IMPROVED APPROXIMATIONS FOR DISTRIBUTED k-CENTER CLUSTERING WITH OUTLIERS AND FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting representatives from distributed data requires geometric coverage, robustness to outliers, and control over group representation among the selected centers. We study k-center clustering with outliers and its fair extension, which imposes bounds on the number of centers selected from each of g disjoint groups. We develop a two-round MAPREDUCE framework for both problems in general metric spaces. The framework uses an unweighted summary that preserves $z+1$ nearby witnesses for every omitted point. Consequently, covering all but at most z summary points allows coverage to be extended to the original dataset without increasing the outlier budget. For k-center with outliers, a greedy merging procedure achieves a 6-approximation using $O\big(k(z+1)+z\big)$ point records per worker. For the fair extension, we additionally preserve nearby same-group candidates and combine LP-guided filtering with maximum-weight bipartite b-matching to obtain a first 11-approximation using $O\big(k(z+g)+z\big)$ point records per worker. These factors hold when the optimal radius is supplied. By guessing the radius, our method yields $(6+\varepsilon)$- and $(11+\varepsilon)$-approximations, respectively, with the number of guesses charged explicitly in communication. Finally, we complement our theoretical results with experiments on six real-world datasets and a synthetic dataset with known optimal solutions, demonstrating substantial improvements in clustering utility over state-of-the-art methods.

1 INTRODUCTION

Clustering is a fundamental task for selecting a small set of representative points in data summarization (Huang et al., 2019) and an important tool for scalable unsupervised learning (Czumaj et al., 2025). A natural formulation is the k-center problem, which selects a set $C \subseteq P$ of at most k centers to minimize $\max_{p\in P} d(p,C)$, where $d(p,C)=\min_{c\in C} d(p,c)$. However, this maximum-distance objective is sensitive to outliers: a few isolated or noisy observations can dominate the covering radius and divert centers from the rest of the dataset. To mitigate this sensitivity, the k-center with outliers problem (Charikar et al., 2001) allows a set $O \subseteq P$ of at most z points to be excluded and minimizes $\max_{p\in P\setminus O} d(p,C)$ subject to $|C| \le k$. This formulation has received sustained attention, leading to approximation and coreset algorithms (Ding et al., 2019), streaming algorithms (Matthew McCutchen & Khuller, 2008), and distributed methods for large-scale datasets (Malkomes et al., 2015; Ceccarello et al., 2019; Huang et al., 2023; 2026).

For large-scale datasets, distributed computation offers a natural way to construct this data summarization without gathering all input points on one machine (Malkomes et al., 2015; Ceccarello et al., 2019). When points also belong to demographic or application-defined groups, geometric coverage alone does not control which groups supply the selected representatives. This motivates center-selection constraints in addition to fairness (Kleindessner et al., 2019; Chiplunkar et al., 2020; Guo et al., 2026a). These constraints govern group representation among the selected centers. In contrast, other fairness formulations impose bounds on group proportions within each cluster (Chierichetti et al., 2017; Wu et al., 2024) or control exclusions through group-specific coverage requirements and outlier budgets (Jia et al., 2022). These notions differ from the center-selection constraints in this work.

In this paper, we study two related problems in general metric spaces: distributed k-center with outliers and its extension with center-selection fairness. Both problems seek a set of at most k input points as centers that minimizes the maximum distance from any remaining point to its nearest center

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

after excluding at most z outliers. In the fair variant, the input is partitioned into disjoint groups $G_1, \ldots, G_g$, and the selected center set C must satisfy $|C \cap G_\ell| \leq k_\ell$ for every $\ell \in [g]$, where $k = \sum_{\ell=1}^{g} k_\ell$.

Distributed k-center with outliers has been studied under different approximation and resource guarantees (Malkomes et al., 2015; Ceccarello et al., 2019; Huang et al., 2023; 2026). Malkomes et al. (Malkomes et al., 2015) give a two-round MapReduce algorithm with a 13-approximation guarantee. Ceccarello et al. (Ceccarello et al., 2019) achieve a two-round $(3+\varepsilon)$-approximation in general metric spaces, but their coreset-size and memory bounds depend exponentially on the doubling dimension D. Their approach is therefore particularly space-efficient when D is small.

Motivated by this approximation-space tradeoff, we develop a two-round MAPREDUCE framework with local summary-size bounds, inspired by Matthew McCutchen & Khuller (2008), that are independent of the doubling dimension. With geometric radius guessing over a supplied positive radius interval, our algorithm achieves a $(6+\varepsilon)$-approximation for k-center with outliers. We further extend the framework to data summarization fairness Yuan et al. (2023), obtaining an $(11+\varepsilon)$-approximation while preserving all group upper bounds and the global outlier budget.

Our Result is summarized as follows:

- By ensuring that every point omitted from a local summary has at least $z+1$ nearby representatives in that summary, we obtain a $(6+\varepsilon)$-approximation for the distributed k-center with outliers (DkCO) problem when the optimal radius is supplied.
- We extend the framework by preserving nearby same-group candidates and combining established LP-guided filtering with maximum-weight bipartite b-matching. The resulting algorithm achieves an $(11+\varepsilon)$-approximation for the distributed fair k-center with outliers (DFkCO) problem when the optimal fair radius is supplied. Each worker transmits at most $k(z+g+1)+z$ point records.
- Experiments on synthetic and real-world datasets show that our algorithms achieve clustering costs comparable to those of the baseline methods for both algorithms.

