. Two hubs are opened by the master problem: h and g. Since total demand is 5, master problem solution allocates 3 units of capacity to h and 2 units of capacity to g due to constraint (41). However, the allocated capacities are not enough to route the traffic demand of all commodities. For example, if i_3–g and i_4–g are routed, the capacity of hub g is totally used, and h–g cannot be routed (since it requires 1 more unit of capacity at hub g).



(a) An infeasible MP solution for a complete network

(b) An infeasible MP solution for an incomplete network

Figure 1: Examples of MP solutions infeasible for a scenario subproblem

4.3.2. Incomplete Networks. Similarly, when the network is incomplete, adding constraint (41) is necessary but not sufficient to ensure feasibility. On top of the infeasibility issues encountered for complete networks as discussed above, further infeasibilities may be encountered. For instance, consider a network with three potential hub locations and two OD pairs illustrated in Figure 1b, where numbers next to rectangles and circles show the available hub capacities and flow amounts, respectively. Hubs h_1 and h_2 are opened, both with available capacity of 20 units, whereas h_3 is closed. Assume that the demand for commodities o_1–d_1 and o_2–d_2 are 20 and 10 units, respectively. Although the total available capacity is bigger than the total demand (i.e., $40 > 30$), there exists no feasible solution to SP due to limited h_2 capacity. Here, h_2 is a bottleneck node for routing the demand of commodities. A feasible solution can be obtained if h_2 capacity is sufficiently large, or h_3 is also opened with enough capacity.

To address this issue, one could derive feasibility cuts using dual information from the subproblems. However, it is a well-known fact that checking for the feasibility of the subproblems and solving the master problem every time new feasibility cuts are added until the feasibility of the subproblems is ensured, worsens the efficiency of the Benders decomposition algorithm to a great extent.

An alternative approach, the one we adopt, is to enhance the master problem with the linear inequalities that characterize the capacity decisions that guarantee feasible solutions in the routing subproblems. In the following sections, we explain the details of this approach and introduce the branch-and-cut algorithm we develop to solve the resulting master problem with exponentially many constraints.

4.3.3. Reformulation of the Master Problem. We first make the following observation that motivates the reformulation of the master problem. Note that in MP, the max-flow min-cut duality results (Ahuja, Magnanti, and Orlin 1988a) cannot be used to detect infeasibilities and add cut-set inequalities to eliminate them since the hub capacities are commonly used by multiple commodities leading to a multi-commodity-flow structure. We introduce a reformulation of the master problem to address this issue.

Let $\mathcal{K}_i$ be the set of commodities with destination $i \in N$ (i.e., $\mathcal{K}_i = \{k \in \mathcal{K} : d(k) = i\}$), which we call as "super commodity". We define $\tilde{z}_{hi}$ as the amount of capacity reserved for destination node $i \in N$ at hub $h \in \mathcal{H}$. Below, we present the reformulation of MP:

$$\begin{aligned}
\text{MPF: Minimize} \quad & (36) & & \\
\text{subject to:} \quad & (37)\text{–}(40) & & \\
& \sum_{i \in N} \tilde{z}_{hi} = \sum_{\ell \in L} q^{\ell} y_h^{\ell} & \forall h \in \mathcal{H}, & \quad (42)\\
& \tilde{z}_{hi} \geq 0 & \forall h \in \mathcal{H}, i \in N. & \quad (43)
\end{aligned}$$

Note that, unlike MP, MPF determines the allocations of hub capacities to super commodities with the equalities (42). Next, we describe how we make use of this lifted formulation to address the infeasibilities in the subproblems. We first introduce the following definitions to facilitate the technical discussions.

Definition 1. A set of hubs $\bar{H} \subseteq \mathcal{H}$ is called a *minimal-cut-set* of a *super commodity* $\mathcal{K}_i$ if all the paths of all the commodities $k \in \mathcal{K}_i$ has to use at least one hub from $\bar{H}$ and no proper subset of $\bar{H}$ has the same property. The set of all minimal-cut-sets of a destination $i \in \mathcal{N}$ is denoted by $\mathcal{H}_i$.

Definition 2. Let $(\bar{y}, \bar{z})$ be a solution for the capacity decisions in HNDC. We call $(\bar{y}, \bar{z})$ a capacity-feasible solution if there exists a feasible solution $(\bar{y}, \bar{z}, v, u)$ for HNDC.

The following proposition presents a full characterization of the capacity-feasible solutions for the master problem when the underlying graph is κ-connected.

Proposition 1. *Assume $\mathcal{G}$ is a κ-connected. Then, MPF is capacity-feasible if and only if the following inequalities hold.*

$$\sum_{h \in \bar{H}} \tilde{z}_{hi} \geq \sum_{k \in \mathcal{K}_i} w_k(s) \quad \forall i \in N, \bar{H} \subseteq \mathcal{H}_i, s \in S. \qquad (44)$$

Proof: Let $(y, \tilde{z})$ be a solution for the master problem. For each super-commodity $\mathcal{K}_i, i \in N$ and scenario $s \in S$, we define a *feasibility graph* $G_i^s = (\mathcal{N}_i^s, \mathcal{A}_i^s)$. The node set $\mathcal{N}_i^s$ contains an artificial source node σ_i, node i, two copies h and $\bar{h}$ for each potential hub location $h \in \mathcal{H}$, and the set of nodes $N_i = \{j \in N : w_{ji}(s) > 0\}$. The arc set $\mathcal{A}_i^s$ contains the following four groups of arcs.

- *origin arcs* A_1: for each $j \in N_i$ A_1 contains the arc (σ_i, j) with capacity $\kappa_{(\sigma_i,j)} = w_{ij}^s$.
- *capacity arcs* A_2: for each hub candidate h, A_2 contains the arc $(h, \bar{h})$ with capacity $\kappa_{(h,\bar{h})} = \tilde{z}_{ih}$.
- *hub access arcs* A_3: for each node $j \in J$, the arc set A_3 contains the arcs (j, h) for all $h \in \mathcal{H}$, if $(j, h) \in \mathcal{A}$ (i.e., the node j can reach to hub h in the original graph $\mathcal{G}$). Arcs in A_3 have infinite capacity.
- *destination arcs* A_4: for each candidate hub location $h \in \mathcal{H}$, the arc set A_4 contains an arc $(\bar{h}, i)$, if the node i is connected to the hub location h in the original graph. The capacities for the destination arcs are defined to be infinite.

Now assume (44) holds, then simply due to the well-known maxflow-mincut duality theorem (Ahuja, Magnanti, and Orlin 1988b) for every scenario $s \in S$ and super commodity $\mathcal{K}_i$, the maximum-flow from σ_i to i is equal to $w_i = \sum_{k \in \mathcal{K}_i} w_k(s)$, hence the solution is capacity-feasible. Similarly, if a given solution (y, z) is capacity-feasible, then the maximum-flow from σ_i to i must be equal to w_i for any super-commodity $\mathcal{K}_i$ and scenario $s \in S$. Then, by the maxflow-mincut duality theorem for any set of potential hub locations $\bar{H} \subseteq \mathcal{H}$ (44) must hold. Since we assume $\mathcal{G}$ is κ-connected, the max-flow solution can be always decomposed into paths that visit no more than κ hubs, hence the result follows. □

As a direct result of Proposition 1, adding (44) to MCF ensures capacity-feasible solutions if $\mathcal{G}$ is κ-connected, and they are valid cuts, otherwise. Here we want to underline that, in practice, hub locations are not chosen arbitrarily. They are typically well-connected nodes in the network and the underlying graph becomes κ-connected for even small κ values (i.e., $\kappa \geq 3$). Moreover, even for those graphs that do not satisfy this condition, one can still use (44) to eliminate a large number of infeasible solutions and handle the remaining few cases by including artificial paths (with arbitrarily large costs) at the start of the column generation algorithm to solve the subproblems.

