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

Vedat Bayram
Industrial Engineering Department, TED University, Ankara, Turkey, vedat.bayram@tedu.edu.tr,

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

Mohammad Saleh Farham
CIRRELT, University of Montreal, Montreal, Quebec, CANADA, pharham@tuta.io,

We introduce 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 decision 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 taking the expected routing and congestion costs into account. 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 performance of the proposed approach and to 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), and mobility as a service (Jittrapirom et al. 2017), which are poised to revolutionize the transport sector. 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 cost 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 number 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 require 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
(Elhedhli and Wu 2010). The introduced problem has the following key characteristics: (i) Hubs are
capacitated, i.e., the total flow assigned to 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 $\kappa$ number of hubs, where $\kappa$ is simply a problem parameter to limit
the number of hub visits for operational convenience. (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.

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). 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 literature.
The largest problem instance solved by our approach contains 81 nodes (40 of which are potential
hub locations), 64,800 OD pairs.

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

Hubs generally serve as switching, transshipment and sorting facilities for transportation, distribution, and telecommunication systems that require transportation/distribution of people/goods/data from multiple origins to multiple destinations (Alumur and Kara 2008). This allows consolidating flows to achieve economies of scale rather than choosing the costlier way of sending these flows directly from their origin to their destination.

Although hub networks allow exploitation of economies of scale, minimizing hub installation and transportation costs, and consolidation and dissemination of flows at hubs tend to generate solutions with a small number of overloaded hubs creating congestion as one major side effect (de Camargo et al. 2009, Elhedhli and Wu 2010). Congestion is a function of the capacity and the total flow assigned to hubs and leads to delays, thus to an increase in total travel/service times. Therefore, capacity building decisions directly affect the congestion at the hubs and the travel time in the network. There is a tradeoff between the cost of installing hubs and building capacities and the congestion level at the hubs. Demand uncertainty is another factor leading to congestion at the hubs (Marianov and Serra 2003). Ignoring congestion and the factors that lead to it at the design of hub-and-spoke networks may lead to high overall costs and possibly infeasible solutions. Studies in the literature that take into account all these considerations simultaneously are lacking. Therefore, strategies that jointly consider congestion cost, the location, number, and capacity of hubs, the routing decisions, and uncertainty in demand must be developed.

Ever since the introduction of the first model by O’Kelly (1986), a vast body of research has flourished on hub location problems (HLP) with contributions on 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 decision making process and for which path based formulations are proposed to deal with incomplete networks and to allow use of more than two hubs in routing decisions.

2.1. Capacity Acquisition and Congestion

O’Kelly (1986) was the first to point out the negative effects of congestion due to traffic consolidation, which may result in schedule delays and increased service 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), Merakh 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 use of more than two hubs on a route.

Modeling congestion through capacity constraints, however, do 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 congestion effect increases more rapidly. Therefore, nonlinear modeling of the congestion yields more realistic results. There exist a limited number of studies (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, Alumur et al. 2018, Najy and Diabat 2020, Bütün, Petrovic, and Muyldermans 2021) in the literature that explicitly consider congestion cost in the HLP. The congestion costs for HLPs in the literature are commonly of two types. 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 congestion cost functions, called Kleinrock average delay functions. In such studies, it is assumed that each hub acts as an M/M/1 queue and under steady-state conditions. Therefore, a nonlinear congestion cost function of the form $bu/(z - u)$ is used, 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.

The first study in the literature that explicitly considers congestion cost in the objective function is by Guldmann and Shen (1997). They employ piecewise linearizations 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), 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 Lagrangean 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 $GI/G/1$ 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.

Most of the studies in the literature that explicitly consider nonlinear congestion costs in their formulation, approximate 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 considers 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 number 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 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 the strategic location and capacity acquisition decisions as well as the 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 Marianov and Serra (2003), Contreras, Cordeau, and Laporte (2011b), Alumur, Nickel, and Saldanha-da Gama (2012), Merakh and Yaman (2016, 2017), de Sá, Morabito, and de Camargo (2018), Taherkhani, Alumur, and Hosseini (2020). The first study in the literature that considers uncertainty in demand is by Marianov and Serra (2003). By considering a Kleinrock type congestion cost function, they limit the probability that more than a given number of airplanes will be in queue is smaller than or equal to a threshold value. Contreras, Cordeau, and Laporte (2011b) propose a two-stage stochastic optimization formulation for uncapacitated,
Bayram, Yıldız, and Farham: *The HLPCC*

Article submitted to : manuscript no. (Please, provide the manuscript number!)

multiple allocation HLP and a solution methodology based on Benders decomposition and sample average approximation to solve the problem. Merakh 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, except for the study by Marianov and Serra (2003), these studies do not consider congestion costs and none considers capacity acquisition decisions. They all assume a complete network structure, triangular inequality assumption 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 which are very well-known to be too restrictive. 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 is 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 master problem may be infeasible for the subproblem. To guarantee that 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 $G = (\mathcal{N}, \mathcal{A})$ 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 $G$ can be complete or incomplete. Let $H \subseteq \mathcal{N}$ be the set of potential hub locations. We say a graph $G$ is $\kappa$-connected, for some $\kappa \geq 2$, if for any simple walk $w = \{h_1, \ldots, h_n\}$ in the hub sub-graph $G(H)$ 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 $K$ be the set of commodities corresponding to origin-destination pairs in $\mathcal{N}$. The origin and destination of a commodity $k \in K$ is 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 $A_H \subseteq \mathcal{A}$.

