An improved approximation algorithm for $k$-median problem using a new factor-revealing LP

Chenchen Wu*  Dachuan Xu†‡  Donglei Du§  Yishui Wang‡

October 15, 2014

Abstract

The $k$-median problem is a well-known strongly NP-hard combinatorial optimization problem of both theoretical and practical significance. The previous best approximation ratio for this problem is $2.611+\epsilon$ (Bryka et al. 2014) based on an $(1, 1.95238219)$ bi-factor approximation algorithm for the classical facility location problem (FLP). This work offers an improved algorithm with an approximation ratio $2.592 + \epsilon$ based on a new $(1, 1.93910094)$ bi-factor approximation algorithm for the FLP.

Keywords: facility location problem; $k$-median; approximation algorithm

1 Introduction

The $k$-median problem is closely related to the facility location problem (FLP), one of the most important problems in the field of operations research and computer science. Formally, in the FLP, we are given a facility set $F$ and a client set $D$. Opening facility $i \in F$ incurs an
opening cost of \( f_i \), and connecting client \( j \) to facility \( i \) incurs a connection cost of \( c_{ij} \). The connection costs form a metric (that is, nonnegative, symmetric, and satisfying the triangle inequality). The objective is to open some facilities in \( F \) and connect each client in \( D \) to an opened facility such that the total opening and connection cost is minimized.

The \( k \)-median problem is the most important variant of the FLP. In the \( k \)-median problem, there is no facility open cost, but no more than \( k \) facilities are allowed to be opened.

The FLP can be formulated as an integer linear program (ILP). Let binary variable \( x_{ij} \) represent whether client \( j \) is connected to facility \( i \) or not and binary variable \( y_i \) represent whether facility \( i \) is opened or not.

\[
\begin{align*}
\min & \quad \sum_{i \in F} f_i y_i + \sum_{i \in F, j \in D} c_{ij} x_{ij} \\
\text{s. t.} & \quad \sum_{i \in F} x_{ij} \geq 1, \quad \forall j \in D, \\
& \quad x_{ij} \leq y_i, \quad \forall i \in F, j \in D, \\
& \quad x_{ij}, y_i \in \{0, 1\}.
\end{align*}
\]

The first constraint above represents that each client must be connected to at least one facility, while the second constraint specifies that clients are connected to only opened facilities.

On the other hand, the \( k \)-median problem can be formulated as the following ILP:

\[
\begin{align*}
\min & \quad \sum_{i \in F, j \in D} c_{ij} x_{ij} \\
\text{s. t.} & \quad \sum_{i \in F} y_i \leq k, \\
& \quad \sum_{i \in F} x_{ij} \geq 1, \quad \forall j \in D \\
& \quad x_{ij} \leq y_i, \quad \forall i \in F, j \in D, \\
& \quad x_{ij}, y_i \in \{0, 1\}, \quad \forall i \in F, j \in D.
\end{align*}
\]

The program (2) differs from (1) in two aspects. First, the objective function only involves the connection cost since there is no facility open cost. Second, the first constraint in (2) guarantees that the number of opened facilities is no more than \( k \).

An established concept in studying the FLP is the bi-factor approximation ratio. The
The purpose of this concept is to balance the facility and connection cost in the optimal solution to create more room for an improved solution.

**Definition 1.1 Bi-factor:** \((\gamma_f, \gamma_c)\) is called a bi-factor of an algorithm for the FLP, if the cost obtained by the algorithm is no more than \(\gamma_f \text{Cost}^* + \gamma_c \text{Cost}^*_c\), where \(\text{Cost}^*\) and \(\text{Cost}^*_c\) respectively are the opening and connection costs in an optimal solution. The algorithm is called the \((\gamma_f, \gamma_c)\)-approximation algorithm.

The obvious connection between bi-factor and the normal approximation ratio is that any \((\gamma_f, \gamma_c)\) bi-factor approximation algorithm implies an approximation ratio of \(\max\{\gamma_f, \gamma_c\}\) for the FLP. There is another important concept, so-called **bi-point solution**, for the \(k\)-median problem.

**Definition 1.2 Bi-point solution:** Given a \(k\)-median instance \(I\), a bi-point solution is a pair of facility sets \(\hat{F}_1, \hat{F}_2 \subseteq F\) with \(|\hat{F}_1| \leq k \leq |\hat{F}_2|\), along with two reals number \(a\) and \(b\) with \(a + b = 1\) such that \(a|\hat{F}_1| + b|\hat{F}_2| = k\).

The connection cost of the bi-point solution is defined as \(aD_1 + bD_2\), where \(D_1\) and \(D_2\) are the connection costs of opening facilities in \(\hat{F}_1\) and \(\hat{F}_2\), respectively.

