# Analysis of a Local Search Heuristic for Facility Location Problems*

*Madhukar R. Korupolu*[†] *C. Greg Plaxton*[‡] *Rajmohan Rajaraman*[§]

January 16, 2000

**Abstract**

In this paper, we study approximation algorithms for several NP-hard facility location problems. We prove that a simple local search heuristic yields polynomial-time constant-factor approximation bounds for metric versions of the uncapacitated $k$-median problem and the uncapacitated facility location problem. (For the $k$-median problem, our algorithms require a constant-factor blowup in the parameter $k$.) This local search heuristic was first proposed several decades ago, and has been shown to exhibit good practical performance in empirical studies. We also extend the above results to obtain constant-factor approximation bounds for the metric versions of capacitated $k$-median and facility location problems.

## 1 Introduction

Suppose that you plan to build a new chain of widget stores in a given city, and you have identified potential store sites in a number of different neighborhoods. Further assume that the demand for widgets in each neighborhood of the city is known. If you want to build exactly $k$ stores, where should you locate them in order to minimize the average traveling distance of your customers? If instead you are willing to build any number of stores, and the cost of building a store at each potential site is known, where should you build stores in order to minimize the sum of the total construction cost and the average traveling distance of your customers?

The preceding questions are examples of *facility location problems.* Facility location problems model applications in diverse fields including public policy (e.g., locating fire stations in a city), telecommunications (e.g., locating base stations in wireless networks), and information retrieval (e.g., clustering of documents), and have been studied extensively over the past three decades. For more information on the applications of facility location problems, we refer the reader to [21, 22].

This paper is concerned with metric versions of the following specific facility location problems: (i) the uncapacitated $k$-median problem (UKM), (ii) the uncapacitated facility location problem (UFL), (iii) the capacitated $k$-median problem with unsplittable demands (CKMU), (iv) the capacitated facility location problem with unsplittable demands (CFLU), (v) the capacitated $k$-median problem with splittable demands (CKMS), and (vi) the capacitated facility location problem with splittable

*A preliminary version of this paper appears in Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 1–10.

†Department of Computer Science, University of Texas, Austin, TX 78712. Email: madhukar@cs.utexas.edu. Supported by the National Science Foundation under Grant No. CCR–9504145.

‡Department of Computer Science, University of Texas, Austin, TX 78712. Email: plaxton@cs.utexas.edu. Supported by the National Science Foundation under Grant No. CCR–9504145.

§College of Computer Science, Northeastern University, Boston MA 02115. Email: rraj@ccs.neu.edu. Part of this work was done when the author was at DIMACS, Rutgers University, Piscataway, NJ 08854. DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999 and partially supported by the New Jersey Commission on Science and Technology.

demands (CFLS). Formal definitions of these problems, all of which are known to be NP-hard, are given in Section 2.

Informally, UKM (resp., UFL) corresponds to the first (resp., second) of the two questions stated above, where we assume that there is no upper bound on the demand that can be satisfied by a given store, and hence the traveling distance of a customer is just the distance from that customer's neighborhood to the nearest store. In the capacitated versions of these problems, there is an upper bound on the demand that can be satisfied by a given store, and we need to exhibit an assignment of customers to stores that minimizes the average traveling distance while respecting the capacity bound of each store. In assigning the demand of customers to stores, there are two natural variations to consider: (i) the demand of a customer must be met by a single store (unsplittable demands), and (ii) the demand of a customer may be divided across any number of stores (splittable demands). The assumption of unsplittable demands yields the problems CKMU and CFLU, while the assumption of splittable demands yields the problems CKMS and CFLS.

The input to a facility location problem includes a distance matrix specifying the traveling distance from node $i$ to node $j$, for all $i$ and $j$. (The nodes correspond to the neighborhoods in our widget chain example.) In the *metric version* of a facility location problem, it is assumed that these distances are nonnegative, symmetric, and satisfy the triangle inequality. All of our results are for metric facility location problems.

The local search heuristic for facility location problems, described more formally in Section 3, is extremely straightforward. The idea is to start with any feasible solution (set of stores) and then to iteratively improve the solution by repeatedly moving to the best "neighboring" feasible solution, where one solution is a neighbor of another if it can be obtained by either adding a facility (store), deleting a facility, or changing the location of a facility. This heuristic was proposed by Kuehn and Hamburger [18], and was subsequently shown to exhibit good practical performance in empirical studies (see, e.g., [9, 26]).

For the $k$-median problems we consider, namely, UKM, CKMU, and CKMS, we define an $(a, b)$-*approximation algorithm* as a polynomial-time algorithm that computes a solution using at most $bk$ facilities and with cost at most $a$ times the cost of an optimal solution using at most $k$ facilities. For facility location variants UFL, CFLU, and CFLS, we define an *$a$-approximation algorithm* as a polynomial-time algorithm that computes a solution with cost at most $a$ times optimal. Using this terminology, and letting $n$ denote the number of nodes in the input instance, we now summarize the previous approximation results known for the facility location problems considered in this paper: (i) Hochbaum [12] showed that a simple greedy algorithm is an $O(\log n)$-approximation algorithm for UFL with general distances (i.e., not restricted to metric distances); (ii) Lin and Vitter [20] gave a $(1+\varepsilon, (1+1/\varepsilon)(\ln n + 1))$-approximation algorithm for UKM with general distances (their algorithm can also be adapted to match Hochbaum's result); (iii) Lin and Vitter [19] gave a $(2(1+\varepsilon), 1+(1/\varepsilon))$-approximation algorithm for metric UKM; (iv) Shmoys et al. [24] gave a 3.16-approximation algorithm for metric UFL, a 7-approximation algorithm for metric CFLS that requires a blowup of 7/2 in the capacity of each facility, and a 9-approximation algorithm for metric CFLU that requires a blowup of 4 in the capacity of each facility. More recently, the result of [24] for metric UFL has been improved by Guha and Khuller [11] who give a 2.41-approximation algorithm. It has also been recently shown [11, 25] that there is no polynomial-time 1.463-approximation algorithm for metric UFL unless P = NP.

**Our results**. The approximation algorithms of Lin and Vitter [19, 20] and Shmoys et al. [24] are based on careful rounding of fractional solutions to *linear programming relaxations*. From a practical standpoint, one drawback of this approach is that the best known polynomial-time optimal algorithms for linear programming [15, 16] are quite complicated. Recently, however, fully polynomial-time approximation schemes (FPTAS) have been derived for fractional packing/covering linear programs [10], which can be applied to certain uncapacitated facility location problems, as shown by Young [28] for

| *Problem* | *Bound* | *Reference* | Approach |
|---|---|---|---|
| general UKM | $(1+\varepsilon, (1+1/\varepsilon)(\ln n+1))$ | [20] | rounding of LP-relaxation |
| | $(1+\varepsilon, \ln(n+n/\varepsilon))$ | [28]* | oblivious rounding & greedy |
| metric UKM | $(2(1+\varepsilon), 1+(1/\varepsilon))$ | [19] | rounding of LP-relaxation |
| | $(1+\varepsilon, 3+5/\varepsilon)$ | this paper | local search |
| | $(1+5/\varepsilon, 3+\varepsilon)$ | this paper | local search |
| | $(20/3, 1)$ | [4]* | rounding of LP-relaxation |
| | $(6, 1)$ | [14]* | primal-dual method |
| | $(4, 1)$ | [3]* | hybrid |
| Euclidean UKM | $(1+\varepsilon, 1)$ | [2]* | dynamic programming |
| general UFL | $O(\log n)$ | [12] | greedy algorithm |
| metric UFL | $\approx 3.16$ | [24] | rounding of LP-relaxation |
| | $\approx 2.41$ | [11] | LP-relaxation & greedy |
| | $5+\varepsilon$ | this paper | local search |
| | $1+2/e \approx 1.74$ | [5]* | rounding of LP-relaxation |
| | $3$ | [14]* | primal-dual method |
| | $\approx 1.728$ | [3]* | hybrid |
| | $\approx 2.414+\varepsilon$ | [3]* | extended local search |

Table 1: Approximation bounds for uncapacitated facility location problems. The notation "$\approx$" used in the entries for metric UFL means "approximately equal to". All of the results marked with an asterisk have been obtained subsequent to our results.

UKM with general distances[1]. In this paper, we focus our attention on the simple *combinatorial* paradigm of local search. We prove that the local search heuristic yields simple algorithms for UKM and UFL with approximation bounds matching (to within constant factors) the corresponding results in (iii) and (iv) above. In particular, for metric UKM, we prove that for any constant $\varepsilon > 0$, local search yields $(a, b)$-approximation bounds with $(a, b) = (1+\varepsilon, 3+5/\varepsilon)$ and $(a, b) = (1+5/\varepsilon, 3+\varepsilon)$. For metric UFL, we prove that for any constant $\varepsilon > 0$, local search yields a $(5+\varepsilon)$-approximation algorithm.

Our results for the capacitated facility location problems are similar. For metric CKMS, we prove that for any constant $\varepsilon > 0$, local search yields $(a, b)$-approximation bounds with $(a, b) = (1+\varepsilon, 5+5/\varepsilon)$ and $(a, b) = (1+5/\varepsilon, 5+\varepsilon)$ with no blowup in the capacity of each facility. For metric CFLS, we prove that for any constant $\varepsilon > 0$, local search yields an $(8+\varepsilon)$-approximation algorithm with no blowup in the capacity of each facility. Using the reduction of [24] that is based on solving a special case of the generalized assignment problem [23], the above results for the capacitated problems with splittable demands imply constant-factor approximation bounds for the capacitated problems with unsplittable demands (CKMU and CFLU) assuming a factor of 2 blowup in the capacity of each facility.

**Recent related work**. Subsequent to the publication of our results, several new approximation algorithms have been obtained for the $k$-median and the facility location problems. For the $k$-median problem, one shortcoming of the approximation bounds that we have discussed thus far is that while the optimal solution uses at most $k$ facilities, the number of facilities used in the approximate solutions may exceed $k$ by a constant factor. In recent work, Charikar et al. [4] have obtained the first constant-factor approximation algorithm for metric UKM without incurring a blowup in the number of facilities.

[1]Another noteworthy result in the class of LP-based algorithms is an *oblivious rounding* technique of Young [27], in which the fractional solution to the LP is rounded without solving the LP. This technique has yielded improved approximations for the general $k$-median problem [28].

| *Problem* | *Bound* | *Capacity Blowup* | *Reference* | *Approach* |
|---|---|---|---|---|
| metric CKMS | $(1+\varepsilon, 5+5/\varepsilon)$ | none | this paper | local search |
| | $(1+5/\varepsilon, 5+\varepsilon)$ | none | this paper | local search |
| | $(16,1)$ | 3 | [4] | rounding of LP-relaxation |
| metric CFLS | 7 | 7/2 | [24] | rounding of LP-relaxation |
| | $8+\varepsilon$ | none | this paper | local search |
| | $6+\varepsilon$ | none | [7]* | local search |
| | $3+2\sqrt{2}+\varepsilon$ | none | [3]* | hybrid |
| metric CKMU | $(1+\varepsilon, 5+5/\varepsilon)$ | 2 | this paper | local search |
| | $(1+5/\varepsilon, 5+\varepsilon)$ | 2 | this paper | local search |
| | $(16,1)$ | 4 | [4]* | rounding of LP-relaxation |
| metric CFLU | 9 | 4 | [24] | rounding of LP-relaxation |
| | $16+\varepsilon$ | 2 | this paper | local search |

Table 2: Approximation bounds for capacitated facility location problems. All of the results marked with an asterisk have been obtained subsequent to our results.

Their algorithm, which relies on rounding the fractional solution of an appropriate linear programming relaxation, yields a $(20/3, 1)$-approximation. Arora et al. [2] give a polynomial-time approximation scheme for the special case of metric UKM in which the distances are Euclidean. The best known approximation ratio for UFL is achieved by the recent 1.728-approximation algorithm due to Charikar and Guha [3] that uses a hybrid approach by combining an LP-based 1.74-approximation algorithm due to Chudak [5] with a new local improvement algorithm in which multiple facilities may be dropped in a single local search step. It is also shown in [3] that this more general local search paradigm, by itself, yields a $(2.414+\varepsilon)$-approximation for UFL. Recent work also includes fast $(6,1)$- and 3-approximation algorithms for UKM and UFL, respectively, that Jain and Vazirani [14] have obtained using an elegant primal-dual approach.

Improved approximation algorithms have been recently obtained for the capacitated versions as well. Charikar et al. [4] give $(16,1)$-approximation algorithms to both CKMS and CKMU while incurring a blowup in capacity by factors of 3 and 4, respectively.

All of the approximation bounds discussed above are summarized in Tables 1 and 2. We remark that the analysis presented in this paper is somewhat sharper than that provided in the preliminary version [17]; as a result, we have obtained constant-factor improvements in all of our bounds. In a recent paper, Chudak and Williamson [7] have simplified and improved our analysis of the local search heuristic for CFLS to show that the same heuristic yields a $(6+\varepsilon)$-approximation for CFLS, for any constant $\varepsilon > 0$. More recently, Charikar and Guha [3] have shown that the same heuristic combined with a scaling of facility costs improves the approximation bound to $3+2\sqrt{2}+\varepsilon$.

The facility location problems that we consider in this paper are closely related to a number of other optimization problems (e.g., the $k$-center problem) that have been studied in the computer science and operations research literature. We refer the reader to [8, 13] for more information about other problems related to facility location.

The rest of the paper is organized as follows. Section 2 formally defines the six facility location problems considered in this paper. Section 3 defines the local search heuristic that we analyze. Section 4 introduces preliminary definitions for the analysis of uncapacitated problems. Sections 5 and 6 analyze the uncapacitated $k$-median and facility location problems, respectively. Section 7 introduces additional definitions for the capacitated problems. Sections 8 and 9 analyze the capacitated $k$-median problem and the capacitated facility location problem with splittable demands, respectively. Section 10 extends these results to capacitated problems with unsplittable demands.

## 2 Facility Location Problems

Let $N = \{1, \ldots, n\}$ be a set of locations and $F \subseteq N$ be a set of locations at which we may open a facility. Each location $j$ in $N$ has a demand $d_j$ that must be shipped to $j$. For any two locations $i$ and $j$, let $c_{ij}$ denote the cost of shipping a unit of demand from $i$ to $j$. We assume that the costs are nonnegative, symmetric, and satisfy the triangle inequality. In each problem studied in this paper, we wish to determine a set of open facilities and an assignment of locations to open facilities such that a given objective function is minimized.

**The uncapacitated $k$-median problem (UKM).** In UKM, we seek a set of at most $k$ open facilities and an assignment of locations to open facilities such that the shipping cost of the solution is minimized, where the shipping cost associated with a set $S$ of open facilities and an assignment $\sigma : N \mapsto S$ is given by $\sum_{j \in N} d_j c_{j\sigma(j)}$. Given a set $S$ of open facilities, an assignment that minimizes the shipping cost may be obtained by assigning each location $j$ in $N$ to the closest open facility in $S$. Thus, any solution to UKM is completely characterized by the set of open facilities. For any set $S$ of open facilities, let $C_s(S)$ denote the shipping cost obtained by assigning each location to its closest open facility.