Adding (44) to MPF, we formally present MP-CUT as follows:

$$\begin{aligned} &\text{MP-CUT: Minimize (36)} \\ &\qquad \text{s.t.:} \qquad \sum_{h \in \bar{H}} \tilde{z}_{hi} \geq \sum_{k \in \mathcal{K}_i} w_k(s) \qquad \forall i \in N, \bar{H} \subseteq \mathcal{H}_i, s \in S, \qquad (45) \\ &\qquad\qquad\qquad (37)\text{–}(40), (42)\text{–}(43). \end{aligned}$$

The number of inequalities in (45) grows exponentially with the number of alternative locations for the hubs ($|\mathcal{H}|$). Therefore, for realistic size problems, it is not practical to solve MP-CUT directly. We develop a branch-and-cut approach (B&C) to address this difficulty.

We start solving the master problem without considering *capacity constraints* (45). For every incumbent integer solution (found during the branch-and-bound search), we check whether there are any violated capacity constraints to include in the model to cut off capacity-infeasible solutions. For a given integer solution $(y, \tilde{z})$ of the relaxed master problem, we solve the separation problem by using the feasibility graphs introduced in the proof of Proposition 1. For any feasibility graph G_i^s, if the maximum-flow solution is less than w_i, we find the arcs in the min-cut $C_i^s \subset \mathcal{A}_i^s$ and consider the hub set $\bar{H} = \{h \in \mathcal{H} : (h, \bar{h}) \in C_i^s\}$ to detect the violated inequality $\sum_{h \in \bar{H}} \tilde{z}_{hi} \geq \sum_{k \in \mathcal{K}_i} w_k(s)$.

5. Solving the Subproblem by Column Generation

The PCSP (20)–(29) assumes that the complete set of candidate paths for each commodity, i.e. $\mathcal{P}_k$, is provided. However, it is impractical to generate and include all possible paths in the problem. In such cases, a column generation approach (CG) can be used to prevent enumerating all possibilities. CG is an optimization technique used to solve large combinatorial problems such as the cutting-stock problem and the vehicle routing problem. The reader is referred to Barnhart et al. (1998) and Desaulniers, Solomon, and Desrosiers (2005) for further information on CG and its applications. CG relies on the Dantzig–Wolfe decomposition (Dantzig and Wolfe 1960) of the problem into two problems, namely the master problem and the subproblem. The master problem is the original problem that only contains a meaningful subset of its columns (i.e., the restricted master problem or the RMP). The idea behind CG is to add new columns to the master problem when needed. In each iteration of CG, the current RMP is optimized in order to calculate dual multipliers. Next, a pricing subproblem is solved to find the reduced costs of the nonbasic variable(s) and add the eligible ones to the RMP. These steps repeat until the current basic feasible solution of the RMP is optimal.

In order to solve PCSP by CG, we start with a small set $\mathcal{P}_k$ including an artificial path $p_k = \{o_k, d_k\}$ for each commodity $k \in \mathcal{K}$ with an arbitrarily large cost. A new path $p \in \mathcal{P}_k$ for commodity $k \in \mathcal{K}$ can be added to the current PCSP if it has negative reduced cost. The reduced cost of a path variable $v_p(s)$, $s \in S$, $p \in \mathcal{P}_k$, denoted by $\tilde{c}_p(s)$, in the PCSP is calculated in (46).

$$\tilde{c}_p(s) = c_p - \gamma_k(s) - \sum_{h \in \mathcal{H}: h \in p} w_k \delta_h(s). \tag{46}$$

To price out candidate paths, a pricing subproblem (PP) is defined and solved for each $k \in \mathcal{K}$ and $s \in S$. The PP for commodity k and scenario s seeks a path p from o_k to d_k with the most negative $\tilde{c}_p(s)$ on the subgraph $\mathcal{G}_k$ of $\mathcal{G}$, which contains o_k, d_k, and all candidate hub locations $\mathcal{H}$. For the origin o_k the pricing graph contains only its outgoing arcs from the original graph, and for the destination d_k the pricing graph contains only the incoming arcs. Then, PP translates into solving an elementary shortest path problem with resource constraint (ESPPRC) on $\mathcal{G}_k^s$, where the number of arcs traversed on the path is the resource-bounded by $\kappa - 1$. We consider the following arc costs in the pricing graph $\mathcal{G}^s_k$.

- The cost of an outgoing arc $(o_k, h) \in \mathcal{A}$ of the source node o_k is defined as $\tilde{c}_{o_k h} = c_{o_k h} - w_k \delta_h(s)$.
- The cost of an interhub arc $(h, \bar{h}) \in \mathcal{A}$ is defined as $\tilde{c}_{h\bar{h}} = \alpha c_{h\bar{h}} - w_k \delta_{\bar{h}}(s)$.
- The cost of an incoming arc $(h, d_k) \in \mathcal{A}$ of the destination node d_k is defined as $-\gamma_k(s)$.

Observe that, the cost of path $p \in P_k$ in the pricing graph G^s_k is equal to the reduced cost $\tilde{c}_p(s)$ of a path variable $v_p(s)$. Therefore, in a column generation iteration, there exists a path variable with a negative reduced cost if and only if the shortest (cheapest) path with $\kappa - 1$ arcs in a pricing graph G^s_k, $k \in K$ $s \in S$ has a negative cost.

Also, note that the ESPPRC we define on the pricing graphs is essentially a hop-constrained shortest path problem. As a significant computational advantage for the column generation approach we propose, one can solve the hop-constrained shortest path problems efficiently (in polynomial time), using the well-known shortest path algorithms such as the Bellman-Ford shortest path algorithm (Ahuja, Magnanti, and Orlin 1988a) that can work with graphs with negative cost arcs and take the hop constraints into account. Note that, although the parameter κ (the maximum number of hubs that can be visited by a path) would restrict encountering them, pricing graphs can contain negative cost cycles. However, this does not invalidate the column generation procedure we use, since such path variables with cycles would be simply priced out in the following column generation iterations.

6. Computational Experiments

We perform our computational tests on a 64-bit Linux-operated workstation with two Intel Xeon Gold 6134 processors at 3.20 GHz and 96 GB of RAM. The algorithm is coded in Java v11.0 using ILOG CPLEX v12.10 (IBM 2019) as the mathematical programming solver. We use the lazy constraint callback function of CPLEX to add the Benders cuts. We used JGraphT library (Michail et al. 2019) Bellman-Ford shortest path algorithm implementation to solve the pricing problems. We employed a time limit (TL) of 10 hours to solve problem instances.

All PCSPs are initiated with a set of starting columns. We add the initial columns for any commodity $k \in \mathcal{K}$ and its possible path types in three ways: (i) single-hub paths, limiting the number of possible connections from origin to hubs to at most eight, selected based on least cost, (ii) two-hub paths, again limiting the number of possible connections from origin to hubs and from hubs to destinations to at most eight, each selected based on least connection cost, and (iii) artificial direct connections with arbitrarily large costs, to ensure feasibility at the start of the column generation algorithm. Note that since the feasibility is guaranteed, as mentioned, one can arbitrarily limit the number of possible connections to generate the initial set of paths, considering the tradeoff between starting with a good set of path variables and the increasing computational complexity in solving the restricted problem with a larger number of variables. Our preliminary experiments showed that for the problem instances we studied, eight would be a good choice regarding computational efficiency.