### 1.1 Literature Review

Shomys et al.[11] offer the first constant approximation ratio for the FLP via the linear programming rounding technique, which was extended many times subsequently with improved approximation ratios (c.f. [4, 12]). Note that the FLP is a special case of the set cover problem. Based on the greedy scheme for the set cover problem, Jain et al.[6] propose an algorithm in which they revise the greedy algorithm into a dual ascent process where the dual variable can be viewed as the contribution for opening facility and paying the connection cost. In the analysis of the algorithm, they use the dual-fitting scheme to obtain factor-revealing LPs with approximation ratio of 1.61 and bi-factor of \((1, 2)\). Mahdian et al. [10] obtain an approximation ratio of 1.52 and bi-factor of \((1.11, 1.78)\) by using the technique of scaling up the facility cost before running the 1.61-approximation algorithm by Jain et
al. [6]. Bryka et al. [2] scale up the contribution of the clients for opening facility to obtain an improved bi-factor \((1, 1.95238219)\). Li [8] introduces a non-uniform random variable into the process of linear programming rounding to obtain the currently best approximation ratio 1.488 for the FLP. Relevant to this work is the 3-approximation algorithm (Jain and Vazirani [7]) based on the standard primal-dual scheme with Lagrangian multiplier persevering (LMP) property, which states that the facility opening cost plus three times the connection cost for the solution obtained by the algorithm is no more than three times the optimal value.

The first constant approximation algorithm for the \(k\)-median problem is given by Charikar et al. [3] in which they propose a \(6\frac{2}{3}\)-approximation algorithm based on linear programming rounding technique. Arya et al. [1] design a \(\left(3 + \frac{2}{p}\right)\)-approximation algorithm through local search scheme with the so-called multi-swapping operation (that is, use \(p\) facilities to swap \(p\) facilities opened in the current solution). The algorithm iteratively finds a local optimal solution by this swap operation, starting from any feasible solution. Jain and Vazirani [7] give a 6-approximation algorithm using the LMP property. Li and Svensson [9] show that if there is a solution with \(k + \epsilon\) facilities (where \(\epsilon\) is a positive constant), there will be a solution with \(k\) facilities and the cost only increases by a factor of \(1 + \epsilon\). This observation leads to an \((1 + \sqrt{3} + \epsilon)\)-approximation algorithm for the \(k\)-median problem. Bryka et al. [2] further improve the approximation ratio for the \(k\)-median problem to \(2.611 + \epsilon\) based on an \((1, 1.95238219)\) bi-factor approximation for the FLP.

As a negative result, Guha and Khuller [5] show that the existence of an \(\alpha\)-approximation algorithm for the FLP with \(\alpha < 1.463\) is impossible unless \(NP \subseteq DTIME(n_c^{O(\log \log n_c)})\), where \(n_c\) is the number of the clients.

### 1.2 High level idea of our algorithm

We give an improved bi-factor approximation for the FLP which results in an improved approximation ratio for the \(k\)-median problem. The main idea leading to this improvement is that the worst-case instances in the factor-revealing LP with different parameters may
be different. Therefore we should expect to have an improved solution if we run the same algorithm with different parameters and choose the best (cheapest) one. However, analyzing this solution through factor-revealing LP directly, we can not identify the clients in this solution corresponding to that in the optimal solution. In order to overcome this difficulty, the main technique in showing the bi-factor of the parametric algorithm is to focus on the convex combination of the solutions, which is much more amenable to the tool of factor-revealing LP, compared to directly handling the cheapest solution.

We present an improved \((1, 1.93812708)\) bi-factor approximation algorithm for the FLP and an improved \((2.592 + \epsilon)\)-approximation algorithm for the \(k\)-median problem in Section 2 and 3 respectively. Some conclusions are given in Section 4.

## 2 Improved bi-factor approximation algorithm for the FLP

### 2.1 Algorithm

We first recall the algorithm proposed by Bryka et al.\([2]\). In the algorithm, denote \(\hat{F}\) as the opened facility set, and \(A\) as the active client set in which the client is unconnected. They introduce a parameter \(\theta\) to scale up the contribution of the active client. For brevity, we denoted the algorithm as BPRS \((\theta)\).

**Algorithm 2.1 BPRS\((\theta)\) \(([2])\)**

**Step 0** Initially, all facilities are un-opened and all clients are un-connected, that is, set \(\hat{F} := \emptyset\) and \(A := D\).

