Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in DISCRETE MATHEMATICS

Electronic Notes in Discrete Mathematics 62 (2017) 183–188 www.elsevier.com/locate/endm

An Approximation Algorithm for the p-Hub Median Problem

Camile Frazão Bordini [1] André Luís Vignatti [1]

DINF - Federal University of Paraná (UFPR), Curitiba-PR, Brasil

Abstract

In the *p-hub median problem* we are given a set of clients V, a set of demands $D \subseteq V \times V$, a cost function $\rho : V \times V \to \mathbb{R}^+$, and an integer $p > 0$. The objective is to select terminals $T \subseteq V$, where $|T| \leq p$, and assign each demand to a terminal, in order to minimize the total cost between demands and terminals. We present the first approximation bounds for the problem: a $1 + 2/e$ lower bound if **NP** $\subset$ **DTIME**$(n^{O(\log \log n)})$, and a (4α)-approximation algorithm if we are allowed to open at most $\left(\frac{2\alpha}{2\alpha-1}\right) p$ terminals, where $\alpha > 1$ is a trade off parameter.

Keywords: approximation algorithms, linear programming, hub location problems, p-hub median problem.

1 Introduction

In hub location problems (HLP), hubs or terminals are facilities that connect source and destination points and commute flows between them. The objective is to open terminals in order to make these connections cheaper or more efficient rather than connecting them directly. This article deals with a variant of HLPs, which we call *p-hub median problem* (pHM).

[1] Email: {cfbordini,vignatti}@inf.ufpr.br

https://doi.org/10.1016/j.endm.2017.10.032
1571-0653/© 2017 Elsevier B.V. All rights reserved.

Definition 1.1 In the *p-hub median problem* (pHM), we have a set of *clients* V, a cost function $\rho : V \times V \to \mathbb{R}^+$, a set of *demands* $D \subseteq V \times V$, and an integer $p > 0$. The objective is to select a subset $T \subseteq V$ of *terminals*, where $|T| \leq p$, and an assignment $\phi : D \to T$ that minimizes the total connection cost between demands and terminals, i.e. $\sum_{(u,v)\in D} \rho(u, \phi(u, v)) + \rho(v, \phi(u, v))$. Also, clients are points in a metric space, and ρ obeys the triangle inequality.

In the 1980's, particularly due to O'Kelly [9], the first articles on hub location problems have appeared. O'Kelly presents the first formulations and solutions for HLPs. Since then many papers have been published on the subject. According to Farahani et al. [5], while several approximation algorithms are already studied for location problems in general, in the case of HLPs, the efforts are on heuristics, metaheuristics or exacts methods. After O'Kelly, Campbell [2] proposes multiple mathematical formulations for HLPs in order to consider objective functions similar to the several classical facility location problems. More recently, Alumur and Kara [1] analyze and categorize some research articles among whose, we can consider the efforts of Campbell et al. [3] and Farahani and Hekmatfar [4] as the main references to the fundamental definitions, classifications, mathematical models, and solution methods for HLPs. Regarding the classification of HLPs, different versions are considered, and we deal with the version where all nodes are terminal candidates. The objective is to minimize the total cost of connection. We have a limited number of terminals to be opened, each demand must be assigned to a single terminal, and the terminals capacity is unlimited. Also, the most common way to deal with HLPs is when all nodes are demands ([1]), but instead we consider the case where the demands are a subset of pair of nodes. E.g., an application of our case is choosing cities where airports will be built (terminals) to connect flights between pairs of cities (demands). We show a lower bound on the approximation ratio by a reduction from the metric k-median problem, and a (4α)-approximation algorithm that opens at most $\left(\frac{2\alpha}{2\alpha-1}\right) p$ terminals, where $\alpha > 1$ is a trade off parameter. It is worth noting that, as far as we know, our results are the first approximation bounds for this problem.

2 Lower Bound and LP Formulation for pHM

Theorem 2.1 presents a reduction from the well-known *metric k-median problem* (which is NP-Hard, see [7]) to pHM.

Theorem 2.1 *The metric k-median problem is reducible to the pHM problem.*

Proof (sketch). For each vertex s of the k-median problem, we create vertices

v, v' on the pHM problem, such that $(v, v') \in D$ and $\rho(v, v') = 0$. This way, the pHM problem can solve any instance of the k-median problem. □

Theorem 2.1 shows that k-median is a particular case of pHM. Therefore, using the results by Jain et al. [6] on hardness of approximating for the k-median problem, we enunciate Corollary 2.2.

