Improved Algorithms for Distributed Balanced Clustering

Kian Mirjalali, Hamid Zarrabi-Zadeh

Improved Algorithms for Distributed Balanced Clustering

Kian Mirjalali and Hamid Zarrabi-Zadeh

Sharif University of Technology, Tehran, Iran
mirjalali@ce.sharif.edu
zarrabi@sharif.edu

Abstract. We study a weighted balanced version of the k-center problem, where each center has a fixed capacity, and each element has an arbitrary demand. The objective is to assign demands of the elements to the centers, so as the total demand assigned to each center does not exceed its capacity, while the maximum distance between centers and their assigned elements is minimized. We present a deterministic $O(1)$-approximation algorithm for this generalized version of the k-center problem in the distributed setting, where data is partitioned among a number of machines. Our algorithm substantially improves the approximation factor of the current best randomized algorithm available for the problem. We also show that the approximation factor of our algorithm can be improved to $5+\varepsilon$, when the underlying metric space has a bounded doubling dimension.

1 Introduction

Clustering is a well-known problem with various applications. The problem is in particular important in distributed environments, where we are dealing with big amounts of data. In these settings, no single machine can store the whole data, and hence, data in partitioned among several nodes.

The *k-center* problem is a popular formulation of clustering, consisting of a set S of n elements in a metric space (U, d), and an integer k. The objective is to select k elements from S as centers and assign each element of S to one of the centers, so as the maximum distance between elements and their assigned centers is minimized. Naturally, each element is assigned to its nearest center.

In the *balanced k-center* problem (also known as *capacitated k-center*), each center has a fixed capacity L, bounding the number of elements that can be covered by that center. Obviously, when centers have capacity, elements are not necessarily covered by their nearest centers. The *weighted balanced k-center* is a generalization of this problem where each element x has a weight/demand $w(x)$. When an element x is assigned to one center, it uses $w(x)$ units of the center's capacity. However, an element can be assigned to more than one center in general, each covering a portion of its demand. The problem in its general form has natural applications in facility location scenarios where resources have capacities, and users have specific demands.

Related Work. The k-center problem is known to be NP-hard. The *furthest-point* greedy algorithm proposed by Gonzalez [14], yields a 2-approximate solution. Hochbaum and Shmoys [15] proposed another 2-approximation algorithm based on parametric pruning. It is known that no $2-\varepsilon$ approximation algorithms is possible for the k-center problem, unless P = NP.

The capacitated version of k-center was first introduced by Barilan *et al.* [3]. They presented a 10-approximation for the problem. Khuller and Sussmann [18] improved the approximation factor to 6, and showed that the factor can be further improved to 5 for soft capacities, i.e., when elements can be selected as center more than once. Cygan *et al.* [8] studied a non-uniform variant of the problem, where each element has a specific capacity if selected as center, and presented an $O(1)$-approximation algorithm for this problem. The constant factor was later improved to 9 by An *et al.* [2]. Other variants of the capacitated k-center problem have been also studied in the literature (see, e.g., [6, 9–11, 13]).

The k-center problem is also studied in the distributed environments when dealing with big data. Several variations and approximation algorithms have been proposed in this context [5, 12, 16, 19–21]. Bateni *et al.* [4] studied the balanced version of the problem in the distributed environment. and presented a randomized algorithm with approximation guarantee 32β, where β is the approximation factor of the corresponding centralized algorithm for weighted balanced k-center. Using the current best bound of $\beta = 5$ [18], their algorithm yields an approximation factor of 160.

Our Results. We present a new deterministic approximation algorithm for the weighted balanced k-center problem in distributed environments, achieving an approximation guarantee of $9\beta + 4$, where β is the approximation factor of the corresponding centralized algorithm for weighted balanced k-center. This substantially improves the current best approximation factor of 32β due to Bateni *et al.* [4]. Our algorithm can be implemented in constant number of rounds in massively parallel computation (MPC) models, such as MapReduce. Moreover, our algorithm uses a small amount of communication, which we show is optimal under fair assumptions. We further show that the approximation factor of our algorithm can be improved to $5+\varepsilon$ if the underlying metric space has a bounded doubling dimension.

Our algorithm uses the "composable coresets" framework introduced by Indyk *et al.* [17]. In this framework, a small subset of data (so-called a coreset) is carefully extracted from each machine, in such a way that the union of coresets contains a good approximation of the whole data set. This framework has been successfully used to devise approximation algorithms for several other optimization problems in distributed settings (see, e.g., [1, 7, 22, 23]).