1.1 RELATED WORKS

DISTRIBUTED AND STREAMING k-CENTER WITH OUTLIERS. To the best of our knowledge, Malkomes et al. (2015) give a two-round MapReduce 13-approximation using local representatives and multiplicity weights, with outlier handling based on the classical approach of Charikar et al. (2001). Ceccarello et al. (2019) develop coreset-based MapReduce and streaming algorithms, including a two-round $(3+\varepsilon)$-approximation for k-center with outliers. Their algorithms apply to general metric spaces, but their representation-size and memory bounds depend on the doubling dimension. Recently, Wu et al. (2025) give the first fully scalable MPC algorithm in constant-dimensional Euclidean spaces, achieving an $O(\log^* n)$ approximation with at most $(1+\varepsilon)z$ outliers in $O(\log\log n)$ rounds and $\Theta(n^\delta)$ local space per machine. In the streaming model, Matthew McCutchen & Khuller (2008) obtain a $(4+\varepsilon)$-approximation for k-center with outliers, while Kale (2019) constructs small-space summaries for matroid center, including its robust variant. These results offer different approximation-space tradeoffs. In contrast, our unweighted summaries preserve the outlier budget on the original input through multiple nearby witnesses, trading additional redundancy for a dimension-independent local size bound.

LARGE-SCALE FAIR k-CENTER FOR DATA SUMMARIZATION. Center-selection fairness constrains group representation among the selected centers. An early algorithmic foundation is the matroid-center formulation of Chen et al. (2016), whose 3-approximation applies to group upper bounds through a partition matroid. Kleindessner et al. (2019) explicitly study fair k-center for data summarization and develop an efficient algorithm with approximation factor $3 \cdot 2^{g-1} - 1$. Jones et al. (2020) subsequently obtain a 3-approximation in $O(nk)$ time using maximum matching. In the streaming setting, Kale (2019) gives a two-pass $(3+\varepsilon)$-approximation for matroid center, together with an efficient one-pass $(17+\varepsilon)$-approximation. Chiplunkar et al. (2020) develop more efficient two-pass streaming and distributed algorithms for center-selection fairness. Subsequent work improves the one-pass approximation guarantee to $7+\varepsilon$ (Lin et al., 2024) and then $5+\varepsilon$ (Guo et al., 2026b) in general metrics, while Guo et al. (2026a) studies one-pass algorithms for Euclidean

inputs. For distributed computation, Wu et al. (2024) gives a two-round $(4+\varepsilon)$-approximation for data-summarization fairness without violating the center quotas, with communication bounds depending exponentially on the doubling dimension.

FAIR k-CENTER WITH OUTLIERS. For Robust Matroid Center, Harris et al. (2019) give an LP-based 3-approximation, while Chakrabarty & Negahbani (2019) develop a more general framework based on weighted optimization under an additional partition constraint. Anegg et al. (2023) provide a combinatorial 5-approximation using Rado matroids. For large-scale computation, Ceccarello et al. (2023) develop coreset-based MapReduce and streaming algorithms for Robust Matroid Center, obtaining a $(3+\varepsilon)$-approximation with an appropriate offline solver in two MapReduce rounds and one streaming pass, respectively; their compact space bounds depend exponentially on the doubling dimension. Recently, Yuan et al. (2023) study both centralized and distributed algorithms. And Amagata (2024) propose a randomized centralized $(3+\gamma)$-approximation together with a sampling-based acceleration, allowing a relaxed outlier budget of $(1+\eta)z$.

2 PRELIMINARY

Let (P, d) be a finite metric space with n points, with m machines with disjoint partition $P = P_1 \dot\cup \cdots \dot\cup P_m$. Independently, let $P = G_1 \dot\cup \cdots \dot\cup G_g$ be a partition into labeled groups. For a point $p \in P$ and a center set $C \subseteq P$, define $d(p, C) = \min_{c\in C} d(p, c)$. Let $B(p, a) = \{q \in P : d(p, q) \le a\}$ denote the closed ball of radius $a \ge 0$ centered at p.

FAIR k-CENTER WITH OUTLIERS. (FkCO) Given the fairness constraint $\{k_\ell\}_{\ell=1}^g$ with $k = \sum_{\ell=1}^g k_\ell$ and outlier budget z, a feasible solution is (C, O) with $C, O \subseteq P$, $|O| \le z$, and $|C \cap G_\ell| \le k_\ell$ for every $\ell \in [g]$. The fair k-center with outliers problem asks for a feasible solution minimizing

$$\mathrm{cost}(P \setminus O, C) = \max_{p\in P\setminus O} \min_{c\in C} d(p, c).$$

The k-**CENTER WITH OUTLIERS** (kCO) problem is the special case $g = 1$, $G_1 = P$, and $k_1 = k$.

We denote the optimal value of kCO and FkCO by r_f and r_f^* respectively. Note that we set an optimal solution (C^*, O^*) for analysis only.

DISTRIBUTED MODEL. The input dataset P is distributed across m workers according to the disjoint partition $P = P_1 \dot\cup \cdots \dot\cup P_m$. Each worker i constructs a summary $\mathcal{S}_i$ using only its local subset P_i and sends it to a designated coordinator. The coordinator aggregates these summaries into $\mathcal{S} = \bigcup_{i=1}^m \mathcal{S}_i$ and performs the global optimization to obtain the final solution.

We assume that the coordinator has enough memory to store the aggregated summary and the workspace required by the optimization subroutines. Both algorithms follow this two-stage structure: parallel local summarization followed by coordinator optimization. Under the adopted MapReduce round-counting convention, these stages correspond to two rounds. The missing proofs can be found in Appendix.

3 k-CENTER WITH OUTLIERS UNDER MAPREDUCE SETTING

3.1 THE TWO-STAGE ALGORITHM

Recall that $P = P_1 \dot\cup \cdots \dot\cup P_m$ be a finite metric instance, with the k centers and z outliers. Note that r^* is the optimum of the k-center problem with at most z outliers, and fix an optimal solution (C^*, O^*) with $|C^*| \le k$ and $|O^*| \le z$. The optimal solution is used only in the analysis. Let $z_i^* = |P_i \cap O^*|$, $L = k(z+1) + z$, and $S = \bigcup_{i=1}^m S_i$ collects all points saving at local stages from the m machines.