6.1. Problem Instances and the Design of Experiments

We use two data sets to generate the problem instances we study in our experiments: (i) the Turkish (TR) data set (Yaman, Kara, and Tansel 2007), retrieved from Kara (2011), and (ii) Australia Post (AP) data set (Ernst and Krishnamoorthy 1996), retrieved from Beasley (2018). The TR data set contains the distance, travel time, unit transportation cost, flow (demand) between all pairs of 81 Turkish cities, and fixed costs for opening hubs at them. No capacity information for the hubs is given in this data set. We assume a maximum capacity of 30K and generate three alternative capacity levels (small, medium, and large) equal to $\frac{1}{3}$, $\frac{2}{3}$, and 1 times the maximum capacity at each location. The details of the data we use for candidate hub locations are presented in online Appendix B. The AP data set contains up to 200 nodes and includes hub capacity information. Fixed hub opening costs and capacities are classified as either tight (T) or loose (L). We consider the 200-node AP data with L configuration for both opening costs and capacities. Both complete (C) and incomplete (IC) networks are considered in our experiments. To derive the incomplete networks for the TR data set, we first calculate a ranking index $\theta_{ij} = \frac{pop_i \times pop_j}{d_{ij}}$ for every origin-destination pair (i, j), where pop_i and pop_j indicates the populations of the cities i and j, respectively, and d_{ij} is the shortest path distance between the two cities. Ordering the origin-destination pairs from largest to smallest by their ranking index, we take the top 33% pairs to connect in the IC problem instances. For the AP data set, we do not have information on node densities. Therefore, we rank origin-destination pairs based on their flow and choose the top 33% pairs with the highest flow to establish connections in the IC instances. More information on how AP instances are used is provided in online Appendix D.

We consider two scenarios for demand between origin and destination pairs in our problem instances. The first scenario represents the expected flow quantities on typical days, whereas the second scenario represents the high-demand days, such as holidays. For the first scenario, we use the flow values given in TR and AP data sets. For the second scenario, we consider a high demand multiplier (HDM) to inflate the flow values. Considering the proportion of holidays in a year, we use 1/12 for the probability of the *high-demand* scenario.

In our numerical study, we investigate the effect of the following factors on the HNDC solution: (i) complete vs. incomplete networks, (ii) HDM, (iii) the discount factor, and (iv) cost of congestion (the multiplier for congestion cost function). To this end, we design four groups of experiments (E1-E3) as we present in Table 1. The first group (denoted by BC) has the base (default) parameter settings. In the other three groups, we consider variants of a parameter in the BC to inspect its impact. For each configuration, we generate five problem instances with $b_h \in \{0, 500, 1000, 1500, 2000\}$ and use capacity costs $f_h^\ell \in \{0.1, 0.09, 0.08\}$. In all our experiments, we allow maximum of five hub visits for a transfer (i.e., we use $\kappa = 5$).

Table 1: Experimental design.

Parameter	BC	E1	E2	E3
HDM	1.5	1, 1.25, 1.75, 2		
α	0.75		0, 0.25, 0.5, 1	
Network	IC			C

Empty cells indicate the same values as BC.

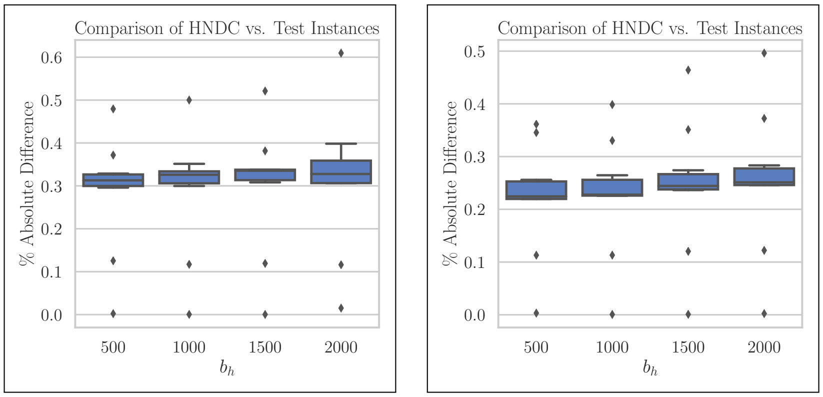
6.2. Algorithmic Efficiency

For our suggested algorithm, the detailed computational performance metrics for the TR and AP instances are presented in Tables 2 and 3, respectively, in online Appendix C. In particular, we investigate the number of feasibility cuts added to ensure that the master problem generates solutions feasible to subproblems, the number of optimality cuts added, the number of columns generated by the subproblems, the number of nodes searched in branch and bound tree not including the root node, solution time, and the percent relative gap at termination, for different networks (incomplete and complete), multiplier (HDM) for the worst case (high demand) scenario, inter-hub transportation discount factor α, and congestion cost coefficient b_h. We also test the deterministic low demand (DLD), mean demand (DMD), and high demand (DHD) scenarios, where a demand multiplier of 1, 1.25, and 1.5 is used, respectively, to generate the demand in the single scenario.

Our results show that the majority (68%) of the scenario-based stochastic TR instances are solved to optimality. The percentage of deterministic TR instances solved to optimality is much higher (87%). Our algorithm was able to find the optimal solution to 74% of the scenario-based stochastic and all deterministic AP instances in the designated time. Instances that cannot be solved to optimality are generally solved with small gaps in the designated time limit, the highest gap being around 18% for TR and 12% for AP instances. The results with larger gaps generally correspond to instances with bigger congestion cost coefficients and higher multiplier levels for high-demand scenarios. As b_h and HDM increase and as α decreases the number of nodes searched in the branch and bound tree, the number of feasibility and optimality cuts added, and hence the CPU times increase. Instances corresponding to a complete network setting, are solved relatively faster than the instances corresponding to an incomplete network setting.

We also investigate the effect of increasing the number of potential hubs in the network. As the number of potential hubs increases in TR instances, the solution times significantly increase. For example, although the instance on an IC network with 30 potential hubs, $b_h = 0$, $\alpha = 0.75$, and an HDM of 1.5 can be solved to optimality in 13,739.62 seconds when we increase the number of potential hubs to 40, the solution time to solve the same instance to optimality increases to 278,269.90 seconds. However, when the number of potential hubs in the TR data set is increased from 20 to 30 or 40, the solutions generally do not change since locations with higher populations are ranked higher in our potential list of hubs and typically have much larger demand to make them more preferable to be a hub. We observed similar behavior for AP instances. When the number of potential hubs increases from 15 to 20, the solution times increase significantly, but the optimal hub locations do not change. Therefore, the number of potential hub locations in TR and AP instances are considered as 20 and 15, respectively, selected among the nodes with the largest total incoming and outgoing flows. Please notice the large number of feasibility cuts added in all the instances. The characterization we provide for capacity-feasible solutions allows us to handle infeasibilities efficiently, solving a series of maximum flow problems.