**The uncapacitated facility location problem (UFL).** In UFL, for each location $i$ in $F$, we are given a nonnegative cost $f_i$ of opening a facility at $i$. The problem is to determine a set of open facilities and an assignment of locations to facilities such that the sum of the cost of opening the facilities and the shipping cost is minimized. (The shipping cost is defined as in UKM.) As in UKM, given a set $S$ of open facilities, an assignment that minimizes the total cost is to assign each location $j$ in $N$ to the closest open facility in $S$. Therefore, any solution to this problem is characterized by the set of open facilities. Given any solution $S$ of open facilities, we let $C_s(S)$, $C_f(S)$, and $C(S)$ denote the shipping cost, facility cost, and the total cost, respectively, obtained by assigning each location to its closest open facility. (Formally, $C_f(S) = \sum_{i \in S} f_i$ and $C(S) = C_s(S) + C_f(S)$.)

**The capacitated versions.** We now define *capacitated* variants of the above two problems in which there is a bound $M$ on the total demand that can be shipped from any facility. The capacitated variants are defined in the same manner as the uncapacitated variants except that any solution must satisfy the additional constraint imposed by the capacity $M$. That is, a solution given by a set $S$ of open facilities and an assignment $\sigma$ is feasible only if the capacity constraints are respected at all facilities in $S$. We note that we allow at most one facility to be opened at any location. For recent work that considers capacitated facility location problems in which multiple facilities may be opened at any location, see [6, 7].

There are two natural variants of capacitated location problems: (i) *splittable demands*, where the demand of each location can be split across more than one facility, and (ii) *unsplittable demands*, where the demand of each location has to be shipped from a single facility. For the capacitated problems with splittable demands (that is, CKMS and CFLS), given a set $S$ of open facilities, an assignment is given by a function $\sigma : N \times S \mapsto \mathbf{R}$, where $\sigma(j, i)$ denotes the amount of demand shipped from facility $i$ to location $j$. For the capacitated problems with unsplittable demands (that is, CKMU and CFLU), an assignment is given by a function $\sigma : N \mapsto S$, as in the uncapacitated problems.

In the uncapacitated problems, given a set of open facilities, an optimal assignment is obtained by simply assigning each location to its closest open facility. In a capacitated problem, however, such an assignment may violate the capacity constraint. Fortunately, for the capacitated problems with splittable demands, given a set $S$ of open facilities, an optimal assignment can be computed in polynomial time by solving an appropriately defined instance of the transportation problem [24]. Thus, for the capacitated variants with splittable demands, any solution is completely characterized by the set $S$ of open facilities. For the capacitated problems with unsplittable demands, however, it is NP-hard to compute an optimal assignment for a given set $S$ of open facilities. Our results for the capacitated problems with unsplittable demands follow from a reduction of [24] to the corresponding capacitated problem with splittable demands. Finally, we note that in the $k$-median problems, the total cost for a solution $S$ is same as the shipping cost for $S$. Hence, we use $C(S)$ and $C_s(S)$ interchangeably for the $k$-median problems.

## 3 The Local Search Paradigm

All of the algorithms considered in this paper are based on the following framework. We start with an arbitrary feasible solution and repeatedly improve the solution by performing a local search step. We show that within a polynomial number of local search steps, we arrive at a solution achieving the desired approximation factor. We now describe the local search step. The problem of finding an initial feasible solution is discussed at the end of this section.

Let us consider the uncapacitated problems UKM and UFL and the capacitated problems with splittable demands CKMS and CFLS. Recall that any solution to these problems is completely characterized by the associated set $S$ of open facilities. For the $k$-median problem, we define the *neighborhood* of $S$ as $\{T \subseteq F : |S - T| = |T - S| = 1\}$. For the facility location problem, we define the neighborhood of $S$ as $\{T \subseteq F : |S - T| \leq 1 \text{ and } |T - S| \leq 1\}$. Given a current solution corresponding to a set $S$ of open facilities, the local search step sets the new solution to be a minimum-cost set $T$ in the neighborhood of $S$. Since the neighborhood contains $O(n^2)$ solutions, a local search step can be performed in polynomial time.

The main motivation behind using the local search step is the following key property of the neighborhood of any solution. Let $S^*$ be an optimal solution. Then, for a sufficiently large constant $\delta$, given any solution $S$ such that $C(S) > (1 + \delta)C(S^*)$ (for the $k$-median problems, we also require that $|S| = (1 + \alpha)k$ for some sufficiently large constant $\alpha$), there exists a solution $T$ in the neighborhood of $S$ such that $C(T) \leq C(S)(1 - 1/p(n))$, where $p(n)$ is a fixed polynomial in $n$. (This property is proved in Theorems 5.1, 6.1, 8.1, and 9.1 for UKM, UFL, CKMS, and CFLS, respectively.)

Thus, if we start with a solution $S_0$ and perform $p(n)\log(C(S_0)/C(S^*))$ local search steps, we will arrive at a solution that has cost at most $(1 + \delta)C(S^*)$. Since $\log(C(S_0))$ is polynomial in the input size, the number of local search steps involved is also polynomial in the input size. Moreover, each local search step takes polynomial time. We thus obtain a polynomial-time algorithm achieving an approximation factor of $(1 + \delta)$, for a sufficiently large constant $\delta$.

**Finding an Initial Feasible Solution**. We now show how to obtain an initial feasible solution for the uncapacitated problems and the capacitated problems with splittable demands. (Recall that our results for the capacitated problems with unsplittable demands are obtained via a reduction to the corresponding capacitated problem with splittable demands.) As mentioned in Section 2, an initial feasible solution for an uncapacitated problem or a capacitated problem with splittable demands is completely specified by a set $S$ of open facilities. For a facility location problem, we select an arbitrary subset $S$ of $F$. For a $k$-median problem, we require that $|S| = (1 + \alpha)k$ for some sufficiently large constant $\alpha$. In the extreme case where the total number of available facilities, $|F|$, is less than $(1+\alpha)k$, our algorithm trivially returns $F$ as the solution. Hence throughout this paper, we assume without loss of generality that $|F|$ is at least $(1 + \alpha)k$ for all $k$-median problems. Given this assumption, we obtain an initial feasible solution for the $k$-median problems by selecting an arbitrary set $S$ of $(1+\alpha)k$ facilities.

## 4 Preliminaries for Uncapacitated Problems

In this section we introduce some notations and definitions that are useful for our analysis of UKM in Section 5 and UFL in Section 6. Given a set $S$ of open facilities and an assignment $\sigma$ for $S$, let $N_i(S, \sigma)$ denote the set of locations assigned to a facility $i$ in $S$. A location $j$ is called a *client* of facility $i$, under $\sigma$, if $\sigma$ assigns $j$ to $i$. Let $D_i(S, \sigma)$ denote the total demand shipped from a facility $i$ in $S$ under assignment $\sigma$. For notational convenience, the parameters $S$ and/or $\sigma$ may be dropped from the expressions $N_i(S,\sigma)$ and $D_i(S,\sigma)$ when there is no ambiguity.

Let $S^*$ (resp., $S$) denote an optimal (resp., current) set of open facilities for the problem instance under consideration, and let $\sigma^*$ (resp., $\sigma$) denote an optimal assignment for $S^*$ (resp., $S$). Given $S$, $S^*$, $\sigma$, and $\sigma^*$, we define $g_i$ and $g_i^*$ (for each facility $i$ in $S$) as follows: $g_i$ denotes the shipping cost paid by the clients of $i$, that is, $g_i = \sum_{j \in N_i(S,\sigma)} d_j c_{ji}$. Correspondingly, $g_i^*$ denotes the shipping cost paid by the same set of clients (i.e., $N_i(S,\sigma)$) in the optimal solution, that is, $g_i^* = \sum_{j \in N_i(S,\sigma)} d_j c_{j\sigma^*(j)}$. Note that $C_s(S) = \sum_{i \in S} g_i$ and $C_s(S^*) = \sum_{i \in S} g_i^*$.

## 5 Uncapacitated $k$-Median

In this section, we analyze the local search heuristic for UKM. Throughout this section, we assume that $p(n)$ is a polynomial in $n$ and that $\alpha$ and $\beta$ are positive constants satisfying the following condition:

$$(1+\beta)\left(1 - \frac{2}{\alpha} - \frac{k}{p(n)}\right) \geq 1 + \frac{2}{\alpha}. \tag{1}$$

One possible choice, for example, is $\alpha = 4$, $\beta = 3$, and $p(n) = 8k$. The main result of this section is that, given a solution $S$ of $(1+\alpha)k$ facilities with cost greater than $(1+\beta)C(S^*)$, we can perform a swapping of facilities to get another solution with the same size but significantly reduced cost.

**Theorem 5.1 (Swapping Facilities)** *Let $S$ be any subset of $F$ such that $|S| = (1+\alpha)k$ and $C(S) > (1+\beta)C(S^*)$. Then there exist $u$ in $S$ and $v$ in $F$ such that $C(S) - C(S+v-u) \geq \frac{C(S)}{p(n)}$.*

We prove the theorem in two stages. First, in Lemma 5.2 we show that we can add a facility to $S$ and get a significant reduction in the cost. Then, in Lemma 5.3, we show that we can drop a facility from $S$ without increasing the cost too much. Combining these two lemmas then yields the above theorem.

**Lemma 5.2 (Adding a Facility)** *Let $S$ be any subset of $F$ such that $C(S) > C(S^*)$. Then there exists $v$ in $F$ such that $C(S) - C(S+v) \geq \frac{C(S)-C(S^*)}{k}$.*

**Proof:** Recall that $\sigma$ and $\sigma^*$ denote the optimal assignments for $S$ and $S^*$, respectively. We have

$$C(S) - C(S^*) = \sum_{j \in N} d_j (c_{j\sigma(j)} - c_{j\sigma^*(j)}) = \sum_{i \in S^*} \sum_{j \in N_i(S^*)} d_j (c_{j\sigma(j)} - c_{ji}).$$

Since $|S^*| = k$, it follows from a simple averaging argument that there exists $v$ in $S^*$ for which

$$\sum_{j \in N_v(S^*)} d_j (c_{j\sigma(j)} - c_{jv}) \geq \frac{C(S) - C(S^*)}{k}.$$

Now consider the set $S+v$ along with the assignment $\sigma'$ defined as follows: Set $\sigma'(j)$ to $v$ if $j$ belongs to $N_v(S^*)$, and to $\sigma(j)$ otherwise. Using the fact that $C(S+v,\sigma') \geq C(S+v)$, we obtain

$$C(S) - C(S+v) \geq C(S) - C(S+v,\sigma') = \sum_{j \in N_v(S^*)} d_j (c_{j\sigma(j)} - c_{jv}) \geq \frac{C(S)-C(S^*)}{k}.$$

■

**Lemma 5.3 (Dropping a Facility)** *Let $S$ be any subset of $F$ such that $|S| = (1+\alpha)k$. Then there exists $u$ in $S$ such that $C(S-u) - C(S) \leq 2 \cdot \frac{C(S)+C(S^*)}{\alpha k}$.*

Before proceeding to prove this lemma, we first classify the facilities in $S$ as "primary" or "secondary" as follows. For each facility $i^*$ in $S^*$, consider those facilities of $S$ that are assigned to $i^*$ by $\sigma^*$ and pick the closest one among these, breaking ties arbitrarily. The closest facility thus chosen is said to be the *primary* facility of $i^*$. Let $S_P$ denote the set of all primary facilities. Facilities which are not primary are said to be *secondary*. For each secondary facility $i$ in $S$, we define the associated primary facility as the primary facility of $\sigma^*(i)$. Note that the number of secondary facilities is at least $|S| - |S^*|$, which is $\alpha k$. We will show that one of the secondary facilities can be dropped from $S$ without incurring a "large" increase in the shipping cost. As a first step in this direction, the following claim gives an upper bound on the increase in the shipping cost if a client of a secondary facility $i$ is reassigned to the primary facility associated with $i$. We note that similar arguments bounding the increase in shipping cost due to reassignment of demand have also been used in the rounding process in LP-based algorithms for facility location [5, 6, 24].

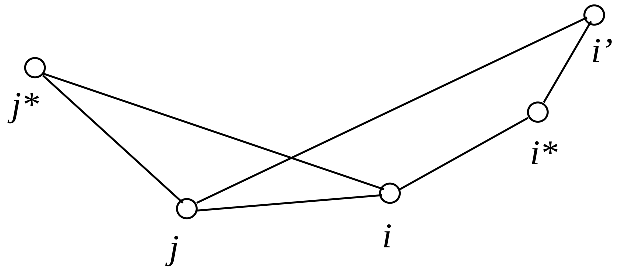


Figure 1: Illustration for Claim 5.4. Here, $i^* = \sigma^*(i)$, $j^* = \sigma^*(j)$, and $i'$ is the primary facility of $i^*$.

**Claim 5.4** *Consider a secondary facility $i$ in $S$, and let $i'$ be the primary facility associated with $i$. Then for all $j$ in $N$, we have $c_{ji'} - c_{ji} \leq 2(c_{ji} + c_{j\sigma^*(j)})$.*

**Proof:** Let $i^*$ denote $\sigma^*(i)$, the facility in $S^*$ that is closest to $i$. (See Figure 1.) Using the triangle inequality, we upper bound $c_{ji'}$ as follows:

$$c_{ji'} \leq c_{ji} + c_{ii^*} + c_{i^*i'} \leq c_{ji} + 2c_{ii^*} \leq c_{ji} + 2c_{i\sigma^*(j)} \leq c_{ji} + 2c_{ij} + 2c_{j\sigma^*(j)}.$$

(The second step uses the fact that $i'$ is the primary associated with $i^*$, i.e., $c_{i^*i'} \leq c_{ii^*}$. The third step holds since $i^*$ is the closest to $i$ among all facilities in $S^*$. The fourth step uses the triangle inequality.) The claim now follows. ■

**Claim 5.5** *Consider a secondary facility $i$ in $S$ and let $i'$ be the primary facility associated with $i$. If we drop the facility $i$ from $S$ and reassign all clients of $i$ to $i'$, then the increase in the shipping cost is at most $2(g_i + g_i^*)$.*

**Proof:** It is easy to verify that the increase in shipping cost due to this reassignment is $\sum_{j \in N_i(S)} d_j(c_{ji'} - c_{ji})$. Using the upper bound on $c_{ji'} - c_{ji}$ given by Claim 5.4, and then plugging in the definitions of $g_i$ and $g_i^*$, completes the proof. ■

The following claim is needed only for the proof of Theorem 6.1 in Section 6, but is in the same spirit as the previous claim.