**Step 1** For each \(i \in F \setminus \hat{F}\), calculate the root \(\bar{t}_i^1\) of the following equation with respect to \(t\)

\[
\sum_{j \in A} \theta(t - c_{ij})_+ + \sum_{j \in D \setminus A} (c_{\phi(j)} - c_{ij})_+ = f_i,
\]

where \((x)_+ := \max\{x, 0\}\). Set

\[
i^*_1 := \arg \min_{i \in F \setminus \hat{F}} \bar{t}_i^1.
\]
Step 2 For each $i \in \hat{F}$, calculate $\bar{t}_i^2 = \min_{j \in A} c_{ij}$. Set

$$i^*_2 := \arg \min_{i \in \hat{F}} \bar{t}_i^2.$$  

Step 3 Set $i^* = \arg \min \{t_1^1, t_2^2\}$.

Case 1 If $i^* = i^*_1$, open $i^*$, that is, $\hat{F} := \hat{F} \cup \{i^*\}$. Set $S_{i^*} = \{j : t_1^1 \geq c_{i^*j} \text{ or } c_{\phi(j)j} \geq c_{i^*j}\}$. Set or update $c_{\phi(j)j} := c_{i^*j}$ for each $j \in S_{i^*}$, that is, all clients in $S_{i^*}$ are connected to the facility $i^*$. Define $v_j = t_1^1$ for each $j \in S_{i^*} \cap A$, and update $A := A \setminus S_{i^*}$.

Case 2 If $i^* = i^*_2$, connect $j^* = \arg \min_{j \in A} c_{i^*_2j}$ to the facility $i^*_2$, that is, set $c_{\phi(j)j} := i^*_2$ and $A := A \setminus \{j^*\}$. Set $v_{j^*} := \bar{t}_2^2$.

Step 4 Repeat Steps 1-3 above until all clients are connected, that is, $A = \emptyset$.

In our algorithm, we run BPRS($\theta_l$) for a list of pre-selected parameters $\theta_l$ ($l = 1, 2, \cdots, L$), and output the solution with the smallest cost.

Algorithm 2.2

Step 0 Fix parameters $\theta_l$ ($l = 1, 2, \cdots, L$).

Step 1 For each $l$ ($l = 1, 2, \cdots, L$), run BPRS($\theta_l$) to obtain the solution $\text{ALG}(\theta_l)$ and the dual solution (maybe infeasible) $\{v^l_j\}$. Denote the corresponding cost as $\text{cost}(\text{ALG}_l)$.

Step 2 Output $\text{ALG} = \arg \min_{l=1, \ldots, L} \text{cost}(\text{ALG}_l)$, namely the solution with smallest cost among $\text{ALG}(\theta_l)$.

2.2 Analysis: a new factor-revealing LP

2.2.1 A new factor-revealing LP

In the algorithm, we define $r_{ji}^l$ as the connection cost for client $j$ when client $i$ is connected to an opened facility if client $j$ is connected at time $v_i^l - \epsilon$, otherwise $r_{ji}^l = v_i^l$ in BPRS($\theta_l$). Note that $r_{jj}^l$ is the connection cost for the client $j$ when the state of $j$ changes from active to
in-active. Moreover, the cost obtained by BPRS(\(\theta_l\)) is denoted as \(\text{cost}(\text{BPRS}(\theta_l)) := \sum_{j \in D} \theta_l(v_j^l - r_{jj}^l) + r_{jj}^l\). Thus, the cost obtained by our algorithm is

\[
\text{cost}(\text{Alg}) = \min_l \text{cost}(\text{Alg}_l) \leq \sum_{l=1}^L p_l \left( \sum_{j \in D} \theta_l(v_j^l - r_{jj}^l) + r_{jj}^l \right),
\]

where \(\{p_l\}\) is a combination coefficient, that is, \(p_l \geq 0 (l = 1, 2, \ldots, L)\) and \(\sum_{l=1}^L p_l = 1\).

Our purpose is to prove a bi-factor of \((\gamma_f, \gamma_c)\) such that:

\[
\sum_{l=1}^L p_l \left( \sum_{j \in D} \theta_l(v_j^l - r_{jj}^l) + r_{jj}^l \right) \leq \gamma_f \text{Cost}^*_f + \gamma_c \text{Cost}^*_c.
\]

(4)

Note that \(\text{Cost}^*_f\) and \(\text{Cost}^*_c\) are the facility opening cost and the connection cost in an optimal solution, respectively. Assume that the opened facility set is \(F^*\) in the optimal solution, and the set of clients connected to \(i \in F^*\) is \(N_i\) in the optimal solution.

Evidently (4) follows if we can prove the following inequality for each \(i \in F^*\):

\[
\sum_{l=1}^L p_l \left( \sum_{j \in N_i} \theta_l(v_j^l - r_{jj}^l) + r_{jj}^l \right) \leq \gamma_f f_i + \gamma_c \sum_{j \in N_i} c_{ij}.
\]

Or equivalently,

\[
\gamma_c \geq \frac{\sum_{l=1}^L p_l \left( \sum_{j \in N_i} \theta_l(v_j^l - r_{jj}^l) + r_{jj}^l \right) - \gamma_f f_i}{\sum_{j \in N_i} c_{ij}}.
\]

For convenience, we omit the subscript \(i\) since all clients in \(N_i\) are all connected to \(i\). Assume that \(|N_i| = n\). Denote \(c_{ij} = d_j, f_i = f\). Thus, we can use the subscript \(i\) to denote a client. The following program (5) gives \(\gamma_c\) for any given \(\gamma_f\).