Following common practice in Malkomes et al. (2015), we assume that the optimal radius is known throughout the paper and show how to handle the unknown case in the Appendix. Let $R = r^*$ We propose that the key idea of Alg. 1 consists of two stages as follows:

1. PARALLEL STAGE: Each machine i constructs an unweighted summary S_i on its dataset P_i, such as any $p \in P_i$ has $|B(p, 2R) \cap S_i| \ge z + 1$.

Algorithm 1: REDUNDANTMR

Input: Dataset P, parameter k, outlier budget z and radius R.
Output: Centers $C \subseteq P$ and outliers $O \subseteq P$.

Send the subdataset to each machine as P_i; // Stage 1: PARALLEL STAGE
foreach $i \in [m]$ in parallel **do**
  $S_i \leftarrow \varnothing$
  foreach $p \in P_i$ **do**
    if $|S_i \cap B(p, 2R)| < z + 1$ **then**
      $S_i \leftarrow S_i \cup \{p\}$
  Send (R, i, true, S_i), with distinct point identities
$S \leftarrow \bigcup_{i=1}^{m} S_i$; $U \leftarrow S$; $C \leftarrow \varnothing$; // Stage 2: COORDINATOR STAGE
while $|U| > z$ ***and*** $|C| < k$ **do**
  Choose $c \in \arg\max_{u \in U} |U \cap B(u, 2R)|$
  $C \leftarrow C \cup \{c\}$
  $U \leftarrow U \setminus B(c, 4R)$
return C; $O = U$

2. COORDINATOR STAGE: The coordinator machine collects $S = \bigcup_{i=1}^{m} S_i$ and repeatedly selects a point c whose $\arg\max_{s \in S} |S \cap B(s, 2R)|$, removing all points within $B(s, 4R)$.

3.2 THEORETICAL ANALYSIS

We first show that Stage 1 retains enough points. We then analyze the coordinator stage, which selects centers based on $2R$-ball counts and removes points within $4R$-balls.

Lemma 1. *After Stage 1 of Alg. 1 completes, every point $p \in P_i \setminus S_i$ satisfies $|S_i \cap B(p, 2R)| \geq z+1$.*

Lemma 2. *For every $R \geq r^*$, Stage 2 of Alg. 1 returns a center set $C \subseteq S$ and a remaining set $U \subseteq S$ satisfying $|C| \leq k$, $|U| \leq z$, and $d(s, C) \leq 4R$ for every $s \in S \setminus U$.*

The following lemma transfers the summary-level coverage guarantee to all original points without increasing the outlier budget.

Lemma 3. *The output of Alg. 1 has $|C| \leq k$, $|U| \leq z$, and $d(p, C) \leq 6R$ for every $p \in P \setminus U$. In other words, $cost(P \setminus U, C) \leq 6R$.*

Proof. By Lem. 2, the coordinator returns $|C| \leq k$ and $|U| \leq z$, and every $s \in S \setminus U$ satisfies $d(s, C) \leq 4R$.

Now consider any point $p \in P \setminus S$, and let P_i be the local partition containing p. Since $S_i \subseteq S$, we have $p \in P_i \setminus S_i$. By Lem. 1, $|S_i \cap B(p, 2R)| \geq z + 1$. Because U contains at most z points, these $z + 1$ points cannot all belong to U. Hence, there exists $s \in (S_i \setminus U) \cap B(p, 2R)$. Since $s \in S \setminus U$, we have $d(s, C) \leq 4R$, and thus $d(p, C) \leq d(p, s) + d(s, C) \leq 2R + 4R = 6R$. Therefore, every point in $S \setminus U$ is covered within $4R$, while every point in $P \setminus S$ is covered within $6R$. Since $(S \setminus U) \cup (P \setminus S) = P \setminus U$, every point in $P \setminus U$ is covered within $6R$. Thus, $\{p \in P : d(p, C) > 6R\} \subseteq U$, and consequently $\text{cost}(P \setminus U, C) \leq 6R$. □

Theorem 4. *There exist Alg. 1 achieves a 6-approximation for* DkCO. *Each worker transmits at most $L = k(z + 1) + z$ points, and the total communication is at most $mk(z + 1) + z$ points.*

4 EXTENSION TO FAIR CENTER SELECTION

Based on the structure of the distributed k-center with the outlier problem (DKCO), we combine it with the fairness of data summarization to propose an improved approximation ratio for the distributed fair k-center with the outlier (DKFCO) problem.

Algorithm 2: LOCALFAIRSUMMARY

Input: A dataset $P = P_1 \dot\cup \cdots \dot\cup P_m$ with group labels χ; parameter $\{k_\ell\}_{\ell=1}^g$, outliers, z radius $R = r_f^*$.
Output: $S_i \cup T_i$.

```
Send the subdataset to each machine as P_i
Initialization S_i ← ∅, T_{i,ℓ} ← ∅ and A_{i,ℓ} ← ∅ for every ℓ ∈ [g]
foreach p ∈ P_i in a fixed order do
    ℓ ← χ(p)
    if |S_i ∩ B(p, 2R)| < z + 1 then S_i ← S_i ∪ {p}
    if k_ℓ > 0 then
        if T_{i,ℓ} ∩ B(p, 2R) = ∅ then T_{i,ℓ} ← T_{i,ℓ} ∪ {p}
T_i ← ⋃_{ℓ=1}^g T_{i,ℓ}
return S_i ∪ T_i.
```