**Claim 5.6** *Consider any facility $i$ in $S$ and let $i^*$ denote $\sigma^*(i)$. If we replace facility $i$ in $S$ by $i^*$ and reassign all clients of $i$ to $i^*$, then the increase in the shipping cost is at most $g_i + g_i^*$.*

**Proof:** For every client $j$ of $i$, we can upper bound $c_{ji^*} - c_{ji}$ by $c_{ji} + c_{j\sigma^*(j)}$ using the same approach as in Claim 5.4. The remainder of the proof is similar to that of Claim 5.5. ∎

We are now ready to prove Lemma 5.3.

**Proof of Lemma 5.3:** Recall that $S_P$ is the set of primary facilities of $S$. Using the definitions of $g_i$ and $g_i^*$, we have

$$\sum_{i \in S - S_P} (g_i + g_i^*) \leq \sum_{i \in S} (g_i + g_i^*) = C_s(S) + C_s(S^*).$$

By an averaging argument, it follows that there exists a secondary facility $v$ for which $g_v + g_v^*$ is at most $\frac{C_s(S)+C_s(S^*)}{\alpha k}$ (since $|S - S_P| \geq \alpha k$). Applying Claim 5.5 for this facility $v$, it follows that $v$ satisfies the requirements of this dropping lemma. ∎

**Proof of Theorem 5.1:** We first apply Lemma 5.3 to drop a facility $u$ from $S$. We then apply Lemma 5.2 to add a facility $v$ to $S - u$. Combining the two operations, we have

$$C(S) - C(S - u + v) \quad \geq \quad \frac{C(S-u) - C(S^*)}{k} - 2 \cdot \frac{C(S) + C(S^*)}{\alpha k}.$$

It now suffices to show that the term on the right side is at least $\frac{C(S)}{p(n)}$. Since $C(S - u) \geq C(S)$, it suffices to show that

$$C(S)\left(1 - \frac{2}{\alpha} - \frac{k}{p(n)}\right) \geq C(S^*)\left(1 + \frac{2}{\alpha}\right).$$

Since $C(S) \geq (1+\beta)C(S^*)$ and $\alpha$, $\beta$, and $p(n)$ satisfy Equation (1), this inequality always holds. This completes the proof of the theorem. ∎

The corollary below is related to the question of how small the constants $\alpha$ and $\beta$ can be. Note that any choice of $\alpha$ and $\beta$ subject to the constraint of Equation (1) is valid. It is shown below that $\beta$ can be made arbitrarily small, while on the other hand, $\alpha$ can be made arbitrarily close to 2.

**Corollary 5.7** *For any constant $\varepsilon > 0$, there is a constant $\delta$ such that the following can be computed in polynomial time using the local search heuristic: (i) a solution $S$ with $C(S) \leq (1+\varepsilon)C(S^*)$ and $|S| \leq (1+\delta)k$, and (ii) a solution $S$ with $|S| \leq (3+\varepsilon)k$ and $C(S) \leq (1+\delta)C(S^*)$.*

**Proof:** It can be verified that the following choices of the constants satisfy the constraint of Equation (1): (i) $\beta = \varepsilon$, $\delta = \alpha = \frac{5}{\varepsilon} + 2$, and $p(n) = \alpha k(1+\varepsilon)$, and (ii) $\alpha = 2 + \varepsilon$, $\delta = \beta = 1 + \frac{5}{\varepsilon}$, and $p(n) = \frac{5\alpha k}{\varepsilon}$. The desired claims now follow from Theorem 5.1. ∎

The following corollary addresses the issue of making $\alpha$ arbitrarily small. The local improvement method requires $\alpha$ to satisfy the constraint of Equation (1) and hence it has to be above 2. However if we postprocess the solution from the local improvement phase by dropping facilities greedily, we can achieve arbitrarily small $\alpha$ while keeping the cost within a constant factor of optimal.

**Corollary 5.8** *For any constant $\varepsilon > 0$, there is a constant $\delta$ such that a solution $S$ with $|S| \leq (1+\varepsilon)k$ and $C(S) \leq (1+\delta)C(S^*)$ can be computed in polynomial time using the local search heuristic.*

**Proof:** Suppose we have a solution $S$ with $|S| = (1+\gamma)k$ and $C(S) \leq (1+\mu)C(S^*)$. For example, by part (ii) of Corollary 5.7, we can choose $\gamma = 2 + \varepsilon'$ and $\mu = 1 + \frac{5}{\varepsilon'}$ for any $\varepsilon' > 0$.

Lemma 5.3 implies that there exists a facility $u$ in $S$ that can be dropped without increasing the cost by more than $\frac{2C(S)+2C(S^*)}{\gamma k}$, which is at most $\frac{4C(S)}{\gamma k}$ assuming $C(S^*) \leq C(S)$. Therefore, the cost of $S - u$ is at most $C(S)(1 + \frac{4}{\gamma k})$. We repeat this dropping process until the number of facilities reduces to $(1+\varepsilon)k$. Let $\hat{S}$ be the set of facilities remaining at the end of this process. Since the $(t+1)$th facility dropped increases the cost by at most a multiplicative factor of $(1+\frac{4}{\gamma k-t})$, it follows that

$$\begin{aligned}
C(\hat{S}) &\leq C(S)\left[\left(1+\frac{4}{\gamma k}\right)\left(1+\frac{4}{\gamma k-1}\right)\cdots\left(1+\frac{4}{\varepsilon k+1}\right)\right] \\
&= C(S)\frac{(\gamma k+4)(\gamma k+3)(\gamma k+2)(\gamma k+1)}{(\varepsilon k+4)(\varepsilon k+3)(\varepsilon k+2)(\varepsilon k+1)} \\
&\leq C(S)\left(\frac{\gamma}{\varepsilon}\right)^4.
\end{aligned}$$

(The second step follows by expanding and canceling out the common terms in the numerator and the denominator. The third step uses the fact that $\frac{\gamma k+i}{\varepsilon k+i} \leq \frac{\gamma}{\varepsilon}$, whenever $i \geq 0$ and $\gamma \geq \varepsilon$.) Thus $\hat{S}$ is a solution with $|\hat{S}| = (1+\varepsilon)k$, and if $\gamma$ is chosen to be $2+\varepsilon'$ (for any $\varepsilon' > 0$) we have

$$C(\hat{S}) \leq C(S^*)\left(2+\frac{5}{\varepsilon'}\right)\left(\frac{2+\varepsilon'}{\varepsilon}\right)^4.$$

The corollary follows if we set the constant $\delta$ to $\left(2+\frac{5}{\varepsilon'}\right)\left(\frac{2+\varepsilon'}{\varepsilon}\right)^4 - 1$, for any $\varepsilon' > 0$.

■

## 6 Uncapacitated Facility Location

The main result of this section is Theorem 6.1, which shows that a local search step for UFL gives a significant improvement in the cost of the current solution if the cost of the current solution is sufficiently larger than the optimal cost. Throughout this section, we assume that $p(n)$ is a sufficiently large polynomial in $n$.

**Theorem 6.1** *Let $S$ be any subset of $F$ such that $C(S) > (5+\varepsilon)C(S^*)$, for some constant $\varepsilon > 0$. Then there exists $T \subseteq F$ such that $|S-T| \leq 1$, $|T-S| \leq 1$, and $C(S) - C(T) \geq \frac{C(S)}{p(n)}$.*

**Corollary 6.2** *For any constant $\varepsilon > 0$, the local search heuristic yields a $(5+\varepsilon)$-approximate solution in polynomial time.* ■

We prove the theorem in two stages. First, in Lemma 6.3, we show that if the shipping cost of $S$ is sufficiently large then we can add a facility to $S$ and obtain a significant improvement in the cost. Second, in Lemma 6.4, we show that if the facility cost of $S$ is sufficiently large then we can either drop a facility from $S$ or substitute a facility in $S$ to obtain a significant improvement in the cost. If none of adding, dropping, or substituting gives a good improvement in the cost, then we combine these two lemmas to obtain a contradictory upper bound on the total cost of $S$.

**Lemma 6.3 (Adding a Facility)** *Let $S$ be any subset of $F$ such that $C_s(S) \geq C(S^*) + \frac{nC(S)}{p(n)}$. Then there exists $v$ in $F$ such that $C(S) - C(S+v) \geq \frac{C(S)}{p(n)}$.*

**Proof:** It suffices to show that for any set $S \subseteq F$, there exists $v$ in $F$, such that $C(S) - C(S+v) \geq \frac{C_s(S)-C(S^*)}{n}$. The quantity $C_s(S) - C(S^*)$ can be expanded as follows:

$$\begin{aligned}
C_s(S) - C_s(S^*) - C_f(S^*) &= -C_f(S^*) + \sum_{j\in N} d_j(c_{j\sigma(j)} - c_{j\sigma^*(j)}) \\
&= \sum_{i\in S^*} -f_i + \sum_{i\in S^*}\sum_{j\in N_i(S^*)} d_j(c_{j\sigma(j)} - c_{ji}) \\
&= \sum_{i\in S^*}\left(-f_i + \sum_{j\in N_i(S^*)} d_j(c_{j\sigma(j)} - c_{ji})\right).
\end{aligned}$$

By a simple averaging argument, it follows that there exists a facility $v$ in $S^*$ such that

$$-f_v + \sum_{j \in N_v(S^*)} d_j (c_{j\sigma(j)} - c_{jv}) \geq \frac{C_s(S) - C(S^*)}{n}. \tag{2}$$

Now consider the solution $S + v$ along with the assignment $\sigma'$ defined as follows: If $j$ is in $N_v(S^*)$, then set $\sigma'(j)$ to $v$; otherwise set $\sigma'(j)$ to $\sigma(j)$. Since $C(S+v, \sigma')$ is at least $C(S+v)$, Equation (2) implies that

$$C(S) - C(S+v) \quad \geq \quad C(S) - C(S+v, \sigma') = -f_v + \sum_{j \in N_v(S^*)} d_j (c_{j\sigma(j)} - c_{jv}) \geq \frac{C_s(S) - C(S^*)}{n},$$

completing the proof of the lemma. ■

**Lemma 6.4 (Dropping/Swapping Facilities)** *Let $S$ be any subset of $F$. If no facility in $S$ can be dropped or substituted (by some facility not in $S$) to improve the cost by at least $\frac{C(S)}{p(n)}$, then the following upper bound on the facility cost of $S$ holds:*

$$C_f(S) \left(1 - \frac{n^2}{p(n)}\right) < 2C(S^*) + 2C_s(S) + \frac{C(S)}{n}.$$

We start by developing some useful tools for proving this lemma. A facility $i$ in $S$ is called *cheap* if $f_i \leq \frac{C(S)}{n^2}$, and *expensive* otherwise. Let $S_C$ denote the set of cheap facilities in $S - S^*$. Each facility in $S$ is also classified as *primary* or *secondary* in the same fashion as in Section 5. We let $S_{EP}$ (resp., $S_{ES}$) denote the set of expensive facilities in $S - S^* - S_C$ that are classified as primary (resp., secondary). Thus $S \cap S^*$, $S_C$, $S_{EP}$, and $S_{ES}$ partition the set $S$. Note that for each $i$ in $S_{EP}$, the facility $i^* = \sigma^*(i)$ cannot be in $S$; if $i^*$ belonged to $S$, then $i^*$, not $i$, would have been designated as the primary facility for $i^*$.

We now construct a set of *candidate operations*, one for each facility in $S$, as follows:

1. For $i$ in $S \cap S^*$, do nothing.
2. For $i$ in $S_C$, do nothing.
3. For $i$ in $S_{ES}$, drop $i$ and reassign the clients of $i$ to the primary facility $i'$ associated with $i$. Note that $i'$ is already in $S$.
4. For $i$ in $S_{EP}$, substitute $i$ with $i^* = \sigma^*(i)$, and reassign all clients of $i$ to $i^*$. Recall that $i^*$ does not belong to $S$.

A candidate operation is called *good* if it reduces the cost of the solution by at least $\frac{C(S)}{p(n)}$, and *bad* otherwise. By the hypothesis of Lemma 6.4, we can assume that all of the candidate operations are *bad*. Our main tools for proving Lemma 6.4 are Claims 6.5 and 6.6, which establish appropriate upper bounds on the facility cost of $i$ if the candidate operation for the facility is bad.

**Claim 6.5** *If the candidate operations for all of the expensive secondary facilities are bad, then*

$$C_f(S_{ES}) \left(1 - \frac{n^2}{p(n)}\right) \leq \sum_{i \in S_{ES}} 2(g_i + g_i^*).$$

**Proof:** Consider an expensive secondary facility $i$ in $S_{ES}$. The decrease in the facility cost caused by the associated candidate operation is $f_i$ and the increase in the shipping cost, by Claim 5.5, is at most $2(g_i + g_i^*)$. So the net decrease in the cost is at least $f_i - 2(g_i + g_i^*)$. If this quantity is greater than $\frac{n^2 f_i}{p(n)}$, then since $i$ is expensive, this candidate operation will be good. Therefore, if this candidate operation is not good, the following upper bound on the facility cost of $i$ holds: $f_i(1 - \frac{n^2}{p(n)}) < 2(g_i + g_i^*)$. Summing over all of the expensive secondary facilities yields the required upper bound on the facility cost of $S_{ES}$. ■

**Claim 6.6** *If the candidate operations for all of the expensive primary facilities are bad, then*

$$C_f(S_{EP})\left(1-\frac{n^2}{p(n)}\right) \leq C_f(S^*-S)+\sum_{i\in S_{EP}}(g_i+g_i^*).$$

**Proof:** Consider an expensive primary facility $i$ in $S_{EP}$. The decrease in the facility cost caused by its candidate operation is $f_i - f_{i^*}$ and the increase in the shipping cost, by Claim 5.6, is at most $g_i + g_i^*$. So the net decrease in the cost is at least $f_i - (g_i + g_i^* + f_{i^*})$. If this quantity is greater than $\frac{n^2 f_i}{p(n)}$, then since $i$ is expensive, this candidate operation is good. Therefore, if this candidate operation is not good, the following upper bound on the facility cost of $i$ holds: $f_i(1-\frac{n^2}{p(n)}) < g_i + g_i^* + f_{i^*}$. Summing over all of the expensive primary facilities yields

$$C_f(S_{EP})\left(1-\frac{n^2}{p(n)}\right) \quad < \quad \sum_{i\in S_{EP}}(g_i+g_i^*+f_{i^*}) < C_f(S^*-S)+\sum_{i\in S_{EP}}(g_i+g_i^*).$$

(The second step holds since $\sigma^*(i)$ belongs to $S^*-S$ for all $i$ in $S_{EP}$ and $\sigma^*(i)\neq\sigma^*(j)$ for all $i$ and $j$ in $S_{EP}$ such that $i \neq j$. Hence, each facility of $S^*-S$ occurs at most once in the summation.) ■

