General Contraction Method for Uncapacitated Single Allocation p-hub Median Problems

Weibin Dai[1], Jun Zhang[2], Xiaoqian Sun[3]*, Sebastian Wandelt[4]
Beijing Key Laboratory for Network-based Cooperative ATM
School of Electronic and Information Engineering, Beihang University
Beijing, China
daiweibin@buaa.edu.cn[1],buaazhangjun@vip.sina.com[2],sunxq@buaa.edu.cn[3],wandelt@buaa.edu.cn[4]

***Abstract*—Hub location problems have been studied by researchers for three decades, yet, most algorithms do not perform well for large-scale networks because of their high computational complexities. Methods that scale up to large networks are usually tailored specifically towards a particular hub location problem instance and cannot be adapted easily to other problems without significant efforts.**

In this paper, we propose a General Contraction Method (GCM), which explores and exploits the idea of efficiently computing hub locations on a reduced network instance, so-called contracted network, and then rewriting the obtained solution back to the original network. If the contracted network preserves major flows and spatial properties, it can be used as a boilerplate for finding good solutions to the original network. In order to evaluate the performance of the contraction methods, three commonly-used datasets (CAB, TR and AP) are used as case studies. We find that GCM provides high-quality initial solutions within a few seconds even for very large-scale problems. GCM can be combined with specifically tailored solution techniques/heuristics and better solutions can be provided within much shorter time further. Moreover, we show that by varying the size of the contracted network, we can nicely explore a fine trade-off between highly efficient computation and close-to-optimal solutions. We believe that GCM can be adapted to many different types of hub location problems, and thus, our work contributes towards the development of scalable transportation network design.

***Index Terms*—Hub location problems, Contraction, Scalability**

I. Introduction

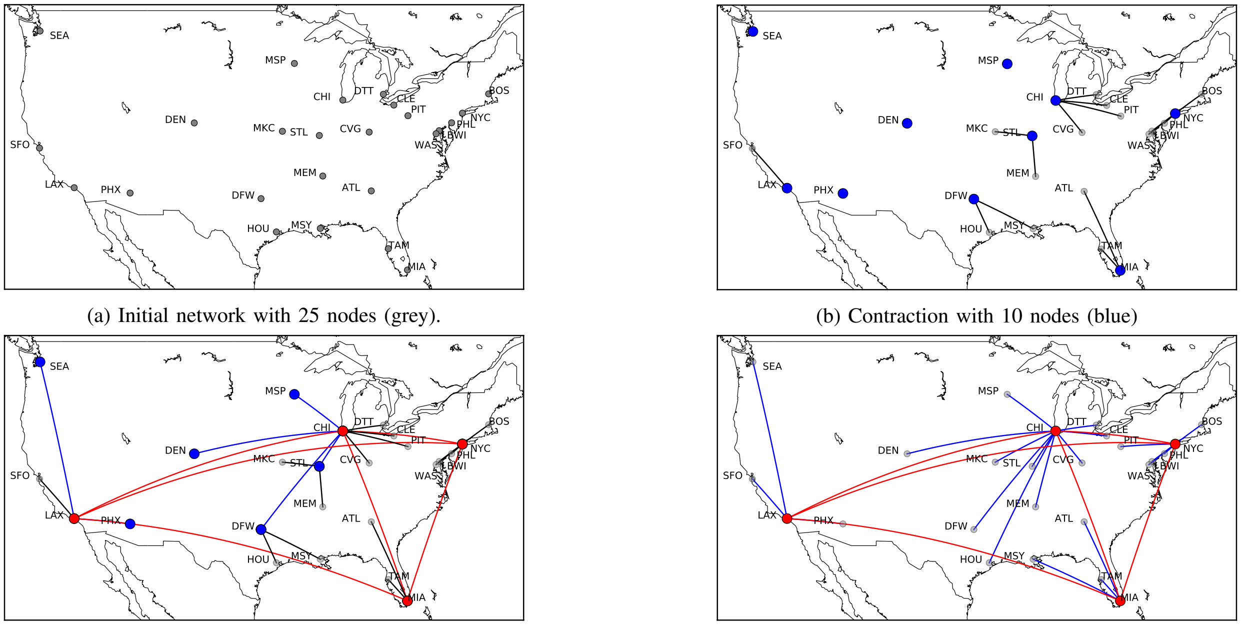
Hub location problems involve the optimal location of hub facilities in a network [1]. They are successfully applied in several fields, including transportation [2]–[5] and telecommunication [6], [7]. In many standard hub location problems, flows are collected from origin nodes by their hubs first; then, flows are transferred to hubs of destination nodes through the hub network; and finally, flows are distributed to their destinations. Economies of scale provide cost discounts for the transportation between hubs, and thus, provide a strong incentive for the research of hub location problems [8], [9]. Since the seminal work by [10], hub location problems have been studied intensively for three decades. A number of solution techniques have been proposed, such as Lagrangian relaxation [11], Benders decomposition [12], variable neighborhood search [13], branch-and-price [14], branch-and-cut [15], clustering-based methods [16] and genetic algorithms [17]. Apart from few exceptions, hub location problems are NP-hard [8], [18]. Therefore, most algorithms do not perform well on large networks. On a network with hundreds of nodes, state-of-the-art techniques usually require hours of run time, without the guarantee of finding the optimal solutions. Although some algorithms may provide good results for a specific problem, the adaption to other problems is often onerous and requires a significant amount of work.

In this paper, we propose General Contraction Method (GCM), a novel and general heuristic which aims at computing solutions for hub location problems in very large networks within short computation time. We propose to transform the input network, for which we are seeking an assignment, into a smaller network with similar topological and flow properties. We refer to this process as *contraction*: Nodes in a network are merged into representative nodes, preserving the structure and flow demands. Solving the hub location problem on the smaller network is significantly faster than for the original network, and the running time and solution quality can be controlled by a single parameter nicely: The size of the contracted network. We show that the solution for the smaller network can be easily rewritten to a solution for the original network. Our experiments on real-world datasets reveal that the solutions obtained are highly competitive with state-of-the-art techniques. The major steps of GCM are visualized in Fig. 1 for the CAB dataset. Although GCM is used to solve uncapacitated single allocation p-hub median problems (USApHMP), the algorithm can be easily adapted to many different types of hub location problems.