We further test the quality of the solutions of the algorithm against randomly generated 100 new demand scenarios. Considering the high-demand scenario probability we use in our study (1/12), we generate nine scenarios with high demand and 91 scenarios with medium demand. Generating different scenarios with high and medium demand, we sample origin-destination demand volumes from a normal distribution with the mean equal to the values we consider in our high and medium demand scenarios and a standard deviation equal to 20% of the mean. We observe that the percentage absolute value differences between the optimal values of the solutions generated by HNDC and the results of the 100 test instances are insignificant for both data sets (Figure 2). The reason behind this result is twofold. First, the number of origin-destination pairs that visit a given hub is typically large (i.e., more than 200 for all hubs in our experiments). Therefore, individual fluctuations of demand volumes between each origin-destination pair cancel out at hubs, as long as there is no significant overall difference in the total demand between different scenarios. Therefore, considering different scenarios with different demand structures (like the primary and high-demand scenarios we study) generally suffices. Moreover, in our model, we allow the optimizer to change routing decisions after observing the demand, providing significant flexibility to accommodate fluctuations in demand. This observation corroborates that the use of carefully selected demand scenarios would suffice in hub network design considerations rather than using a large number of scenarios, which could have a trivial effect to improve the quality of solutions but would deteriorate the effectiveness of the algorithm and increase the solution time and hence result in worse solution quality within the same time limit.



(a) TR instances

(b) AP instances

Figure 2: Percentage absolute value differences between optimal values of HNDC and 100 test scenario instances

6.3. Managerial Insights

We explore the impact of congestion costs and demand uncertainty on hub network design and operational decisions. Our detailed results presented in online Appendix E, demonstrate that the inclusion of congestion costs influences hub network design decisions for certain instances. As expected, this effect is emphasized for higher levels of HDM, i.e., when we expect higher demand levels in the worst-case scenario. We also see that beyond a threshold, the higher increases in demand during the high times justify opening new hubs to take advantage of the discounts for inter-hub transfers. Similarly, the change in optimal design, compared to the solution with no congestion (i.e., $b_h = 0$), is observed for TR instances at smaller levels of discount factor α, due to the trade-off between achieving economies of scale and incurring congestion costs as a result of consolidating traffic flow.

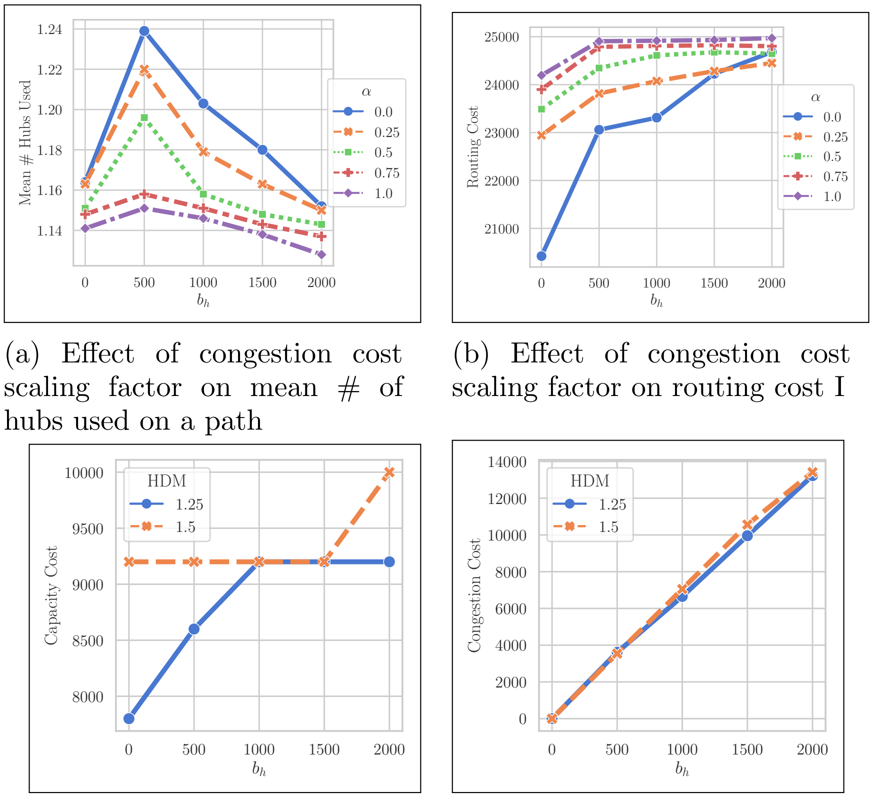
In Table 2, we compare the deterministic planning for hub network design in TR instances, where we use a single scenario corresponding to low demand (DLD, HDM $= 1$), medium (expected) demand (DMD, HDM $= 1.04$) and high demand (DHD, HDM $= 1.5$) cases against stochastic planning (SP) for various congestion cost scaling factors, when discount factor $\alpha = 0.75$. The results indicate that accounting for congestion cost and the uncertainty in demand are both important in hub network design. For all the TR instances where deterministic planning with a low or medium demand scenario is assumed and congestion is not considered (DLD0), the resulting hub network design cannot account for uncertainties in demand and is unable to accommodate enough capacity for routing operations and is therefore infeasible. However, in such a deterministic planning setting, considering congestion costs with a higher scaling factor, i.e., $b_h \geq 2,000$ and $b_h \geq 1,000$ for low and medium demand scenarios, respectively, hedges against uncertainties and results in a feasible hub network design. As we show in Table 9, in online Appendix E, our results for the AP instances present similar insights.

Table 2: Deterministic vs. stochastic planning (TR instances)

Method	b_h	Open Hubs	Hub Capacities
DLD0	500	Infeasible	
	1,000	Infeasible	
	1,500	Infeasible	
	2,000	Infeasible	
DLD500	500	Infeasible	
DLD1000	1,000	Infeasible	
DLD1500	1,500	Infeasible	
DLD2000	2,000	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
DMD0	500	Infeasible	
	1,000	Infeasible	
	1,500	Infeasible	
	2,000	Infeasible	
DMD500	500	Infeasible	
DMD1000	1,000	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
DMD1500	1,500	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
DMD2000	2,000	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
DHD0	500	[65 1 34 35 6 25]	[10000.0 10000.0 30000.0 30000.0 30000.0 10000.0]
	1,000	[65 1 34 35 6 25]	[10000.0 10000.0 30000.0 30000.0 30000.0 10000.0]
	1,500	[65 1 34 35 6 25]	[10000.0 10000.0 30000.0 30000.0 30000.0 10000.0]
	2,000	[65 1 34 35 6 25]	[10000.0 10000.0 30000.0 30000.0 30000.0 10000.0]
DHD500	500	[65 1 34 35 6 25]	[10000.0 20000.0 30000.0 30000.0 30000.0 10000.0]
DHD1000	1,000	[65 1 34 35 6 25]	[10000.0 30000.0 30000.0 30000.0 30000.0 20000.0]
DHD1500	1,500	[65 1 34 35 6 25]	[10000.0 30000.0 30000.0 30000.0 30000.0 10000.0]
DHD2000	2,000	[65 1 34 35 6 25]	[10000.0 20000.0 30000.0 30000.0 30000.0 10000.0]
SP	0	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
	500	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
	1,000	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
	1,500	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 10000.0]
	2,000	[65 1 34 6 25]	[10000.0 30000.0 30000.0 30000.0 20000.0]