**Proof of Lemma 6.4:** The total facility cost $C_f(S)$ equals the sum of $C_f(S_C)$, $C_f(S_{EP})$, $C_f(S_{ES})$, and $C_f(S\cap S^*)$. Claims 6.5 and 6.6 place upper bounds on $C_f(S_{ES})$ and $C_f(S_{EP})$, respectively. We now consider $S_C$, the set of cheap facilities. Since the total number of facilities is at most $n$ and since the cost of each cheap facility is at most $C(S)/n^2$, it follows that

$$C_f(S_C) \leq \frac{C(S)}{n}. \tag{3}$$

Using the upper bounds from Equation (3), Claims 6.5 and 6.6, and adding $C_f(S\cap S^*)$ to both sides, we obtain

$$\begin{aligned} C_f(S)\left(1-\frac{n^2}{p(n)}\right) &< \frac{C(S)}{n}+C_f(S^*)+2\cdot\sum_{i\in S_{EP}\cup S_{ES}}(g_i+g_i^*) \\ &< \frac{C(S)}{n}+C_f(S^*)+2C_s(S)+2C_s(S^*) \\ &< \frac{C(S)}{n}+2C_s(S)+2C(S^*). \end{aligned}$$

(The second step follows from the definition of $g_i$ and $g_i^*$, and the fact that the sets $S_{EP}$ and $S_{ES}$ are disjoint.) ■

**Proof of Theorem 6.1:** If no facility can be added to $S$ to get a cost improvement of at least $\frac{C(S)}{p(n)}$, then by Lemma 6.3, the shipping cost of $S$ is upper-bounded as follows: $C_s(S) < C(S^*)+\frac{nC(S)}{p(n)}$. Similarly, if no facility of $S$ can be dropped or substituted to get a cost improvement of at least $\frac{C(S)}{p(n)}$, then by Lemma 6.4 we have the following upper bound on the facility cost of $S$:

$$C_f(S)\left(1-\frac{n^2}{p(n)}\right) \quad < \quad 2C(S^*)+2C_s(S)+\frac{C(S)}{n} < 4C(S^*)+nC(S)\left(\frac{2}{p(n)}+\frac{1}{n^2}\right).$$

(The second step follows from the upper bound on $C_s(S)$.) Adding the upper bounds on the shipping cost and the facility cost, we have

$$C(S)\left(1-\frac{n^2}{p(n)}\right) < 5C(S^*)+nC(S)\left(\frac{3}{p(n)}+\frac{1}{n^2}\right).$$

Rearranging, we obtain

$$C(S)\left(1-\frac{n^2}{p(n)}-\frac{3n}{p(n)}-\frac{1}{n}\right)<5C(S^*).$$

If $p(n)$ is chosen to be sufficiently large, then this contradicts the hypothesis (of Theorem 6.1) that $C(S)>(5+\varepsilon)C(S^*)$. Specifically, we need

$$(5+\varepsilon)\left(1-\frac{n^2}{p(n)}-\frac{3n}{p(n)}-\frac{1}{n}\right)>5.$$

This completes the proof of the theorem. ■

## 7 Preliminaries for Capacitated Problems with Splittable Demands

In this section we introduce some notations and definitions that are useful in our analysis of the capacitated problems with splittable demands, CKMS (see Section 8) and CFLS (see Section 9).

Recall that for the capacitated problems with splittable demands, given a set $S$ of open facilities, an assignment is given by a function $\sigma : N \times S \mapsto \mathbf{R}$, where $\sigma(j,i)$ denotes the amount of demand shipped from facility $i$ to location $j$. Given $S$ and an assignment $\sigma$, let $D_v(S,\sigma)$ denote the amount of demand shipped from a facility $v$ in $S$ under the assignment $\sigma$, that is, $D_v(S,\sigma)=\sum_{j\in N}\sigma(j,v)$. For notational convenience, we sometimes drop the parameters $S$ and/or $\sigma$ when there is no ambiguity.

Also recall that given a set $S$ of open facilities, the corresponding optimal splittable assignment $\sigma$ can be computed in polynomial time by a reduction to the transportation problem [24]. Hence, any solution to CKMS or CFLS is completely characterized by the set $S$ of open facilities.

We let $S^*$ (resp., $S$) denote an optimal (resp., current) solution for a given instance of CKMS or CFLS, and let $\sigma^*$ (resp., $\sigma$) denote an optimal assignment for $S^*$ (resp., $S$). Recall that $M$ denotes the upper bound on the capacity of any facility.

## 8 Capacitated $k$-Median with Splittable Demands

In this section we consider the capacitated $k$-median problem with splittable demands (CKMS) and show in Theorem 8.1 that, as in the uncapacitated case, each step of the local search heuristic yields a significant improvement in cost whenever the current cost is sufficiently higher than the optimal cost. Throughout this section, we assume that $p(n)$ is a polynomial in $n$ and that $\alpha$ and $\beta$ are positive constants satisfying the following constraint,

$$(1+\beta)\left(1-\frac{2}{\alpha}-\frac{k}{p(n)}\right)\geq 1+\frac{2}{\alpha}. \tag{4}$$

Note that this constraint is the same as that of Equation (1). Hence, as before, one possible choice for the parameters is $\alpha=4$, $\beta=3$, and $p(n)=8k$.

**Theorem 8.1 (Swapping Facilities)** *Let $S$ be any subset of $F$ such that $|S|=(3+\alpha)k$ and $C(S)>(1+\beta)C(S^*)$. Then there exist $u$ in $S$ and $v$ in $F$ such that $C(S)-C(S+v-u)\geq\frac{C(S)}{p(n)}$.*

As in the uncapacitated case, we prove this theorem in two stages. First, in Lemma 8.2, we show that we can add a facility to $S$ and get a significant reduction in the cost. Due to the capacity constraints, the proof of this lemma is more involved than in the uncapacitated case (Lemma 5.2). Section 8.1 gives the proof of this lemma after developing the required technical machinery. Then, in Lemma 8.3, we show that we can drop a facility from $S$ without increasing the cost too much. The proof of this lemma is given in Section 8.2.

**Lemma 8.2 (Adding a Facility)** *Let $S$ be any subset of $F$ such that $C(S)>C(S^*)$. Then there exists $v$ in $F$ such that $C(S)-C(S+v)\geq\frac{C(S)-C(S^*)}{k}$.*

**Lemma 8.3 (Dropping a Facility)** *Let $S$ be any subset of $F$ such that $|S| = (3+\alpha)k$. Then there exists $u$ in $S$ such that $C(S-u) - C(S) \leq 2 \cdot \frac{C(S)+C(S^*)}{\alpha k}$.*

Theorem 8.1 follows directly from Lemmas 8.2 and 8.3; the proof is along the same lines as the uncapacitated case (Theorem 5.1). We now turn to the questions of how small the constants $\alpha$ and $\beta$ can be. Recall that the constraint of Equation (4) to be satisfied by these constants is the same as in the uncapacitated case. Hence, we obtain the following corollaries to Theorem 8.1, which are analogous to Corollaries 5.7 and 5.8 in the uncapacitated case. The only difference is an additive term of $2k$ in the number of facilities used.

**Corollary 8.4** *For any constant $\varepsilon > 0$, there is a constant $\delta$ such that the following can be computed in polynomial time using the local search heuristic: (i) a solution $S$ with $C(S) \leq (1+\varepsilon)C(S^*)$ and $|S| \leq (1+\delta)k$, and (ii) a solution $S$ with $|S| \leq (5+\varepsilon)k$ and $C(S) \leq (1+\delta)C(S^*)$.* ■

**Corollary 8.5** *For any constant $\varepsilon > 0$, there is a constant $\delta$ such that a solution $S$ with $|S| \leq (3+\varepsilon)k$ and $C(S) \leq (1+\delta)C(S^*)$ can be computed in polynomial time.* ■

### 8.1 Adding a Facility

In this section, we prove Lemma 8.2. We begin by developing some technical machinery that is useful in analyzing the differences between the two solutions $S$ and $S^*$. Let $\sigma$ and $\sigma^*$ be any optimal (splittable) assignments for $S$ and $S^*$, respectively. In any solution to the uncapacitated problem UKM, each client is assigned to a single facility, namely, the closest facility in the given solution. The preceding property is useful in providing a simple expression for the difference in the shipping costs of the two solutions $S$ and $S^*$ as a summation over all facilities in $S^*$ (see the proof of Lemma 5.2). The same is is not true, however, for CKMS since a splittable assignment may split the demand of a client among multiple facilities. To overcome this hurdle, we partition the demand of each client into smaller entities, which we refer to as *chunks*, such that each chunk is shipped from a single facility in $S$ (resp., $S^*$) under assignment $\sigma$ (resp., $\sigma^*$). We now describe this partitioning in Lemma 8.6.

**Lemma 8.6 (Splitting into Chunks)** *Let $\sigma$ and $\sigma^*$ be any assignments for $S$ and $S^*$ respectively. Then for each client $j$, there exists a partition of the demand $d_j$ into a finite number of chunks, each of which has an associated size, such that the following four properties hold:*

A1. *Each chunk is shipped from a single facility in $S$ under assignment $\sigma$.*

A2. *For each facility $i$ in $S$, the total size of all chunks of $d_j$ shipped from $i$ under assignment $\sigma$ is $\sigma(j,i)$.*

A3. *Each chunk is shipped from a single facility in $S^*$ under assignment $\sigma^*$.*

A4. *For each facility $i^*$ in $S^*$, the total size of all chunks of $d_j$ shipped from $i^*$ under assignment $\sigma^*$ is $\sigma^*(j,i^*)$.*

**Proof:** Fix a client $j$. The assignment $\sigma$ splits the demand $d_j$ among the facilities in $S$. Thus, for each facility $i$ in $S$ that satisfies $\sigma(j,i) > 0$, $\sigma$ defines a chunk of size $\sigma(j,i)$ that is shipped from $i$. Let $X$ denote the set of these chunks. While the set $X$ satisfies properties A1 and A2, $X$ may not satisfy properties A3 and A4. We now further partition each chunk in $X$ into smaller chunks and obtain the desired set $Y$ of chunks satisfying all of the four properties.

Consider any chunk $\alpha$ of client $j$ in $X$. Let $i$ be the facility that ships $\alpha$ in $\sigma$. For each $i^*$ in $S^*$ such that $\sigma^*(j,i^*) > 0$, we include a chunk $\alpha_{i^*}$ of size $s\sigma^*(j,i^*)/d_j$ in $Y$, where $s$ is the size of $\alpha$. We stipulate that chunk $\alpha_{i^*}$ is shipped from $i$ in $\sigma$ and from $i^*$ in $\sigma^*$. Since $Y$ is obtained by separately dividing each chunk of $X$ into smaller chunks, $Y$ still satisfies properties A1 and A2 above. Moreover, by construction, $Y$ satisfies properties A3 and A4. ■

We note that while Lemma 8.6 states that each chunk is shipped from a single facility under $\sigma$ (resp., $\sigma^*$), it does not preclude two distinct chunks associated with the same client from being shipped from the same facility under $\sigma$ (resp., $\sigma^*$).

Using the above notion of chunks, one can attempt to extend the proof of Lemma 5.2 for the uncapacitated problem UKM to apply for the capacitated problem CKMS. We now illustrate why such a direct argument fails. For each chunk $\alpha$, let $size(\alpha)$ denote the size of the chunk, $cl(\alpha)$ denote the client associated with the chunk, and $f(\alpha)$ (resp., $f^*(\alpha)$) denote the facility that ships $\alpha$ in assignment $\sigma$ (resp., $\sigma^*$). We can now write the difference between $C_s(S)$ and $C_s(S^*)$ as follows:

$$C_s(S) - C_s(S^*) = \sum_{j \in N} \sum_{\alpha : cl(\alpha) = j} size(\alpha)(c_{jf(\alpha)} - c_{jf^*(\alpha)}) = \sum_{i \in S^*} \sum_{\alpha : f^*(\alpha) = i} size(\alpha)(c_{cl(\alpha)f(\alpha)} - c_{cl(\alpha)i}).$$

Furthermore, by a simple averaging argument, we obtain that there exists a $v$ in $S^*$ for which

$$\sum_{\alpha : f^*(\alpha) = v} size(\alpha)(c_{cl(\alpha)f(\alpha)} - c_{cl(\alpha)v}) \geq \frac{C_s(S) - C_s(S^*)}{k}.$$

If $v$ is not in $S$, then $v$ can be added to $S$ and the demands reassigned to obtain a new feasible assignment, as in Lemma 5.2, without violating any facility capacities. In this case, we have the desired claim: $C_s(S) - C_s(S+v) \geq (C_s(S) - C_s(S^*))/k$. If $v$ is in $S$, however, then assigning more demand to $v$ may violate its capacity constraint, and hence the argument of Lemma 5.2 no longer applies. In order to address this problem, we introduce the notion of a "difference graph" that captures the differences in the assignments $\sigma$ and $\sigma^*$. We then use this difference graph to show that there exists a facility in $S^* - S$ that satisfies the requirement of the lemma.

Given the set $\tau$ of all chunks, taken over all of the clients, we construct a directed (multi)graph $G(\sigma, \sigma^*)$ (or simply $G$, when the context is clear) in which we associate a label, a cost, and a size with each arc. The node set is $S \cup S^*$. The arc set is in one-to-one correspondence with the set $\tau$, and is constructed as follows: For each chunk $\alpha$ that is shipped from facilities $i$ and $i^*$ in $\sigma$ and $\sigma^*$, respectively, we add a directed arc $\hat{a}$ from $i$ to $i^*$. This arc $\hat{a}$ is labeled by the corresponding chunk $\alpha$. The size of $\hat{a}$, denoted $size(\hat{a})$, is set to the $size(\alpha)$. The cost of $\hat{a}$, denoted $cost(\hat{a})$, is set to $c_{ji} - c_{ji^*}$, where $j$ is $cl(\alpha)$. The directed graph $G$ thus constructed is called the *difference graph.*

Note that the difference graph $G$ may contain cycles, self-loops, as well as multiple arcs between the same pair of nodes. Furthermore, there are no outgoing arcs for nodes in $S^* - S$, and no incoming arcs for nodes in $S - S^*$.

For notational convenience, throughout this section we interpret $D_v(S, \sigma)$ as zero whenever $v$ does not belong to $S$. (Recall that for $v$ in $S$, $D_v(S, \sigma)$ denotes the total demand that is shipped from facility $v$ in solution $S$.) The following lemma is immediate from Lemma 8.6 and the definition of the difference graph.

**Lemma 8.7** *Let $\sigma$ and $\sigma^*$ be any assignments for $S$ and $S^*$, respectively. Let $G(\sigma, \sigma^*)$ be the difference graph for assignments $\sigma$ and $\sigma^*$. Then, for each $v$ in $S \cup S^*$, the total size of all arcs coming into $v$ is $D_v(S^*, \sigma^*)$, and the total size of all arcs going out of $v$ is $D_v(S, \sigma)$.* ∎

We can view the difference graph as a flow network in which the capacity of every arc is simply equal to the size of the arc. Now consider a flow that saturates every arc in the network; that is, let the flow along any arc equal to the size (hence, capacity) of the arc. Using network flow terminology, we say that a node $v$ in the difference graph is a *surplus* node if the total size of all arcs coming into $v$ is greater than the total size of all arcs going out of $v$. For a surplus node $v$, the quantity $D_v(S^*, \sigma^*) - D_v(S, \sigma)$ is called the *excess* at $v$. We now invoke the well-known flow decomposition theorem (see, for example, Theorem 3.5 of [1]) to decompose the difference graph into a set $\mathcal{C}$ of directed cycles and a set $\mathcal{P}$ of directed paths that satisfy some useful properties.