4.1 ALGORITHMIC FRAMEWORK

Assume m disjoint machines partition the dataset, denoted as $P = P_1 \dot\cup \cdots \dot\cup P_m$, and $P = G_1 \dot\cup \cdots \dot\cup G_g$ be a partition into fairness groups, with label function $\chi : P \to [g]$. Then, we propose that the key idea of the fairness setting consists of two stages as follows:

1. PARALLEL STAGE (ALG. 2): Each machine i constructs two sets, including an unweighted set S_i on its dataset P_i, such that any $p \in P_i \setminus S_i$ has $|B(p, 2R) \cap S_i| \geq z+1$, and the $2R$-independence center set $\Gamma_i = \bigcup_{j=1}^g \Gamma_{i,j}$ based on each color j, such that any $p \in P_i \cap G_j$ has $d(p, \Gamma_i) \leq 2R$.
2. COORDINATOR STAGE (ALG. 3): The coordinator collects $S = \bigcup_i S_i$ and $T = \bigcup_i \Gamma_i$, and uses an LP at $\tau = 3R$ to form 2τ-radius blocks. A maximum-weight b-matching, with block-size weights and color capacities k_ℓ, selects centers within τ of matched points. Finally, at most z points remain uncovered within 3τ and are returned as outliers.

Then, we analyze the approximation guarantee of the algorithm through these two stages.

Assume that an optimal fair solution (C^*, O^*) with radius r_f^*, where $|C^* \cap G_\ell| \leq k_\ell$ for every $\ell \in [g]$ and $|O^*| \leq z$. Let $T = \bigcup_{i=1}^m T_i$ be the candidate center set constructed by Alg. 2 with $R = r_f^*$. Note that the optimal sets C^* and O^* are used only in the analysis and are not required by the algorithm.

4.2 STAGE 1: FAIR CENTER CANDIDATES

Lemma 5. *The sets constructed by Alg. 2 satisfy $|S_i \cap B(p, 2R)| \geq z + 1$ for every $p \in P_i \setminus S_i$. Moreover, for every $p \in P_i \cap G_\ell$, we have $T_{i,\ell} \cap B(p, 2R) \neq \varnothing$; in other words, $d(p, T_{i,\ell}) \leq 2R$ for $T_{i,\ell} \subseteq T_i$.*

Lemma 6. *There exists $C \subseteq T$ such that $|C \cap G_\ell| \leq k_\ell, \forall \ell \in [g]$ and $d(p, C) \leq 3R$ for $p \in P \setminus O^*$.*

4.3 STAGE 2: LP-GUIDED CENTER SELECTION AND WEIGHTED MATCHING

Let $\tau = 3R = 3r_f^*$, $T_\ell = T \cap G_\ell$, and $T_s = T \cap B(s, \tau)$. Let x_s denote fractional coverage for $s \in S$, and y_t be fractional center selection for $t \in T$. The coordinator solves the following LP (I):

$$
\begin{aligned}
\min_{x,y} \quad & 0 \\
\text{s.t.} \quad & x_s \leq \sum_{t \in T_s} y_t && \forall s \in S && (1) \\
& \sum_{s \in S} x_s \geq |S| - z && && (2) \\
& \sum_{t \in T_\ell} y_t \leq k_\ell && \forall \ell \in [g] && (3) \\
& 0 \leq x_s \leq 1 && \forall s \in S \\
& 0 \leq y_t \leq 1 && \forall t \in T
\end{aligned}
$$

Algorithm 3: FAIRMERGE

Input: set $S = \bigcup_{i=1}^{m} S_i$; candidate centers $T = \bigcup_{i=1}^{m} T_i$ with group labels χ; parameter $k = \sum_{\ell=1}^{g} k_\ell$ and z; and threshold $\tau = 3R = 3r_f^*$.
Output: Centers $C \subseteq T$ and outliers $U \subseteq S$ with $|U| \leq z$.

1 $T_\ell \leftarrow \{t \in T : \chi(t) = \ell\}$ for every $\ell \in [g]$
2 Find a feasible solution (x, y) to LP (1)–(3) regarding to $S \cup T$ at threshold τ with $k = \sum_{\ell=1}^{g} k_\ell$
3 $V \leftarrow S$; $\Gamma \leftarrow \varnothing$
4 **while** $V \neq \varnothing$ **do**
5 Choose $v \in \arg\max_{s \in V} x_s$
6 $D_v \leftarrow V \cap B(v, 2\tau)$; $w_v \leftarrow |D_v|$; $T_v \leftarrow T \cap B(v, \tau)$
7 $\Gamma \leftarrow \Gamma \cup \{v\}$; $V \leftarrow V \setminus D_v$
8 $\mathcal{E} \leftarrow \{(v, \ell) \in \Gamma \times V_D : T_v \cap T_\ell \neq \varnothing\}$
9 Compute a maximum-weight bipartite b-matching $\mathcal{M}$ by Def. 7 on $\mathcal{H}(\Gamma, V_D; \mathcal{E})$, with weight w_v for each edge (v, ℓ), representative point capacity 1, and group capacity k_ℓ
10 $C \leftarrow \varnothing$
11 **foreach** $(v, \ell) \in \mathcal{M}$ **do**
12 Choose $t_v \in T_v \cap T_\ell$
13 $C \leftarrow C \cup \{t_v\}$
14 $U \leftarrow \{s \in S : d(s, C) > 3\tau\}$
15 **return** C and U.

Constraint (1) requires fractional points coverage to be supported by nearby candidate centers. Constraint (2) limits the total fractional uncovered points to z. Constraint (3) imposes an upper bound on the fractional number of selected centers in each group. Since the groups partition T and $k = \sum_{\ell=1}^{g} k_\ell$, the group upper bounds also imply $\sum_{t \in T} y_t \leq k$.