The remainder of this paper is organized as follows. We provide a literature review on hub location problems in Section II. The formulation of uncapacitated single allocation p-hub median problem is provided in Section III. The rationale and process of GCM are proposed in Section IV. To evaluate the performance of the contraction methods, the CAB dataset, the TR dataset and the AP dataset are used as case studies in Section V. The paper concludes with Section VI.

II. Literature Overview

Hub location problems were introduced by [10], together with his first mathematical formulation for the p-hub median problem (pHMP) [1]. Four fundamental hub location problems (p-hub median, uncapacitated hub location, p-hub center and



(a) Initial network with 25 nodes (grey).

(b) Contraction with 10 nodes (blue)

(c) The optimal assignment of hubs links (red) and spoke links (blue).

(d) Contracted solution rewritten on the original network

Fig. 1: The process of the General Contraction Method (GCM) on the USApHMP ($\alpha = 0.3, p = 4$) for the CAB dataset: The original network with 25 nodes (a) is contracted into a smaller network with only 10 nodes are selected (b). The USApHMP is solved optimally in the contracted network (c) and the results for the contracted network are rewritten on the original network (d). The computation time for original network with 25 nodes is 2.8 seconds; while GCM needs less than 0.1 seconds.

hub covering) were defined by [19] and the corresponding formulations were also proposed. Since then, hub location problems have been studied by many researchers and a number of methods have been proposed. Standard methods deal with hub location problems mostly by solving integer programmings and they perform well in particular size of problems. However, the computation time grows fast with larger networks. Some researchers found that the assignment for nodes can be easily encoded into chromosomes in genetic algorithm. The near-optimal solutions could also be accepted if they could be obtained within a short computation time. The capacitated case was solved by [20]. In addition, reference [13] proposed a new general variable neighborhood search method that searches good solutions in three types of neighborhoods. Their algorithms were very efficient and could solve uncapacitated single allocation p-hub median problems with up to 1000 nodes and 20 hubs within hundreds of seconds. Reference [21] proposed a two-stage method to simplify the complexities of hub location problems by solving a p-median problem before the original problem. Reference [16] proposed a method that, based on the spatial and flow properties of nodes, generates a clustering-based potential hub set, which can help narrow the solution set and reduce the computational complexity of the problem. [22] proposed new models considering the reliability of hubs for single and multiple allocation p-hub median problems, given that air transportation networks are often vulnerable to disruptions [23],

III. Problem Formulations

In this section, we show the formulations of uncapacitated single allocation p-hub median problems (USApHMP).

In the USApHMP, each pair of hubs are connected with a link and each node is allocated to a single hub. Let $G = (V, E)$ be a network. Here V and E are the set of nodes and links between nodes, separately. The number of nodes and hubs are n and p. For each pair of nodes *(i,j)*, let c_{ij} and w_{ij} be the cost and flow between them. Let $O_i = \sum_{j \in V} w_{ij}$ be the total flows from the source node i and $D_i = \sum_{j \in V} w_{ji}$ be the total flows to the destination node i [24]. The USApHMP is formulated as follows [25]:

$$\min \sum_{i \in V} \sum_{k \in V} c_{ik} Y_{ik} (\delta_1 O_i + \delta_2 D_i) + \sum_{i \in V} \sum_{k \in V} \sum_{m \in V} \alpha c_{km} X^i_{km} \quad (1)$$

$$\text{subject to} \sum_{k \in V} Y_{ik} = 1, \forall i \in V \quad (2)$$

$$Y_{ik} \leq Y_{kk}, \forall i,k \in V \quad (3)$$

$$\sum_{m \in V, m \neq k} X^i_{km} - \sum_{m \in V, m \neq k} X^i_{mk} = O_i Y_{ik} - \sum_{j \in V} w_{ij} Y_{jk}, \forall i,k \in V \quad (4)$$

$$\sum_{k \in V} Y_{kk} = p \quad (5)$$

$$Y_{ik} \in \{0, 1\}, \forall i, k \in V \quad (6)$$

$$X^i_{km} \geq 0, \forall i, k, m \in V \quad (7)$$

Here, the objective function (1) is divided into two terms: The first term is the costs of flows between hubs and non-hub nodes; the second term is the costs of flows through the hub network. Parameters $\alpha < 1, \delta_1 > \alpha, \delta_2 > \alpha$ are the cost coefficients for transporting flows between hubs, from spoke nodes to hubs and from hubs to spoke nodes, respectively. Variables X^i_{km} represent the flows routed on hub link *(k,m)* originating from node *i*. Binary decision variables Y_{ik} take the value of 1, if node *i* is assigned to hub *k* and 0 otherwise.

Equation (2) and Equation (3) ensure that each node is assigned to exactly one hub. Equation (5) shows the number of hubs *p*. Equation (4) ensures the flow equilibrium from each node *i* for each hub *k*.

IV. GCM: Solving hub location problems by network contraction

Given models and formulations from Section III, many algorithms have been proposed to solve hub location problems in the past. Most algorithms are designed and tuned to obtain close-to-optimal solutions on small network instances. For larger networks, however, existing solution techniques either need long computation time or provide solutions with low quality. Moreover, scalable solution techniques are tailored towards a specific problem instance, making it not directly usable for another; for instance, the highly-efficient variable neighborhood search for the uncapacitated single allocation p-hub median problems [13]: Incrementally adapting the flows throughout the exploration of neighborhood solutions is a difficult task for many hub location problems.

In our study, we propose GCM, which first contracts an input network, then solves the hub location problem in the contracted network, and finally rewrites the solution back to the original network. The rationale of GCM is further described in Section IV-A. Four types of strategies for contraction are proposed in Section IV-B. Finally, a strategy for rewriting based on neighborhood search is presented in Section IV-C.