**Lemma 8.8 (Decomposition of $G$)** *There exists a decomposition of the difference graph $G$ into a set of cycles $\mathcal{C}$ and a set of paths $\mathcal{P}$, with an associated positive size for each path and each cycle, such that:*

B1. *For each arc $\hat{a}$ in $G$, the sum of the sizes of all paths and cycles that contain the arc $\hat{a}$ is equal to the size of arc $\hat{a}$ in $G$.*

B2. *All paths terminate at a surplus node.*

B3. *For each surplus node $v$, the sum of the sizes of all paths terminating at $v$ is equal to the excess at $v$.*

Properties B1 and B2 follow immediately from the flow decomposition theorem, while Property B3 follows from the definition of a surplus node.

We define the cost of a path or cycle $P$ to be $size(P) \cdot \sum_{\hat{a} \in P} cost(\hat{a})$. (We note that since the cost of an arc may be negative, the cost of a path or cycle may also be negative.) We define the cost of a set of paths (resp., cycles) to be the sum of the costs of the paths (resp., cycles) in the set.

**Lemma 8.9** *Let $\mathcal{C}$ and $\mathcal{P}$ be the set of cycles and paths, respectively, obtained in the decomposition of $G$. Then $cost(\mathcal{C}) + cost(\mathcal{P}) = C_s(S) - C_s(S^*)$.*

**Proof:** Consider any arc $\hat{a}$ with label $\alpha$ in $G$. From property B1 of Lemma 8.8, it follows that the total contribution from arc $\hat{a}$ to the cost of $\mathcal{C}$ and $\mathcal{P}$ equals $size(\hat{a})cost(\hat{a})$, which is the difference between the shipping cost of the chunk $\alpha$ in assignment $\sigma$ and the shipping cost of the same chunk in assignment $\sigma^*$. Thus, by summing $size(\hat{a})cost(\hat{a})$ over all arcs, and thus accounting for all chunks, we obtain that $cost(\mathcal{C}) + cost(\mathcal{P}) = C_s(S) - C_s(S^*)$. ∎

By the hypothesis of Lemma 8.2, it follows that $C_s(S)$ is greater than $C_s(S^*)$, and hence that the set $S^* - S$ is nonempty. We partition the set of paths $\mathcal{P}$ on the basis of the terminating nodes of the paths, all of which are in $S^*$. Let $\ell$ denote $|S^* - S|$ and let $u_1, \ldots, u_\ell$ be the facilities in $S^* - S$. Let $\mathcal{P}_i$, for $1 \le i \le \ell$, denote the set of paths in $\mathcal{P}$ with terminating node $u_i$. Let $\mathcal{P}_0$ denote the set of paths in $\mathcal{P}$ that terminate at a node in $S \cap S^*$. We will prove Lemma 8.2 by showing that there exists an $i$, $1 \le i \le \ell$, such that the cost of the solution $S + u_i$ satisfies the upper bound stated in the lemma. We determine $u_i$, along with an associated assignment, by modifying the original assignment $\sigma$ through a sequence of "augmentation" operations, which we define below.

Let $\sigma'$ be an intermediate assignment obtained during the sequence of augmentations. Let $\mathcal{P}' \subseteq \mathcal{C} \cup \mathcal{P}$ denote the set of paths and cycles along which the augmentations were performed to obtain $\sigma'$. Given an assignment $\sigma'$, an *augmentation* along a path or cycle $P$ in $\mathcal{C} \cup \mathcal{P} \setminus \mathcal{P}'$ yields a new assignment $\sigma''$ that is defined as follows. We initially set $\sigma''$ to $\sigma'$. Then, for each arc $(u, v)$ in $P$ with label $\alpha$, we decrease $\sigma''(j, u)$ by $size(P)$ and increase $\sigma''(j, v)$ by $size(P)$, where $j$ is $cl(\alpha)$, the client associated with chunk $\alpha$. That is, a portion ($size(P)$ units of demand) of the chunk $\alpha$ which is shipped from $u$ in $\sigma'$ is now shipped from $v$ in $\sigma''$. This completes the description of an augmentation operation.

The validity of an intermediate assignment $\sigma'$ obtained during a sequence of augmentations follows from a simple inductive argument. We only state the inductive claims here: (i) for every arc $\hat{a}$ in $G$ from $u$ to $v$ with associated label $\alpha$, the amount of chunk $\alpha$ that is shipped from facility $u$ to client $cl(\alpha)$ in $\sigma'$ equals the sum of the sizes of all the paths and cycles not in $\mathcal{P}'$ that contain $\hat{a}$, and (ii) the excess at a surplus node $v$ under $\sigma'$ equals the sum of the sizes of all paths that terminate at $v$ and are not in $\mathcal{P}'$. (Recall that $\mathcal{P}'$ denotes the set of paths and cycles along which the augmentations were performed to obtain $\sigma'$.) This yields Lemma 8.10.

**Lemma 8.10 (Augmentations)** *Let $\sigma'$ be an assignment obtained from $\sigma$ by a sequence of augmentations along a subset $\mathcal{Q}$ of distinct paths and cycles drawn from $\mathcal{C} \cup \mathcal{P}$. Then, $\sigma'$ is a valid assignment and has shipping cost $C_s(S) - cost(\mathcal{Q})$.* ∎

We now apply Lemma 8.10 to the paths in $\mathcal{P}_i$ and the paths and cycles in $\mathcal{C} \cup \mathcal{P}_0$ to obtain Lemmas 8.11 and 8.12, respectively.

**Lemma 8.11** *For each $u_i$ in $S^* - S$, there exists an assignment for $S + u_i$ with a shipping cost of $C_s(S) - cost(\mathcal{P}_i)$.*

**Proof:** The validity and the shipping cost of the new assignment follows directly from Lemma 8.10. Moreover, since $\sigma$ uses only the facilities in $S$, and since all paths of $\mathcal{P}_i$ terminate at $u_i$, the final assignment uses only the facilities in $S + u_i$. This completes the proof. ■

**Lemma 8.12** *Each path or cycle in $\mathcal{C} \cup \mathcal{P}_0$ has a nonpositive cost.*

**Proof:** First, note that each node of any path or cycle in $\mathcal{C} \cup \mathcal{P}_0$ belongs to $S$. Now let $P$ be a path or a cycle in $\mathcal{C} \cup \mathcal{P}_0$ with positive cost, if one exists. By starting with $\sigma$ and augmenting along $P$, we obtain a new assignment for $S$ with a shipping cost that is strictly less than the shipping cost of the current assignment $\sigma$. This contradicts the optimality of $\sigma$. ■

We are now ready to prove Lemma 8.2.

**Proof of Lemma 8.2:** From Lemma 8.9 and Lemma 8.12, it follows that $\sum_{i=1}^{\ell} cost(\mathcal{P}_i) \geq C_s(S) - C_s(S^*)$. By averaging, there exists an $i$ such that $cost(\mathcal{P}_i) \geq \frac{C_s(S)-C_s(S^*)}{\ell}$. Applying Lemma 8.11 with this value of $i$, we obtain $C_s(S + u_i) = C_s(S) - cost(\mathcal{P}_i)$. Combining these bounds, we have

$$\begin{aligned} C(S) - C(S+u_i) &= C_s(S) - C_s(S+u_i) \\ &= cost(\mathcal{P}_i) \\ &\geq \frac{C_s(S) - C_s(S^*)}{\ell} \\ &\geq \frac{C_s(S) - C_s(S^*)}{k}, \end{aligned}$$

since $\ell \leq k$. This completes the proof of the lemma. ■

### 8.2 Dropping a Facility

The simple approach used in the uncapacitated case (Lemma 5.3) does not suffice for the proof of Lemma 8.3. The main reason is that the operation of dropping a facility $i$ and assigning all of its clients to another facility $i'$ may violate the capacity constraint at $i'$. To get around this hurdle, we use the notion of "heavy" and "light" facilities. A facility $i$ in $S$ is called *heavy* if $D_i(S, \sigma)$ is greater than $M/2$, and *light* otherwise. The motivation behind this definition is that dropping a light facility $i$ and reassigning the demand $D_i(S, \sigma)$ to another light facility does not violate the capacity constraints.

Let $S_L$ denote the set of these light facilities in $S$. Our strategy will be to focus on $S_L$ and proceed essentially as in the uncapacitated case. (Note that the number of light facilities is at least $(1+\alpha)k$. This is because the total demand of all clients cannot exceed $kM$ and hence the number of heavy facilities can be at most $2k$.) The basic idea is the same as used in Section 5 (see Lemma 5.3 and its proof): classify the facilities of $S_L$ as "primary" and "secondary" and then show that a secondary facility of $S_L$ can be dropped without incurring a "large" increase in the shipping cost. However, some differences do arise in the analysis due to the fact that a client's demand may be split across several facilities.

Let $\hat{\sigma} : N \mapsto S^*$ be a function that maps each location $i$ to the closest facility in $S^*$, breaking ties arbitrarily. (If there were no capacity constraints, then $\sigma^*$ would have been exactly the same as this function.) Now using $\hat{\sigma}$, we classify the facilities in $S_L$ as "primary" or "secondary", as follows: For each facility $i^*$ in $S^*$, consider those facilities of $S_L$ that are assigned to $i^*$ by $\hat{\sigma}$ and pick the closest one among these, breaking ties arbitrarily. The closest facility thus chosen is said to be the *primary* facility of $i^*$. Let $S_{LP}$ denote the set of these primary facilities. (Note that all primary facilities are light.) Facilities of $S_L$ that are not primary are said to be *secondary.* For each secondary facility $i$ in $S_L$, we define the associated primary facility as the primary facility of $\hat{\sigma}(i)$. Note that the number of secondary facilities is given by $|S_L| - |S_{LP}|$, and is at least $\alpha k$. We will show that one of these secondary facilities can be dropped without incurring a "large" increase in the shipping cost.

**Claim 8.13** *Consider a secondary facility $i$ in $S_L$, and let $i'$ be its associated primary facility. Then for all $j$ in $N$, we have $c_{ji'} - c_{ji} \leq 2(c_{ji} + c_{j\hat{\sigma}(j)})$.*

**Proof:** Let $i^*$ denote $\hat{\sigma}(i)$, the facility in $S^*$ that is closest to $i$. (Note that $\hat{\sigma}(i')$ is also $i^*$.) Now using the same reasoning as in the proof of Claim 5.4, we obtain:

$$c_{ji'} \leq c_{ji} + c_{ii^*} + c_{i^*i'} \leq c_{ji} + 2c_{ii^*} \leq c_{ji} + 2c_{i\hat{\sigma}(j)} \leq c_{ji} + 2c_{ij} + 2c_{j\hat{\sigma}(j)}.$$

■

Let $\Delta_i$ denote the increase incurred in the shipping cost when a secondary facility $i$ is dropped and the demand $D_i(S,\sigma)$ is reassigned to the primary facility associated with $i$. (Note that this reassignment does not violate any capacity constraints since both these facilities are light.)

**Claim 8.14** *For any secondary facility $i$ in $S_L$, we have $\Delta_i \leq \sum_{j \in N} 2\sigma(j,i)(c_{ji} + c_{j\hat{\sigma}(j)})$.*

**Proof:** Let $i'$ denote the primary facility associated with $i$. The reassignment of demand $D_i(S,\sigma)$ to $i'$ causes $\sigma(j,i)$ units of demand that was originally shipped from $i$ to a client $j$ to now be shipped from $i'$ to $j$. Hence the increase $\Delta_i$ in shipping cost is equal to $\sum_{j \in N} \sigma(j,i)(c_{ji'} - c_{ji})$. Applying the upper bound of Claim 8.13, the result follows. ■

**Proof of Lemma 8.3:** It suffices to show that there exists a secondary facility $i$ for which $\Delta_i \leq 2 \cdot \frac{C_s(S)+C_s(S^*)}{\alpha k}$. Recall that $S_L - S_{LP}$ is the set of secondary facilities, and that $|S_L - S_{LP}| \geq \alpha k$. We have

$$\begin{aligned}
\sum_{i \in S_L - S_{LP}} \Delta_i &\leq 2 \cdot \sum_{i \in S_L - S_{LP}} \sum_{j \in N} \sigma(j,i)(c_{ji} + c_{j\hat{\sigma}(j)}) \\
&\leq 2 \cdot \sum_{i \in S} \sum_{j \in N} \sigma(j,i)(c_{ji} + c_{j\hat{\sigma}(j)}) \\
&= 2 \cdot C_s(S) + 2 \cdot \sum_{j \in N} \sum_{i \in S} \sigma(j,i) c_{j\hat{\sigma}(j)} \\
&= 2C_s(S) + 2 \cdot \sum_{j \in N} d_j c_{j\hat{\sigma}(j)} \\
&\leq 2C_s(S) + 2C_s(S^*).
\end{aligned}$$

(The first step uses Claim 8.14. The fourth step uses the fact that $\sum_{i \in S} \sigma(j,i) = d_j$, the demand at $j$. The last step uses the fact that $\hat{\sigma}(j)$ is the closest facility to $j$ among all facilities in $S^*$.) By a simple averaging argument, it follows that the required secondary facility exists. ■

## 9 Capacitated Facility Location with Splittable Demands

In this section we consider the capacitated facility location problem with splittable demands (CFLS) and show in Theorem 9.1 that, as in the uncapacitated case, each step of the local search heuristic yields a significant improvement in cost whenever the current cost is sufficiently higher than the optimal cost. Throughout this section, we assume that $p(n)$ is a sufficiently large polynomial in $n$.

**Theorem 9.1 (Swapping Facilities)** *Let $S$ be any subset of $F$ such that $C(S) > (8+\varepsilon)C(S^*)$. Then there exists $T \subseteq F$ such that $|S - T| \leq 1$, $|T - S| \leq 1$, and $C(S) - C(T) \geq \frac{C(S)}{p(n)}$.*

**Corollary 9.2** *For any constant $\varepsilon > 0$, the local search heuristic yields a $(8+\varepsilon)$-approximate solution in polynomial time.* ∎

We prove the theorem in two stages, as in the uncapacitated case. First, in Lemma 9.3, we show that if the shipping cost of $S$ is sufficiently large, then we can add a facility to $S$ and obtain a significant improvement in the cost. Due to the capacity constraints, the proof of this lemma is more involved than in the uncapacitated case (Lemma 6.3). Section 9.1 gives the proof of this lemma using much of the technical machinery developed in Section 8.1, where the operation of adding a facility for CKMS is analyzed. Second, in Lemma 9.4, we show that if the facility cost of $S$ is sufficiently large then we can either drop a facility from $S$ or substitute a facility in $S$ to obtain a significant improvement in the cost. The proof of this lemma is given in Section 9.2.