The rest of this paper is organized as follows. In Section 2, the basic definitions and formulation of the problem is given. In Section 3, we describe our distributed approximation algorithm for the weighted balanced k-center problem and analyze its approximation factor. In Section 4, we show how the approximation factor of our algorithm can be improved to $5 + \varepsilon$ in metric spaces with bounded doubling dimension.

2 Preliminaries

Given two sets A and B, a *relation* from A to B is a subset of the Cartesian product of A to B. We can generalize this concept by adding a multiplicity to each pair $a \in A$ and $b \in B$, denoting the weight of relation between a and b.

Definition 1. *Given two sets A and B, a* weighted relation *R from A and B is a function $R : A \times B \longrightarrow \mathbb{N}_0$, where $R(a, b)$ stands for the number of relationships between $a \in A$ and $b \in B$. In particular, $R(a, b) = 0$ means that there is no relation between a and b.*

In the weighted balanced k-center problem, we are given as input a set S of n elements in a metric space (U, d), an integer k denoting the number of centers, and an integer L representing the capacity of each center. For each element $x \in S$, a demand (weight) $w(x)$ is also given, representing the number of times required for x to be covered by centers. A *clustering* $C = (D, A)$ consists of a set $D = \{c_1, c_2, \ldots, c_k\}$ of (not-necessarily distinct) centers selected from S, and a weighted relation A from S to D representing the assignment of elements to the centers. More precisely, $A(x, c_i)$ represents the number of times $x \in S$ is covered by center c_i. A clustering $C = (D, A)$ is *feasible*, if the following two conditions hold:

$$\forall x \in S : \sum_{1 \le i \le k} A(x, c_i) = w(x),$$

and

$$\forall c_i \in D : \sum_{x \in S} A(x, c_i) \le L.$$

The latter is called *capacity constraint*, and the former is called *demand constraint*. Note that we are considering the soft version of the problem, where an element can be selected as a center more than once. The objective of the weighted balanced k-center problem is to find a feasible clustering $C = (D, A)$ minimizing the cost

$$R_{S,w}(C) := \max_{\substack{x \in S,\, c_i \in D \\ A(x,c_i) > 0}} d(x, c_i).$$

Obviously, the problem has no feasible solution if $\sum_{x \in S} w(x) > kL$.

3 Distributed Weighted Balanced k-Center

In this section, we present our distributed algorithm for the weighted balanced k-center problem. We assume that the input data set S is partitioned into m subsets $S_1, S_2, \ldots, S_m$, each stored in a separate machine. A good point about our algorithm is that we have no specific assumption on the partitioning of data, such as a particular ordering or a random partitioning.

Our algorithm uses the following two centralized approximation algorithms as subroutines:

– Algorithm $\mathcal{A}$: an α-approximation algorithm for the k-center problem,
– Algorithm $\mathcal{B}$: a β-approximation algorithm for the weighted balanced k-center problem.

The pseudo-code of our algorithm is presented in Algorithm 1. In this algorithm, we first run algorithm $\mathcal{A}$ separately in each machine to obtain m coresets each of size k. The coresets are then composed into a single set T in the central machine, and algorithm $\mathcal{B}$ is applied on this set to obtain a set D of k centers, along with its corresponding assignment. The set D is then sent back to the machines to obtain the final assignment. A general schema of the algorithm is illustrated in Figure 1.

Algorithm 1 Distributed Weighted Balanced k-Center

Input: Data sets $S_1, \ldots, S_m$, an integer k, a capacity L, and a weight function w
Output: A feasible k-clustering of the set $S = \bigcup_{i=1}^{m} S_i$

1: For each $1 \leq i \leq m$, run algorithm $\mathcal{A}$ on S_i to obtain a clustering $C_i = (D_i, A_i)$.
2: For each center $c \in D_i$, define $w'(c) = \sum_{x \in S_i} A_i(x, c)$
3: Send D_i's and their demands w' to the central machine. Let $T = \bigcup_{i=1}^{m} D_i$.
4: Run algorithm $\mathcal{B}$ on $\langle T, w' \rangle$ to obtain a clustering $C = (D, A)$.
5: Send C back to the machines containing S_i's.
6: In machine i, assign demands covered by $c \in D_i$ to the centers in D based on A. Call the new assignment A'.
7: **return** clustering $C' = (D, A')$.