Lem. 6 guaranties the existence of a candidate set $C \subseteq T$ satisfying the group upper bounds and covering every point in $P \setminus O^*$ within τ. Setting $y_t = \mathbf{1}_{\{t \in C\}}$ and $x_s = \mathbf{1}_{\{s \notin O^*\}}$ therefore gives a feasible solution, since $|S \cap O^*| \leq z$. Then, the algorithm computes a feasible fractional solution (x, y) in step 2 of Alg. 3. Let S be the demand summary and T the candidate center set, with disjoint groups $G_1, \ldots, G_g$ and (x, y) be a feasible solution to the coverage LP at threshold τ. Then, we define the auxiliary graph and compute the maximum-weight b-matching to select the final centers C.

Definition 7 (AUXILIARY BIPARTITE GRAPH)**.** *The auxiliary bipartite graph is $\mathcal{H} = (\Gamma, V_D; \mathcal{E})$, where Γ is the left vertex set and $V_D = \{d_\ell : \ell \in [g]\}$ is a disjoint set of dummy vertices, with d_ℓ representing group G_ℓ. The edge set is $\mathcal{E} = \{(v, d_\ell) : v \in \Gamma,\ \ell \in [g],\ T_v \cap T_\ell \neq \varnothing\}$. Each edge $(v, d_\ell) \in \mathcal{E}$ has weight $\omega(v, d_\ell) = w_v$. The vertex capacities are $b(v) = 1$ for every $v \in \Gamma$ and $b(d_\ell) = k_\ell$ for every $\ell \in [g]$.*

Definition 8 (MAXIMUM-WEIGHT b-MATCHING)**.** *Let $\mathcal{H} = (\Gamma, V_D; \mathcal{E})$ be the auxiliary bipartite graph in Def. 7, with edge-weight function ω and vertex-capacity function b. A b-matching is a subset $\mathcal{M} \subseteq \mathcal{E}$ such that $\deg_{\mathcal{M}}(v) \leq b(v) = 1, \forall v \in \Gamma$, and $\deg_{\mathcal{M}}(d_\ell) \leq b(d_\ell) = k_\ell, \forall d_\ell \in V_D$. The weight of $\mathcal{M}$ is $W(\mathcal{M}) = \sum_{(v,d_\ell) \in \mathcal{M}} \omega(v, d_\ell) = \sum_{(v,d_\ell) \in \mathcal{M}} w_v$. A maximum-weight b-matching is a feasible b-matching $\mathcal{M}^\star$ satisfying*

$$W(\mathcal{M}^\star) = \max_{\mathcal{M} \text{ is a feasible } b\text{-matching}} W(\mathcal{M}).$$

Let $\mathcal{M}$ denote a maximum-weight b-matching in $\mathcal{H}$. For each matched edge $(v, d_\ell) \in \mathcal{M}$, we select one candidate $t_v \in T_v \cap G_\ell$. Since each point in Γ has capacity one and each dummy vertex d_ℓ has capacity k_ℓ, the resulting center set C satisfies $|C \cap G_\ell| \leq k_\ell$ for every $\ell \in [g]$. Moreover, since $\omega(v, d_\ell) = w_v = |D_v|$, maximizing $W(\mathcal{M})$ maximizes the total number of points contained in the selected blocks.

Lemma 9. *Let (x, y) be a feasible solution to LP (1), and let $\mathcal{M}$ be a maximum-weight b-matching in the auxiliary graph $\mathcal{H}$ constructed by Alg. 3. Then $W(\mathcal{M}) = \sum_{(v,d_\ell) \in \mathcal{M}} w_v \geq |S| - z$.*

Proof. When a point v is selected by step 5 in Alg. 3, it has the largest fractional coverage value among the current demands. Hence, for every $s \in D_v$, we have $x_s \leq x_v$, and thus $\sum_{s \in D_v} x_s \leq |D_v| x_v = w_v x_v$. Since steps 4-7 in Alg. 3 remove every demand exactly once, the sets $\{D_v : v \in \Gamma\}$

partition S, and we obtain

$$\sum_{s\in S} x_s = \sum_{v\in\Gamma}\sum_{s\in D_v} x_s \le \sum_{v\in\Gamma} w_v x_v.$$

Hence, $\sum_{v\in\Gamma} w_v x_v \ge \sum_{s\in S} x_s \ge |S| - z$.

For every selected point $v \in \Gamma$, since $v \in S$ and $T_v = T \cap B(v,\tau)$, Constraint (1) in LP (I) gives $x_v \le \sum_{t\in T_v} y_t$. Hence, we can choose values $0 \le q_{vt} \le y_t$ for $t \in T_v$ such that $\sum_{t\in T_v} q_{vt} = x_v$. For every group ℓ, define $q_{v\ell} = \sum_{t\in T_v\cap G_\ell} q_{vt}$.

We first verify the representative point capacities. Since the groups partition T, $\sum_{\ell=1}^{g} q_{v\ell} = \sum_{t\in T_v} q_{vt} = x_v \le 1$, where the last inequality follows from the variable bounds of LP (I).

We next verify the group capacities. For any two distinct representative points $u, v \in \Gamma$, the filtering procedure guarantees $d(u,v) > 2\tau$. Hence, $T_u \cap T_v = \varnothing$; otherwise, some $t \in T_u \cap T_v$ would satisfy $d(u,v) \le d(u,t) + d(t,v) \le 2\tau$, a contradiction. Thus, for every group ℓ,

$$\sum_{v\in\Gamma} q_{v\ell} = \sum_{v\in\Gamma}\sum_{t\in T_v\cap T_\ell} q_{vt} \le \sum_{t\in T_\ell} y_t \le k_\ell,$$

where the last inequality follows from Constraint (3) of LP (I).