**Theorem 2.3** The value of \(\gamma_c\) can be obtained by the following factor-revealing LP for any given \(\gamma_f\):

\[
\eta_n = \max \frac{\sum_{l=1}^L p_l \left( \sum_{j=1}^n \left( \theta_l(v_j^l - r_{jj}^l) + r_{jj}^l \right) \right) - \gamma_f f}{\sum_{j=1}^n d_j}
\]

s. t. \(\sum_{j=1}^n d_j, v_j^l \leq v_{j+1}^l, \forall j = 1, \ldots, n, l = 1, \ldots, L\)

(6)

\(r_{ji}^l \geq r_{j+1,i}^l, \forall 1 \leq j \leq i \leq n - 1, l = 1, \ldots, L\)

(7)
\[ v^l_j \leq r^l_{ji} + d_i + d_j, \quad \forall 1 \leq j \leq i \leq n, l = 1, \ldots, L \]  

\[ \sum_{j=1}^{i-1} (r^l_{ji} - d_j) + \theta_i \sum_{j=i}^{n} (v^l_i - d_j) \leq f, \quad \forall i = 1, \ldots, n, l = 1, \ldots, L, \]  

\[ r^l_{jj} \leq v^l_j, \quad \forall j = 1, \ldots, n, l = 1, \ldots, L \]  

\[ v^l_j \geq 0, \quad \forall j = 1, \ldots, n, l = 1, \ldots, L \]  

\[ r^l_{ji} \geq 0, \quad \forall 1 \leq j \leq i \leq n, l = 1, \ldots, L \]  

\[ f, d_j \geq 0, \quad \forall j = 1, \ldots, n. \]  

The proof available in the appendix of this Theorem is similar to that in [6]. Moreover, calculating \( \gamma_c \) requires knowing the number of the clients in the star centered at facility \( f \) in Theorem 2.3. Without knowing the number, we need to consider all possible numbers as the following result dictates.

**Corollary 2.4** For the FLP, there is a bi-factor of \((\gamma_f, \gamma_c)\) for a fixed \( \gamma_f \), where

\[ \gamma_c = \sup_n \{ \eta_n \}. \]

### 2.2.2 Upper bound of the factor-revealing LP

Finding the exact value of \( \gamma_c \) in Corollary 2.4 involves all \( n \) and turns out to be highly computationally intensive. Instead, we give an upper bound indexed by a parameter \( t \) in the factor-revealing LP (5)-(13) for each \( n \). One property of this upper bound useful later is that it is increasing with respect to the parameter \( t \). For proving the upper bound of the factor-revealing LP (5)-(13), we need two lemmas first.

**Lemma 2.5** \( \eta_n \leq \eta_{n-m} \), for any positive integer \( m \).

**Lemma 2.6** For each \( t \) and \( n \), we have

\[ \eta_{tn} \leq \rho_t, \]

where

\[ \rho_t = \max \left\{ \frac{\sum_{l=1}^{L} p_l \left( \frac{\theta_i (\bar{v}^l_i - \bar{r}^l_{ji}) + \bar{r}^l_{jj}}{\sum_{l=1}^{l} d_j} \right) - \gamma_f \bar{f}}}{\sum_{l=1}^{L} d_j} \right\} \]  

(14)
Note that (15)-(16) and (19)-(22) are the same as the corresponding constraints in program (5). Actually these two programs differ only by two constraints: (8) and (9) are replaced by (17) and (18), respectively. Lemmas 2.5-2.6 imply the following result.

**Lemma 2.7** For each integer $t$ and $n$,  
$\eta_n \leq \rho_t$.

**Proof.** For any integer $t$ and $n$, we have  
$\eta_n \leq \eta_{tn} \leq \rho_t$.

An upper bound for $\gamma_c$ in our algorithm is the minimum $\rho_t$ over the convex combination of coefficients $\{p_l\}$ for any fixed $t$. However, finding the optimal $\{p_l\}$ is computationally expensive. For our purpose, it suffices to calculate $\gamma_c$ only for some selected $t$, $p_l$, $\theta_l$, and $\gamma_f$.

The factor-revealing LP and the upper bound in Bryka et al. [2] correspond to $l = 1$ in programs (5) and (14) in our algorithm, respectively. Note that our factor-revealing LP and the upper bound are separable in terms of $l$. Therefore, the proofs (available in the appendix) for Lemmas 2.5-2.6 are similar to those in Bryka et al. [2].