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 $h \in H$ is given by $f^\ell_h$. The capacity associated with a hub size $\ell$ is denoted by $q^\ell$ and the maximum possible capacity for a hub (the capacity of the largest size hub) is shown as $Q$.

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

- $\{n^1_p, \ldots, n^m_p\} \subseteq H$, i.e., all intermediate nodes of a path are hubs and no direct transshipment of commodities are allowed.
- $\{(o_k, n^1_p), (n^m_p, d_k)\} \cup \{(n^i_p, n^{i+1}_p) : i = 1, \ldots, m_p - 1\} \subseteq \mathcal{A}$, 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. Binary variable $y^\ell_h$ indicates whether a hub of size $\ell \in L$ is opened at $h \in H$ 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 \( s \in S \). Variable \( u_h(s) \) represents the total flow on hub \( h \) in scenario \( s \in S \).

The fraction of flow from an origin to a destination on path \( p \) is denoted by continuous variable \( v_p(s) \), \( s \in S \). To model the congestion effect, we use the congestion cost function as in Elhedhli and Wu (2010) inspired by \( M/M/1 \) queues, which is used by majority of the studies in the literature to model congestion effect at hubs. 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 to 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

\[ b_h \frac{u_h}{\sum_{\ell \in L} q^\ell y^\ell_h - u_h + \epsilon}, \quad h \in \mathcal{H}, \quad (1) \]

where \( b_h \) is a scaling factor used to calculate the congestion cost of hub \( h \) and \( \epsilon \) 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.

Minimize \[ \sum_{h \in \mathcal{H}} \sum_{\ell \in L} f^\ell_h y^\ell_h + \sum_{s \in S} \varphi(s) \left( \sum_{h \in \mathcal{H}} b_h \frac{u_h(s)}{\sum_{\ell \in L} q^\ell y^\ell_h - u_h(s) + \epsilon} + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k} c_p v_p(s) \right) \quad (2) \]

subject to:

\[ \sum_{p \in \mathcal{P}_k} v_p(s) = 1 \quad \forall k \in \mathcal{K}, s \in S, \quad (3) \]

\[ \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k \cap \mathcal{P}_h} w_k(s)v_p(s) = u_h(s) \quad \forall h \in \mathcal{H}, s \in S, \quad (4) \]

\[ u_h(s) \leq \sum_{\ell \in L} q^\ell y^\ell_h \quad \forall h \in \mathcal{H}, s \in S, \quad (5) \]

\[ \sum_{\ell \in L} y^\ell_h \leq 1 \quad \forall h \in \mathcal{H}, \quad (6) \]

\[ y^\ell_h \in \{0, 1\} \quad \forall h \in \mathcal{H}, \ell \in L, \quad (7) \]

\[ v_p(s) \geq 0 \quad \forall k \in \mathcal{K}, p \in \mathcal{P}_k, s \in S, \quad (8) \]

\[ u_h(s) \geq 0 \quad \forall h \in \mathcal{H}, s \in S, \quad (9) \]

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 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 \( \epsilon \) may cause numerical issues when one tries to solve it in practice. To address these issues, we make the following modifications in the model.
• We define \( u_\ell^h(s) \) as the total flow in a hub at location \( h \in H \) with capacity \( \ell \in L \).
• We replace the objective (2) with the following convex function.

\[
\text{Minimize} \quad \sum_{h \in H} \sum_{\ell \in L} f_\ell^h y_\ell^h + \sum_{s \in S} \varphi(s) \left( \sum_{h \in H} \sum_{\ell \in L} b_h \frac{u_\ell^h(s)}{q^\ell - u_\ell^h(s)} + \sum_{k \in K} \sum_{p \in P_k} c_p v_p(s) \right)
\]  

(10)

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

\[
\sum_{k \in K} \sum_{p \in P_k,h \in p} w_k(p) v_p(s) = \sum_{\ell \in L} u_\ell^h(s) \quad \forall h \in H, s \in S,
\]

(11)

\[
u_\ell^h(s) \leq q^\ell y_\ell^h \quad \forall h \in H, \ell \in L, s \in S,
\]

(12)

\[
u_\ell^h(s) \geq 0 \quad \forall h \in H, s \in S,
\]

(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 \( r_\ell^h(s) \), for each \( h \in H, s \in S, \ell \in L \) as follows.

\[
r_\ell^h(s) \geq 0 \quad \forall h \in H, s \in S, \ell \in L,
\]

(14)

\[
r_\ell^h(s) \geq \frac{u_\ell^h(s)}{q^\ell - u_\ell^h(s)} \quad \forall h \in H, s \in S, \ell \in L.
\]

(15)

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

\[
(u_\ell^h(s))^2 \leq (q^\ell r_\ell^h(s) - u_\ell^h(s))(q^\ell - u_\ell^h(s)) \quad \forall h \in H, s \in S, \ell \in L.
\]

(16)

Constraint (16) is a hyperbolic inequality of the form \( \zeta^2 \leq \xi_1 \xi_2 \) where \( \zeta, \xi_1, \xi_2 \geq 0 \). The constraint \( \zeta^2 \leq \xi_1 \xi_2 \) can be transformed into the quadratic form \( \| (2\zeta, \xi_1 - \xi_2) \| \leq \xi_1 + \xi_2 \), 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):

\[
\| (2u^f_h(s), q^f r^f_h(s) - q^f) \| \leq q^f r^f_h(s) + q^f - 2u^f_h(s) \quad \forall h \in \mathcal{H}, s \in S, \ell \in L. \tag{17}
\]