Finally, q is a feasible fractional b-matching in $\mathcal{H}$. Its total weight is $\sum_{v\in\Gamma}\sum_{\ell=1}^{g} w_v q_{v\ell} = \sum_{v\in\Gamma} w_v x_v \ge |S| - z$. Then, the bipartite b-matching polytope with integral vertex capacities is integral. Hence, a maximum-weight integral b-matching has weight at least that of the feasible fractional b-matching constructed above. Therefore, $W(\mathcal{M}) \ge |S| - z$. □

4.4 Correctness of Alg. 3

We first show that Alg. 3 respects the group upper bounds and covers all but at most z summary points within 3τ. We then extend this guarantee to the original dataset using the redundant center set retained in Stage 1.

Lemma 10. *According to LP (I) including the variable bounds, let $k = \sum_{\ell=1}^{g} k_\ell$. Then Alg. 3 returns $C \subseteq T$ and $U = \{s \in S : d(s,C) > 3\tau\}$ satisfying $|C \cap G_\ell| \le k_\ell$ for every $\ell \in [g]$, $|C| \le k$, and $|U| \le z$.*

With $\tau = 3R$, Lem. 10 guarantees a $9R$ cover of $S \setminus U$. The following lemma transfers this coverage to the original dataset without increasing the outlier budget.

Lemma 11. *Let $S = \bigcup_{i=1}^{m} S_i$. Alg. 3 returns $C \subseteq T$ and $U \subseteq S$ with $|U| \le z$ and $d(s,C) \le 9R$ for every $s \in S \setminus U$, then $\{p \in P : d(p,C) > 11R\} \subseteq U$. In the other words, $\mathrm{cost}(P \setminus O, C) \le 11R$.*

Theorem 12. *The proposed MPC algorithm achieves an $(11+\varepsilon)$-approximation for fair k-center with outliers. With exact LP and matching optimization, it communicates $O(HmL)$ point records and requires $O(\max\{n/m, HmL + W_{\mathrm{opt}}\})$ memory per machine, where $L = k(z+g+1)+z$, $H = O(1+\varepsilon^{-1}\log(R_{\max}/R_{\min}))$, and W_{opt} are the peak additional coordinator workspace when processing radius trials sequentially.*

For every $R \ge r_f^*$, Lem. 6 guarantees LP feasibility at $\tau = 3R$. By Lem. 10, the coordinator returns centers satisfying all group upper bounds and covers all but at most z summary points within $3\tau = 9R$. Combining Lems. 5 and 11, the output $(C, O = U)$ satisfies $|C \cap G_\ell| \le k_\ell$ for every $\ell \in [g]$, $|C| \le k$, $|O| \le z$, and $\mathrm{cost}(P \setminus O, C) \le 11R$. For a fixed optimal solution (C^*, O^*), let $z_i^* = |P_i \cap O^*|$ and $L = k(z+g+1)+z$. The local insertion rules give $|S_i| \le k(z+1)+z_i^*$ and $|T_i| \le kg + z_i^*$ throughout each local scan. Counting the distinct optimal outliers only once in the union yields $|S_i \cup T_i| \le k(z+g+1)+z_i^* \le L$. Since $\sum_i z_i^* = |O^*| \le z$, the total communication for a sufficient radius is at most $mk(z+g+1)+z$ point records. To handle the unknown optimum, set $\eta = \varepsilon/11$ and construct a geometric grid with ratio $1+\eta$ over a supplied interval $0 < R_{\min} \le r_f^* \le R_{\max}$, including the first grid value at least $R_{\max}$. Each worker prepares capped summaries for all $H = O(1+\varepsilon^{-1}\log(R_{\max}/R_{\min}))$ radii and transmits them in one batch. The coordinator then searches locally as described in Section B, obtaining an accepted radius $\widehat{R} \le (1+\eta) r_f^*$. Every accepted trial retains the $11R$ coverage guarantee, so $\mathrm{cost}(P \setminus O, C) \le 11\widehat{R} \le (11+\varepsilon) r_f^*$. The total communication is $O(HmL)$ point records, with no

Table 1: Dataset summary.

Dataset	#Records	#Features	Group attribute (#groups)
ADULT	32,561	6	Sex (2); Race (5)
SKIN	245,057	3	–
COVERTYPE	581,012	54	Wilderness (4)
POWER	2,075,259	7	–
HIGGS	11,000,000	7	–
PHONES-3D	13,062,475	3	Activity (7)

additional worker communication during the search. Assuming balanced input partitions and constant-size point records, sufficient per-machine memory is $O(\max\{n/m, HmL + W_{\text{opt}}\})$, where W_{opt} denotes the peak additional coordinator workspace, including the LP representation and optimization subroutines, when radius trials are processed sequentially.

5 EXPERIMENTAL EVALUATION

We evaluate the clustering performance of our MPC algorithms on six real-world datasets. Our comparison includes both fair and unconstrained algorithms for k-center clustering with outliers. Further details are given in Appendix.

DATASETS. We evaluate our algorithms on six real-world datasets: ADULT, SKIN, COVERTYPE, HIGGS, and PHONES-3D. For the fair experiments, we use SEX and RACE separately on ADULT, WILDERNESS on COVERTYPE, and ACTIVITY on PHONES-3D to define the center groups. SKIN and HIGGS are used in the unconstrained experiments. Table 1 summarizes the dataset sizes, features, and group attributes.