### 2.2.3 Numerical results

In the numerical experiments, we set $l = 2$, $p_1 = 0.93$, $p_2 = 0.07$, $\theta_1 = 1$, $\theta_2 = 1.35$, and $\gamma_f = 1$. The following table compares the values of $\rho_t$ between our algorithm and that by
Bryka et al. [2] for different $t$.

Table 1: The values of $\rho_t$.

| $t$ | Byrka et al. | our algorithm |
|-----|--------------|---------------|
| 25  | 1.97005293   | 1.95603908    |
| 50  | 1.96071418   | 1.94677717    |
| 75  | 1.95735066   | 1.94361565    |
| 100 | 1.95551828   | 1.94201089    |
| 175 | 1.95327445   | 1.93993533    |
| 200 | 1.95291564   | 1.93958933    |
| 250 | 1.95238219   | 1.93910914    |
| 300 | 1.95202855   | 1.93877777    |
| 400 | 1.95158711   | 1.93837073    |
| 500 | 1.95132412   | 1.93812708    |

Moreover, the table below lists the corresponding $\eta_n$ in our algorithm and that in Bryka et al. [2] for different $n$:

Table 2: The values of $\eta_n$.

| $n$ | Byrka et al. | our algorithm |
|-----|--------------|---------------|
| 25  | 1.92701290   | 1.91680198    |
| 50  | 1.93915391   | 1.92711657    |
| 75  | 1.94295757   | 1.93048145    |
| 100 | 1.94483254   | 1.93216420    |
| 175 | 1.94717425   | 1.93431534    |
| 200 | 1.94756190   | 1.93467105    |
Tables 1-2 for \( t = 500 \) and \( n = 500 \) show that our upper bound \( 1.93812708 \) is strictly smaller than the lower bound \( 1.94918216 \) of the algorithm in Bryka et al. [2], therefore leading to an improved bi-factor approximation algorithm which is summarized in the following lemma.

**Lemma 2.8** Algorithm 2.2 is an \((1, 1.93812708)\)-approximation algorithm for the FLP.

**Proof.** The lemma is concluded by Lemma 2.7 and Table 1 for \( t = 500 \). \( \Box \)

## 3 Improved \((2.592 + \epsilon)\)-approximation for the \( k \)-median problem

Obtaining the bi-point solution of the \( k \)-median problem involves the dual program of the LP-relaxation.

\[
\begin{align*}
\text{max} & \quad \sum_{j \in D} v_j - k \cdot l \\
\text{s. t.} & \quad v_j \leq w_{ij} + c_{ij}, \quad \forall i \in F, j \in D, \\
& \quad \sum_{j \in D} w_{ij} \leq l, \quad \forall i \in F, \\
& \quad v_j, w_{ij}, l \geq 0, \quad \forall i \in F, j \in D.
\end{align*}
\]

If we fix the value of \( l \), the dual of (23) can be viewed as an instance of the FLP with the facility cost \( l \). Changing the value of \( l \) adjusts the number of the opened facilities. The \((1, 1.93812708)\) bi-factor approximation in Lemma 2.8 to the FLP, together with the

| 250 | 1.94810094 | 1.93517247 |
|-----|------------|------------|
| 300 | 1.94846654 | 1.93550473 |
| 400 | 1.94891300 | 1.93591696 |
| 500 | 1.94918216 | 1.93616294 |
approach of Jain et al. [6], imply a bi-point solution for the $k$-median problem whose cost is within $1.93812708 \cdot OPT$, where $OPT$ is the optimal value of the $k$-median problem.

We now call upon the following result from Byrka et al. [2] to obtain an improved approximation ratio for the $k$-median problem. A bi-point solution for the $k$-median problem can be converted to a solution with $k + f(\epsilon)$ facilities and with a cost no more than $1.3371\alpha$ times the optimal solution, where $f(\epsilon)$ is a constant dependent on $k$, and $\alpha$ is the approximation ratio to obtain the bi-point solution. Moreover, the solution with $k + f(\epsilon)$ facilities can be converted into a feasible integer solution for the $k$-median problem with a cost scaled by no more than a factor of $(1 + \epsilon)$.

**Theorem 3.1** ([2]) Suppose that we have a bi-factor $(1, \alpha)$-approximation algorithm for the FLP. For any instance of the $k$-median problem with the input size $N$ and a fixed constant $\epsilon > 0$, there exists a $(1.3371\alpha + \epsilon)$-approximation algorithm which runs in $O(N^{O(1/\epsilon \log(1/\epsilon))})$.

From Lemma 2.8, we have $\alpha = 1.93812708$, leading to our main result.