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

\[
\text{(HNDC) Minimize } \sum_{h \in \mathcal{H}} \sum_{\ell \in \mathcal{L}} f^f_{h\ell} y^f_{h\ell} + \sum_{s \in S} \varphi(s) \left( \sum_{h \in \mathcal{H}} \sum_{\ell \in \mathcal{L}} b_h r^f_{h\ell}(s) + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k} c_p v_p^k(s) \right) \tag{18}
\]

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

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).

\[
\text{PCSP}(\bar{y}, s): \text{Minimize } \sum_{h \in \mathcal{H}} \sum_{\ell \in \mathcal{L}} b_h r^f_{h\ell}(s) + \sum_{k \in \mathcal{K}} \sum_{p \in \mathcal{P}_k} c_p v_p^k(s) \tag{20}
\]

subject to: \[
\sum_{p \in \mathcal{P}_k} v_p(s) = 1 \quad \forall k \in \mathcal{K},
\]
\[
\sum_{k \in \mathcal{K}, p \in \mathcal{P}_k : h \in p} w_k(s)v_p(s) = \sum_{\ell \in \mathcal{L}} u^\ell_h(s) \quad \forall h \in \mathcal{H},
\]
\[
u^\ell_h(s) \leq \gamma^\ell_h \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L}
\]
\[
2u^\ell_h(s) - t^\ell_h(s) = 0 \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L}
\]
\[
q^\ell r^\ell_h(s) - t^\ell_h(s) = q^\ell \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L}
\]
\[
q^\ell r^\ell_h(s) - 2u^\ell_h(s) - t_h^\ell(s) = -q^\ell \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L},
\]
\[
(t_h^\ell(s))^2 + (t_h^\ell(s))^2 \leq (t_h^\ell(s))^2 \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L},
\]
\[
u^\ell_h(s), t^\ell_h(s) \geq 0 \quad \forall k \in \mathcal{K}, p \in \mathcal{P}_k,
\]

where \( \gamma^\ell_h \) 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 \( \gamma(s), \delta(s), \eta^h(s), \lambda^h(s), \mu^h(s), \) and \( \nu^h(s) \) be the dual variables associated with constraints (21)–(26), respectively for a given scenario subproblem \( s \in \mathcal{S} \). Then, we can formulate the dual conic subproblem (DCSP) as follows.

\[
\text{DCSP}(\gamma, s): \text{Maximize} \quad \sum_{k \in \mathcal{K}} \gamma_k(s) - \sum_{h \in \mathcal{H}, \ell \in \mathcal{L}} q^\ell \left( \eta^\ell_h(s) \gamma^\ell_h - \mu^\ell_h(s) + \nu^\ell_h(s) \right)
\]

subject to:
\[
\gamma_k(s) + \sum_{h \in \mathcal{H}, h : p \in \mathcal{P}_k} w_k(s) \delta_h(s) \leq c_p \quad \forall k \in \mathcal{K}, p \in \mathcal{P}_k,
\]
\[
-\delta_h(s) - \eta^h(s) + 2\lambda^h(s) - 2\mu^h(s) \leq 0 \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L},
\]
\[
q^\ell \mu^\ell_h(s) + q^\ell \nu^\ell_h(s) \leq b_h \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L},
\]
\[
(\lambda^h(s))^2 + (\mu^h(s))^2 \leq (\nu^h(s))^2 \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L},
\]
\[
\eta^\ell_h(s), \nu^\ell_h(s) \geq 0 \quad \forall h \in \mathcal{H}, \ell \in \mathcal{L}.
\]

The DCSP is also a SOCP problem. Note that for any feasible PCSP\((\gamma, s)\), there exists a strictly feasible point. Since this problem is also bounded, the corresponding DCSP\((\gamma, s)\) 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 \( \gamma \). 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 \( \Omega \) to represent congestion and transportation costs of the subproblem. The MP of the HNDC is given below.
MP: Minimize \[ \sum_{h \in H} \sum_{\ell \in L} f_{h}^{\ell} y_{h}^{\ell} + \sum_{s \in S} \Omega(s) \] (36)

s.t.: \[ \Omega(s) \geq \varphi(s) \left\{ \sum_{k \in K} \gamma_{k}^{j} \left( s \right) - \sum_{h \in H} \sum_{\ell \in L} q_{\ell}^{j} \left( s \right) y_{h}^{\ell} \right\} \] \quad \forall s \in S, j \in J (37)
\[ \sum_{\ell \in L} y_{h}^{\ell} \leq 1 \quad \forall h \in H, \] (38)
\[ y_{h}^{\ell} \in \{0, 1\} \quad \forall h \in H, \ell \in L \] (39)
\[ \Omega(s) \geq 0, \quad \forall s \in S \] (40)

where \( \gamma_{k}^{j} \), \( \eta_{h}^{j} \), \( \mu_{h}^{j} \), and \( \nu_{h}^{j} \) are the values of \( \gamma, \eta, \mu \) and \( \nu \) 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 \( \pi \), 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 (Rahmanian 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.
\[ \sum_{h \in H} \sum_{\ell \in L} y_{h}^{\ell} q_{\ell}^{j} \geq \sum_{k \in K} w_{k}(s), s \in S. \] (41)

However, this constraint does not guarantee feasibility. See Figure 1a for such an example. Suppose there are 5 commodities, \( a-h, b-h, c-g, d-g, h-g \), 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 $c-g$, $d-g$ 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 $g$).