Corollary 2.2 *The pHM problem does not admit an algorithm with an approximation ratio better than* $1 + 2/e$*, unless* $\boldsymbol{NP} \subset \boldsymbol{DTIME}(n^{O(\log \log n)})$.

The goal of pHM is to minimize the total cost between demands and clients, but function ρ relates two clients. So we preprocess the input, defining a new cost function $\hat{\rho}$ such that $\hat{\rho}(d, i) = \rho(u, i) + \rho(i, v), \forall d = (u, v) \in D, i \in V$.

As our algorithm uses linear programming (LP) rounding, we present the integer program (IP) formulation for pHM:

$$
\begin{array}{lll}
\text{minimize} & \sum_{d \in D} \sum_{i \in V} x_{di}\, \hat{\rho}(d, i) & \text{(1a)} \\
\text{subject to} & \sum_{i \in V} y_i \leq p & \text{(1b)} \\
& \sum_{i \in V} x_{di} = 1, \qquad \forall d \in D & \text{(1c)} \\
& x_{di} \leq y_i, \qquad \forall d \in D, i \in V & \text{(1d)} \\
& x_{di} \in \{0, 1\}, \qquad \forall d \in D, i \in V & \text{(1e)} \\
& y_i \in \{0, 1\}, \qquad \forall i \in V. & \text{(1f)}
\end{array}
\tag{1}
$$

In IP (1), $y_i = 1$ if $i \in V$ is chosen to be in T, 0 otherwise, and $x_{di} = 1$ if $i \in V$ is assigned to demand $d \in D$, 0 otherwise. Note that the objective function is now defined in term of $\hat{\rho}$. Constraints (1b), (1c), (1d) ensure that at most p terminal are open, each demand is assigned to exactly one terminal, and each demand is assigned to an open terminal, respectively.

Finally, our algorithm uses the LP relaxation of IP (1), where $x_{di} \geq 0$ and $y_i \geq 0$, $\forall d \in D, i \in V$.

3 Rounding Algorithm

Our algorithm uses the filtering technique of Lin and Vitter [8]. As a consequence of their technique, an approximation factor is obtained, but not without violating some constraint of the LP. Nevertheless, this same technique allows to quantifies the amount of such violation. So, in our case, we obtain a 4α-approximation solution such that at most $\left(\frac{2\alpha}{2\alpha - 1}\right) p$ terminals are opened. Here, $\alpha > 1$ works as a trade off parameter, as increasing it causes

an increasing on the approximation factor and a decreasing in the number of open terminals, and vice-versa.

For each $d \in D$, let $C_d = \sum_{i \in V} x_{di}\ \hat{\rho}(d,i)$. By constraint (1c) of the LP formulation, the values x_{di} can be interpreted as a probability distribution for each d. Therefore, C_d can be seen as the *expected* cost from clients to d. Note that the LP objective function is equal to $\sum_{d \in D} C_d$.

For each $u \in V$, let $B(u, \alpha C_d) = \{u' \in V : \rho(u,u') \leq \alpha C_d\}$. For each $d = (u,v) \in D$, let $I_d = \{u' \in V : u' \in B(u, \alpha C_d) \cap B(v, \alpha C_d)\}$, we call I_d the *neighborhood* of d. By the definition of I_d, we have the following.

Lemma 3.1 *If d=$(u,v) \in D$, then $\forall i \in I_d$, $\rho(u,i) \leq \alpha C_d$ and $\rho(v,i) \leq \alpha C_d$.*

Finally, for each $d \in D$, let $\overline{V_d} = \{(u',v') \in V^2 : (u',v') \in D$ and $I_d \cap I_{(u',v')} \neq \emptyset\}$. We call $\overline{V_d}$ the *extended neighborhood* of d, motivated by the following idea: for $d = (u,v)$, if we select u or v as a terminal, we can use it to cover any other demand d'=$(u',v') \in D$ such that $(u',v') \in \overline{V_d}$, because d and d' are "close" to each other, and also, if we consider only the neighborhoods I, we do not have disjoint neighborhoods, a fact that is useful in Theorem 3.3. Also, notice that if $\alpha \leq 1$, we may have $B(u, \alpha C_d) \cap B(v, \alpha C_d) = \emptyset$ and there is no extended neighborhood for demand d.