**Corollary 3.2** There exists a $(2.592 + \epsilon)$-approximation algorithm for the $k$-median problem.

## 4 Conclusions

In this work, we offer an improved approximation algorithm for one of the most important problems in location theory. A natural question is to see whether our multi-parameter idea can be extended to other problems such as the classical FLP. Our preliminary experiments showed negative result: calculating the factor-revealing LP in Bryka et al. [2] for $\gamma_f = 1.11$ revealed that $\gamma_c$ is increasing with respect to $\theta$; and hence the multi-parameter idea cannot improve the bi-factor.

**Acknowledgements** We thank Xujin Chen and Rolf H. Möhring for helpful discussions on the paper. The research of the second and fourth authors are supported by NSF of China (No. 11371001). The third author’s research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant 283106.
References

[1] Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., Pandit, V. Local Search Heuristics for k-Median and Facility Location Problem. SIAM Journal on Computing, 33: 544–562, 2004.

[2] Byrka, J., Pensyl, T., Rybicki, B., Srinivansan, A., Trinh, K. An improved approximation for k-median, and positive correlation in budgeted optimization. Accepted by Proceedings of SODA, 2015.

[3] Charikar, M., Guha, S., Tardos, E., and Shmoys, D.B. A constant-factor approximation algorithm for the k-median problem. Journal of Computer and System Sciences, 65: 129–149, 2002.

[4] Chudak, F.A., Shmoys, D.B. Improved approximation algorithms for the uncapacitated facility location problem. SIAM Journal on Computing, 33: 1–25, 2003.

[5] Guha, S., Khuller, S. Greedy strikes back: improved facility location algorithms. Journal of Algorithm. 31: 228–248, 1999.

[6] Jain, K., Mahdian, M., Markakis, E., Saberi, A., Vazirani, V.V. Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP. Journal of the ACM. 50: 795–824, 2003.

[7] Jain, K., Vazirani, V.V. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. Journal of the ACM. 48: 274–296, 2001.

[8] Li, S. A 1.488-approximation algorithm for the uncapacitated facility location problem. In Proceedings of ICALP, pp. 77–88, 2011.

[9] Li, S., Svensson O. Approximating k-median via pseudo-approximation. In STOC, pp. 901-910, 2013.
Appendix

Proof of Theorem 2.3. The lemma follows if we show that $v^l_j$ and $r^l_{ji}$ $(l = 1, 2, \cdots, L)$ together with the input $f$ and $d_j$ satisfy the above factor-revealing LP. It suffices to show the desired for each $l = 1, 2, \cdots, L$ due to the separability of the constraints in LP (5)-(13).

Obviously (6) holds for each $l = 1, 2, \cdots, L$ since we can re-order the clients by increasing the value of $v^l_j$ obtained by BPRS($\theta_l$).

For (7), $r^l_{ji}$ is the connection cost of the client $j$ at the time $v^l_i$. By BPRS($\theta_l$), $j$ just switches to a closer opened facility, that is, $r^l_{ji}$ is non-increasing with $i$ increasing.

For (8), we consider the time $t = v^l_i - \epsilon$ of Algorithm 2.2. If the constraint is not satisfied at that time, then one of the client $i, i+1, \cdots, n$ is connected to the facility $f$. This last claim contradicts the fact that $i, i+1, \cdots, n$ is not connected to any facility at that time.

For (10), $r^l_{jj}$ is the connection cost when the client $j$ is connected and $v^l_j$ is the time when client $j$ changes state from active to in-active. Obviously, $r^l_{jj}$ is no more than $v^l_j$.

The constraints of (11), (12), and (13) hold obviously. ⌜
Proof of Lemma 2.5. Assume that the \((v'_j, r'_{ji}, d_j, f)\) is the optimal solution with exactly \(n\) clients. We now construct a feasible solution with \(n \cdot m\) clients as follows:

- \(\bar{v}'_{(j-1)m+a} = \frac{v'_j}{m}, \forall j = 1, \ldots, n, a = 1, \ldots, m, l = 1, \ldots, L;\)
- \(\bar{r}'_{(j-1)m+a,(i-1)m+b} = \frac{r'_{ji}}{m}, \forall 1 \leq j \leq i \leq n, a, b = 1, \ldots, m, l = 1, \ldots, L;\)
- \(\bar{d}_{(j-1)m+a} = \frac{d_j}{m}, \forall j = 1, \ldots, n, a = 1, \ldots, m, l = 1, \ldots, L;\)
- \(\bar{f} = f.\)

Then, we show that \((\bar{v}'_j, \bar{r}'_{ji}, \bar{d}_j, \bar{f})\) satisfies the constraints (6)-(13) when the number of the clients is \(nm\). We only show the constraint (9) holds since the others are obviously true.