A. Rationale for GCM

With the analysis of the optimal solutions for a number of hub location problems, several observations for the hub selection and spoke allocation can be obtained, as extracted from [16]: The hubs and their spokes are often distributed in clusters. These distributions would not be changed with different cost coefficients. In addition, those nodes with larger flows and closer to the spatial centers of clusters are more likely to be selected as hubs in the optimal solutions. With these insights, contraction methods are proposed to reduce the computation time for large-scale problems. The major steps of GCM are visualized in Fig. 1 in Section I.

1) **Contraction:** For a given number *k*, we define a contraction function *f:* $V \to V$ on the set of nodes V such that $|f(V)| = k$. Each node $i \in V$ is mapped to a node $s \in V$. Let V^* be the image of function f:

$$V^* = f(V)$$

The elements in set V^* are called contraction nodes. Then, a contracted network is constructed with these contraction nodes. For a contraction node *s*, its flow in the contracted network is the sum flow of nodes that are mapped to node *s*. Thus, the flow from contraction node *r* to contraction node *s* after contraction is calculated as follows:

$$w^*_{rs} = \sum_{i \in f^{-1}(r), j \in f^{-1}(s)} w_{ij}, \forall r, s \in V^*$$

where $f^{-1}(s) = \{j \in V : f(j) = s\}$ is the inverse of *f*.

2) **Solving the contracted network:** We define the cost for any pair of contraction nodes *(r,s)*:

$$c^*_{rs} = c_{rs}, \ \forall r, s \in V^*, r \neq s$$

Then the contracted hub location problems are formulated by replacing *c,w* and *V* with c^*, w^* and V^* in Equations (1–7). We solve the hub location problems for the contracted networks using an existing algorithm and obtain the optimal solution for the contracted network. From the solution, the hub set and spoke assignment for the contracted network are obtained.

3) **Rewriting:** Based on the solutions of the contracted network, we reassign the remaining nodes in $V \backslash V^*$ for the original network. The solution for the original problem is obtained afterwards.

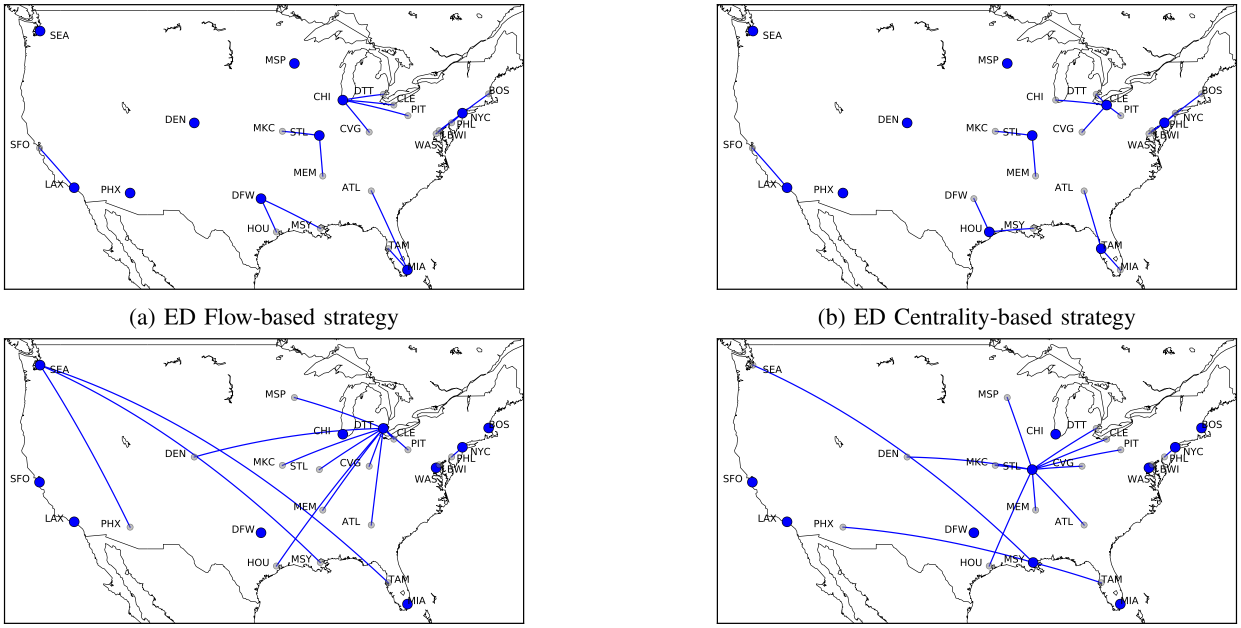
B. Strategies for contraction

For a given *k*, we first generate *k* cluster, using agglomerative clustering [26]. Nodes are distributed in these clusters depending on the distance among each pair of them. Here, the distance could have different definitions. In this paper, in addition to the Euclidean Distance (**ED**), we define a new distance $\hat{c_{ij}}$ with the consideration of flows between nodes (**FD**) as follows: $\hat{c_{ij}} = c_{ij} * w_{ij}$.

Note that if the flow between two nodes is small, these two nodes are closer to each other by FD. They would be more likely to be in one cluster. Therefore, most large flows between nodes in the original network would be transformed into flows between clusters, but not between nodes in the same cluster in this case.

For each cluster, one contraction node is selected based on the properties of nodes and the remaining nodes in the cluster are mapped to this contraction node. The contracted network is constructed then. Two strategies for selecting contraction nodes with the consideration of flows and spatial positions are introduced as follows. Four contraction strategies (ED flow-based, ED centrality-based, FD flow-based, FD centrality-based strategies) are visualized in Fig. 2.