Next, we present Algorithm 1. Note that, if we obtain T, then the assignment ϕ is already defined since, without loss of generality, a demand is always assigned to the nearest open terminal in an optimal solution.

Algorithm 1
Solve the LP and use it to compute the C_d values; Set $T := \{\}$ and $\overline{D} := D$
while $\overline{D} \neq \emptyset$ **do**
  Choose $d = (u,v) \in \overline{D}$ with the lowest value of C_d; Set $T := T \cup \{u\}$
  for $(u',v') \in \overline{D}$ **do**
    if $(u' \in \overline{V_d})$ **and** $(v' \in \overline{V_d})$ **then** $\overline{D} := \overline{D} \setminus (u',v')$
  Set $\overline{D} := \overline{D} \setminus (u,v)$
return T

Theorem 3.2 *Algorithm 1 is a (4α)-approximation algorithm, with $\alpha > 1$.*

Proof. Let OPT and OPT_{LP} be the optimal values, resp., for an instance of the problem, and an instance of the (relaxed) LP formulation. Let d=$(u,v) \in D$, we abuse notation, using $u \in d$ to express that u is one of the vertices of d. Assume that $d = (u,v) \in D$ is the demand selected by the algorithm in a given step with the lowest value C_d and $u \in d$ is chosen to be included in T. Let $d' = (u',v')$ such that $(u',v') \in \overline{V_d}$; so d' is removed from $\overline{D}$. Denote the

terminal u to which d' has been assigned by $k^{d'}$. At the time d is selected, both d and d' are in $\overline{D}$, and $C_d \leq C_{d'}$. Let C(alg) be the cost of Algorithm 1, $\mathrm{C}_{d'}^{alg}$ the cost incurred by a demand d' in Algorithm 1 and x^* the optimal solution of the LP. Thus, $\mathrm{C(alg)} = \sum_{d' \in D} \mathrm{C}_{d'}^{alg} = \sum_{d' \in D} \hat{\rho}(d', k^{d'})$, which is

$$\begin{aligned}
&= \sum_{d' \in D} \left[\rho(u', k^{d'}) + \rho(v', k^{d'})\right] \\
&\leq \sum_{d' \in D} \left[\left(\rho(u', s) + \rho(s, k^{d'})\right) + \left(\rho(v', s) + \rho(s, k^{d'})\right)\right] \\
&\leq \sum_{d' \in D} \left[(\alpha C_{d'} + \alpha C_d) + (\alpha C_{d'} + \alpha C_d)\right] \leq \sum_{d' \in D} [2\alpha C_{d'} + 2\alpha C_{d'}] \\
&= 4\alpha \sum_{d' \in D} \sum_{i \in V} x^*_{d'i} \hat{\rho}(d', i) = (4\alpha)\mathrm{OPT_{LP}} \leq (4\alpha)\mathrm{OPT}
\end{aligned}$$

where the first inequality uses the triangle inequality and, because $(u', v') \in \overline{V_d}$, implies that there is some s in both $I_{d'}$ and I_d, and the second inequality follows from Lemma 3.1. □

In order to bound $|T|$, we count the number of iterations until $\overline{D} = \emptyset$, since each iteration opens one terminal. We claim that, for each $k \in d$ included in T, I_d contains at least $\frac{2\alpha-1}{2\alpha}$ "fractional terminals" corresponding to the LP values, that is, the sum of the y_i for all $i \in I_d$ is at least $\frac{2\alpha-1}{2\alpha}$. Note that neighborhoods I_d are all disjoint (for different $k \in T$). Indeed, suppose that $k_i \in d_i$ and $k_j \in d_j$ such that $k_i, k_j \in T$, with k_i chosen earlier than k_j by the algorithm. If $I_{d_i} \cap I_{d_j} \neq \emptyset$ then $k_j \in \overline{V_{d_i}}$ and k_j could not be in T. Besides that, the sum of all the y_i is at most p. Theorem 3.3 concludes this fact.

Theorem 3.3 *Algorithm 1 produces a solution with* $|T| \leq \left(\frac{2\alpha}{2\alpha-1}\right) p$.