![Image](a) An infeasible MP solution for a complete network

![Image](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 Similary, 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 feasibility of the subproblems and solving the master problem every time new feasibility cuts are added until 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 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 \( K_i \) be the set of commodities with destination \( i \in N \) (i.e., \( K_i = \{ k \in K : d(k) = i \} \)), 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 \( h \in H \). Below we present the reformulation of MP:

\[
\text{MPF: Minimize } \quad \text{(36)} \\
\text{subject to: } \quad \text{(37)} - (40) \\
\sum_{i \in N} \tilde{z}_{hi} = \sum_{\ell \in L} q^\ell y^\ell_h \quad \forall h \in H, \tag{42}
\]
\[
\tilde{z}_{hi} \geq 0 \quad \forall h \in H, i \in N, \tag{43}
\]

Note that different than MP, MPF determines the allocations of hub capacities to super commodities with the equalities (42). Next, we show how to take advantage of this lifted formulation to address the infeasibilities in the sub-problems. We first introduce the following definitions to facilitate the technical discussions.

**Definition 1.** A set of hubs \( \bar{H} \subseteq H \) is called a minimal-cut-set of a super commodity \( K_i \) if all the paths of all the commodities \( k \in K_i \) 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 \( i \in N \) is denoted by \( H_i \).

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

The following proposition presents a full characterization of the capacity-feasible solutions for the master problem, when the underlying graph is \( \kappa \)-connected.

**Proposition 1.** Assume \( G \) is a \( \kappa \)-connected. Then, MPF is capacity feasible if and only if the following inequalities hold.

\[
\sum_{h \in \bar{H}} \tilde{z}_{hi} \geq \sum_{k \in K_i} w_k(s) \quad \forall i \in N, \bar{H} \subseteq H_i, s \in S \tag{44}
\]

Proof: Let \((y, z)\) be a solution for the master problem. For each super-commodity \( K_i, i \in N \) and scenario \( s \in S \), we define a feasibility graph \( G^s_i = (N^s_i, A^s_i) \). The node set \( N^s_i \) contains an artificial source node \( \sigma_i \), node \( i \), two copies \( h \) and \( \bar{h} \) for each potential hub location \( h \in H \), and the set of nodes \( N_i = \{ j : w_{ji}(s) > 0 \} \). The arc set \( A^s_i \) contains the following four groups of arcs.
• origin arcs $A_1$: for each $j \in N$, $A_1$ contains the arc $(\sigma_i, j)$ with capacity $\kappa(\sigma_i, j) = w_{ij}$.

• 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 $h \in H$, if $(j, h) \in A$ (i.e., the node $j$ can reach to hub $h$ in the original graph $G$). Arcs in $A_3$ have infinite capacity.

• destination arcs $A_4$: for each candidate hub location $h \in H$, 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 $K_i$, the maximum-flow from $\sigma_i$ to $i$ is equal to $w_i = \sum_{k \in K_i} w_k(s)$, hence the solution is capacity-feasible. Similarly, if a given solution $(y, z)$ is capacity-feasible then the maximum-flow from $\sigma_i$ to $i$ must be equal to $w_i$ for any super-commodity $K_i$ and scenario $s \in S$. Then, by the maxflow-mincut duality theorem for any set of potential hub locations $\bar{H} \subseteq H$ (44) must hold. Since we assume $G$ is $\kappa$-connected, the max-flow solution can be always decomposed into paths that visit no more than $\kappa$ hubs, hence the result follows. □

As a direct result of the Proposition 1, adding (44) to MCF ensures capacity feasible solutions if $G$ is $\kappa$-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 become $\kappa$-connected for even small $\kappa$ values (i.e., $\kappa \geq 3$). Moreover, even for those graphs that does 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 column generation algorithm to solve the sub-problems.

Adding (44) to MPF, we formally present MP\textsubscript{CUT} as follows:

MP\textsubscript{CUT}: Minimize (36)

\[
\text{s.t.: } \sum_{h \in \bar{H}} z_{hi} \geq \sum_{k \in K_i} w_k(s) \quad \forall i \in N, \bar{H} \subseteq H, s \in S, \quad (45)
\]

\[
(37) - (40), (42) - (43)
\]

The number of inequalities in (45) grows exponentially with the number of alternative locations for the hubs ($|H|$). Therefore, for realistic size problems it is not practical to solve MP\textsubscript{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 a 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 \(C_i^s \subset A_i^s\) and consider the hub set \(\bar{H} = \{h \in H : (h, \bar{h}) \in C_i^s\}\) to detect the violated inequality \(\sum_{h \in \bar{H}} \tilde{z}_{hi} \geq \sum_{k \in K_i} w_k(s)\).