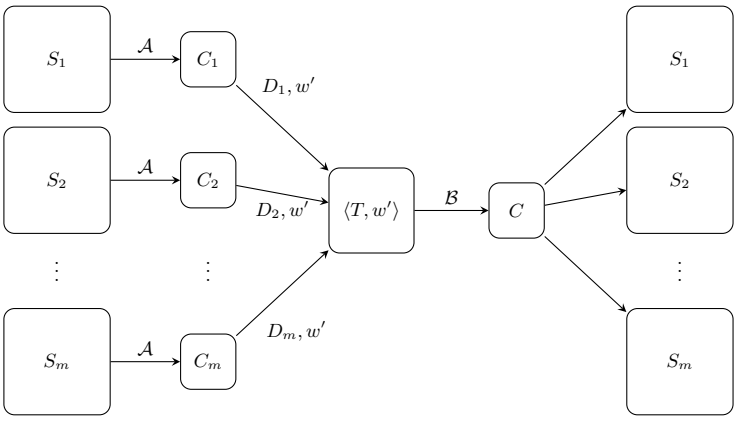

Fig. 1: A general schema of Algorithm 1.

Clearly, if the runtime of the algorithms $\mathcal{A}$ and $\mathcal{B}$ is polynomial, the whole algorithm runs in polynomial time. Now we analyze the approximation factor of Algorithm 1. In the following, we assume that $C^* = (D^*, A^*)$ is an optimal solution for the whole input set $S = \bigcup_{i=1}^{m} S_i$.

Lemma 1. *The cost of clustering computed for each subset in the first step of Algorithm 1 is not greater than 2α times the cost of optimal solution for the whole set. In other words, for all $1 \leq i \leq m$:*

$$R_{S_i,w}(C_i) \leq 2\alpha R_{S,w}(C^*).$$

Proof. Let $C^\dagger$ and $C_i^\dagger$ be optimal solutions for the uncapacitated k-center problem on inputs S and S_i, respectively. As the uncapacitated k-center problem is a relaxed form of its capacitated version, we have:

$$R_{S,w}(C^\dagger) \leq R_{S,w}(C^*) \tag{1}$$

Now, we build a feasible solution $\hat{C} = (\hat{D}, \hat{A})$ for the uncapacitated k-center problem on input S_i, based on $C^\dagger = (D^\dagger, A^\dagger)$ in the following way. If $D^\dagger = \{q_1, \ldots, q_k\}$, we set $\hat{D} = \{\hat{q}_1, \ldots, \hat{q}_k\}$, where $\hat{q}_j$ be the nearest member of S_i to q_j. Therefore:

$$\forall t \in S_i : d(q_j, \hat{q}_j) \leq d(q_j, t) \tag{2}$$

We also define $\hat{A}(x, \hat{q}_j)$ the same as $A^\dagger(x, q_j)$. Since $C_i^\dagger$ is the optimal solution and $\hat{C}$ is a feasible solution for the uncapacitated k-center problem on input S_i:

$$R_{S_i,w}(C_i^\dagger) \leq R_{S_i,w}(\hat{C}) \tag{3}$$

Assume that element $x \in S_i$ has the maximum distance from its covering center $\hat{q}_j$ in $\hat{C}$. Thus:

$$\begin{aligned}
R_{S_i,w}(\hat{C}) &= d(x, \hat{q}_j) && \text{(by definition of } R_{S_i,w}) \\
&\leq d(x, q_j) + d(q_j, \hat{q}_j) && \text{(by triangle inequality)} \\
&\leq d(x, q_j) + d(q_j, x) && \text{(by (2))} \\
&\leq 2R_{S,w}(C^\dagger) && (\hat{A}(x, \hat{q}_j) > 0 \implies A^\dagger(x, q_j) > 0)
\end{aligned} \tag{4}$$

Putting all together, we get:

$$\begin{aligned}
R_{S_i,w}(C_i) &\leq \alpha R_{S_i,w}(C_i^\dagger) && \text{(by approximation factor of algorithm } \mathcal{A}) \\
&\leq \alpha R_{S_i,w}(\hat{C}) && \text{(by inequality (3))} \\
&\leq 2\alpha R_{S,w}(C^\dagger) && \text{(by inequality (4))} \\
&\leq 2\alpha R_{S,w}(C^*) && \text{(by inequality (1))}
\end{aligned}$$