Flow-based strategy: In this strategy, we compute the flow $F_i = O_i + D_i$ for each node *i*. The node with the largest flow



(a) ED Flow-based strategy

(b) ED Centrality-based strategy

(c) FD Flow-based strategy

(d) FD Centrality-based strategy

Fig. 2: Visualization of four contraction strategies in the CAB dataset: (a,b) Clusters are generated with the Euclidean distance (ED); (c,d) Clusters are generated with a new distance based on flows (FD). Note that those nodes with small flows between them are more likely to be in the same cluster in FD, although they might be far away from each other by the Euclidean distance.

is selected to be the contraction node for each cluster and the remaining nodes in the cluster are mapped to this node. The visualization of 10 clusters and contraction nodes with largest flows in the CAB dataset is shown in Fig. 2(a,c). Because the spatial properties of nodes are not considered in this strategy, the selected contraction nodes could be far away from the spatial centres of clusters.

Centrality-based strategy: We assume that *Cl* is the set of nodes in each cluster. For node $i \in Cl$, its centrality is computed with $Ce_i = 1/\sum_{j \in Cl, j \neq i} c_{ij}$. If a node is closer to the centre of the cluster, the average distance between this node and other nodes in the cluster would be shorter. Thus, the node with the highest centrality *Ce* is selected as contraction node with this strategy. The visualization of 10 clusters and contraction nodes with the largest centrality in the CAB dataset is shown in Fig. 2(b,d).

C. Rewriting strategy based on neighborhood search

After solving problems in the contracted network using an existing algorithm, the hub set *H* and spoke assignment *ASSI* for the contracted network are obtained. In order to rewrite the solution to the original network, each remaining spoke is allocated to its closest hub initially. Then, a strategy based on neighborhood search is presented to find the best assignment. This strategy does not change the locations of hubs. For each spoke node, we try to change its allocation. Assume that the hub of spoke i is swapped from hub k^o to hub k^c. Let $Win[i][k]$ and $Wout[i][k]$ be the flow originating from and targeting to node *i* through hub *k*, separately. The change of total cost is shown as follows:

$$\begin{aligned} dC = & \delta_1 (c_{ik^c} - c_{ik^o}) w_{ii} + \delta_2 (c_{k^c i} - c_{k^o i}) D_i \\ & + \delta_2 (c_{k^c i} - c_{k^o i}) w_{ii} + \delta_1 (c_{ik^c} - c_{ik^o}) O_i \\ & + \sum_{k \in H} \alpha \left[(c_{kk^c} - c_{kk^o}) Win[i][k] + (c_{k^c k} - c_{k^o k}) Wout[i][k] \right] \end{aligned}$$

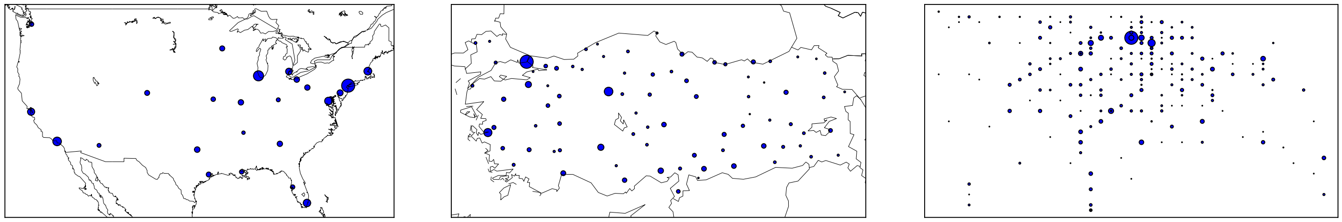
In each iteration, we do the operation for one spoke node and update $Win[i][k]$ and $Wout[i][k]$ in *O(n)* if the current solution is improved. The program is terminated if the solution cannot be improved within a certain number of consecutive iterations.

V. Evaluation

In this section, we report the results of our experiments for GCM. In Section V-A, three datasets used in our study are introduced. The experiment setups are also provided in this section. In Section V-B, we compare the performance of four contraction strategies of GCM. The evaluations of GCM compared with state-of-the-art methods are presented in Section V-C. The visualization of the results is presented in Section V-D.

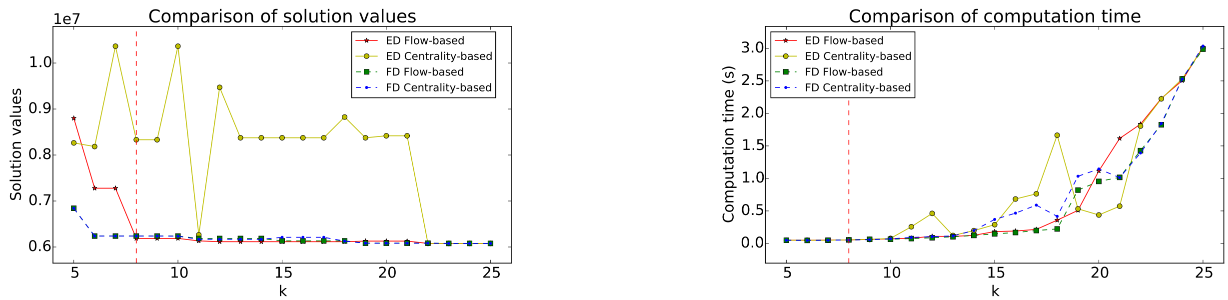
A. Datasets and experimental setup

To evaluate the performance of GCM proposed in Section IV, three well-known datasets are used as case studies. The CAB (Civil Aeronautics Board) dataset that is based on the airline passenger interactions between cities in the



(a) CAB dataset with $n = 25$ (b) TR dataset with $n = 81$ (c) AP dataset with $n = 200$

Fig. 3: Visualization of the three datasets with flows of nodes: CAB dataset with 25 nodes, TR dataset with 81 nodes and AP dataset with 200 nodes. The size of each node is proportional to its flow ($F_i = O_i + D_i$).



(a) Solution values for the USApHMP (b) Computation time for the USApHMP

Fig. 4: The comparison of four contraction strategies for solving USApHMP with $\alpha = 0.3, p = 4$ in the CAB dataset.