Figure 3 illustrates the effect of congestion cost scaling factor b_h on different network design and operational cost factors for TR data set. It plots the mean number of hubs used on a path (Figure 3a) and routing cost (Figure 3b) versus b_h for various levels of discount factor α. Figures 3c and 3d illustrate the effect of b_h on capacity and congestion costs for HDM $= 1.25$ and 1.5, when discount factor $\alpha = 0.75$. We observe that the maximum number of hubs used on a path is three and that there is a nonlinear relationship between the congestion cost scaling factor b_h and the mean number of hubs used on a path. As the congestion cost scaling factor increases, the model generates solutions with longer paths using more hubs, hence tolerating larger routing costs (Figure 3b) to evade congestion costs at the hubs up to a point (scaling factor 500 in Figure 3a). After that point, due to the higher contribution of congestion costs to overall costs, the model tries to use paths with as few hubs as possible, since every hub used on a path causes an increase in congestion cost nonlinearly relative to the flow on the path and hence the mean number of hubs used on a path decreases. Another effect of increasing congestion cost scaling factor is an increase in routing costs as expected. One also expects to see increasing routing costs as discount factor α increases, i.e., as economies of scale due to consolidation decreases, as also illustrated in Figure 3b. However, as the congestion cost scaling factor increases, the question of which discount factor is used becomes unimportant, i.e., the difference between various transportation means used for inter-hub transportation with different capacities disappears in the TR network. Increasing congestion cost scaling factors also results in higher capacity levels to hedge against a nonlinear increase in congestion costs. When HDM is small (Figure 3c), an increase in congestion cost scaling factor results in higher capacities, whereas when HDM is larger, larger capacities are allocated, and congestion cost scaling factor begins to affect capacity decisions only at higher levels as the model generates hub network design solutions with large capacities to hedge against uncertain high-level demands and uses that extra capacity against congestion up to a point. This is further illustrated in Figure 3d. This figure shows how allocating more capacity for higher HDM values helps to prevent a significant increase in the congestion cost. However, despite this hedging and having the flexibility to reroute the flow to reduce excessive congestion at hubs, disregarding the congestion costs when designing the hub networks can result in significant losses. As an example, Figure 4a compares the resulting expected costs of the network designs that consider the congestion costs (with congestion) and that disregard it (without congestion) for various values of congestion cost scaling factors for the problem instances from TR network, with the high demand multiplier of 1.25. To calculate the expected costs, we draw 100 random demand realizations and find the average cost for the fixed network design decisions with and without congestion costs. As expected, we see that as the congestion cost increases, the cost of disregarding the congestion may reach significant amounts (up to three percent of the total design and operation costs for TR and four percent for AP instances in this experiment).

Figure 5 illustrates the effect of discount factor α on the mean number of hubs (Figure 5a) for various levels of congestion, on congestion cost (Figure 5b), on routing cost (Figure 5c), on location cost (Figure 5d), on capacity cost (Figure 5e), and on total cost savings (Figure 5f). Due to decreasing economies of scale advantages between two hubs, as discount factor α increases, the mean number of hubs used on a path decreases, as well. There is a trade-off between consolidating flows at hubs to benefit from inter-hub discounts due to economies of scale, hence decreasing routing costs and increasing congestion costs at the hubs due to consolidation. This trade-off is illustrated in Figures 5b and 5c. As α decreases, consolidating flows at the hubs and inter-hub routing becomes more advantageous at the expense of congestion costs at the hubs (Figure 5b). As α increases, the advantage of routing flows between hubs is lost, and routing costs increase (Figure 5c). We can also see from Figures 5d and 5e that when there is no congestion cost and $\alpha = 0$, the solution offers a bigger number of open hubs and higher capacity to benefit from economies of scale and save on routing cost (Figure 5c). However, with higher values of b_h and α, we face larger total congestion and routing costs, therefore, the solution provides a different network topology with less number of hubs and smaller capacities to save on the total cost. Although the discount factor has an impact



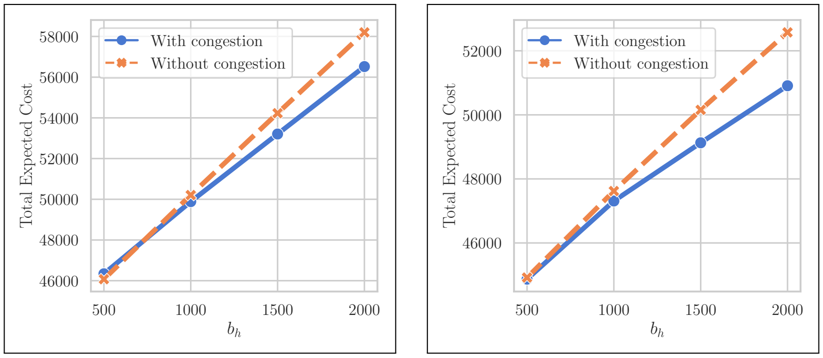
(a) Effect of congestion cost scaling factor on mean # of hubs used on a path

(b) Effect of congestion cost scaling factor on routing cost I

(c) Effect of congestion cost scaling factor on capacity cost ($\alpha = 0.75$)

(d) Effect of congestion cost scaling factor on the total congestion cost ($\alpha = 0.75$)

Figure 3: Effect of congestion cost scaling factor on hub network design and operational costs in TR instances

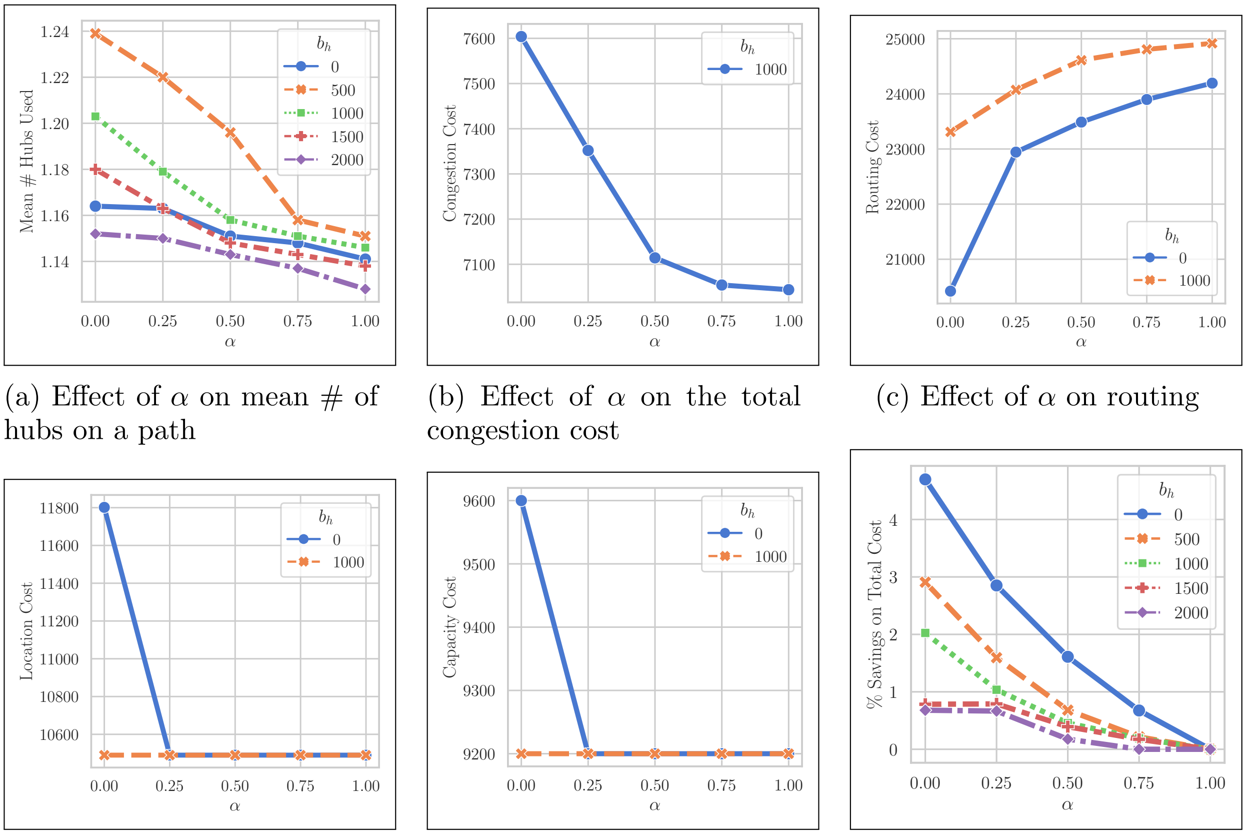


(a) TR instances (HDM = 1.25)