□

Lemma 2. *In Algorithm 1, the cost of clustering C computed by algorithm $\mathcal{B}$ for input $\langle T, w' \rangle$ is not greater than $\beta(4\alpha + 1)$ times the cost of optimal solution for the whole input. In other words:*

$$R_{T,w'}(C) \leq \beta(4\alpha + 1) R_{S,w}(C^*).$$

Proof. Algorithm $\mathcal{A}$ used in the first phase of Algorithm 1 solves an uncapacitated version of k-center, and thus, each element of S is assigned by $\mathcal{A}$ to a single center in T. We can model this assignment with a function $f : S \longrightarrow T$ and its reverse $F : T \longrightarrow P(S)$, so that for each $x \in S$, $f(x)$ is the covering center of x, and for each $t \in T$, $F(t)$ is the set of elements covered by t. Therefore, the demand $w'(t)$ for $t \in T$ (computed in step 2 of Algorithm 1) can be written as

$$w'(t) = \sum_{x \in F(t)} w(x). \tag{5}$$

Let $D^* = \{q_1^*, q_2^*, \ldots, q_k^*\}$ be the set of centers in an optimal clustering $C^* = (D^*, A^*)$ for the weighted balanced k-center problem on the whole input $\langle S, w\rangle$. We build a feasible solution $\hat{C} = (\hat{D}, \hat{A})$ for the same problem on input $\langle T, w'\rangle$ using C^* and functions f and F. The set of centers in $\hat{C}$ is $\hat{D} = \{\hat{q}_1, \ldots, \hat{q}_k\}$ where $\hat{q}_j$ is $f(q_j^*)$ (for $1 \le j \le k$), and the coverage assignment $\hat{A} : T \times \hat{D} \longrightarrow \mathbb{N}_0$ is defined as

$$\hat{A}(t, \hat{q}_j) := \sum_{x \in F(t)} A^*(x, q_j^*) \qquad (\text{for } t \in T, 1 \le j \le k).$$

To prove the feasibility of $\hat{C}$, we only need to verify that its coverage assignment satisfies the following two constraints:

– Demand constraint: for each element $t \in T$, we have:

$$\begin{aligned}
\sum_{1 \le j \le k} \hat{A}(t, \hat{q}_j) &= \sum_{1 \le j \le k} \sum_{x \in F(t)} A^*(x, q_j^*) && \text{(by definition of } \hat{A}) \\
&= \sum_{x \in F(t)} \sum_{1 \le j \le k} A^*(x, q_j^*) \\
&= \sum_{x \in F(t)} w(x) && \text{(due to demand constraint in } C^*) \\
&= w'(t) && \text{(by (5))}
\end{aligned}$$

– Capacity constraint: for each center $\hat{q}_j$ $(1 \le j \le k)$, we have:

$$\begin{aligned}
\sum_{t \in T} \hat{A}(t, \hat{q}_j) &= \sum_{t \in T} \sum_{x \in F(t)} A^*(x, q_j^*) && \text{(by definition of } \hat{A}) \\
&= \sum_{x \in S} A^*(x, q_j^*) && \text{(Each } x \in S \text{ is in exactly one } F(t).) \\
&\le L && \text{(due to capacity constraint in } C^*)
\end{aligned}$$

Now, assume that the element $t \in T$ has the maximum distance from its covering center $(\hat{q}_j)$ in $\hat{C}$, as shown in Figure 2. Since $\hat{A}(t, \hat{q}_j) > 0$, there exists an element $x \in F(t)$ with $A^*(x, q_j^*) > 0$. Let x and q_j^* be members of S_i and $S_{i'}$ for some $i, i' \in \{1, \ldots, m\}$, respectively. So, we can say:

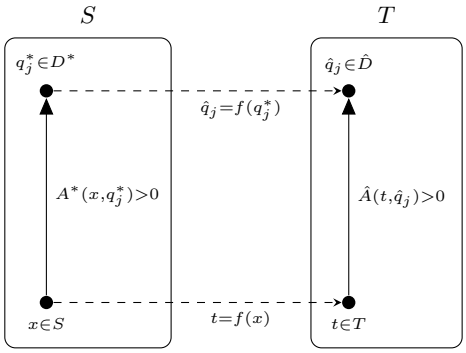

Fig. 2: Element $t \in T$ with the maximum distance from its covering center ($\hat{q}_j$).

$$\begin{aligned}
R_{T,w'}(\hat{C}) &= d(t, \hat{q}_j) && \text{(by definition of } R_{T,w'}) \\
&\leq d(t, x) + d(x, q_j^*) + d(q_j^*, \hat{q}_j) && \text{(by triangle inequality)} \\
&\leq R_{S_i,w}(C_i) + R_{S,w}(C^*) + R_{S_{i'},w}(C_{i'}) && \text{(by definition of } R_{S_i,w} \text{ and } R_{S,w}) \\
&\leq 2\alpha R_{S,w}(C^*) + R_{S,w}(C^*) + 2\alpha R_{S,w}(C^*) && \text{(by Lemma 1)} \\
&= (4\alpha + 1) R_{S,w}(C^*) && (6)
\end{aligned}$$