**Lemma 9.3 (Adding a Facility)** *Let $S$ be any subset of $F$ such that $C_s(S) \geq C(S^*) + \frac{nC(S)}{p(n)}$. Then there exists a $v$ in $F$ such that $C(S) - C(S+v) \geq \frac{C(S)}{p(n)}$.*

**Lemma 9.4 (Dropping/Swapping Facilities)** *Let $S$ be any subset of $F$. If no facility of $S$ can be dropped or substituted (by a facility not in $S$) to improve the cost by at least $\frac{C(S)}{p(n)}$, then the following upper bound on the facility cost of $S$ holds:*

$$C_f(S)\left(1 - \frac{n^2}{p(n)}\right) < 5C(S^*) + 2C_s(S) + \frac{C(S)}{n}.$$

Theorem 9.1 follows directly from Lemmas 9.3 and 9.4.

**Proof of Theorem 9.1:** This is similar to the proof of Theorem 6.1, except that we use Lemmas 9.3 and 9.4 instead of Lemmas 6.3 and 6.4. In this case, we obtain

$$C(S)\left(1 - \frac{n^2}{p(n)} - \frac{3n}{p(n)} - \frac{1}{n}\right) < 8C(S^*),$$

which is a contradiction if $p(n)$ is chosen sufficiently large. ∎

## 9.1 Adding a Facility

In this section, we prove Lemma 9.3. By the hypothesis of the lemma it follows that $C_s(S)$ is greater than $C_s(S^*)$, and hence that the set $S^* - S$ is nonempty. Let $\ell$ denote the size of $S^* - S$ and let $u_1, \ldots, u_\ell$ be the facilities in $S^* - S$. We will now show that there exists a facility $u_i$ in $S^* - S$ that satisfies the requirement of the lemma.

We use the technical machinery developed in Section 8.1 for CKMS. As in that case, we construct a difference graph based on the sets $S$ and $S^*$ and the assignments $\sigma$ and $\sigma^*$. Since Lemmas 8.6 through 8.12 only concern the difference graph and the shipping costs associated with the solutions $S$ and $S^*$, all of these claims also hold for CFLS. Lemma 9.3 now follows from Lemmas 8.9, 8.11, and 8.12.

**Proof of Lemma 9.3:** We are given that $C_s(S) - C(S^*) \geq nC(S)/p(n)$. From Lemmas 8.9 and 8.12, it follows that $\sum_{i=1}^{\ell} \mathit{cost}(\mathcal{P}_i) \geq C_s(S) - C_s(S^*)$. Subtracting $C_f(S^*)$ from both sides, we get

$$\sum_{i=1}^{\ell}(\mathit{cost}(\mathcal{P}_i) - f_{u_i}) \geq C_s(S) - C_s(S^*) - C_f(S^*).$$

By averaging, there exists an $i$ such that $cost(\mathcal{P}_i) - f_{u_i} \geq \frac{C_s(S)-C_s(S^*)-C_f(S^*)}{\ell}$. Applying Lemma 8.11 with this value of $i$, we obtain $C_s(S+u_i) = C_s(S) - cost(\mathcal{P}_i)$. Combining these bounds, we have

$$\begin{aligned} C(S) - C(S+u_i) &= C_s(S) - C_s(S+u_i) - f_{u_i} \\ &= cost(\mathcal{P}_i) - f_{u_i} \\ &\geq \frac{C_s(S) - C_s(S^*) - C_f(S^*)}{\ell} \\ &\geq \frac{C_s(S) - C(S^*)}{n} \\ &\geq \frac{C(S)}{p(n)}, \end{aligned}$$

since $C_s(S) - C(S^*) \geq nC(S)/p(n)$ by assumption. This completes the proof of the lemma. ■

### 9.2 Dropping or Swapping Facilities

It remains to prove Lemma 9.4. Due to the capacity constraint, the simple approach used in the uncapacitated case (Lemma 6.4) does not suffice here. In particular, the third candidate operation may be invalid because reassigning the demand associated with a facility $i$ to another facility $j$ may violate the capacity constraint at $j$. We overcome this difficulty by studying a problem that we call the auxiliary allocation problem, the solution of which leads to a construction of suitable candidate operations. More specifically, the solution of the auxiliary allocation problem yields a valid reassignment of the demand associated with a facility that is being dropped. The auxiliary allocation problem is defined in Section 9.2.1. Section 9.2.2 contains the design and the analysis of the candidate operations. Finally, Section 9.2.3 completes the proof of Lemma 9.4.

#### 9.2.1 The Auxiliary Allocation Problem

Given any two solutions $S$ and $S^*$, an assignment $\sigma$ for $S$, and the original distance metric $c$, we define the *auxiliary allocation problem* $R$ in order to capture the task of reassigning demand when a facility from $S$ is dropped or swapped. The problem $R$ is specified as follows:

- The nodes of $S - S^*$ are the *requestors*.
- The nodes of $S^*$ are the *servers*.
- Each requestor node $v$ in $S - S^*$ has a demand of $D_v(S, \sigma)$.
- Each server node $v$ in $S \cap S^*$ has a capacity of $M - D_v(S, \sigma)$.
- Each server node $v$ in $S^* - S$ has a capacity of $M$.
- The distance function between the nodes is given by $c$.

(For ease of notation, we drop $S$ and $\sigma$ and use $D_v$ for $D_v(S, \sigma)$ throughout this section.) We seek an *allocation* $\beta : (S - S^*) \times S^* \mapsto \mathbf{R}$, that satisfies the demands at all of the requestor nodes without violating the capacity constraints at the server nodes. This allocation $\beta$ will be used to guide the reassignment of the demand $D_v$ when a facility $v$ in $S - S^*$ is being considered for dropping or substituting. For a requestor node $v$ in $S - S^*$, let $q_\beta(v)$ denote the cost of satisfying the demand at $v$; that is, $q_\beta(v)$ equals $\sum_{j \in S^*} \beta(v, j) c_{vj}$. Let $q_\beta$ denote the overall cost of the allocation $\beta$, defined as $\sum_{v \in S - S^*} q_\beta(v)$. Figure 2 illustrates an allocation.

We note that an allocation is conceptually the same as a splittable assignment. An allocation and an assignment differ only in their domains. (While the domain of an assignment is the set of all clients, the domain of an allocation is a subset of the facilities.) Indeed, our analysis of the auxiliary allocation problem uses the concept of a difference graph and much of the other technical machinery developed in Section 8 for the analysis of splittable assignments. We use the techniques of Section 8 to establish the following lemma, which states that there is an allocation that is not too expensive.

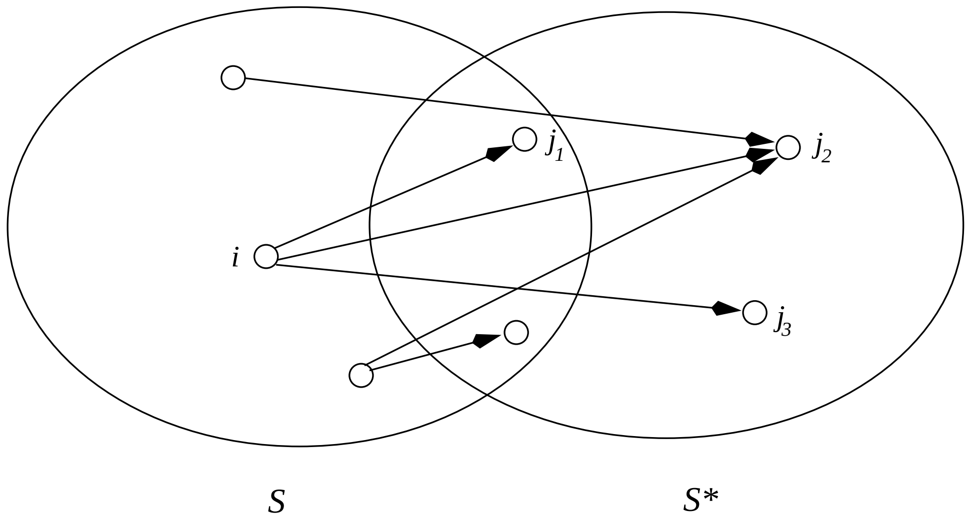


Figure 2: An allocation $\beta$. Only those edges $(i, j)$ are shown for which $\beta(i, j) > 0$. Thus, for example, $\beta$ reassigns the demand of requestor $i$ to servers $j_1$, $j_2$, and $j_3$. Note that every edge is incident from a node in $S - S^*$ to a node in $S^*$.

**Lemma 9.5** *There exists an allocation $\beta$ for the auxiliary allocation problem $R$ with cost $q_\beta$ at most $C_s(S) + C_s(S^*)$.*

For proving Lemma 9.5, we first construct a difference graph $G'(\sigma, \sigma^*)$ (or, simply $G'$, when the context is clear) that is identical to the difference graph $G(\sigma, \sigma^*)$ constructed in Section 8.1, except that the costs on the arcs are different: While the cost of an arc $\hat{a}$ is set to $c_{ji} - c_{ji^*}$ in $G(\sigma, \sigma^*)$, we set it to $c_{ji} + c_{ji^*}$ in $G'(\sigma, \sigma^*)$. The cost of the new difference graph $G'$ is defined as before, namely, $\sum_{\hat{a} \in A} \mathit{size}(\hat{a}) \mathit{cost}(\hat{a})$.

Since $G$ and $G'$ are identical except for the arc costs, Lemmas 8.7 and 8.8 proved for $G$ also hold for $G'$. Furthermore, we can place an upper bound on the cost of the paths and cycles obtained in the decomposition of Lemma 8.8.

**Lemma 9.6** *If $\mathcal{P}$ is the set of paths obtained by the decomposition of $G'$ in Lemma 8.8, then $\mathit{cost}(\mathcal{P})$ is at most $C_s(S) + C_s(S^*)$.*

**Proof:** Let $\mathcal{C}$ denote the set of cycles obtained by the decomposition of $G'$ in Lemma 8.8. Then, $\mathit{cost}(\mathcal{P}) + \mathit{cost}(\mathcal{C})$ is the sum of the costs of all the arcs in $G'$, which equals $C_s(S) + C_s(S^*)$. Since all of the arc costs in $G'$ are positive, it follows that $\mathit{cost}(\mathcal{C}) \geq 0$ and hence the desired claim holds. ■

We are now ready to prove Lemma 9.5.

**Proof of Lemma 9.5:** Using the decomposition $\mathcal{P}$, we construct the allocation $\beta$ as follows: For each $i$ in $S - S^*$ and $j$ in $S \cup S^*$, let $\mathcal{P}_{ij}$ denote the set of paths in $\mathcal{P}$ that start at $i$ and terminate at $j$. Then we set $\beta(i, j) = \sum_{P \in \mathcal{P}_{ij}} \mathit{size}(P)$. By Lemma 8.7, all nodes with an incoming arc in $G'$ belong to $S^*$, since all other nodes $v$ have $D_v(S^*, \sigma^*) = 0$. Thus, the terminating nodes of all paths $P$ in $\mathcal{P}$ belong to $S^*$. This ensures that $\beta$ is a function from $(S - S^*) \times S^*$ to $\mathbf{R}$. We now show that $\beta$ satisfies the capacity constraints and that the desired upper bound on cost holds.

**Validity of $\beta$.** We first show that the demands of all requestor nodes are satisfied. Consider an arbitrary requestor node $v$ in $S - S^*$. By Lemma 8.7, the total size of all arcs going out of $v$ in $G'$ is $D_v(S, \sigma)$. Moreover, since $D_v(S^*, \sigma^*)$ is zero, $v$ can only occur in the decomposition of $G'$ as the starting node of a path in $\mathcal{P}$. It follows from property B1 of Lemma 8.8 that the sum of the sizes of all the paths in $\mathcal{P}$ starting at $v$ is $D_v(S, \sigma)$. Therefore, the total demand shipped to $v$ is $D_v(S, \sigma)$. We next show that $\beta$ observes the capacity constraint at each server node. Consider an arbitrary server node $v$ in $S^*$. The total demand shipped from $v$ under $\beta$ (i.e., $\sum_{i \in S-S^*} \beta(i,v)$) is at most the sum of the sizes of all paths in $\mathcal{P}$ terminating at $v$. By properties B2 and B3 of Lemma 8.8, this sum equals the excess of $v$, which is $D_v(S^*,\sigma^*) - D_v(S,\sigma)$. Since $D_v(S^*,\sigma^*) \leq M$, it follows that the total demand shipped from $v$ under $\beta$ is at most the capacity of $v$ in the allocation problem. Thus, all of the capacity constraints are observed.

**Upper bound on cost**. Consider any arc $\hat{a}$ in $G'$. Let the endpoints of $\hat{a}$ be $u$ and $u^*$, and let the chunk corresponding to $\hat{a}$ be $\alpha$, and let $y$ be $cl(\alpha)$, the client associated with chunk $\alpha$. Recall that $cost(\hat{a})$ is $c_{uy} + c_{yu^*}$. By the triangle inequality, this is at least $c_{uu^*}$. Now consider any path $P$ in $\mathcal{P}$ with starting and terminating nodes $i$ and $j$. Then $\sum_{\hat{a} \in P} cost(\hat{a}) \geq c_{ij}$ by repeated applications of the triangle inequality. Thus, $cost(P)$ is at least $size(P)c_{ij}$. Summing over all paths $P$ starting at $i$ and terminating at $j$, we obtain $\beta(i,j)c_{ij} \leq \sum_{P \in \mathcal{P}_{ij}} cost(P)$. Then, summing over all $i$ and $j$, it follows that the cost of allocation $\beta$ is at most $cost(\mathcal{P})$. The desired upper bound then follows from Lemma 9.6. ■

For any requestor node $v$, we define the servers $j$ in $S^*$ for which $\beta(v,j)$ is nonzero as the *$\beta$-suppliers* of $v$. If every requestor node $v$ in $S - S^*$ has at most one *$\beta$-supplier* in $S^* - S$, then $\beta$ is said to be *refined*. Note that a refined allocation $\beta$ allows requestor nodes to have multiple $\beta$-suppliers in $S \cap S^*$. Figure 3 illustrates a refined allocation. Note that unlike Figure 2, from any node in $S - S^*$ there is only one arc entering $S^* - S$.

For the purpose of constructing the candidate operations, a refined allocation is more useful than an unrefined allocation. To see this, recall that when a facility $v$ in $S - S^*$ is being dropped or substituted, our plan is to reassign the demand $D_v$ among the $\beta$-suppliers of $v$. Since we are allowed to add only one new facility to $S - \{v\}$, it is essential that at most one $\beta$-supplier of $v$ be outside $S$. This condition is satisfied only if $\beta$ is refined.