(b) AP instances (HDM = 1.25)

Figure 4: Cost increase due to ignoring congestion cost

(provides savings) on routing costs, which typically constitute the largest part of the total costs, the net savings that can be obtained through discounted inter-hub transfer costs strongly depend on the congestion cost factor, as shown in Figure 5f. For example, we see that when the congestion



(a) Effect of α on mean # of hubs on a path

(b) Effect of α on the total congestion cost

(c) Effect of α on routing

(d) Effect of α on location cost

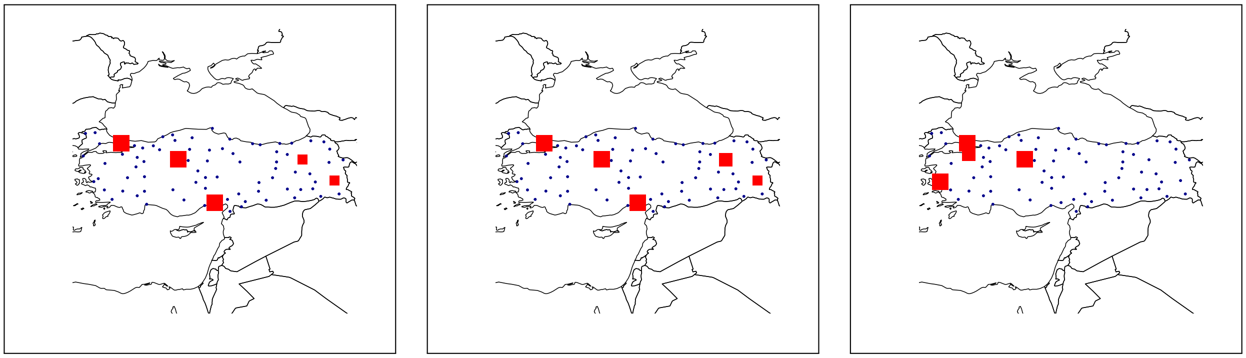
(e) Effect of α on capacity cost

(f) Effect of α on total cost savings

Figure 5: Effect of discount factor α on operational costs in TR instances (HDM = 1.5)

cost scaling factor is 2000, a 25% discount in inter-hub transfer costs ($\alpha = 0.75$) cannot bring any significant net savings, while the total costs can be reduced by almost 1% by using inter-hub links more if there were no congestion cost (i.e., congestion cost scaling factor is equal to zero). In our problem setting, peak demand levels are observed at only small time intervals during the whole planning horizon (e.g., one month of the holiday season in a year). Thus, it is plausible that for certain instances, the achievable savings may be comparatively modest. In contrast, in scenarios where peak demand levels are more prolonged, we anticipate that greater savings could be realized. Similar results we obtain for the AP instance solutions are presented in online Appendix E.

We also investigate the hub network topology differences when the underlying graph is complete and incomplete. When we work with incomplete graphs in TR instances with a discount factor of 0.75 and an HDM of 1.5, the resulting network design has Istanbul, Ankara, Adana, Erzurum, and Van cities as hubs for all congestion cost scaling factors, increasing the capacity of Erzurum hub when scaling factor increases from 1,500 (Figure 3a) to 2,000 (Figure 3b). This topology is able to serve the whole network with hubs spread all around Turkey and having enough capacity. However, when we have a complete network structure, the resulting design opens a smaller number of hubs with bigger capacities, concentrating them on the west side of Turkey in more populated cities Istanbul, Ankara, Izmir, and Bursa (Figure 3c). Similar behavior can be observed in AP network design solutions (see Figure 3 in online Appendix E). The optimal topology resulting from using incomplete networks is a good example for ground cargo logistics systems, whereas the second one is an example of how airlines operate.



(a) Incomplete network, $0 \leq b_h \leq 1500$

(b) Incomplete network, $b_h = 2000$

(c) Complete network, $0 \leq b_h \leq 2000$

Figure 6: Complete vs. incomplete networks (TR instances)

7. Conclusion

This study proposes a new approach to address the congestion issues in hub network design by integrating capacity acquisition decisions and congestion costs into the classical hub location problem. We introduce the hub network design problem with congestion, capacity, and stochastic demand considerations, which allows for dynamic routing and joint decision-making on hub locations and capacity acquisitions based on expected routing and congestion costs. We propose a path-based mixed-integer second-order cone programming formulation of the problem and develop an efficient solution algorithm based on Benders decomposition and column generation.

Our computational experiments demonstrate that accounting for congestion costs and uncertainty in demand leads to the preservation of larger hub capacity levels, and the resulting hub network topology tends to have a higher number of dispersed hubs when the underlying network is incomplete. Ignoring congestion costs, uncertainty in demand, and the network structure, results in higher total costs and may even lead to infeasible solutions. We also observe a trade-off between consolidating flows at hubs for inter-hub economies of scale and increasing congestion costs due to consolidation.

In this study, we assumed a single stage for network design. Our study can be extended for those applications where the network needs to be built in multiple steps with changing demand trends and limited resources to use in each stage. Incorporating user perspectives and service quality measures to generate solutions to satisfy both the network owner and network users is another interesting research direction we suggest for future studies.

Acknowledgement

The authors thank the Scientific and Technological Research Council of Turkey (TÜBİTAK) who provided financial support for this work (Grant No: 218M520).

References

Ahuja RK, Magnanti TL, Orlin JB, 1988a *Network flows* (Cambridge, Mass.: Alfred P. Sloan School of Management, Massachusetts . . .).

Ahuja RK, Magnanti TL, Orlin JB, 1988b *Network flows* .

Alibeyg A, Contreras I, Fernández E, 2016 *Hub network design problems with profits. Transportation Research Part E: Logistics and Transportation Review* 96:40–59.

Alizadeh F, Goldfarb D, 2003 *Second-order cone programming. Mathematical Programming* 95(1):3–51.

Alumur SA, Campbell JF, Contreras I, Kara BY, Marianov V, O'Kelly ME, 2020 *Perspectives on modeling hub location problems. European Journal of Operational Research* .

Alumur SA, Kara BY, 2008 *Network hub location problems: The state of the art. European Journal of Operational Research* 190(1):1–21.

Alumur SA, Kara BY, Karasan OE, 2009 *The design of single allocation incomplete hub networks. Transportation Research Part B: Methodological* 43(10):936–951.