5. Solving the Subproblem by Column Generation

The PCSP (20)–(29) assumes that the complete set of candidate paths for each commodity, i.e. \(P_k\), 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 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 \(P_k\) set including an artificial path \(p_k = \{o_k, d_k\}\) for each commodity \(k \in K\) with an arbitrarily large cost. A new path \(p \in P_k\) for commodity \(k \in K\) can be added to the current PCSP if it has negative reduced cost. The reduced cost of a path variable \(v_p(s), s \in S, p \in P_k\), denoted by \(\tilde{c}_p(s)\), in the PCSP is calculated in (46).

\[
\tilde{c}_p(s) = c_p - \gamma_k(s) - \sum_{h \in H: h \in p} w_k \delta_h(s).
\]  

(46)

To price out candidate paths, a pricing subproblem (PP) is defined and solved for each \(k \in K\) 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 \(G_k\) of \(G\), which contains \(o_k, d_k\), and all candidate hub locations \(H\). 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 \(G_k^s\), 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 \(G_k^s\).
• The cost of an outgoing arc \((o_k, h) \in A\) 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 \((h, \bar{h}) \in A\) 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 \((h, d_k) \in A\) 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 \(G_k^s\) 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 negative reduced cost if and only if the shortest (cheapest) path with \(\kappa - 1\) arcs in a pricing graph \(G_k^s\), \(k \in K\), \(s \in S\) 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 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 \(\kappa\) (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 \(k \in K\) 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 8, 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 8, 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.

6.1. Problem Instances and the design of experiments.

We use the Turkish data set (Yaman, Kara, and Tansel 2007), retrieved from Kara (2011), to generate the problem instances we study in our experiments. This 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 the data set. We assume three alternative capacity levels (small, medium, and large) of 10K, 20K, and 30K at each location. The details of the data we use for candidate hub locations are presented in Appendix A. We use both complete (C) and incomplete (IC) networks in our experiments. To derive the incomplete networks, we first calculate a “weight” between any two nodes by multiplying the populations of the origin and destination cities and dividing this value by the distance between them as indicated in the data set. Ordering the origin-destination pairs by their weight, we take the top 33% of those pairs to connect in the IC problem instances.

We consider two scenarios for demand between origin and destination pairs in our problem instances. The first scenario represents the expected flow quantities in typical days, whereas the second scenario represents the high-demand days, such as holidays. For the first scenario, we use the flow values in Kara (2011). 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 $b_h \in \{0, 500, 1000, 1500, 2000\}$. In all our experiments we allow maximum five hub visits for a transfer (i.e., we use $\kappa = 5$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>BC</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDM</td>
<td>1.5, 1, 1.25, 1.75, 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.75</td>
<td>0, 0.25, 0.5, 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>network</td>
<td>IC</td>
<td></td>
<td></td>
<td>C</td>
</tr>
</tbody>
</table>

Empty cells indicate same values as BC.

6.2. Discussion

6.2.1. Algorithmic Efficiency In Table 2, we report the number of feasibility cuts added (nFeasCuts) to ensure that master problem generates solutions feasible to subproblems, number of optimality cuts added (nOptCuts), number of columns generated by the subproblems (nCol.),
number of nodes searched in branch and bound tree not including the root node (nBBnodes), solution time (Sol.T.), and the relative gap (Gap), for different networks (incomplete and complete), multiplier (HDM) for the worst case (high demand) scenario, inter-hub transportation discount factor $\alpha$, 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, 1.04, and 1.5 is used to generate the demand in the single scenario.

Majority (68%) of the scenario-based stochastic instances are solved to optimality. The percentage of deterministic instances solved to optimality is much higher (87%). Instances that cannot be solved to optimality are generally solved with small gaps in the designated time limit, the highest gap being around 18%. 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 $\alpha$ decreases the number of nodes searched in the branch and bound tree, 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 number of potential hubs increase, 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 a 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 number of potential hubs 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. Please notice the large number of feasibility cuts added in all of the instances. The characterization we provide for capacity-feasible solutions allows us to handle infeasibilities in an efficient way, solving a series of maximum flow problems.

We further test the quality of 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 the 20% of the mean. We observe that the percentage absolute value differences between the optimal values of the solutions generated by HNDC and results of the 100 test instances are insignificant (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
Table 2: Performance of the Algorithm

<table>
<thead>
<tr>
<th>Network</th>
<th>HDM</th>
<th>α</th>
<th>b_α</th>
<th>nFeasCuts</th>
<th>nOptCuts</th>
<th>nCol.</th>
<th>nBBnodes</th>
<th>Sol.T.</th>
<th>Gap (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>1</td>
<td>0.75</td>
<td>1</td>
<td>14,296</td>
<td>500</td>
<td>1,246</td>
<td>6,200.44</td>
<td>2,794</td>
<td>470.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,000</td>
<td>22,299</td>
<td>800</td>
<td>1,096</td>
<td>3,335</td>
<td>95,238</td>
<td>7,285.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,500</td>
<td>25,404</td>
<td>1,496</td>
<td>1,142</td>
<td>187,231</td>
<td>14,509.51</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2,000</td>
<td>32,466</td>
<td>2,608</td>
<td>1,469</td>
<td>281,866</td>
<td>25,937.65</td>
<td>0.00</td>
</tr>
<tr>
<td>0.33</td>
<td>1.25</td>
<td>0.75</td>
<td>1</td>
<td>13,380</td>
<td>600</td>
<td>1,075</td>
<td>6,200.44</td>
<td>2,794</td>
<td>470.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,000</td>
<td>22,835</td>
<td>900</td>
<td>1,061</td>
<td>3,335</td>
<td>95,238</td>
<td>7,285.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,500</td>
<td>25,980</td>
<td>1,496</td>
<td>1,142</td>
<td>187,231</td>
<td>14,509.51</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2,000</td>
<td>33,036</td>
<td>2,608</td>
<td>1,469</td>
<td>281,866</td>
<td>25,937.65</td>
<td>0.00</td>
</tr>
<tr>
<td>0.33</td>
<td>1.5</td>
<td>0</td>
<td>1</td>
<td>13,380</td>
<td>600</td>
<td>1,061</td>
<td>6,200.44</td>
<td>2,794</td>
<td>470.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,000</td>
<td>22,835</td>
<td>900</td>
<td>1,061</td>
<td>3,335</td>
<td>95,238</td>
<td>7,285.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1,500</td>
<td>25,980</td>
<td>1,496</td>
<td>1,142</td>
<td>187,231</td>
<td>14,509.51</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2,000</td>
<td>33,036</td>
<td>2,608</td>
<td>1,469</td>
<td>281,866</td>
<td>25,937.65</td>
<td>0.00</td>
</tr>
</tbody>
</table>

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.