BASELINES. We compare our algorithms with the distributed baselines as follows:

- Ceccarello et al. (2019) (CPP19-DET and CPP19-RAND): deterministic and randomized two-round MapReduce algorithms for k-center with outliers, respectively. Both achieve a $(3+\varepsilon)$-approximation, with the randomized guarantee holding with high probability.
- *Malkomes et al. (2015)* (OUTLIERS-MR): a $13+\varepsilon$-approximation distributed algorithm for k-center clustering with outliers.
- *Ceccarello et al. (2023)* (MRCORESET): a MapReduce algorithm for robust matroid center.
- *Yuan et al. (2023)* (YUAN-DIST): a distributed algorithm for fair k-center clustering with outliers.

EXPERIMENTAL SETTINGS. We evaluate the algorithms on both real-world and synthetic datasets with varying dataset sizes and different values of the center size k, the number of machines m, and the size of the outliers z. For the *real-world datasets*, we report the clustering cost (shown as Section 2) under different experimental settings, with at least 5 runs. Following the previous impact fairness principle of Feldman et al. (2015), we restrict the selection to k_ℓ data points from the ℓ-th group as centers, which sets the number of fair center selections from each group proportional to the group size. In our experiments, we use the Euclidean distance as the default metric, as it is commonly adopted in this setting. All experiments were conducted on a 10-core Apple M1 Max CPU with 32 GB of RAM using Python 3.

5.1 EXPERIMENTAL RESULTS

In Fig. 1, we compare OURALG-kCO (Alg. 1) with CPP19-DET, CPP19-RAND, and OUTLIERS-MR on 10K-point subsets of SKIN, COVERTYPE, POWER and HIGGS. These rows examine the effects of varying the center budget k, the outlier budget z, and the number of machines m, respectively. Unless otherwise varied, we set $k = 50$, $m = 8$, and $z = 256$. Overall, OURALG-kCO achieves the lowest or nearly the lowest clustering cost in most tested configurations. Its improvement over OUTLIERS-MR is consistent with its tighter worst-case approximation guarantee, while its

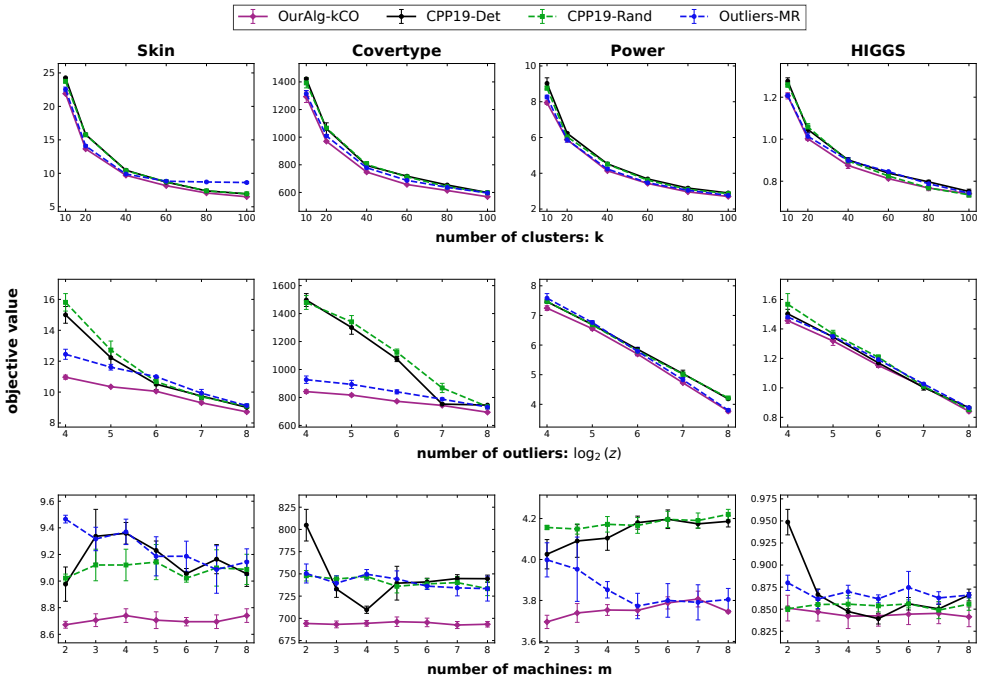


Figure 1: Clustering quality for distributed k-center with outliers.

competitive performance against the CPP19 variants indicates that the proposed summarization and merging strategy is also effective in practice.

As the center budget k increases from 10 to 100, the clustering cost generally decreases across all algorithms. This trend is expected for k-center clustering and k-center with outliers since a larger center budget provides greater flexibility in covering the dataset and can reduce the maximum distance to the selected centers. Similarly, increasing the outlier budget z generally decreases the clustering cost, since a larger number of distant points can be excluded from the coverage objective. These trends are consistent with the expectation of the kCO objective with respect to both k and z.

Then, OURALG-kCO achieves relatively low clustering costs under small outlier budgets across all datasets. In particular, for SKIN and COVERTYPE, our method substantially outperforms the baseline algorithms. We attribute this behavior in part to the $z+1$-witness preservation mechanism: every point omitted from a local summary is supported by at least $z+1$ nearby retained points, so excluding at most z summary points cannot eliminate all of its nearby representatives. This makes the final solution less sensitive to the loss of locally important points when the outlier budget is small.

As m increases, OURALG-kCO objective value remains stable and is generally lower than those of the baseline methods. This suggests that the proposed summarization and merging procedure is relatively insensitive to the number of data partitions over the tested range.

6 CONCLUSION

In this paper, we presented a two-round MapReduce framework for k-center with outliers and its extension with center fairness in general metric. The framework constructs unweighted local summaries that retain $z+1$ nearby witnesses for every omitted point, allowing coverage guarantees to transfer to the original dataset without increasing the outlier budget. A greedy merging procedure handles the unconstrained problem, while same-group candidate preservation and LP-guided maximum-weight bipartite b-matching enforce group upper bounds in the fair variant. With geometric radius guessing over a supplied positive interval, the algorithms achieve $(6+\varepsilon)$- and $(11+\varepsilon)$-approximations, respectively, while respecting the center constraints and the global outlier budget.