Proof. We define the probability space $(\Omega_d, \Pr)$ as $\Omega_d = \{(d, i) : i \in V\}$ and $\Pr[(d, i)] = x_{d_i}$ such that $\sum_{i \in V} \Pr[(d, i)] = 1$, $\forall i \in V$. Let Z be a random variable of the cost between a given $d \in D$ to all $i \in V$, i.e. $Z : \Omega_d \to \mathbb{R}$ where $\forall i \in V$, $Z((d, i)) = \hat{\rho}(d, i)$ and the probability of Z assuming each of these values is x_{di}. Thus, $E[Z] = \sum_{i \in V} x_{di}\ \hat{\rho}(d, i) = C_d$. Therefore, for a demand d=$(u, v) \in D$ such that $u \in T$ or $v \in T$, we have, $\sum_{i \in I_d} y_i \geq \sum_{i \in I_d} x_{di} = \sum_{i \in I_d} \Pr[Z = \hat{\rho}(d, i)] = \bigcup_{i \in I_d} \Pr[Z = \hat{\rho}(d, i)] = \Pr[Z \leq 2\alpha C_d] = 1 - \Pr[Z > 2\alpha E[Z]]$, where the inequality follows from constraint (1d) of the LP, the second equality follows since the events are mutually disjoint, and the third equality follows since the union of these probabilities is equal to the probability of $Z \leq 2\alpha C_d$, because all $i \in I_d$ has cost at most αC_d from both u and v (Lemma 3.1). By Markov's inequality, $\Pr[Z > 2\alpha E[Z]] \leq \frac{1}{2\alpha}$, so, $\sum_{i \in V_d} y_i \geq 1 - \Pr[Z > 2\alpha E[Z]] \geq \frac{2\alpha-1}{2\alpha}$. Thus, for each $k \in d$ included in T by Algorithm 1, the neighborhood I_d contains at least $\frac{2\alpha-1}{2\alpha}$ of a terminal according to the fractional solution of LP. Finally, the upper bound on T is

immediate, since the neighborhood of each open terminal has at least $\frac{2\alpha-1}{2\alpha}$ "fractional terminals" (and these neighborhoods are all disjoint for different $k \in T$). Combining with constraint (1b) of the LP, then $|T| \leq \left(\frac{2\alpha}{2\alpha-1}\right) p$. □

4 Extensions and Future Works

For the pHM problem presented here, an important direction is to ensure that no more than p terminals are opened. We are aware of some research, post the Lin and Vitter [8] article, that deal with this issue. Another obvious direction is to improve the approximation factor, since the bound in Theorem 3.2 is not tight compared to the result of Corollary 2.2. Also, there are many HLPs variants that can be considered, e.g.. with terminal capacities, min-max objective functions, terminal opening cost, specific metrics, and so on.

References

[1] Alumur, S., and Kara, B. Y., *Network hub location problems: The state of the art*, European Journal of Operational Research. **190(1)** (2008), 1–21.

[2] Campbell, J. F., *Integer programming formulations of discrete hub location problems*, European Journal of Operational Research. **72(2)** (1994), 387–405.

[3] Campbell, J. F., Ernst, A. T., and Krishnamoorthy, M., "Facility location: Applications and theory," 2nd Ed., Zvi Drezner & Horst Hamacher, 2002.

[4] Farahani, R. Z., and Hekmatfar, H., *Facilities location: concepts, models, algorithms and case studies*, Heidelberg: Springer-Verlag, 2009.

[5] Farahani, R. Z., Kekmatfar, M., Arabani, A. B., and Nikbakhsh, E., *Hub location problems: A review of models, classification, solution techniques, and applications*, Computers & Industrial Engineering. **64(4)** (2013), 1096–1109.

[6] Jain, K., Mahdian, M. and Saberi, A., *A new greedy approach for facility location problems*, ACM Symposium on Theory of Computing, (2002), 731–740.

[7] Kariv, O., and Hakimi, S. L., *An algorithmic approach to network location problems. ii: The p-medians*, SIAM Journal on Applied Mathematics. **37(3)** (1979), 539–560.

[8] Lin, J. H. and Vitter, J. S., *ϵ-approximations with minimum packing constraint violation*, ACM Symposium on Theory of Computing, (1992), 771–782

[9] O'Kelly, M. E., *The location of interacting hub facilities*, Transportation Science. **20(2)** (1986), 92–106.