Assume that C^+ is an optimal solution for the weighted balanced k-center problem on input $\langle T, w' \rangle$. Since $\hat{C}$ is a feasible solution for this problem, we have:

$$R_{T,w'}(C^+) \leq R_{T,w'}(\hat{C}) \tag{7}$$

Therefore:

$$\begin{aligned}
R_{T,w'}(C) &\leq \beta R_{T,w'}(C^+) && \text{(by approximation factor of alg. } \mathcal{B}) \\
&\leq \beta R_{T,w'}(\hat{C}) && \text{(by inequality (7))} \\
&\leq \beta(4\alpha + 1) R_{S,w}(C^*) && \text{(by inequality (6))}
\end{aligned}$$ □

Theorem 1. *The approximation factor of Algorithm 1 is at most* $2\alpha + \beta(4\alpha + 1)$ *where* α *and* β *are the approximation factor of algorithms* $\mathcal{A}$ *and* $\mathcal{B}$.

Proof. We want to prove that:

$$R_{S,w}(C') \leq (2\alpha + \beta(4\alpha + 1)) R_{S,w}(C^*)$$

We have to show that for each element $x \in S$ and each center $c \in D$ covering x ($A'(x, c) > 0$):

$$d(x, c) \leq (2\alpha + \beta(4\alpha + 1)) R_{S,w}(C^*)$$

Let x be in S_i (for some $i \in \{1, \ldots, m\}$). Since x is covered by c ($A'(x, c) > 0$), the way of constructing A' implies that there exists at least one intermediate

center y in D_i that covers x ($A_i(x, y) > 0$) and is covered by c ($A(y, c) > 0$). This is clarified as an example in Figure 3. A directed edge from a to b here shows that a is covered by b. Note that the separate presentation of sets S_i, D_i, and D in the figure is to give more intuition; we know that D_i is a subset of S_i and might have intersection with D.

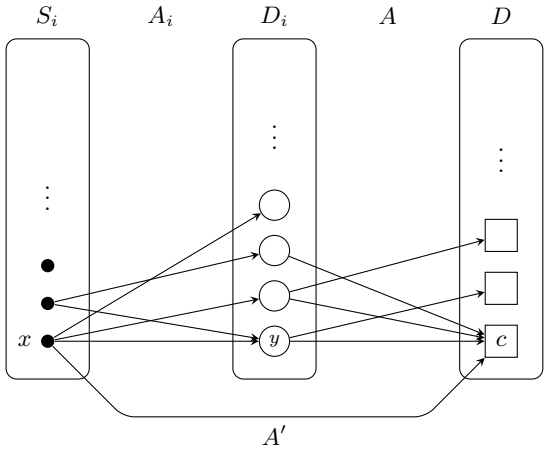

Fig. 3: An example of the covering state of elements in the algorithm

$$
\begin{aligned}
d(x,c) &\le d(x,y) + d(y,c) && \text{(by triangle inequality)}\\
&\le R_{S_i,w}(C_i) + R_{T,w'}(C) && \text{(by definition of } R_{S_i,w} \text{ and } R_{T,w'}\text{)}\\
&\le 2\alpha R_{S,w}(C^*) + \beta(4\alpha+1)R_{S,w}(C^*) && \text{(by Lemmas 1 and 2)}\\
&= (2\alpha + \beta(4\alpha+1))R_{S,w}(C^*)
\end{aligned}
$$

□

Corollary 1. *We have a distributed algorithm for the balanced k-center problem with approximation factor* 49*.*

Proof. We can achieve the 49-approximation by using Algorithm 1 with proper placement of algorithms $\mathcal{A}$ and $\mathcal{B}$. We can use Gonzalez's algorithm [14] as algorithm $\mathcal{A}$, which yields $\alpha = 2$. For algorithm $\mathcal{B}$, we use the centralized 5-approximation algorithm proposed by Khuller and Sussmann [18] which yields $\beta = 5$. Now, by Theorem 1, the approximation factor of the whole algorithm is $2\alpha + \beta(4\alpha + 1) = 49$. □

The communication complexity of our algorithm is $O(mk)$. We show in the next theorem that this complexity is optimal, under the composable coreset framework.