Lemma 9.7 guarantees the existence of an "inexpensive" refined allocation $\hat{\beta}$ provided that we relax the auxiliary allocation problem by increasing the capacity of each server node $i^*$ in $S^* - S$ from $M$ to $2M$. We use $\hat{R}$ to denote this relaxed auxiliary allocation problem. (We note that this blowup in the capacities is for the purposes of the analysis only.)

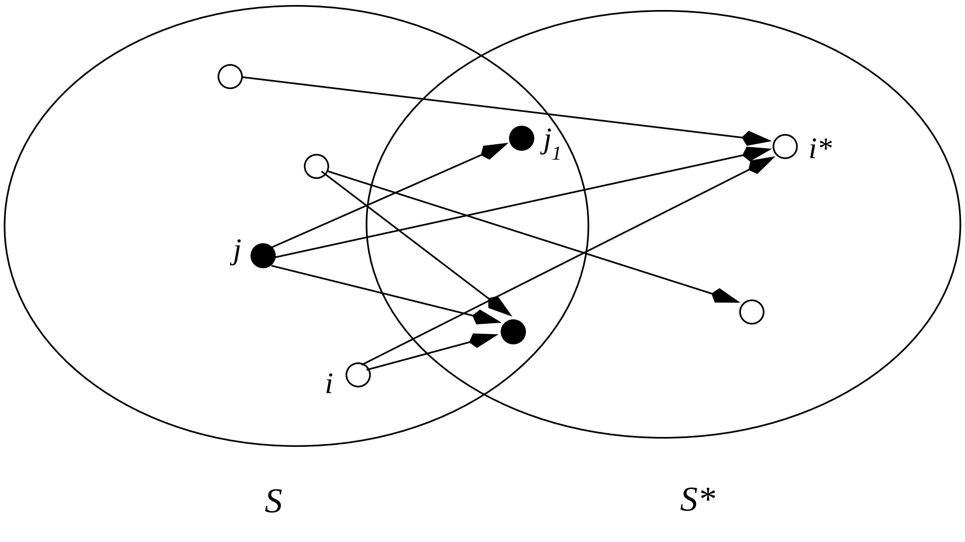


Figure 3: A refined allocation $\hat{\beta}$. Note that for each node $u$ in $S - S^*$, there is at most one node in $S^* - S$ to which the demand of $u$ is reassigned. The figure also illustrates a candidate operation for a facility $i$ in $S_{ELS}$. Facility $i^*$ is the $\hat{\beta}$-foreign-supplier of $i$, and $j$ is the primary facility of $i^*$. The candidate operation for $i$, which is defined in Section 9.2.2, reassigns the demand of $i$ among $j$ and the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$; that is, the demand of $i$ is reassigned among the three black nodes.

**Lemma 9.7** *There exists a refined auxiliary allocation $\hat{\beta}$ for the relaxed auxiliary allocation problem, $\hat{R}$, with cost $q_{\hat{\beta}}$ at most $C_s(S) + C_s(S^*)$.*

**Proof:** Let $\beta$ be the (unrefined) allocation provided by Lemma 9.5 for the allocation problem $R$. Using $R$ and $\beta$, we first construct a restricted allocation problem $R'$, which involves the servers in $S^* - S$ only. We then apply a result of Shmoys and Tardos (Theorem 10.1) to $R'$ to obtain an unsplittable allocation for $R'$, which is then combined with $\beta$ to construct the required refined allocation $\hat{\beta}$ for $R$. (An allocation $\alpha$ is said to be unsplittable if each requestor has at most one supplier.)

Before proceeding, we need some notation to split the demands and shipping costs into two parts, one for servers in $S \cap S^*$, and the other for servers in $S^* - S$. Throughout this proof, for notational convenience, we use $D_v$ to denote $D_v(S, \sigma)$. Let $D'_v(\beta)$ (resp., $D''_v(\beta)$) denote the amount of demand shipped to a requestor node $v$ from servers in $S \cap S^*$ (resp., $S^* - S$) under the allocation $\beta$. That is, $D'_v(\beta)$ equals $\sum_{j \in S \cap S^*} \beta(v, j)$ and $D''_v(\beta)$ equals $\sum_{j \in S^* - S} \beta(v, j)$. Note that $D'_v(\beta) + D''_v(\beta) = D_v$. The precise formulation of $R'$ can now be given as follows:

- The nodes of $S - S^*$ are the requestors.
- The nodes of $S^* - S$ are the servers.
- Each requestor node $v$ in $S - S^*$ has a demand of $D''_v(\beta)$ .
- Each server node $v$ in $S^* - S$ has a capacity of $M$.
- The distance function between the nodes is given by $c$.

Let $q'_\beta$ (resp., $q''_\beta$) denote the cost of shipping the demand from servers in $S \cap S^*$ (resp., $S^* - S$) under allocation $\beta$. More precisely, $q'_\beta$ equals $\sum_{i \in S - S^*} \sum_{j \in S \cap S^*} \beta(i, j) c_{ij}$ and $q''_\beta$ equals $\sum_{i \in S - S^*} \sum_{j \in S^* - S} \beta(i, j) c_{ij}$. Clearly, $q'_\beta + q''_\beta = q_\beta$. Consider the allocation $\beta'$ obtained by restricting $\beta$ to servers in $S^* - S$. Thus, for each $i$ in $S - S^*$ and $j$ in $S^* - S$, set $\beta'(i, j) = \beta(i, j)$. The allocation $\beta'$ is clearly valid for $R'$ and has cost $q''_\beta$. Theorem 10.1 then implies that there exists an unsplittable allocation $\beta''$ of cost at most $q''_\beta$ that is feasible for $R'$ provided that the capacity constraints on the servers are relaxed to $M + \max_{i \in S - S^*} D''_i(\beta)$, which is at most $2M$. Since $\beta''$ is an unsplittable allocation, it ensures that each requestor node has at most one supplier in $S^* - S$. Using $\beta''$ and $\beta$, we construct a refined allocation $\hat{\beta}$ for $\hat{R}$ as follows:

- For each $i$ in $S - S^*$ and $j$ in $S^* - S$, set $\hat{\beta}(i, j) = \beta''(i, j)$.
- For each $i$ in $S - S^*$ and $j$ in $S^* \cap S$, set $\hat{\beta}(i, j) = \beta(i, j)$.

It is straightforward to verify that the above allocation $\hat{\beta}$ is a valid refined allocation for $\hat{R}$. Moreover, the cost of $\hat{\beta}$ is equal to the cost of $\beta''$ plus $q'_\beta$. Hence it is upper-bounded by $q_\beta$, which in turn is at most $C_s(S) + C_s(S^*)$, by Lemma 9.5. This completes the proof. ∎

Let $D'_v$ (resp., $D''_v$) denote the amount of demand shipped to $v$ from servers in $S \cap S^*$ (resp., $S^* - S$) under the refined allocation $\hat{\beta}$. If $D''_v$ is nonzero, then there is a unique $\hat{\beta}$-supplier in $S^* - S$ that we refer to as the *$\hat{\beta}$-foreign-supplier* of $v$. For the sake of uniformity, we extend this notion even to those requestor nodes $v$ for which $D''_v$ is zero. Though such requestor nodes do not have any $\hat{\beta}$-suppliers in $S^* - S$, we pick an arbitrary facility from $S^* - S$ and call it the $\hat{\beta}$-foreign-supplier of $v$.

Table 3 summarizes the notation developed in this section for the auxiliary allocation problem.

### 9.2.2 Candidate Operations

Let $S^*$ denote the optimal solution and $S$ denote a current solution that satisfies the hypothesis of Lemma 9.4. In this section, we define and analyze a set of candidate operations for the facilities in $S$; our analysis leads to the proof of Lemma 9.4 in Section 9.2.3.

We partition $S$ into five classes and define a candidate operation for the facilities of each class. As in the uncapacitated case, a facility $i$ in $S$ is called *cheap* if $f_i \leq C(S)/n^2$, and *expensive* otherwise.

| *Notation* | *What it denotes* |
|---|---|
| $\beta$ | allocation: a function from $S - S^* \times S^*$ to $\mathbf{R}$ that indicates the demand that is satisfied at each requestor through each server |
| $\beta$-supplier of $i$ | any facility $j$ such that $\beta(i,j) > 0$ |
| $q_\beta$ | cost of allocation $\beta$ |
| $\hat{\beta}$ | refined allocation: an allocation in which for each $i \in S - S^*$ there is at most one facility $j$ in $S^* - S$ such that $\hat{\beta}(i,j) > 0$ |
| $D_v$ | demand of requestor $v$ in the allocation problem |
| $D'_v$ | demand shipped to $v$ from servers in $S \cap S^*$ under $\hat{\beta}$ |
| $D''_v$ | demand shipped to $v$ from servers in $S^* - S$ under $\hat{\beta}$ |
| $\hat{\beta}$-foreign-supplier of $i$ | if $D''_i > 0$, then the unique facility $j \in S^* - S$ that satisfies $\hat{\beta}(i,j) > 0$; otherwise, an arbitrary facility in $S^* - S$ |

Table 3: Notation used for the auxiliary allocation problem

Moreover, a facility $i$ in $S - S^*$ is called *light* if $D''_i \leq M/2$, and *heavy* otherwise. Let $S_C$ denote the set of cheap facilities in $S - S^*$. Let $S_{EL}$ (resp., $S_{EH}$) denote the set of expensive facilities in $S - S^* - S_C$ that are light (resp., heavy).

We now partition the set $S_{EL}$ of expensive light facilities into *primary* and *secondary* facilities. Let $i^*$ be any facility in $S^* - S$. If there is at least one facility in $S_{EL}$ for which $i^*$ is the $\hat{\beta}$-foreign-supplier, then we designate one such facility in $S_{EL}$, appropriately chosen, as the primary facility for $i^*$. The precise criterion for selecting the primary facility is deferred to later in this section. The facilities in $S_{EL}$ that are not the primary facility of any facility in $S^* - S$ are referred to as secondary facilities. We let $S_{ELP}$ (resp., $S_{ELS}$) denote the set of expensive light facilities that are classified as primary (resp., secondary). Thus, $S \cap S^*$, $S_C$, $S_{EH}$, $S_{ELP}$, and $S_{ELS}$ form a partition of $S$.

We now define *candidate operations* for facilities in each of these subsets, as follows:

1. For $i$ in $S \cap S^*$, do nothing.

2. For $i$ in $S_C$, do nothing.

3. For $i$ in $S_{EH}$, substitute $i$ with $i^*$, the $\hat{\beta}$-foreign-supplier of $i$. The demand $D_i$ is reassigned among the $\hat{\beta}$-suppliers of $i$.

4. For $i$ in $S_{ELP}$, substitute $i$ with $i^*$, the $\hat{\beta}$-foreign-supplier of $i$. The demand $D_i$ is reassigned among the $\hat{\beta}$-suppliers of $i$.

5. For $i$ in $S_{ELS}$, drop $i$ and let $j$ be the primary facility associated with the $\hat{\beta}$-foreign-supplier of $i$. The demand $D_i$ is reassigned as follows. First, $D'_i$ is reassigned among the $\hat{\beta}$-suppliers of $i$ in $S \cap S^*$. Second, $D''_i$ is reassigned among $j$ and the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$; more specifically, a demand of up to $M - D_j$ is reassigned to $j$, and the remaining demand $(D''_i - M + D_j)$, if any, is reassigned among the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$. The set of nodes involved in the reassignments of demand $D_i$ is illustrated in Figure 3.

We first need to ensure that each candidate operation is valid, i.e., the capacity constraints are not violated when an individual candidate operation is performed. (We note that each candidate operation is performed on the solution $(S, \sigma)$.) The candidate operations for facilities in $S \cap S^*$ and $S_C$ are trivially valid. The validity of a candidate operation for a facility $i$ in $S_{EH} \cup S_{ELP}$ follows from the following two observations: (i) the reassignment respects the capacity constraint at each facility $j$ in $S \cap S^*$ since the capacity of $j$ in the relaxed auxiliary allocation problem is equal to the unused capacity of $j$ in $\sigma$, and (ii) the demand reassigned to the $\hat{\beta}$-foreign-supplier of $i$ is at most $D_i$, which is at most $M$. Finally, we consider a candidate operation for a facility $i$ in $S_{ELS}$. The reassignment of $D'_i$ for $i$ in $S_{ELS}$ respects the capacity constraints since it is performed only among facilities in $S \cap S^*$, and for each facility in $S \cap S^*$, its capacity in the auxiliary allocation problem equals its unused capacity in $\sigma$. But for the reassignment of $D''_i$, we need to ensure that there is sufficient remaining capacity at $j$ and the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$. This is ensured by the particular criterion by which we partition $S_{EL}$ into primary and secondary facilities, which we now describe.

Let $Q(i^*)$ denote the set of requestor nodes for which $i^*$ is the $\hat{\beta}$-foreign-supplier. Note that the sets $Q(i^*)$, for $i^*$ in $S^* - S$, induce a partition of $S - S^*$. For each $i^*$ in $S^* - S$ such that $Q(i^*) \cap S_{EL}$ is nonempty, the primary facility associated with $i^*$ is defined as the facility $j$ in $Q(i^*) \cap S_{EL}$ that minimizes $c_{ji^*} + \theta_j$, where $\theta_j$ is defined as follows: it is the average cost per unit demand for servicing an additional demand of $M/2$ units at $j$, using the remaining capacity (in the current solution $(S, \sigma)$) at facility $j$ and at the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$. (Remark: If there were no capacity constraints, then $\theta_j$ would be zero and the criterion for picking the primary facility would degenerate to that used in the uncapacitated case.) As mentioned earlier, a facility $i$ in $S_{EL}$ is referred to as a secondary facility if $i$ is not the primary facility for any $i^*$ in $S^* - S$.

For the parameter $\theta_j$ to be well-defined, we need to ensure that the total remaining capacity at facility $j$ and at the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$ is at least $M/2$. (Note that this also guarantees the validity of the candidate operation for any facility in $S_{ELS}$.) To see that this condition is satisfied, we first observe that the remaining capacity at $j$ is $M - D_j$. Then, we observe that the remaining capacity at the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$ is at least $D'_j$, since $D'_j$ is the amount of demand shipped from $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$ to the requestor node $j$ in the auxiliary allocation problem. Adding these two bounds, and using the fact that $D_j - D'_j = D''_j \le M/2$ (since $j$ is light), we find that the remaining capacity is at least $M/2$. Hence $\theta_j$ is well-defined.

We now analyze the three non-trivial candidate operations, namely, the ones associated with sets $S_{EH}$, $S_{ELP}$, and $S_{ELS}$. We call a candidate operation *good* if it reduces the cost of the solution by at least $\frac{C(S)}{p(n)}$, and *bad* otherwise. By the hypothesis of the Lemma 9.4, it follows that all of the candidate operations are bad. Claims 9.8, 9.9, and 9.10 show that if the candidate operation for a facility is bad, then its facility cost is "small". The proof of Lemma 9.4 then follows by combining these claims.