![Comparison of HNDC vs. Test Instances](image)

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

6.2.2. Managerial Insights Next, we investigate the effects of congestion costs and demand uncertainty on hub network design and operational decisions. Our results (see Appendix B for details) verify that including congestion costs in hub network design has an effect on the hub capacity levels and for some instances, on the hub network’s topology. 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 the demand during the high times justify opening new hubs to take advantage of the discounts for the inter-hub transfers. Similarly, the change in optimal design, compared to the solution with no congestion (i.e., $b_h = 0$), is observed at smaller levels of discount factor $\alpha$, due to the trade off between achieving economies of scale and incurring congestion costs as a result of consolidating traffic flow.

In Table 3, we compare the deterministic planning for hub network design, where we use a single scenario corresponding to low demand (DLD, HDM = 1), medium demand (DMD, HDM = 1.5) and high demand (DHD, HDM = 2) cases against stochastic planning (SP) for various congestion levels, 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 of the instances where a 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 congestion level, 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. When the worst case scenario (DHD) is employed in deterministic planning, the resulting hub network design is feasible for all congestion levels as expected but at the expense of opening one more hub and using up to 30,000 units of extra capacity compared to the hub network design achieved using stochastic planning.

Table 3: Deterministic vs. Stochastic Planning

<table>
<thead>
<tr>
<th>Method</th>
<th>$b_h$</th>
<th>Open Hubs</th>
<th>Hub Capacities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DLD0</td>
<td>1,000</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,500</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DLD500</td>
<td>500</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DLD1000</td>
<td>1,000</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DLD1500</td>
<td>1,500</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DLD2000</td>
<td>2,000</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DMD0</td>
<td>1,000</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,500</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DMD500</td>
<td>500</td>
<td>INFEASIBLE</td>
<td></td>
</tr>
<tr>
<td>DMD1000</td>
<td>1,000</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>DMD1500</td>
<td>1,500</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>DMD2000</td>
<td>2,000</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 10000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>DHD0</td>
<td>1,000</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 10000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td></td>
<td>1,500</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 10000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 10000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>DHD500</td>
<td>500</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 20000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>DHD1000</td>
<td>1,000</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 20000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>DHD1500</td>
<td>1,500</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 20000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>DHD2000</td>
<td>2,000</td>
<td>[65 1 34 35 6 25]</td>
<td>[10000.0 20000.0 30000.0 30000.0 20000.0]</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td>SP</td>
<td>1,000</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td></td>
<td>1,500</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 10000.0]</td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>[65 1 34 6 25]</td>
<td>[10000.0 30000.0 30000.0 30000.0 20000.0]</td>
</tr>
</tbody>
</table>

Figure 3 illustrates the effect of congestion level $b_h$ on mean number of hubs used on a path (Figure 3a) and routing cost (Figure 3b) for various levels of discount factor $\alpha$ when HDM = 1.5, and on capacity cost for HDM = 1.25 (Figure 3c) and HDM = 1.5 (Figure 3d) when discount
factor $\alpha = 0.75$. We observe that maximum number of hubs used on a path is three and that there is a nonlinear relationship between the congestion level $b_h$ and the mean number of hubs used on a path. As congestion level increases, the model generates solutions with longer paths using bigger number of hubs, hence tolerating bigger routing costs (Figure 3b) to evade congestion costs at the hubs up to a point (congestion level 500 in Figure 3a). After that point, due to higher contribution of congestion costs to overall costs, the model tries to use paths with as small number of 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 mean number of hubs used on a path decreases. Another effect of increasing congestion levels is an increase in routing costs as expected. One also expects to see increasing routing costs as discount factor $\alpha$ increases, i.e., as economies of scale due to consolidation decreases as also illustrated in Figure 3b. However, as congestion level increases, the question of which discount factor is valid becomes unimportant, i.e., the difference between various discount factors disappears. Increasing congestion levels also results in higher capacity levels to hedge against nonlinear increase in congestion costs. When HDM is small (Figure 3c), an increase even in small congestion levels results in higher capacities, whereas when HDM is larger (Figure 3d), congestion level begins to effect 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.

Figure 4 illustrates the effect of discount factor $\alpha$ on mean number of hubs (4a) for various levels of congestion, on congestion cost (4b), and on routing cost (4c) when congestion level $b_h$ is 1000 and HDM = 1.5. Due to decreasing economies of scale advantages between two hubs, as discount factor $\alpha$ increases, mean number of hubs used on a path decreases, as well. There is a trade-off between consolidating flows at hubs to enjoy inter-hub economies of scale advantages, hence decreasing routing costs and increasing congestion costs at the hubs due to consolidation. This trade-off is illustrated by Figure 4b. As discount factor $\alpha$ decreases, consolidating flows at the hubs and inter-hub routing becomes more advantageous at the expense of congestion costs at the hubs. As discount factor increases, the advantage of routing flows between hubs is lost and hence routing costs increase (4c).