Theorem 2. *Any distributed algorithm for the capacitated/uncapacitated k-center problem with a bounded approximation guarantee under the composable coreset framework requires $\Omega(mk)$ communication, where k is the number of centers and m is the number of data partitions.*

Proof. Let S_i (the set of elements in the i-th machine) be a set of k points of pairwise distance F. Consider a scenario in which all the other points in other machines are at a small distance ε from a single element $e \in S_i$. If the i-th machine does not send all of its k elements to the central machine, the cost of the final solution will be F while the cost of the optimal solution is ε and thus, the approximation factor will be unbounded. So, all of the k elements in S_i must be sent to the central machine. Now, consider another scenario in which all the local machines have an input set similar to S_i. With the same argument, all of them will independently send k elements to the central machine resulting a communication of size $\Omega(mk)$. □

4 Metric Spaces with Bounded Doubling Dimension

The approximation factor of the algorithm presented in Section 3 can be potentially improved by reducing the values of α and β. As there is no $2-\varepsilon$ approximation algorithm for the k-center problem (unless P = NP), one may conclude that the effect of factor α cannot be reduced to less than 2. However, we show that for special metric spaces, increasing the number of centers picked from each partition S_i can reduce the effect of α. To clarify this, we first introduce a new parameter of metric spaces, which we call "half-coverage constant".

Definition 2. *Let $\mathrm{OPT}(S,k)$ be the cost of an optimal solution for the k-center problem on an input set S. The* half-coverage constant *of a metric space is the minimum constant c such that for any instance $\langle S,k\rangle$ of the k-center problem in that space,*

$$\mathrm{OPT}(S,ck) \leq \frac{1}{2}\mathrm{OPT}(S,k).$$

According to the above definition, for any arbitrary instance of the k-center problem in a metric space with half-coverage constant H, if the number of centers is multiplied by H, the cost of the optimal solution is reduced by a factor of $1/2$. The following observation is immediate by our definition of half-coverage constant.

Observation 3. *In a metric space with half-coverage constant H, for any input set S and any nonnegative integer t,*

$$\mathrm{OPT}(S,H^t k) \leq \frac{1}{2^t}\mathrm{OPT}(S,k).$$

Not all metric spaces have a bounded half-coverage constant. For example, consider a metric space where the distance of every two distinct elements is 1. If this space has a half-coverage constant H, then for an input set S of size $H+1$,

we have $\mathrm{OPT}(S,1) = \mathrm{OPT}(S,H) = 1$, which contradicts the definition of half-coverage constant.

Although not all metric spaces have a bounded half-coverage constant, the following theorem shows that every metric space with a bounded doubling dimension (including all $\mathbb{R}^d$ spaces, under any ℓ_p metric) has a bounded half-coverage constant.

Theorem 4. *Every metric space with a doubling constant M has a half-coverage constant at most M^2.*

Proof. Recall that the doubling constant M of a metric space denotes the minimum number of balls of radius $r/2$ needed to cover a ball of radius r. Let *quadrupling constant* Q be the minimum number of balls of radius $r/4$ to cover a ball of radius r. Obviously, $Q \le M^2$.

Let C^* be an optimal solution for an arbitrary instance $\langle S,k\rangle$ of the k-center problem. For each center $c \in C^*$, consider its coverage ball of radius $\mathrm{OPT}(S,k)$. The ball can be covered by Q balls of radius $\mathrm{OPT}(S,k)/4$, so we can partition the coverage area of c to Q regions in which the distance between any pair of elements is at most $\mathrm{OPT}(S,k)/2$. For each region R, if $S \cap R$ is not empty, select an arbitrary member as a center covering all the elements in $S \cap R$. This results in a feasible solution with at most Qk centers having cost at most $\mathrm{OPT}(S,k)/2$, and thus $\mathrm{OPT}(S,Qk) \le \mathrm{OPT}(S,k)/2$. So, the half-coverage constant is at most $Q \le M^2$. □

Now, we show how a bounded half-coverage constant in a metric space can help us achieve better approximation factors for the balanced k-center problem.