For each \(i \in \{1, 2, \ldots, nm\}\), there exist some \(h\) and \(a\) such that \(i = hm+a,\) where \(1 \leq a \leq m.\)

Therefore,

\[
\sum_{j=1}^{i-1} (r'_{ji} - \bar{d}_j)_+ + \theta_l \sum_{j=1}^{nm} (v'_i - \bar{d}_j)_+ = \sum_{j=1}^{hm+a-1} (r'_{j,hm+a} - \bar{d}_j)_+ + \theta_l \sum_{j=hm+a}^{nm} (v'_i - \bar{d}_j)_+ \\
= \sum_{j=1}^{hm} (r'_{j,hm+a} - \bar{d}_j)_+ + \theta_l \sum_{j=hm+a}^{nm} (v'_i - \bar{d}_j)_+ + \theta_l \sum_{j=hm+a}^{nm} (v'_i - \bar{d}_j)_+ \\
= \sum_{j=1}^{hm} (r'_{j,hm+a} - \bar{d}_j)_+ + \theta_l \sum_{j=hm+1}^{nm} (v'_i - \bar{d}_j)_+ + \theta_l \sum_{j=hm+a}^{nm} (v'_i - \bar{d}_j)_+ \\
= \sum_{j=1}^{hm} m \cdot \left( \frac{r'_{j,h+1}}{m} - \frac{d_j}{m} \right)_+ + \theta_l \sum_{j=h+1}^{nm} m \cdot \left( \frac{v'_{h+1}}{m} - \frac{d_j}{m} \right)_+ \\
= \sum_{j=1}^{h} (r'_{j,h+1} - d_j)_+ + \theta_l \sum_{j=h+1}^{nm} (v'_i - d_j)_+ \leq f = \bar{f}
\]

The third inequality holds since \(r'_{ji}\) and \(v'_i\) satisfy the constraints (6), (7) and (10), that is, \(r'_{j,hm+a} \leq r'_{ji} \leq v'_{i,hm+a}\) for all \(j \in [hm+1, hm+a-1]\) and \(\theta_l \geq 1.\) The fourth equality follows from the definition of \((v'_i, r'_{ji}, \bar{d}_j, \bar{f})\). The sixth inequality holds since \((v'_j, r'_{ji}, d_j, f)\) is a feasible solution.

Moreover, the value of the solution \((\bar{v}'_j, \bar{r}'_{ji}, \bar{d}_j, \bar{f})\) is the same as that of \((v'_j, r'_{ji}, d_j, f). \square\)
Proof of Lemma 2.6. Assume that \((v'_j, r'_{ji}, d_j, f)\) is an optimal solution for (5). We now construct a feasible solution for program (14)-(22) from \((v'^l_j, r'^l_{ji}, d_j, f)\) when the number of client is \(t \cdot n\).

- \(\bar{v}_j^l = \sum_{s=(j-1)n+1}^{jn} v'^l_s, \forall j = 1, \ldots, t;\)
- \(\bar{r}_{ji}^l = \sum_{s=(j-1)n+1}^{jn} r'^l_{s, in}, \forall 1 \leq j \leq i \leq t;\)
- \(\bar{d}_j = \sum_{s=(j-1)n+1}^{jn} d_s, \forall j = 1, \ldots, t;\)
- \(\bar{f} = f.\)

We only prove that \((\bar{v}_j^l, \bar{r}_{ji}^l, \bar{d}_j, \bar{f})\) satisfies (17) and (18) since it satisfies (15)–(16) and (19)–(22) evidently.

We first show the constraint (17) holds.

For each \(a = 1, \ldots, n\), we have

\[
\sum_{i=1}^{i-1} r'_{(j-1)n+a, (i-1)n+a} + d_{(i-1)n+a} + d_{(j-1)n+a} \leq r'^l_{(j-1)n+a, (i-1)n} + d_{(i-1)n+a} + d_{(j-1)n+a}.
\]

Thus,

\[
\bar{v}_i^l = \sum_{s=(i-1)n+1}^{in} v'^l_s \leq \sum_{a=1}^{n} \left( r'^l_{(j-1)n+a, (i-1)n} + d_{(i-1)n+a} + d_{(j-1)n+a} \right) = \sum_{s=(j-1)n+1}^{jn} r'^l_{s, (i-1)n} + \sum_{s=(i-1)n+1}^{in} d_s + \sum_{s=(j-1)n+1}^{jn} d_s = r'^l_{j, i-1} + \bar{d}_i + \bar{d}_j.
\]