We also investigate the hub network topology differences when the underlying graph is complete and incomplete. When we work with incomplete graphs, with a discount factor of 0.75 and a HDM of 1.5, the resulting network design (5a) has Istanbul, Ankara, Adana, Erzurum, and Van cities as hubs for all congestion levels, increasing capacity of Erzurum hub when congestion level increases from 1,500 to 2,000 (5b). 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 smaller number of hubs with bigger capacities concentrating them on the
Figure 3: Effect of congestion level on hub network design and operational costs

west side of Turkey in more populated cities Istanbul, Ankara, Izmir and Bursa (5c). 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 in Turkey.

7. Conclusion

Hubs are consolidation points in distribution networks to provide economics of scale, they are strategically located facilities in multi-modal transportation networks where the travelers/commodities switch the mode of transportation, or they are recharging/refueling stations for alternative fuel vehicles in green logistics. They are also at the center of many disruptive transportation applications/concepts, such as physical internet, crowd-sourced delivery, and mobility as a service. The classical modeling approaches and solution methodologies in hub network design problems cannot
account for widening use of hub networks in various real-world applications and requires extensions in classical approaches.

In this study, we focus on one such crucial extension and develop a new modeling approach and a state-of-the-art solution methodology to address the congestion issues at hubs, which can impair the system performance and introduce significant costs. In particular, we introduce the hub network design problem with congestion, capacity, and stochastic demand considerations, which generalizes the classical hub location problem in several directions. We extend the existing literature by integrating capacity acquisition decisions and the nonlinear congestion cost effect into the problem and allowing dynamic routing for origin-destination pairs. Connecting strategic and operational level decisions, the introduced problem jointly decides hub locations and capacity acquisitions by taking the expected routing and congestion costs into account.
A path-based mixed-integer second-order cone programming formulation of the HNDC is proposed, which allows to work with both complete and incomplete networks and use of more than two hubs on a path. To solve this challenging problem we exploit SOCP duality results and propose an exact algorithm based on Benders decomposition and column generation. We use a specific characterization of the capacity-feasible solutions and develop an efficient branch-and-cut algorithm to solve the master problem. We conduct extensive computational experiments to test the performance of the proposed approach and to derive managerial insights based on realistic problem instances.

In particular, we found that including congestion costs, accounting for the uncertainty in demand and whether the underlying network is complete or incomplete have an effect on the hub capacity levels and on the topology of the hub network and may lead to infeasible solutions, otherwise. As congestion level increases, longer paths using bigger number of hubs are used tolerating bigger routing costs to evade congestion costs at the hubs. But this effect lingers up to a specific point and after that point, due to higher contribution of congestion costs to overall costs, paths with smaller number of hubs are used, and hence mean number of hubs used on a path decreases. There is a trade-off between consolidating flows at hubs resulting in inter-hub economies of scale advantages, hence decreasing routing costs and increasing congestion costs at the hubs due to consolidation. As discount factor decreases, consolidating flows at the hubs and inter-hub routing becomes more advantageous at the expense of congestion costs at the hubs. As discount factor increases, the advantage of routing flows between hubs is lost and hence mean number of hubs used on a path decreases and routing costs increase.

This study can be extended in a variety of ways. We assumed that the demand will remain the same throughout the planning horizon. In reality, demand may be seasonal and may be affected by economical reasons and regional or global health crises such as Covid-19. For that reason, the proposed model can be extended to account for demand change over a planning horizon. One can also, incorporate user perspectives and service quality to generate solutions to satisfy both the network owner and network users.