Lemma 3. *In a metric space with half-coverage constant H, if we modify the first step of Algorithm 1 to select $H^t k$ (instead of k) centers in each machine, then the approximation factor of Algorithm 1 is reduced to $\beta + \frac{2\alpha+4\alpha\beta}{2^t}$.*

Proof. The proof is quite similar to Theorem 1. Assume that $C^* = (D^*, A^*)$ is an optimal solution for the whole input set $S = \bigcup_{i=1}^{m} S_i$. We first update the result of Lemma 1 based on the modified algorithm. Define $C^\dagger$, $C_i^\dagger$, and $\hat{C}$ in the same way as in the proof of Lemma 1. In addition, let $C_i^\ddagger$ be an optimal solution for the uncapacitated k-center problem on input S_i with $H^t k$ centers. Now,

$$\begin{aligned}
R_{S_i,w}(C_i) &\le \alpha R_{S_i,w}(C_i^\ddagger) && \text{(by approximation factor of algorithm } \mathcal{A}) \\
&\le \frac{\alpha}{2^t} R_{S_i,w}(C_i^\dagger) && \text{(by Observation 3)} \\
&\le \frac{\alpha}{2^t} R_{S_i,w}(\hat{C}) && \text{(by inequality (3))} \\
&\le \frac{2\alpha}{2^t} R_{S,w}(C^\dagger) && \text{(by inequality (4))} \\
&\le \frac{2\alpha}{2^t} R_{S,w}(C^*) && \text{(by inequality (1))} \qquad (8)
\end{aligned}$$

We also show that for the modified version of Algorithm 1, we have a result similar to Lemma 2. We can have the same definitions of $\hat{C}$ and C^+ as its proof and will then have:

$$\begin{aligned}
R_{T,w'}(C) &\le \beta R_{T,w'}(C^+) && \text{(by algorithm } \mathcal{B}) \\
&\le \beta R_{T,w'}(\hat{C}) && \text{(by (7))} \\
&\le \beta(R_{S_i,w}(C_i) + R_{S,w}(C^*) + R_{S_{i'},w}(C_{i'})) && \text{(similar to (6))} \\
&\le \beta(\frac{2\alpha}{2^t}R_{S,w}(C^*) + R_{S,w}(C^*) + \frac{2\alpha}{2^t}R_{S,w}(C^*)) && \text{(by (8))} \\
&= \beta(1 + \frac{4\alpha}{2^t})R_{S,w}(C^*) && (9)
\end{aligned}$$

We finally prove this theorem similar to the proof of Theorem 1. Let $x \in S_i$ (for some $i \in \{1, \ldots, m\}$) have the maximum distance to its covering center c in C'. Therefore,

$$\begin{aligned}
R_{S,w}(C') &= d(x, c) && \text{(by definition of } R_{S,w}) \\
&\le R_{S_i,w}(C_i) + R_{T,w'}(C) && \text{(similar to proof of Theorem 1)} \\
&\le \frac{2\alpha}{2^t}R_{S,w}(C^*) + \beta(1 + \frac{4\alpha}{2^t})R_{S,w}(C^*) && \text{(by inequalities (8) and (9))} \\
&= (\beta + \frac{2\alpha + 4\alpha\beta}{2^t})R_{S,w}(C^*),
\end{aligned}$$

which completes the proof. □

Theorem 5. *There is a distributed algorithm with an approximation factor of* $5 + \varepsilon$ *for the weighted balanced* k*-center problem in metric spaces with bounded doubling dimension.*

Proof. Again, we use Gonzalez's algorithm [14] for algorithm $\mathcal{A}$, and the algorithm of Khuller and Sussmann [18] for algorithm $\mathcal{B}$, yielding $\alpha = 2$ and $\beta = 5$. Therefore, by Lemma 3, the approximation factor of our algorithm is

$$\beta + \frac{2\alpha + 4\alpha\beta}{2^t} = 5 + \frac{44}{2^t} \le 5 + \varepsilon,$$

which holds for $t = \lceil \log_2(\frac{44}{\varepsilon}) \rceil$. Furthermore, the communication complexity of the algorithm, i.e. $O(mk)$, is multiplied by $H^t \approx \left(\frac{44}{\varepsilon}\right)^{\log_2 H}$, which is poly$(\frac{1}{\varepsilon})$. □

5 Conclusion

In this paper, we presented a new approximation algorithm for the weighted balanced k-center problem, improving over the best current algorithm of Bateni *et al.* [4]. We also showed that the approximation factor of our algorithm can be improved to $5 + \varepsilon$ in some metric spaces including those with bounded doubling dimension. The ideas used in this paper seems applicable to other variants of the centroid-based clustering, including k-median and k-means, and hence are open for further investigation.

