BETTER APPROXIMATION FOR DISTRIBUTED DATA SUMMARIZATION FAIR k-CENTER IN THE MPC MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

The massively parallel computation (MPC) model provides a theoretical framework for large-scale data processing systems such as MapReduce. We study data-summarization fair k-center in a one-round, coordinator-based MPC model. The objective is to select at most k centers from the input to minimize the maximum distance from any input point to its nearest center, subject to group upper bounds. Given n points partitioned into m groups and distributed across ℓ workers, each worker independently constructs a subset summary, from which the coordinator computes a fair solution. Our framework combines group-wise local summaries with a compact maximum-flow formulation: the summaries preserve same-group replacements for optimal centers, while geometric separation ensures flow feasibility. This yields a $(7+\varepsilon)$-approximation in general metrics using at most $\min\{n, \ell mk\}$ transmitted point records, improving the previous approximation factor of 17 while retaining $O(km)$ transmitted records per worker. When the optimal fair radius is known to the coordinator, refined summaries yield a $(3+\varepsilon)$-approximation in metrics of bounded doubling dimension and a $(5.84+\varepsilon)$-approximation in Euclidean spaces of arbitrary dimension. For one-way subset-summary protocols with only local distance access, we also establish an $\Omega(n)$ communication lower bound for approximation factors strictly below 5 on instances with $\Theta(n)$ workers. Experiments on three real-world datasets and synthetic inputs demonstrate competitive clustering quality and practical efficiency under simulated parallel execution.

1 INTRODUCTION

Selecting a small set of representatives is a fundamental task in data summarization, with applications in search, recommendation, and content selection (Kleindessner et al., 2019; Mahabadi & Trajanovski, 2023). The k-center objective captures the geometric coverage of such a summary. Given a finite metric space (P, d) with $|P| = n$ and a budget k, it seeks $C \subseteq P$ with $|C| \leq k$ minimizing $\mathrm{cost}(C) = \max_{p \in P} \min_{c \in C} d(p, c)$. Coverage alone, however, does not control which demographic or categorical groups appear among the selected representatives. The importance of representation is illustrated by the gender stereotypes and imbalances documented in occupational image search (Kay et al., 2015).

We study *data-summarization fair k-center* (DSFkC) with group upper bounds, following Chiplunkar et al. (2020). The dataset is partitioned into m disjoint groups $G_1, \ldots, G_m$, and the selected centers must satisfy $|C \cap G_j| \leq k_j$ for every $j \in [m]$, where the nonnegative integer budgets satisfy $\sum_{j=1}^{m} k_j = k$. Thus, fairness constrains the composition of the representative set, and the output may contain fewer than k centers. These constraints couple coverage decisions across the dataset: a geometrically useful candidate may be unavailable once its group's budget has been exhausted.

For massive datasets, these coupled decisions must be made across distributed data. We consider a coordinator-based massively parallel computation (MPC) model in which P is partitioned among ℓ workers. Each worker independently sends a subset of its local points to a coordinator, which selects the final centers from the received summaries. We count this simultaneous transmission as one communication round and measure communication in point records. The central challenge is to preserve both local coverage and the group-specific candidates needed for a globally feasible solution with a small radius.

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.

Our framework constructs a separate geometric summary for each local group, preserving nearby replacements for optimal centers within their respective groups. The coordinator extracts a well-separated representative set and selects centers through an integral maximum flow whose group capacities enforce the budgets. Separation ensures that distinct representatives correspond to distinct optimal centers, certifying flow feasibility. The resulting analysis relates local summary precision to the approximation attainable from a single worker-to-coordinator transmission.

1.1 Related Work

Notions of fairness in clustering. Fair clustering encompasses several different requirements. The fairlet framework of Chierichetti et al. (2017) controls group balance within clusters, while Bera et al. (2019) allow prescribed representation ranges and overlapping protected groups. Schmidt et al. (2019) develop composable coresets and streaming algorithms for fair k-means under cluster-balance constraints. Proportional fairness instead protects sufficiently large subsets of points against a jointly preferable alternative center (Chen et al., 2019). Colorful k-center imposes coverage requirements on colors: Bandyapadhyay et al. (2019) obtain a constant-factor approximation for the two-color problem in the Euclidean plane. Individual fairness bounds each point's service distance relative to its local neighborhood radius (Mahabadi & Vakilian, 2020; Negahbani & Chakrabarty, 2021). These formulations differ from DSFkC, which constrains the groups of the selected centers.