**Acknowledgement**

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

<table>
<thead>
<tr>
<th>Order</th>
<th>Location ID</th>
<th>ID</th>
<th>Set up cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>small</td>
</tr>
<tr>
<td>1</td>
<td>34</td>
<td>İSTANBUL</td>
<td>2148.65</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>ANKARA</td>
<td>2552.19</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>İZMİR</td>
<td>2236.67</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>ANTALYA</td>
<td>2632.30</td>
</tr>
<tr>
<td>5</td>
<td>55</td>
<td>SAMSUN</td>
<td>3771.96</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>VĄN</td>
<td>3834.03</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>ERZURUM</td>
<td>3560.16</td>
</tr>
<tr>
<td>8</td>
<td>63</td>
<td>ŞANLIURA</td>
<td>4234.20</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>BURSA</td>
<td>2481.89</td>
</tr>
<tr>
<td>10</td>
<td>42</td>
<td>KONYA</td>
<td>3394.79</td>
</tr>
<tr>
<td>11</td>
<td>45</td>
<td>MANISA</td>
<td>3267.60</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>ADANA</td>
<td>339.47</td>
</tr>
<tr>
<td>13</td>
<td>61</td>
<td>TRABZON</td>
<td>3493.77</td>
</tr>
<tr>
<td>14</td>
<td>44</td>
<td>MALATYA</td>
<td>3474.43</td>
</tr>
<tr>
<td>15</td>
<td>27</td>
<td>GAZİANTEP</td>
<td>3469.88</td>
</tr>
<tr>
<td>16</td>
<td>41</td>
<td>KOCAELİ</td>
<td>3498.08</td>
</tr>
<tr>
<td>17</td>
<td>38</td>
<td>KAYSERİ</td>
<td>3016.46</td>
</tr>
<tr>
<td>18</td>
<td>9</td>
<td>AYDIN</td>
<td>3748.45</td>
</tr>
<tr>
<td>19</td>
<td>33</td>
<td>İÇEL</td>
<td>2849.66</td>
</tr>
<tr>
<td>20</td>
<td>52</td>
<td>ORDU</td>
<td>4723.71</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>DİYARBAKIR</td>
<td>3573.94</td>
</tr>
<tr>
<td>22</td>
<td>10</td>
<td>BALIKESİR</td>
<td>2809.01</td>
</tr>
<tr>
<td>23</td>
<td>26</td>
<td>ESKİŞEHİR</td>
<td>2892.70</td>
</tr>
<tr>
<td>24</td>
<td>20</td>
<td>DENİZLİ</td>
<td>2734.51</td>
</tr>
<tr>
<td>25</td>
<td>31</td>
<td>HATAY</td>
<td>3652.91</td>
</tr>
<tr>
<td>26</td>
<td>60</td>
<td>TOKAT</td>
<td>3618.47</td>
</tr>
<tr>
<td>27</td>
<td>23</td>
<td>ELAZIĞİ</td>
<td>3603.05</td>
</tr>
<tr>
<td>28</td>
<td>47</td>
<td>MARDİN</td>
<td>3832.10</td>
</tr>
<tr>
<td>29</td>
<td>59</td>
<td>TEKİRDAĞ</td>
<td>3300.32</td>
</tr>
<tr>
<td>30</td>
<td>58</td>
<td>SİVAS</td>
<td>3762.98</td>
</tr>
<tr>
<td>31</td>
<td>48</td>
<td>MUĞLA</td>
<td>3237.30</td>
</tr>
<tr>
<td>32</td>
<td>46</td>
<td>KAHRAMANMARAŞ</td>
<td>4129.03</td>
</tr>
<tr>
<td>33</td>
<td>67</td>
<td>ZONGULDAK</td>
<td>4105.15</td>
</tr>
<tr>
<td>34</td>
<td>4</td>
<td>AĞRI</td>
<td>4358.18</td>
</tr>
<tr>
<td>35</td>
<td>2</td>
<td>ADIYAMAN</td>
<td>4874.21</td>
</tr>
<tr>
<td>36</td>
<td>54</td>
<td>SAKARYA</td>
<td>4150.17</td>
</tr>
<tr>
<td>37</td>
<td>19</td>
<td>ÇORUM</td>
<td>3612.70</td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>AFYON</td>
<td>3369.05</td>
</tr>
<tr>
<td>39</td>
<td>80</td>
<td>OSMANİYE</td>
<td>4416.95</td>
</tr>
<tr>
<td>40</td>
<td>28</td>
<td>GİRESUN</td>
<td>3904.20</td>
</tr>
</tbody>
</table>

In Table 4 we present the list of hubs we use in our experiments. The first column of the table indicates the order in which we include the hub locations in the candidate hub location lists. The second column indicates the city that the respective hub candidate is to be located. To build this order we first divide the candidates
into seven groups (considering the seven geographical regions of Turkey) and rank them in their group by their populations. Then we list the candidate locations in the table starting from highest ranked cities in each region. We use the unique city codes (indicated in (Kara 2011)) for each candidate location, which is indicated in the third column. Last three columns show the set-up costs for different capacity levels, where the capacities for the small, medium and large configurations are 10K, 20K and 30K, respectively.

Appendix B: Managerial Insights

In Tables 5 and 6, we report the locations of the hubs and their capacities in the optimal solution for different congestion and HDM levels. Note that instances with $b_h = 0$ correspond to designing hub networks without considering the effect of congestion.

Table 5: Planning with different congestion levels I

<table>
<thead>
<tr>
<th>HDM</th>
<th>$b_h$</th>
<th>Open Hubs</th>
<th>Hub Capacities</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>560</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 10000.0 30000.0 30000.0 10000.0</td>
</tr>
<tr>
<td>500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 20000.0 30000.0 30000.0 10000.0</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 20000.0 30000.0 30000.0 10000.0</td>
<td></td>
</tr>
<tr>
<td>1,500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 10000.0</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 10000.0</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>0</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 10000.0</td>
</tr>
<tr>
<td>500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td>0</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
</tr>
<tr>
<td>500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1.75</td>
<td>0</td>
<td>[65 33 34 35 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
</tr>
<tr>
<td>500</td>
<td>[65 33 34 35 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>[65 33 34 35 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,500</td>
<td>[65 33 34 35 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>[65 33 34 35 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
</tr>
<tr>
<td>500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>1,500</td>
<td>[65 1 34 6 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 20000.0</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>[65 33 34 6 16 25]</td>
<td>10000.0 30000.0 30000.0 30000.0 30000.0</td>
<td></td>
</tr>
</tbody>
</table>
Table 6: Planning with different congestion levels II

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( b_\alpha )</th>
<th>Open Hubs</th>
<th>Hub Capacities</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>([65, 34, 35, 6, 25, 27])</td>
<td>([10000.0, 30000.0, 20000.0, 30000.0, 10000.0, 10000.0])</td>
</tr>
<tr>
<td>0.25</td>
<td>0</td>
<td>([65, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>([65, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>([65, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>([65, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>([65, 1, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>([65, 1, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
<tr>
<td>0.75</td>
<td>0.75</td>
<td>([65, 1, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>([65, 1, 34, 6, 25])</td>
<td>([10000.0, 30000.0, 30000.0, 30000.0, 10000.0])</td>
</tr>
</tbody>
</table>

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*.


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


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.


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