United States in 1970 [1]. The TR (Turkish Postal) dataset includes 81 cities with distance and flows between each other in Turkish postal system [27]. The AP (Australia Post) dataset provides 200 postcode districts with coordinates and flows in Australia [25]. The visualization of these datasets with flows of nodes is shown in Fig. 3. In order to show the travel demand of nodes, the size of each node is proportional to its flow ($F_i = O_i + D_i$). There are self-flows for nodes in the AP dataset, i.e., the flows from nodes to themselves. We set these flows to zero. The cost coefficients of hub location problems are set to $\alpha \in \{0.3, 0.5, 0.7\}, \delta_1 = \delta_2 = 1$ for the TR dataset and CAB dataset, and $\alpha = 0.75, \delta_1 = 3, \delta_2 = 2$ for the AP dataset.

We solve hub location problems for contracted networks with the commonly-used programming solver CPLEX in Section V-B and Section V-C. Because CPLEX uses multiple threads as default, we use a single thread for all experiments to obtain comparable results. The maximum computation time is set to 2 hours. All experiments were executed on a Dell laptop with Intel Core i5-4310M processor and 16 GB RAM, running Fedora 24.

B. Comparison of different contraction strategies

As presented in Section IV-B, there are four strategies (ED flow-based strategy, ED centrality-based strategy, FD flow-based strategy and FD centrality-based strategy) for contraction. In order to evaluate their performance, the CAB dataset is used as a case study. The problems with $k \in \{p+1, p+2, ..., 25\}$ are solved here. The results for the USApHMP with $\alpha = 0.3, p = 4$ are shown in Fig. 4.

As shown in Fig. 4(a), the solution values obtained by four contraction strategies with different *k* are compared. The performance of two FD-based strategies (green squares and blue points) is better than ED-based ones (red stars and yellow circles). In addition, the solutions of FD Flow-based strategy are slightly better than that of FD Centrality-based strategy. The difference of flows between nodes in the CAB dataset is quite large. Several important nodes have much larger flows than other nodes. Thus, it is a good strategy to select those nodes with large flows as hubs. In addition to the solution values, the computation time for four strategies is shown in Fig. 4(c). It indicates that their computation time is close to each other for most values of *k*, because their computational complexities are similar to each other.

Finally, based on the results obtained in this section, we select the FD Flow-based strategy in the following evaluations because the flows of nodes are distributed quite unevenly in these three datasets. The contraction size *k* plays an important role on the qualities of final solutions and both *n* and *p* should be taken into account. A set including three values of *k* is generated as follows: $L_k = \left\{2p, 2\sqrt{n}, \frac{n}{2}\right\}$.

In Section V-C, we report the results of GCM with three values of *k*.

C. Comparison between state-of-the-art methods and GCM

In this section, the properties of GCM for solving USApHMPs in the TR dataset with $\alpha \in \{0.3, 0.5, 0.7\}, p \in \{4, 6, 8, 16\}$ and the AP dataset with $\alpha = 0.75, p \in \{4, 6, 8, 10, 20, 40\}$ are evaluated. As mentioned in Section IV-A, the contracted problems in GCM can be solved with

TABLE I: Gaps of solutions and computation time for solving the USApHMPs. The symbol "-" represents that no acceptable solution is obtained within 2 hours. The symbol "*" shows that GCM cannot be used if k is less than p.

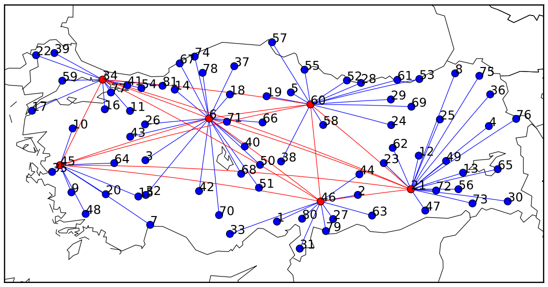
Datasets	n	α	p	Results	GCM_CPLEX with k=2p	GCM_CPLEX with $k=2\sqrt{n}$	GCM_CPLEX with $k=\frac{n}{2}$	CPLEX	GCM_GVNS with k=2p	GCM_GVNS with $k=2\sqrt{n}$	GCM_GVNS with $k=\frac{n}{2}$	GVNS	RCBS
TR	81	0.3	4	gaps	6.88%	1.71%	1.43%	0.11%	6.65%	1.71%	0.10%	0.00%	1.18%
				time (s)	0.45	0.44	20.75	7204.43	0.24	0.21	0.32	0.70	47.12
			6	gaps	15.46%	4.67%	1.63%	0.00%	4.71%	4.67%	1.25%	0.00%	6.75%
				time (s)	0.49	1.05	44.70	6233.72	0.28	0.27	0.41	1.01	82.69
			8	gaps	6.92%	6.59%	1.82%	0.00%	4.60%	4.21%	0.50%	0.05%	3.42%
				time (s)	0.65	1.00	36.54	7205.10	0.31	0.37	0.48	1.96	255.62
			16	gaps	4.71%	14.66%	3.38%	0.00%	1.30%	2.95%	1.24%	0.18%	2.61%
				time (s)	3.44	0.76	10.13	908.68	0.77	0.54	0.81	2.76	6213.79
		0.5	4	gaps	4.80%	4.33%	0.06%	0.00%	4.80%	2.33%	0.49%	0.79%	2.35%
				time (s)	0.24	1.13	57.29	6529.70	0.21	0.26	0.27	0.76	270.14
			6	gaps	8.57%	4.69%	1.52%	0.00%	5.36%	4.19%	0.93%	0.83%	5.26%
				time (s)	0.49	2.05	92.68	7204.21	0.38	0.32	0.39	1.02	883.96
			8	gaps	3.95%	3.21%	0.00%	0.15%	2.15%	2.57%	0.51%	0.40%	2.68%
				time (s)	0.78	1.30	72.47	7204.46	0.40	0.31	0.57	1.45	950.94
			16	gaps	4.37%	6.78%	4.08%	0.00%	1.51%	2.74%	1.26%	0.00%	2.31%
				time (s)	4.50	1.00	15.84	2904.95	0.77	0.54	1.21	4.12	7204.09
		0.7	4	gaps	10.08%	5.13%	1.91%	0.00%	5.11%	2.17%	0.73%	0.37%	3.13%
				time (s)	0.35	2.44	144.52	7201.14	0.21	0.19	0.30	0.61	925.25
			6	gaps	5.77%	5.19%	1.99%	0.22%	4.60%	6.39%	2.34%	0.00%	12.17%
				time (s)	0.39	2.04	296.41	7204.18	0.53	0.29	0.50	0.85	684.96
			8	gaps	5.89%	3.55%	3.80%	0.89%	4.04%	3.37%	0.67%	0.00%	3.44%
				time (s)	1.27	1.47	150.37	7204.15	0.49	0.31	0.61	1.24	1635.38
			16	gaps	2.37%	9.44%	3.20%	0.02%	1.98%	2.76%	1.91%	0.00%	-
				time (s)	5.08	0.96	19.39	7206.82	0.97	0.61	1.02	3.68	-
AP	200	0.75	4	gaps	6.04%	1.36%	-	-	1.34%	1.36%	0.69%	0.00%	5.57%
				time (s)	1.25	4.90	-	-	1.54	0.93	1.96	7.66	1547.07
			6	gaps	5.31%	0.79%	-	-	4.53%	0.79%	0.78%	0.00%	4.57%
				time (s)	1.43	5.56	-	-	1.19	1.57	2.58	9.18	7298.88
			8	gaps	7.84%	2.94%	-	-	5.02%	1.91%	1.16%	0.00%	5.67%
				time (s)	2.12	5.21	-	-	2.50	1.84	3.83	16.65	7297.64
			10	gaps	6.58%	6.54%	-	-	3.95%	1.71%	0.65%	0.00%	5.81%
				time (s)	3.37	3.77	-	-	2.00	2.81	3.48	10.55	7272.42
			20	gaps	7.11%	10.32%	-	-	2.02%	2.04%	1.04%	0.00%	3.58%
				time (s)	9.30	12.93	-	-	4.84	4.75	11.22	34.41	7298.47
			40	gaps	1.84%	*	-	-	1.50%	*	0.91%	0.00%	-
				time (s)	847.84	*	-	-	19.43	*	24.82	116.86	-