References

1. S. Aghamolaei, M. Farhadi, and H. Zarrabi-Zadeh. Diversity maximization via composable coresets. In *Proc. 27th Canad. Conf. Computat. Geom.*, page 43, 2015.
2. H.-C. An, A. Bhaskara, C. Chekuri, S. Gupta, V. Madan, and O. Svensson. Centrality of trees for capacitated k-center. *Math. Program.*, 154(1-2):29–53, 2015.
3. J. Barilan, G. Kortsarz, and D. Peleg. How to allocate network centers. *J. Algorithms*, 15(3):385–415, 1993.
4. M. Bateni, A. Bhaskara, S. Lattanzi, and V. Mirrokni. Distributed balanced clustering via mapping coresets. In *Proc. 27th Annu. Conf. Neural Info. Processing Systems*, pages 2591–2599, 2014.
5. M. Ceccarello, A. Pietracaprina, and G. Pucci. Solving k-center clustering (with outliers) in mapreduce and streaming, almost as accurately as sequentially. In *Proc. 45th Internat. Conf. Very Large Data Bases*, volume 12, pages 766–778, 2019.
6. D. Chakrabarty, R. Krishnaswamy, and A. Kumar. The heterogeneous capacitated k-center problem. In *Proc. 19th Internat. Conf. Integer Prog. and Comb. Opt.*, pages 123–135, 2017.
7. K. Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM J. Comput.*, 39(3):923–947, 2009.
8. M. Cygan, M. Hajiaghayi, and S. Khuller. Lp rounding for k-centers with non-uniform hard capacities. In *Proc. 53rd Annu. IEEE Sympos. Found. Comput. Sci.*, pages 273–282, 2012.
9. M. Cygan and T. Kociumaka. Constant factor approximation for capacitated k-center with outliers. In *Proc. 31st Sympos. Theoret. Aspects Comput. Sci.*, volume 25 of *STACS '14*, pages 251–262, 2014.
10. H. Ding. Balanced k-center clustering when k is a constant. In *Proc. 29th Canad. Conf. Computat. Geom.*, pages 179–184, 2017.
11. H. Ding, L. Hu, L. Huang, and J. Li. Capacitated center problems with two-sided bounds and outliers. In *Proc. 15th Workshop Algorithms Data Struct.*, pages 325–336, 2017.
12. A. Ene, S. Im, and B. Moseley. Fast clustering using mapreduce. In *Proc. 17th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining*, pages 681–689, 2011.
13. C. G. Fernandes, S. P. de Paula, and L. L. C. Pedrosa. Improved approximation algorithms for capacitated fault-tolerant k-center. In *Proc. 12th Latin American Theoret. Inform. Sympos.*, pages 441–453, 2016.
14. T. F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoret. Comput. Sci.*, 38:293–306, 1985.
15. D. S. Hochbaum and D. B. Shmoys. A best possible heuristic for the k-center problem. *Math. Oper. Res.*, 10(2):180–184, 1985.
16. S. Im and B. Moseley. Brief announcement: Fast and better distributed mapreduce algorithms for k-center clustering. In *Proc. 27th ACM Sympos. Parallel Algorithms Architect.*, pages 65–67, 2015.
17. P. Indyk, S. Mahabadi, M. Mahdian, and V. S. Mirrokni. Composable core-sets for diversity and coverage maximization. In *Proc. 33rd Sympos. Principles of Database Systems*, pages 100–108, 2014.
18. S. Khuller and Y. J. Sussmann. The capacitated k-center problem. *SIAM J. Discrete Math.*, 13(3):418–403, 2000.
19. G. Malkomes, M. J. Kusner, W. Chen, K. Q. Weinberger, and B. Moseley. Fast distributed k-center clustering with outliers on massive data. In *Proc. 28th Annu. Conf. Neural Info. Processing Systems*, pages 1063–1071, 2015.

20. J. McClintock and A. Wirth. Efficient parallel algorithms for k-center clustering. In *Proc. 45th Internat. Conf. Parallel Processing*, pages 133–138, 2016.
21. R. M. McCutchen and S. Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. In *Proc. 11th Internat. Workshop Approx. Algorithms*, volume 5171, pages 165–178, 2008.
22. K. Mirjalali, S. A. Tabatabaee, and H. Zarrabi-Zadeh. Distributed unit clustering. In *Proc. 31st Canad. Conf. Computat. Geom.*, pages 236–241, 2019.
23. V. S. Mirrokni and M. Zadimoghaddam. Randomized composable core-sets for distributed submodular maximization. In *Proc. 47th Annu. ACM Sympos. Theory Comput.*, pages 153–162, 2015.