Next, we will show the constraint (18) holds. On one hand, we have

\[
\sum_{j=1}^{i-1} (r'^l_{ji} - \bar{d}_j)_+ = \sum_{j=1}^{i-1} \left( \sum_{s=(j-1)n+1}^{jn} r'^l_{s, in} - \sum_{s=(j-1)n+1}^{jn} d_s \right)_+.
\]
\begin{align*}
\sum_{j=1}^{i-1} \sum_{s=(j-1)n+1}^{jn} (r_{s,in}^l - d_s) + \\
= \sum_{j=1}^{(i-1)n} (r_{j,in}^l - d_j). \quad (24)
\end{align*}

The first equality follows from the definitions of \( r_{ji}^l \) and \( d_j \). The second holds since \( (a+b)_+ \leq (a)_+ + (b)_+ \). On the other hand, we have

\begin{align*}
\theta t \sum_{j=i}^{t} (\bar{v}_{i}^l - \bar{d}_j) + &= \theta t \sum_{j=i}^{t} \left( \sum_{s=(i-1)n+1}^{in} v_{s}^l - \sum_{s=(j-1)n+1}^{jn} d_s \right) + \\
&\leq \theta t \sum_{j=i}^{t} \left( n\bar{v}_{in} - \sum_{s=(j-1)n+1}^{jn} d_s \right) + \\
&= \theta t \sum_{j=i}^{t} \left( \sum_{s=(j-1)n+1}^{jn} (v_{in} - d_s) \right) + \\
&\leq \theta t \sum_{j=i}^{t} \sum_{s=(j-1)n+1}^{jn} (v_{in}^l - d_s) + \\
&= \theta t \sum_{j=(i-1)n+1}^{tn} (v_{in}^l - d_j). \quad (25)
\end{align*}

The first inequality holds since \( v_{s}^l \leq v_{in}^l \) for all \( s \in [(i-1)n+1, in] \). Moreover, from (24) and (25), we also have

\begin{align*}
\sum_{j=1}^{i} (\bar{r}_{ji}^l - \bar{d}_j) + &+ \theta t \sum_{j=i+1}^{t} (\bar{v}_{i}^l - \bar{d}_j) + \\
= &\sum_{j=1}^{i-1} (\bar{r}_{ji}^l - \bar{d}_j) + + (r_{ii}^l - d_i) + + \theta t \sum_{j=i+1}^{t} (\bar{v}_{i}^l - \bar{d}_j) + \\
&\leq \sum_{j=1}^{i-1} (\bar{r}_{ji}^l - \bar{d}_j) + + (\bar{v}_{i}^l - \bar{d}_i) + + \theta t \sum_{j=i+1}^{t} (\bar{v}_{i}^l - \bar{d}_j) + \\
&\leq \sum_{j=1}^{(i-1)n} (\bar{r}_{ji}^l - \bar{d}_j) + + \theta t \sum_{j=1}^{t} (\bar{v}_{i}^l - \bar{d}_j) + \\
&\leq \sum_{j=1}^{(i-1)n} (r_{j,in}^l - d_j) + + \theta t \sum_{j=(i-1)n+1}^{tn} (v_{in}^l - d_j) + \\
&\leq f. \quad (26)
\end{align*}

Thus, \( (\bar{v}_{j}^l, \bar{r}_{ji}^l, \bar{d}_j, \bar{f}) \) is a feasible solution for program (14)-(22).
Now, we prove the desired $\rho_t \geq \eta_{tn}$. Since $r^l_{ji}$ satisfies (7), we obtain that

$$\sum_{s=(i-1)n+1}^{in} r^l_{s, in} \leq \sum_{s=(i-1)n+1}^{in} r^l_{ss}.$$

Taking summation of the above inequality over 1 to $t$ leads to

$$\sum_{i=1}^{t} \sum_{s=(i-1)n+1}^{in} r^l_{s, in} \leq \sum_{j=1}^{tn} r^l_{jj},$$

or equivalently,

$$\sum_{i=1}^{t} r^l_{ii} \leq \sum_{j=1}^{tn} r^l_{jj}.$$

Together with $\sum_{i=1}^{t} \bar{v}^l_i = \sum_{j=1}^{tn} v_j$, we have the desired result:

$$\rho_t \geq \frac{\sum_{l=1}^{L} \sum_{i=1}^{t} \theta_l (\bar{v}^l_i - \bar{r}^l_{ii}) + \bar{r}^l_{ii} - \gamma f \bar{f}}{\sum_{i=1}^{t} d_i} \geq \frac{\sum_{l=1}^{L} \sum_{j=1}^{tn} \theta_l (v^l_j - r^l_{ij}) + r^l_{ij} - \gamma f \bar{f}}{\sum_{i=1}^{tn} d_i} = \eta_{tn}.$$

The first inequality follows since $(\bar{v}^l_j, \bar{r}^l_{ji}, \bar{d}_j, \bar{f})$ is a feasible solution for program (14)-(22). □