any existing techniques. It is shown that solving hub location problems with CPLEX directly is not a good strategy [20], [28], [29]. However, in order to evaluate the performance of the contraction methods completely, CPLEX is used here. The results obtained by solving hub location problems directly with CPLEX are also reported as benchmarks. General variable neighborhood search method (GVNS) [13] performs well for solving large-scale USApHMPs. We have implemented it as a competitor, and we also use it to solve the contracted problems for GCM. Finally, a heuristic called restricted clustering-based potential hub set method (RCBS) proposed by [16] is implemented. Similarly with GCM, this algorithm can be applied to different types of hub location problems. The performance of GCM will be compared with it.

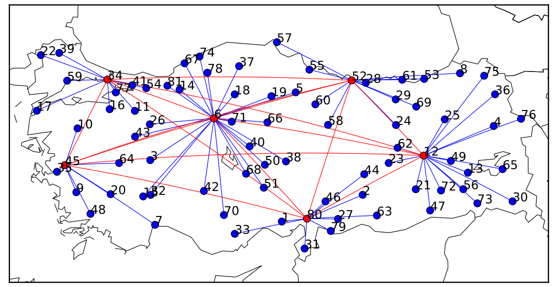
The gaps [30] of solutions (the difference between the obtained solutions and the optimal solutions) and computation time for solving the USApHMPs in the TR dataset and AP dataset with GCM and other methods are presented in Table I. As discussed in Section V-B, three particular values of $k \in \{2p, 2\sqrt{n}, \frac{n}{2}\}$ are used for the GCM. The results obtained by solving contracted problems with CPLEX and GVNS are provided in columns 6–8 and columns 10–12.

It indicates that CPLEX provides good solutions ($gap < 1\%$) within 2 hours for the TR dataset because of low computational complexities. However, we cannot obtain any acceptable solutions for the AP dataset with CPLEX directly. In terms of solution gaps, GCM_CPLEX and GCM_GVNS have similar quality with contraction size k. With increasing values of k, both of them provide better solutions with smaller gaps. However, because of the low efficiency of CPLEX on large-scale problems, the computation time of GCM_CPLEX increases much faster than that of GCM_GVNS. For instance, the former needs hundreds of seconds while the latter needs less than 5 seconds when $k = \frac{n}{2}$ for the AP dataset.

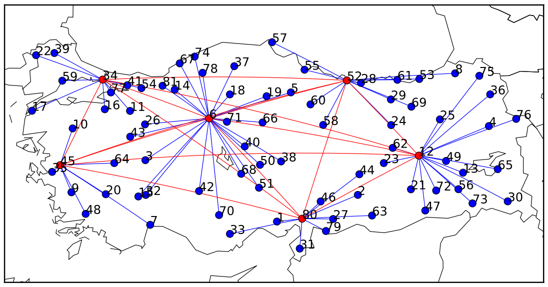
In the twelfth column, GCM_GVNS with $k = \frac{n}{2}$ provides solutions with gaps less than 1% in most cases. Compared with the results in the thirteenth column, it reduces more than a half of computation time and provides close solutions. With increasing sizes of networks and hub sets, the reduction of computation time becomes more significant. The speed of GCM can be faster further if it is combined with other better algorithms. In addition, GCM can be applied to many types of



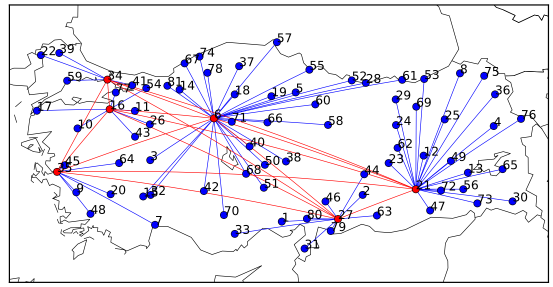
(a) GCM_GVNS with $k = \frac{n}{2}$ (31,305,361,942; 0.41s)