Alumur SA, Nickel S, Rohrbeck B, da Gama FS, 2018 *Modeling congestion and service time in hub location problems. Applied Mathematical Modelling* 55:13–32.

Alumur SA, Nickel S, Saldanha-da Gama F, 2012 *Hub location under uncertainty. Transportation Research Part B: Methodological* 46(4):529–543.

Atamtürk A, Berenguer G, Shen ZJ, 2012 *A conic integer programming approach to stochastic joint location-inventory problems. Operations Research* 60(2):366–381.

Aykin T, 1994 *Lagrangian relaxation based approaches to capacitated hub-and-spoke network design problem. European Journal of Operational Research* 79(3):50–523.

Azizi N, Vidyarthi N, Chauhan SS, 2018 *Modelling and analysis of hub-and-spoke networks under stochastic demand and congestion. Annals of Operations Research* 264(1):1–40.

Barnhart C, Johnson EL, Nemhauser GL, Savelsbergh MWP, Vance PH, 1998 *Branch-and-price: Column generation for solving huge integer programs. Operations Research* 46(3):316–329.

Bayram V, Yaman H, 2018 *Shelter location and evacuation route assignment under uncertainty: A benders decomposition approach. Transportation Science* 52(2):416–436.

Beasley JE, 2018 *OR-library: Hub location.* URL `http://people.brunel.ac.uk/~mastjjb/jeb/orlib/phubinfo.html`, accessed July 20, 2022.

Ben-Tal A, Nemirovski A, 2001 *Lectures on modern convex optimization: Analysis, algorithms, and engineering applications. MPS-SIAM Series on Optimization* .

Benders JF, 1962 *Partitioning procedures for solving mixed-variables programming problems. Numerische Mathematik* 4(1):238–252.

Brimberg J, Mladenović N, Todosijević R, Urošević D, 2019 *A non-triangular hub location problem. Optimization Letters* 1–20.

Bütün C, Petrovic S, Muyldermans L, 2021 *The capacitated directed cycle hub location and routing problem under congestion. European Journal of Operational Research* 292(2):714–734.

Campbell JF, 1992 *Location and allocation for distribution systems with transshipments and transportion economies of scale. Annals of operations research* 40(1):77–99.

Campbell JF, Ernst AT, Krishnamoorthy M, 2005a *Hub arc location problems: part ii—formulations and optimal algorithms. Management Science* 51(10):1556–1571.

Campbell JF, Ernst AT, Krishnamoorthy M, 2005b *Hub arc location problems: part i—introduction and results. Management Science* 51(10):1540–1555.

Campbell JF, O'Kelly ME, 2012 *Twenty-five years of hub location research. Transportation Science* 46(2):153–169.

Contreras I, 2015 *Hub location problems.* Laporte G, Nickel S, Saldanha da Gama F, eds., *Location Science*, 311–344 (Cham: Springer International Publishing).

Contreras I, Cordeau JF, Laporte G, 2011a *Benders decomposition for large-scale uncapacitated hub location. Operations Research* 59(6):1477–1490.

Contreras I, Cordeau JF, Laporte G, 2011b *Stochastic uncapacitated hub location. European Journal of Operational Research* 212(3):518–528.

Contreras I, Cordeau JF, Laporte G, 2012 *Exact solution of large-scale hub location problems with multiple capacity levels. Transportation Science* 46(4):439–459.

Contreras I, Fernández E, Marín A, 2010 *The tree of hubs location problem. European Journal of Operational Research* 202(2):390–400.

Contreras I, O'Kelly M, 2019 *Hub location problems. Location science*, 327–363 (Springer).

Correia I, Nickel S, Saldanha-da Gama F, 2010 *Single-assignment hub location problems with multiple capacity levels. Transportation Research Part B: Methodological* 44(8-9):1047–1066.

Dantzig GB, Wolfe P, 1960 *Decomposition principle for linear programs. Operations Research* 8(1):101–111.

De Camargo RS, de Miranda Jr G, Ferreira RP, 2011 *A hybrid outer-approximation/benders decomposition algorithm for the single allocation hub location problem under congestion. Operations Research Letters* 39(5):329–337.

de Camargo RS, de Miranda Jr G, O'Kelly ME, Campbell JF, 2017 *Formulations and decomposition methods for the incomplete hub location network design problem with and without hop-constraints. Applied Mathematical Modelling* 51:274–301.

de Camargo RS, Miranda G, 2012 *Single allocation hub location problem under congestion: Network owner and user perspectives. Expert Systems with Applications* 39(3):3385–3391.

de Camargo RS, Miranda G, Ferreira RPM, Luna HP, 2009 *Multiple allocation hub-and-spoke network design under hub congestion. Computers & Operations Research* 36(12):3097–3106, new developments on hub location.

de Sá EM, de Camargo RS, de Miranda G, 2013 *An improved benders decomposition algorithm for the tree of hubs location problem. European Journal of Operational Research* 226(2):185–202.

de Sá EM, Morabito R, de Camargo RS, 2018 *Benders decomposition applied to a robust multiple allocation incomplete hub location problem. Computers & Operations Research* 89:31–50.

Desaulniers G, Solomon MM, Desrosiers J, 2005 *Column Generation* (Springer).

Di Crescenzo A, Giorno V, Kumar BK, Nobile AG, 2018 *M/m/1 queue in two alternating environments and its heavy traffic approximation. Journal of Mathematical Analysis and Applications* 465(2):973–1001.

Ebery J, Krishnamoorthy M, Ernst A, Boland N, 2000 *The capacitated multiple allocation hub location problem: Formulations and algorithms. European Journal of Operational Research* 120(3):614–631.

Elhedhli S, Hu FX, 2005 *Hub-and-spoke network design with congestion. Computers & Operations Research* 32(6):1615–1632.

Elhedhli S, Wu H, 2010 *A lagrangean heuristic for hub-and-spoke system design with capacity selection and congestion. INFORMS Journal on Computing* 22(2):282–296.

Ernst A, Krishnamoorthy M, 1999 *Solution algorithms for the capacitated single allocation hub location problem. Annals of Operations Research* 86:141–159.

Ernst AT, Krishnamoorthy M, 1996 *Efficient algorithms for the uncapacitated single allocation p-hub median problem. Location Science* 4(3):139–154, URL `http://dx.doi.org/10.1016/S0966-8349(96)00011-3`, hub Location.

FAA, 1976 *Techniques for determining airport airside capacity and delay.* Technical report, U.S. DEPARTMENT OF TRANSPORTATION, FEDERAL AVIATION ADMINISTRATION, FAA-RD-74-124.

Farahani RZ, Hekmatfar M, Arabani AB, Nikbakhsh E, 2013 *Hub location problems: A review of models, classification, solution techniques, and applications. Computers & Industrial Engineering* 64(4):1096–1109.

Forbes, 2019 *Airlines Struggle To Cope With Rush-Hour-Style Congestion.* URL `https://www.forbes.com/sites/oliverwyman/2019/09/19/airlines-struggle-to-cope-with-rush-hour-style-congestion/?sh=4bec21c12594`, accessed on 2021-01-21.

Geoffrion AM, 1972 *Generalized benders decomposition. Journal of Optimization Theory and Applications* 10(4):237–260.

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

Guldmann JM, Shen G, 1997 *A general mixed integer nonlinear optimization model for hub network design. 44th North American meeting of the Regional Science Association International.*

Günlük O, Linderoth J, 2008 *Perspective relaxation of mixed integer nonlinear programs with indicator variables.* Lodi A, Panconesi A, Rinaldi G, eds., *Integer Programming and Combinatorial Optimization*, 1–16 (Berlin: Springer).