AI USE STATEMENT

We used an AI Model to assist with language editing, proofreading, and literature search. The author(s) take full responsibility for the final content of this work, including all theoretical claims, proofs, code, and reported experimental results.

REFERENCES

Daichi Amagata. Fair k-center clustering with outliers. In *International Conference on Artificial Intelligence and Statistics*, pp. 10–18. PMLR, 2024.

Georg Anegg, Laura Vargas Koch, and Rico Zenklusen. A simple combinatorial algorithm for robust matroid center. In *Symposium on Simplicity in Algorithms (SOSA)*, pp. 96–102. SIAM, 2023.

Matteo Ceccarello, A Pietracaprina, and G Pucci. Solving k-center clustering (with outliers) in mapreduce and streaming, almost as accurately as sequentially. *Proceedings of the VLDB Endowment*, 12(7):766–778, 2019.

Matteo Ceccarello, Andrea Pietracaprina, Geppino Pucci, and Federico Soldà. Scalable and space-efficient robust matroid center algorithms. *Journal of Big Data*, 10(1):49, 2023.

Deeparnab Chakrabarty and Maryam Negahbani. Generalized center problems with outliers. *ACM Transactions on Algorithms (TALG)*, 15(3):1–14, 2019.

Moses Charikar, Samir Khuller, David M Mount, and Giri Narasimhan. Algorithms for facility location problems with outliers. In *SODA*, volume 1, pp. 642–651, 2001.

Danny Z Chen, Jian Li, Hongyu Liang, and Haitao Wang. Matroid and knapsack center problems. *Algorithmica*, 75:27–52, 2016.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5036–5044, 2017.

Ashish Chiplunkar, Sagar Kale, and Sivaramakrishnan Natarajan Ramamoorthy. How to solve fair k-center in massive data models. In *Proceedings of the 37th ICML*, pp. 1877–1886, 2020.

Artur Czumaj, Guichen Gao, Mohsen Ghaffari, and Shaofeng H.-C. Jiang. Fully Scalable MPC Algorithms for Euclidean k-Center. In *52nd International Colloquium on Automata, Languages, and Programming (ICALP 2025)*, volume 334, 2025.

Hu Ding, Haikuo Yu, and Zixiu Wang. Greedy Strategy Works for k-Center Clustering with Outliers and Coreset Construction. In *27th Annual European Symposium on Algorithms (ESA 2019)*, volume 144 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 40:1–40:16, 2019.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Longkun Guo, Chaoqi Jia, and Chao Chen. Fair k-center clustering on massive social network data streams. In *Proceedings of the ACM Web Conference 2026*, pp. 4689–4700, 2026a.

Longkun Guo, Zeyu Lin, Chaoqi Jia, and Chao Chen. Improved streaming algorithm for fair k-center clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36946–36954, 2026b.

David G Harris, Thomas Pensyl, Aravind Srinivasan, and Khoa Trinh. A lottery model for center-type problems with outliers. *ACM Transactions on Algorithms (TALG)*, 15(3):1–25, 2019.

Junyu Huang, Qilong Feng, Ziyun Huang, Jinhui Xu, and Jianxin Wang. Fast algorithms for distributed k-clustering with outliers. In *International Conference on Machine Learning*, pp. 13845–13868. PMLR, 2023.

Junyu Huang, Zhize Li, Zhen Zhang, Xujia Li, Jianxin Wang, and Qilong Feng. New algorithms for fully-dynamic k-center with outliers. In *Forty-third International Conference on Machine Learning*, 2026.

Lingxiao Huang, Shaofeng Jiang, and Nisheeth Vishnoi. Coresets for clustering with fairness constraints. *Advances in neural information processing systems*, 32, 2019.

Xinrui Jia, Kshiteej Sheth, and Ola Svensson. Fair colorful k-center clustering. *Mathematical Programming*, 192(1-2):339–360, 2022.

Matthew Jones, Huy Nguyen, and Thy Nguyen. Fair k-centers via maximum matching. In *International Conference on Machine Learning*, pp. 4940–4949. PMLR, 2020.

Sagar Kale. Small space stream summary for matroid center. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*, 145:20, 2019.

Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*, pp. 3448–3457. PMLR, 2019.

Zeyu Lin, Longkun Guo, and Chaoqi Jia. Streaming fair k-center clustering over massive dataset with performance guarantee. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 105–117. Springer, 2024.

Gustavo Malkomes, Matt J Kusner, Wenlin Chen, Kilian Q Weinberger, and Benjamin Moseley. Fast distributed k-center clustering with outliers on massive data. *Advances in Neural Information Processing Systems*, 28, 2015.

Richard Matthew McCutchen and Samir Khuller. Streaming algorithms for k-center clustering with outliers and with anonymity. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pp. 165–178, 2008.

Di Wu, Qilong Feng, Junyu Huang, Jinhui Xu, Ziyun Huang, and Jianxin Wang. Fully-scalable massively parallel algorithm for k-center with outliers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21519–21527, 2025.

Xiaoliang Wu, Qilong Feng, Ziyun Huang, Jinhui Xu, and Jianxin Wang. New algorithms for distributed fair k-center clustering: Almost accurate as sequential algorithms. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 1938–1946, 2024.

Fan Yuan, Luhong Diao, Donglei Du, and Lei Liu. Fair k-center problem with outliers on massive data. *Tsinghua Science and Technology*, 28(6):1072–1084, 2023.