(b) CPLEX (30,917,872,699; 6233.72s)

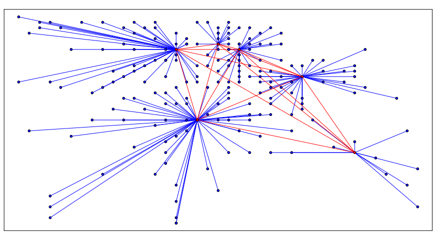


(c) GVNS (30,917,872,699; 1.01s)

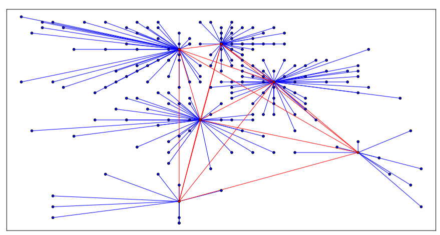


(d) RCBS (33,003,293,094; 82.69s)

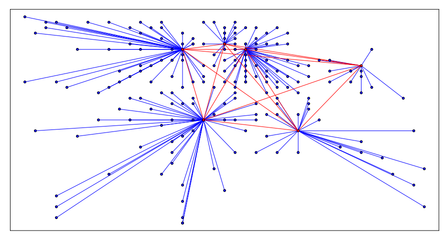
Fig. 5: The visualization of solutions for the USApHMPs with $\alpha = 0.3, p = 6$ in the TR dataset. The hub nodes and links between hubs are represented by red dots and red lines. The numbers at the end of the caption in each sub-figure represent the solution values and corresponding computation time.



(a) GCM_GVNS with $k = \frac{n}{2}$ (132,830; 2.58s)



(b) GVNS (131,797; 9.18s)



(c) RCBS (137,820; 7298.88s)

Fig. 6: The visualization of solutions for the USApHMPs with $\alpha = 0.75, p = 6$ in the AP dataset. The hub nodes and links between hubs are represented by red dots and red lines. The numbers at the end of the caption in each sub-figure represent the solution values and corresponding computation time.

hub location problems, while GVNS can only be used to solve several particular hub location problems, such as USApHMPs.

D. The visualization of solutions and further discussion

In the sections above, we have analyzed the properties of the contraction methods by comparing the solution qualities and computation time. In this section, the solutions for the USApHMPs in the TR dataset and the AP dataset are visualized in Fig. 5 and Fig. 6, respectively. At the end of the caption in each sub-figure, there are two numbers that represent the solution value and computation time. The hub nodes and links between hubs are represented by red points and red lines. Here, the results of GCM_GVNS with $k = \frac{n}{2}$ are reported as the representative of GCM because of its good performance in Section V-C.

The visualization of results for the case of $\alpha = 0.3, p = 6$ for the TR dataset is presented in Fig. 5. CPLEX (5(b)) and GVNS (5(c)) provide the same allocation here. Note that two nodes are selected as hubs in all the cases, i.e., node 6 and node 34. They are Ankara and Istanbul (the capital and the largest city in Turkey), respectively. Large flows enable them to be selected as hub nodes. The selection of FD Flow-based strategy for contraction is identified again. The results with $\alpha = 0.75, p = 6$ for the AP dataset are visualized in Fig. 6.

VI. Conclusions

In this paper, we proposed GCM, which explores and exploits the idea of efficiently solving the large-scale hub location problem in a contracted network, and then rewriting the solution back to the original network. The USApHMPs in the CAB dataset were solved to compare the properties of different contraction strategies for GCM. In order to evaluate the performance of GCM, the TR dataset and the AP dataset were used as case studies. Several state-of-the-art solution techniques (GVNS, RCBS and CPLEX) were also implemented in order to evaluate the effectiveness and efficiency of GCM. CPLEX and GVNS were used to solve the contracted problems for GCM. There are several major observations obtained from our evaluation. First, the FD Flow-based strategy is the best contraction strategy for networks with unevenly-distributed flows of nodes (like CAB, TR, and AP) for GCM. The numbers of nodes (n) and hubs (p) should be considered for selecting an appropriate contraction size (k).

Second, GCM can be used directly by solving the contracted problems with CPLEX for newly designed problems without efficient solution heuristics. If an appropriate method for the problem is known (like GVNS for USApHMP), the computation time could be reduced further with small solution gaps by solving the contracted network with this method. On the other hand, better solutions could be obtained by setting larger contraction size because of the reduced computation time. This further highlights the generality of GCM.

In our study, we used GCM to solve uncapacitated single allocation p-hub median problems. We believe that the idea of contraction can be adapted to many other hub location problems (such as r-allocation p-hub median problems [31], p-hub center problems [32] and capacitated hub location problems [33]) and other types of location problems [34]. Our methodology contributes towards the development of easily scalable solution techniques.

Acknowledgment

This study is supported by the Research Fund from National Natural Science Foundation of China (Grant Nos. 61650110516,. 61601013, 61521091, and 91538204).

References

[1] M. E. O'Kelly, "A quadratic integer program for the location of interacting hub facilities," *European Journal of Operational Research*, vol. 32, no. 3, pp. 393–404, 1987.

[2] S. Gelareh and S. Nickel, "Hub location problems in transportation networks," *Transportation Research Part E: Logistics and Transportation Review*, vol. 47, no. 6, pp. 1092–1111, 2011.

[3] X. Sun, S. Wandelt, M. Hansen, and A. Li, "Multiple airport regions based on inter-airport temporal distances," *Transportation Research Part E: Logistics and Transportation Review*, vol. 101, pp. 84–98, 2017.

[4] X. Sun, S. Wandelt, and X. Cao, "On node criticality in air transportation networks," *Networks and Spatial Economics*, pp. 1–25, 2017.

[5] X. Sun, S. Wandelt, and F. Linke, "On the topology of air navigation route systems," in *Proceedings of the Institution of Civil Engineers-Transport*, vol. 170, no. 1, 2017, pp. 46–59.