Hübl A, Altendorfer K, 2015 *State probabilities for an m/m/1 queuing system with two capacity levels. 2015 Winter Simulation Conference (WSC)*, 2219–2226 (IEEE).

IBM, 2019 *IBM ILOG CPLEX Optimization Studio V12.10.0.* URL `https://www.ibm.com/products/ilog-cplex-optimization-studio`, accessed on 2019-12-22.

Ignaccolo M, 2003 *A simulation model for airport capacity and delay analysis. Transportation Planning and Technology* 26(2):135–170.

Ishfaq R, Sox CR, 2012 *Design of intermodal logistics networks with hub delays. European Journal of Operational Research* 220(3):629–641.

Jittrapirom P, Caiati V, Feneri AM, Ebrahimigharehbaghi S, González M, Narayan J, 2017 *Mobility as a service: A critical review of definitions, assessments of schemes, and key challenges. Urban Planning* 2(2):13–25.

Kara BY, 2011 *Hub location* URL `https://ie.bilkent.edu.tr/~bkara/hub_location.php`, accessed on 2019-09-04.

Kian R, Kargar K, 2016 *Comparison of the formulations for a hub-and-spoke network design problem under congestion. Computers & Industrial Engineering* 101:504–512.

Kızıl KU, Yıldız B, 2022 *Public transport-based crowd-shipping with backup transfers. Transportation Science* .

Kleinrock L, 1964 *Communication nets: Stochastic message flow and delay* (McGraw-Hill, New York).

Klincewicz JG, 1998 *Hub location in backbone/tributary network design: a review. Location Science* 6(1):307–335.

Little JD, 1961 *A proof for the queuing formula: L=* λ *w. Operations research* 9(3):383–387.

Lobo MS, Vandenberghe L, Boyd S, Lebret H, 1998 *Applications of second-order cone programming. Linear algebra and its applications* 284(1):193–228.

Macrina G, Pugliese LDP, Guerriero F, Laporte G, 2020 *Crowd-shipping with time windows and transshipment nodes. Computers & Operations Research* 113:104806.

Marianov V, Serra D, 2003 *Location models for airline hubs behaving as M/D/c queues. Computers & Operations Research* 30(7):983–1003.

Marín A, Nickel S, Schöbel A, Sonneborn T, 2002 *Extensions of the uncapacitated hub location problem for applications in intermodal public transportation. 13th mini-EURO conference and IX meeting of the EURO Working Group on Transportation* (Bari, Italy).

Mayer C, Sinai T, 2003 *Network effects, congestion externalities, and air traffic delays: Or why not all delays are evil. American Economic Review* 93(4):1194–1215.

Meraklı M, Yaman H, 2016 *Robust intermodal hub location under polyhedral demand uncertainty. Transportation Research Part B: Methodological* 86:66–85.

Meraklı M, Yaman H, 2017 *A capacitated hub location problem under hose demand uncertainty. Computers & Operations Research* 88:58–70.

Meuffels WJM, 2015 *The design of road and air networks for express service providers.* Ph.D. thesis, Tilburg University, The Netherlands.

Michail D, Kinable J, Naveh B, Sichi JV, 2019 *Jgrapht–a java library for graph data structures and algorithms. arXiv preprint arXiv:1904.08355* .

Montreuil B, 2011 *Toward a physical internet: meeting the global logistics sustainability grand challenge. Logistics Research* 3(2-3):71–87.

Najy W, Diabat A, 2020 *Benders decomposition for multiple-allocation hub-and-spoke network design with economies of scale and node congestion. Transportation Research Part B: Methodological* 133:62–84.

Nesterov Y, Nemirovskii AS, Ye Y, 1994 *Interior-point polynomial algorithms in convex programming*, volume 13 (SIAM).

Nickel S, Schöbel A, Sonneborn T, 2001 *Hub location problems in urban traffic networks. Mathematical methods on optimization in transportation systems*, 95–107 (Springer).

Odoni AR, Roth E, 1983 *An empirical investigation of the transient behavior of stationary queueing systems. Operations Research* 31(3):432–455.

O'Kelly ME, 1986a *Activity levels at hub facilities in interacting networks. Geographical Analysis* 18(4):343–356.

O'Kelly ME, 1986b *The location of interacting hub facilities. Transportation science* 20(2):92–106.

O'Kelly ME, Campbell JF, de Camargo RS, de Miranda Jr G, 2015 *Multiple allocation hub location model with fixed arc costs. Geographical Analysis* 47(1):73–96.

Pérez JMQ, Lange JC, Tancrez JS, 2018 *A multi-hub express shipment service network design model with flexible hub assignment. Transportation Research Part E: Logistics and Transportation Review* 120:116–131.

Potra FA, Wright SJ, 2000 *Interior-point methods. Journal of Computational and Applied Mathematics* 124(1):281–302, numerical Analysis 2000. Vol. IV: Optimization and Nonlinear Equations.

Pyrgiotis N, Malone KM, Odoni A, 2013 *Modelling delay propagation within an airport network. Transportation Research Part C: Emerging Technologies* 27:60–75.

Rahmaniani R, Crainic TG, Gendreau M, Rei W, 2017 *The benders decomposition algorithm: A literature review. European Journal of Operational Research* 259(3):801–817.

Rodríguez-Martín I, Salazar-González JJ, 2008 *Solving a capacitated hub location problem. European Journal of Operational Research* 184(2):468–479.

Rothenbächer AK, Drexl M, Irnich S, 2016 *Branch-and-price-and-cut for a service network design and hub location problem. European Journal of Operational Research* 255(3):935–947.

Salimian M, 2013 *A mixed integer second order cone programming reformulation for a congested location and capacity allocation problem in supply chain network design.* Master's thesis, Middle East technical University, Turkey.

Schwarz JA, Selinka G, Stolletz R, 2016 *Performance analysis of time-dependent queueing systems: Survey and classification. Omega* 63:170–189.

Şen A, Atamtürk A, Kaminsky P, 2018 *A conic integer optimization approach to the constrained assortment problem under the mixed multinomial logit model. Operations Research* 66(4):994–1003.

Taherkhani G, Alumur SA, Hosseini M, 2020 *Benders decomposition for the profit maximizing capacitated hub location problem with multiple demand classes. Transportation Science* 54(6):1446–1470.

TAM, 1964 *Traffic Assignment Manual. Bureau of Public Roads, U.S. Department of Commerce* .

Tanash M, Contreras I, Vidyarthi N, 2017 *An exact algorithm for the modular hub location problem with single assignments. Computers & Operations Research* 85:32–44.

van Essen JT, 2009 *Heuristics for the hub location and network design problem with a mixed vehicle fleet.* Master's thesis, Delft University of Technology, The Netherlands.

Yaman H, 2008 *Star p-hub median problem with modular arc capacities. Computers & Operations Research* 35(9):3009–3019.

Yaman H, Carello G, 2005 *Solving the hub location problem with modular link capacities. Computers & Operations Research* 32(12):3227–3245.

Yaman H, Kara BY, Tansel BC, 2007 *The latest arrival hub location problem for cargo delivery systems with stopovers. Transportation Research Part B: Methodological* 41(8):906–919.

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

Yildiz B, Yaman H, Karasan O, 2021 *Hub location, routing and route dimensioning: Strategic and tactical intermodal transportation hub network design. Transportation Science* .