**Claim 9.8** *If the candidate operations for all of the expensive heavy facilities are bad, then*

$$C_f(S_{EH}) \left(1 - \frac{n^2}{p(n)}\right) \le 4C_f(S^* - S) + \sum_{i \in S_{EH}} q_{\hat{\beta}}(i).$$

**Proof:** Consider an expensive heavy facility $i$ in $S_{EH}$. The decrease in the facility cost caused by its candidate operation is $f_i - f_{i^*}$, and by the triangle inequality, the increase in the shipping cost is at most $q_{\hat{\beta}}(i)$. So the net decrease in cost is at least $f_i - (q_{\hat{\beta}}(i) + f_{i^*})$. If this quantity is at least $\frac{n^2 f_i}{p(n)}$, then since $i$ is expensive, this candidate operation is good. Therefore, if this candidate operation is bad, the following upper bound on the facility cost of $i$ holds: $f_i(1 - \frac{n^2}{p(n)}) < f_{i^*} + q_{\hat{\beta}}(i)$.

Since any facility $i^*$ in $S^* - S$ can supply a demand of at most $2M$ under the refined auxiliary allocation $\hat{\beta}$ and each heavy facility receives at least $M/2$ units from its unique supplier, it follows that $i^*$ can be the candidate substitute for at most four heavy facilities. Hence, summing over all of the expensive heavy facilities yields the required upper bound on $C_f(S_{EH})$. ■

**Claim 9.9** *If the candidate operations for all of the expensive light primary facilities are bad, then*

$$C_f(S_{ELP}) \left(1 - \frac{n^2}{p(n)}\right) \le C_f(S^* - S) + \sum_{i \in S_{ELP}} q_{\hat{\beta}}(i).$$

**Proof:** The net cost decrease caused by the substitution of an expensive light primary facility $i$ in $S_{ELP}$ is $f_i - (q_{\hat{\beta}}(i) + f_{i^*})$. If this substitution is bad then, by proceeding as in the previous claim, we obtain the following upper bound on the facility cost of $i$: $f_i(1 - \frac{n^2}{p(n)}) < f_{i^*} + q_{\hat{\beta}}(i)$.

Summing over all of the facilities of $S_{ELP}$, and using the fact that each $i^*$ in $S^* - S$ has at most one associated primary facility, we obtain the required upper bound on $C_f(S_{ELP})$. ■

**Claim 9.10** *If the candidate operations for all of the expensive light secondary facilities are bad, then*

$$C_f(S_{ELS})\left(1 - \frac{n^2}{p(n)}\right) \leq \sum_{i \in S_{ELS}} 2q_{\hat{\beta}}(i).$$

**Proof:** The heart of our proof is to establish the following claim: The net cost decrease caused by the candidate operation for an expensive light secondary facility $i$ in $S_{ELS}$ is at least $f_i - 2q_{\hat{\beta}}(i)$. We note that the preceding claim implies that if the candidate operation for a facility $i$ in $S_{ELS}$ is bad, then $f_i(1 - \frac{n^2}{p(n)}) \leq 2q_{\hat{\beta}}(i)$. Summing over all expensive light secondary facilities then yields the desired upper bound on $C_f(S_{ELS})$.

We now prove the desired lower bound on the decrease in cost caused by the candidate operation for an expensive light secondary facility $i$ in $S_{ELS}$. Let $i^*$ be the $\hat{\beta}$-foreign-supplier of $i$ and let $j$ be the primary facility associated with $i^*$. Let $q'_{\hat{\beta}}(i)$ denote the cost of shipping the demand $D'_i$ from the $\hat{\beta}$-suppliers of $i$ in $S \cap S^*$, and let $q''_{\hat{\beta}}(i)$ denote the cost of shipping the demand $D''_i$ from $i^*$. Thus $q_{\hat{\beta}}(i) = q'_{\hat{\beta}}(i) + q''_{\hat{\beta}}(i)$.

When $i$ is dropped, the demand $D'_i$ is reassigned to the $\hat{\beta}$-suppliers of $i$ in $S \cap S^*$. This increases the shipping cost by at most $q'_{\hat{\beta}}(i)$. It now suffices to bound the shipping cost increase due to the reassignment of $D''_i$ by $q'_{\hat{\beta}}(i) + 2q''_{\hat{\beta}}(i)$.

We know that the candidate operation of $i$ reassigns the demand $D''_i$ to $j$ and the $\hat{\beta}$-suppliers of $j$ in $S \cap S^*$. By the definition of $\theta_j$, the cost of servicing an additional demand of $D''_i$ originating at $j$ is at most $D''_i\theta_j$ (Recall that $D''_i \leq M/2$ since $i$ is light). Therefore, the increase in the shipping cost due to this reassignment of $D''_i$ is at most

$$\begin{aligned} D''_i(c_{ij} + \theta_j) &\leq D''_i(c_{ii^*} + c_{ji^*} + \theta_j) \\ &\leq D''_i(c_{ii^*} + c_{ii^*} + \theta_i) \\ &\leq 2q''_{\hat{\beta}}(i) + \frac{M}{2} \cdot \theta_i \\ &\leq 2q''_{\hat{\beta}}(i) + q'_{\hat{\beta}}(i). \end{aligned}$$

The second step in the above argument uses the fact that $j$, and not $i$, was chosen as the primary for $i^*$; and hence $c_{ji^*} + \theta_j \leq c_{ii^*} + \theta_i$. The third step follows since $q''_{\hat{\beta}}(i) = D''_i c_{ii^*}$ and $D''_i \leq M/2$. The fourth step needs some elaboration. The goal here is to prove that $q'_{\hat{\beta}}(i)$, is an upper bound on $M\theta_i/2$. Note that $q'_{\hat{\beta}}(i)$ is the cost of shipping an additional demand of $D'_i$ units at $i$ using the remaining capacities at the $\hat{\beta}$-suppliers of $i$ in $S \cap S^*$, while $M\theta_i/2$ is the cost of servicing an additional demand of $M/2$ units at $i$ using the remaining capacities at $i$ and the $\hat{\beta}$-suppliers of $i$ in $S \cap S^*$. We consider two cases depending on whether $D'_i > M/2$. If $D'_i > M/2$, then it directly follows from the definitions of $q'_{\hat{\beta}}(i)$ and $\theta_i$ that $q'_{\hat{\beta}}(i) \geq M\theta_i/2$. If $D'_i \leq M/2$, then we observe that the remaining capacity at $i$ is $M - (D'_i + D''_i)$, which is at least $\frac{M}{2} - D'_i$ since $D''_i \leq M/2$. Thus, out of $M/2$ units of demand, at least $M/2 - D'_i$ units can be satisfied by $i$ itself. The cost of satisfying the remaining at most $D'_i$ units of demand using the $\hat{\beta}$-suppliers of $i$ in $S \cap S^*$ is at most $q'_{\hat{\beta}}(i)$. Therefore, $M\theta_i/2$ is at most $q'_{\hat{\beta}}(i)$. This completes the proof of the claim. ■

### 9.2.3 Proof of Lemma 9.4

Using Claims 9.8, 9.9, and 9.10, and the inequality $C_f(S_C) \le C(S)/n$, we obtain

$$\begin{aligned}
C_f(S)\left(1-\frac{n^2}{p(n)}\right) &= (C_f(S\cap S^*)+C_f(S_C)+C_f(S_{EH})+C_f(S_{ELP})+C_f(S_{ELS}))\left(1-\frac{n^2}{p(n)}\right) \\
&\le C_f(S\cap S^*)+\frac{C(S)}{n}+5C_f(S^*-S)+\sum_{i\in S-S^*}2q_{\hat{\beta}}(i) \\
&\le \frac{C(S)}{n}+2q_{\hat{\beta}}+5C_f(S^*) \\
&\le \frac{C(S)}{n}+2C_s(S)+2C_s(S^*)+5C_f(S^*) \\
&\le \frac{C(S)}{n}+2C_s(S)+5C(S^*).
\end{aligned}$$

(The third step follows from the definition of $q_{\hat{\beta}}$. The fourth step follows from Lemma 9.7.) This completes the proof of the lemma.

## 10 Capacitated Problems with Unsplittable Demands

In this section we consider the capacitated problems with unsplittable demands, namely, CKMU and CFLU. We obtain constant-factor approximation algorithms for these problems by combining our results for CKMS and CFLS, respectively, with a result of Shmoys and Tardos on the generalized assignment problem. In [23], Shmoys and Tardos show how to obtain an unsplittable assignment from a splittable assignment. Their result was presented in the context of job scheduling. Here we give an adaptation of their result to our context.

**Theorem 10.1 (Generalized Assignment Problem [23])** *Let $S$ be a set of open facilities and $N$ be a set of clients, and let $d_i$ denote the demand at client $i$. Suppose there is a splittable assignment from $N$ to $S$ with shipping cost $C$ that assigns at most $M$ units of demand to any facility. Then there exists a polynomial-time computable unsplittable assignment from $N$ to $S$ with shipping cost at most $C$ that assigns at most $M + \max_{i\in N} d_i$ units of demand to any facility.*

We now show how to obtain constant-factor approximation algorithms for CKMU and CFLU. We first relax the unsplittability requirement, and use the local search heuristic to solve the relaxed problem. By Corollary 8.4 (resp., Corollary 9.2), it follows that the splittable solution $(S,\sigma)$ thus obtained is within a constant factor of the optimal for the CKMS (resp., CFLS). We then use Theorem 10.1 to obtain an unsplittable assignment $\sigma'$ for $S$ with shipping cost at most that of the splittable assignment $\sigma$ for $S$. However, Theorem 10.1 requires the capacities of all facilities to be increased by a factor of at most two. If we assume, as in [24], that a blowup in the capacity incurs a proportionate increase in the facility cost, then the facility cost for $S$ in the unsplittable case is at most twice that in the splittable case. Therefore, the total cost of the unsplittable solution $(S,\sigma')$ is at most twice the total cost of the splittable solution $(S,\sigma)$. Thus, the cost of $(S,\sigma')$ is within a constant factor of the optimal unsplittable solution. Theorems 10.2 and 10.3 below summarize our results for CKMU and CFLU, respectively.

**Theorem 10.2** (CKMU) *For any constant $\varepsilon > 0$, there is a constant $\delta$ such that the following can be computed in polynomial time, assuming a factor two blowup in the capacity of each facility: (i) a solution $S$ with $C(S) \le (1+\varepsilon)C(S^*)$ and $|S| \le (1+\delta)k$, and (ii) a solution $S$ with $|S| \le (5+\varepsilon)k$ and $C(S) \le (1+\delta)C(S^*)$.* ■

**Theorem 10.3** (CFLU) *For any constant $\varepsilon > 0$, the local search heuristic yields a $(16+\varepsilon)$-approximate solution in polynomial time, assuming a factor two blowup in the capacity of each facility.* ■

## Acknowledgments

We would like to thank the anonymous referees for their many valuable comments and suggestions.

## References

[1] R. K. Ahuja, T. L Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice-Hall, Upper Saddle River, NJ, 1993.

[2] S. Arora, P. Raghavan, and S. Rao. Approximation schemes for Euclidean $k$-medians and related problems. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing*, pages 106–113, May 1998.

[3] M. Charikar and S. Guha. Improved combinatorial algorithms for the facility location and $k$-median problems. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 378–388, October 1999.

[4] M. Charikar, S. Guha, D. Shmoys, and É. Tardos. A constant-factor approximation algorithm for the $k$-median problem. In *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pages 1–10, May 1999.

[5] F. A. Chudak. Improved approximation algorithms for the uncapacitated facility location problem. In *Proceedings of the 6th Conference on Integer Programming and Combinatorial Optimization*, pages 180–194, June 1998.

[6] F. A. Chudak and D. Shmoys. Improved approximation algorithms for a capacitated facility location problem. In *Proceedings of the 10th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 875–876, January 1999.

[7] F. A. Chudak and D. Williamson. Improved approximation algorithms for capacitated facility location problems. In *Proceedings of the 7th Conference on Integer Programming and Combinatorial Optimization*, pages 99–113, June 1999.

[8] G. Cornuéjols, G. L. Nemhauser, and L. A. Wolsey. The uncapacitated facility location problem. In P. Mirchandani and R. Francis, editors, *Discrete Location Theory*, pages 119–171. Wiley, New York, NY, 1990.

[9] G. Diehr. An algorithm for the $p$-median problem. Technical Report No. 191, Western Management Science Institute, UCLA, June 1972.

[10] N. Garg and J. Könemann. Faster and simpler algorithms for multicommodity flow and other fractional packing problems. In *Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science*, pages 300–309, October 1998.

[11] S. Guha and S. Khuller. Greedy strikes back: Improved facility location algorithms. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 649–657, January 1998. To appear in the *Journal of Algorithms.*

[12] D. S. Hochbaum. Heuristics for the fixed cost median problem. *Mathematical Programming*, 22:148–162, 1982.

[13] D. S. Hochbaum. Various notions of approximations: Good, better, best, and more. In D. S. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, pages 346–398. PWS Publishing Company, Boston, MA, 1995.

[14] K. Jain and V. Vazirani. Primal-dual approximation algorithms for metric facility location and $k$-median problems. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 1–10, October 1999.

[15] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4:373–395, 1984.

[16] L. G. Khachiyan. A polynomial algorithm for linear programming. *Soviet Math. Doklady*, 20:191–194, 1979.

[17] M. Korupolu, C. G. Plaxton, and R. Rajaraman. Analysis of a local search heuristic for facility location problems. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1–10, January 1998.

[18] A. A. Kuehn and M. J. Hamburger. A heuristic program for locating warehouses. *Management Science*, 9:643–666, 1963.

[19] J.-H. Lin and J. S. Vitter. Approximation algorithms for geometric median problems. *Information Processing Letters*, 44:245–249, 1992.

[20] J.-H. Lin and J. S. Vitter. $\varepsilon$-Approximations with minimum packing constraint violation. In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, pages 771–782, May 1992.

[21] R. F. Love, J. G. Morris, and G. O. Wesolowsky. *Facilities Location: Models and Methods.* North-Holland, New York, NY, 1988.

[22] P. Mirchandani and R. Francis. *Discrete Location Theory.* Wiley, New York, NY, 1990.

[23] D. B. Shmoys and É. Tardos. An approximation algorithm for the generalized assignment problem. *Mathematical Programming*, 62:461–474, 1993.

[24] D. B. Shmoys, É. Tardos, and K. Aardal. Approximation algorithms for facility location problems. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, pages 265–274, May 1997.

[25] M. Sviridenko, December 1999. Personal communication.

[26] M. B. Teitz and P. Bart. Heuristic methods for estimating the generalized vertex median of a weighted graph. *Operations Research*, 16:955–961, 1968.

[27] N. Young. Randomized rounding without solving the linear program. In *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 170–178, January 1995.

[28] N. Young. Greedy approximation algorithms for $k$-medians by randomized rounding. Technical Report PCS-TR99-344, Department of Computer Science, Dartmouth College, March 1999.