[6] H. Yaman and G. Carello, "Solving the hub location problem with modular link capacities," *Computers & Operations Research*, vol. 32, no. 12, pp. 3227–3245, 2005.

[7] H. Kim and M. E. O'Kelly, "Reliable p-hub location problems in telecommunication networks," *Geographical Analysis*, vol. 41, no. 3, pp. 283–306, 2009.

[8] J. F. Campbell and M. E. O'Kelly, "Twenty-five years of hub location research," *Transportation Science*, vol. 46, no. 2, pp. 153–169, 2012.

[9] M. Sasaki, J. F. Campbell, M. Krishnamoorthy, and A. T. Ernst, "A stackelberg hub arc location model for a competitive environment," *Computers & Operations Research*, vol. 47, pp. 27–41, 2014.

[10] M. E. O'Kelly, "The location of interacting hub facilities," *Transportation science*, vol. 20, no. 2, pp. 92–106, 1986.

[11] H. Pirkul and D. A. Schilling, "An efficient procedure for designing single allocation hub and spoke systems," *Management Science*, vol. 44, no. 12-part-2, pp. S235–S242, 1998.

[12] R. S. de Camargo, G. d. Miranda, and H. Luna, "Benders decomposition for the uncapacitated multiple allocation hub location problem," *Computers & Operations Research*, vol. 35, no. 4, pp. 1047–1064, 2008.

[13] A. Ilić, D. Urošević, J. Brimberg, and N. Mladenović, "A general variable neighborhood search for solving the uncapacitated single allocation p-hub median problem," *European Journal of Operational Research*, vol. 206, no. 2, pp. 289–300, 2010.

[14] I. Contreras, J. A. Díaz, and E. Fernández, "Branch and price for large-scale capacitated hub location problems with single assignment," *INFORMS Journal on Computing*, vol. 23, no. 1, pp. 41–55, 2011.

[15] I. Rodríguez-Martín, J.-J. Salazar-González, and H. Yaman, "A branch-and-cut algorithm for the hub location and routing problem," *Computers & Operations Research*, vol. 50, pp. 161–174, 2014.

[16] M. Peker, B. Y. Kara, J. F. Campbell, and S. A. Alumur, "Spatial analysis of single allocation hub location problems," *Networks and Spatial Economics*, pp. 1–27, 2015.

[17] N. Azizi, S. Chauhan, S. Salhi, and N. Vidyarthi, "The impact of hub failure in hub-and-spoke networks: Mathematical formulations and solution techniques," *Computers & Operations Research*, vol. 65, pp. 174–188, 2016.

[18] T. Meyer, A. T. Ernst, and M. Krishnamoorthy, "A 2-phase algorithm for solving the single allocation p-hub center problem," *Computers & Operations Research*, vol. 36, no. 12, pp. 3143–3151, 2009.

[19] J. F. Campbell, "Integer programming formulations of discrete hub location problems," *European Journal of Operational Research*, vol. 72, no. 2, pp. 387–405, 1994.

[20] Z. Stanimirović, "A genetic algorithm approach for the capacitated single allocation p-hub median problem," *Computing and Informatics*, vol. 29, no. 1, pp. 117–132, 2012.

[21] R. M. A. Figueiredo, M. E. O'Kelly, and N. D. Pizzolato, "A two-stage hub location method for air transportation in Brazil," *International Transactions in Operational Research*, vol. 21, no. 2, pp. 275–289, 2014.

[22] Y. An, Y. Zhang, and B. Zeng, "The reliable hub-and-spoke design problem: Models and algorithms," *Transportation Research Part B: Methodological*, vol. 77, pp. 103–122, 2015.

[23] S. Wandelt, X. Sun, and X. Cao, "Computationally efficient attack design for robustness analysis of air transportation networks," *Transportmetrica A: Transport Science*, vol. 11, no. 10, pp. 939–966, 2015. [Online]. Available: http://dx.doi.org/10.1080/23249935.2015.1089953

[24] I. Contreras, E. Fernández, and A. Marín, "Tight bounds from a path based formulation for the tree of hub location problem," *Computers & Operations Research*, vol. 36, no. 12, pp. 3117–3127, 2009.

[25] A. T. Ernst and M. Krishnamoorthy, "Efficient algorithms for the uncapacitated single allocation p-hub median problem," *Location science*, vol. 4, no. 3, pp. 139–154, 1996.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] S. Çetiner, "An iterative hub location and routing problem for postal delivery systems," Ph.D. dissertation, Citeseer, 2003.

[28] J. F. Campbell, "Hub location for time definite transportation," *Computers & Operations Research*, vol. 36, no. 12, pp. 3107–3116, 2009.

[29] A. K. Singh, N. Subramanian, K. S. Pawar, and R. Bai, "Cold chain configuration design: location-allocation decision-making using coordination, value deterioration, and big data approximation," *Annals of Operations Research*, pp. 1–25, 2016.

[30] M. Caserta, S. Schwarze, and S. Voß, "A mathematical formulation and complexity considerations for the blocks relocation problem," *European Journal of Operational Research*, vol. 219, no. 1, pp. 96–104, 2012.

[31] J. Peiró, Á. Corberán, and R. Martí, "GRASP for the uncapacitated r-allocation p-hub median problem," *Computers & Operations Research*, vol. 43, pp. 50–60, 2014.

[32] A. T. Ernst, H. Hamacher, H. Jiang, M. Krishnamoorthy, and G. Woeginger, "Uncapacitated single and multiple allocation p-hub center problems," *Computers & Operations Research*, vol. 36, no. 7, pp. 2230–2241, 2009.

[33] A. Hoff, J. Peiró, Á. Corberán, and R. Martí, "Heuristics for the capacitated modular hub location problem," *Computers & Operations Research*, 2017.

[34] D. Kress and E. Pesch, "Sequential competitive location on networks," *European Journal of Operational Research*, vol. 217, no. 3, pp. 483–499, 2012.