Centralized fair center selection. For data-summarization fairness, Kleindessner et al. (2019) give a 5-approximation for two groups and a $(3 \cdot 2^{m-1} - 1)$-approximation for m groups. Group upper bounds define a partition matroid, connecting DSFkC to the 3-approximation for matroid center of Chen et al. (2016), who also give a 7-approximation for matroid center with outliers. Jones et al. (2020) obtain a 3-approximation using maximum matching in $O(nk)$ time. Angelidakis et al. (2022) give a 15-approximation in $O(nk^2 + k^5)$ time for a richer formulation that additionally requires lower bounds on cluster sizes. For fair center selection with prescribed group counts, Thejaswi (2026) establishes that a $(3 - \varepsilon)$-approximation is NP-hard for every $\varepsilon > 0$, even with two disjoint groups.

Related representative-selection problems. In fair k-supplier, centers are chosen from a designated facility set rather than necessarily from the client set. Chen et al. (2024) give a 5-approximation under group representation requirements. Gadekar et al. (2025) improve the factor to 3 for both disjoint and overlapping groups; their algorithms run in polynomial time for disjoint groups and in fixed-parameter time, parameterized by the numbers of groups and centers, for overlapping groups. Adding capacities leads to further complexity: Gadekar & Thejaswi (2025) study hardness for capacitated fair-range clustering and, for a constant number of groups, give fixed-parameter $(3 + \varepsilon)$- and $(9 + \varepsilon)$- approximations for k-median and k-means, respectively. These guarantees concern different objectives and constraints from the uncapacitated center-selection problem studied here.

Clustering with limited memory and communication. Classical k-center admits a 2-approximation via farthest-first traversal (Gonzalez, 1985). Malkomes et al. (2015) analyze a distributed 4-approximation based on local summaries, while Im & Moseley (2015) obtain factor 2 in three MapReduce rounds when the optimal radius is known and four rounds otherwise. Haqi & Zarrabi-Zadeh (2023) further study k-center and diversity maximization in the MPC model. For k-center with at most z outliers, Matthew McCutchen & Khuller (2008) give a one-pass $(4 + \varepsilon)$-approximation using $O(kz/\varepsilon)$ memory. Ceccarello et al. (2019) obtain a MapReduce $(2 + \varepsilon)$-approximation without outliers and $(3 + \varepsilon)$-approximations with outliers in both MapReduce and one-pass streaming. For matroid center, Kale (2018) establishes streaming space lower bounds and gives efficient $(17 + \varepsilon)$- and $(3 + \varepsilon)$-approximations using one and two passes, respectively.

Streaming fair center selection. For DSFkC, Chiplunkar et al. (2020) give a two-pass $(3 + \varepsilon)$-approximation. In one-pass streaming over general metrics, Lin et al. (2024) obtain a $(7 + \varepsilon)$-approximation, improved to $(5 + \varepsilon)$ by Guo et al. (2026b). Guo et al. (2026a) further investigate fair center selection on massive social-network data streams. These algorithms exploit sequential access to the input, whereas our workers must construct their summaries independently.

Distributed fair center selection. Chiplunkar et al. (2020) obtain a $(17 + \varepsilon)$-approximation in general metrics; their radius-given analysis achieves factor 17 with $O(km)$ point records per worker.

For metrics of doubling dimension D, Wu et al. (2024) obtain a $(4+\varepsilon)$-approximation with total communication $O(\ell mk(4/\varepsilon)^D)$. Both approaches combine geometric representatives with group-specific candidates to support feasible center selection at the coordinator. Our framework instead summarizes each group separately and uses them to certify the coordinator's quota-constrained flow.

Other distributed fairness constraints. For fairness within clusters, Bera et al. (2022) give a two-round MPC 9-approximation and a two-pass streaming $(7+\varepsilon)$-approximation, both allowing constant additive violations of the fairness constraints. Ceccarello et al. (2024) develop coreset-based sequential, streaming, and MapReduce algorithms for this cluster-balance formulation, with efficiency analyzed through doubling dimension. For lower-bounded k-center, where every cluster must contain sufficiently many points, Liang et al. (2025) obtain two-round approximations of $9+\varepsilon$ in general metrics and $3+\varepsilon$ in bounded-doubling metrics. These cluster-composition and cluster-size constraints are distinct from the group budgets on centers in DSFkC.

1.2 Contributions

The first three results are deterministic one-round algorithms assuming that the coordinator knows the optimal fair radius r^*; their local summaries do not require this value. For general metrics, coordinator-only radius guessing extends the 7-approximation to a $(7+\varepsilon)$-approximation when r^* is unknown, without changing the point-record communication bound.

- **General metrics.** We obtain a 7-approximation by retaining at most k farthest-first representatives per local group, with total communication at most $\min\{n, \ell mk\}$ point records, improving the previous state-of-art approximation ratio from 17 while retaining $O(km\ell)$ transmitted point records.
- **Extension to other metrics.** For doubling dimension D and $\varepsilon \in (0,1]$, our algorithm extends by using finer local summaries and yields a $(3+\varepsilon)$-approximation using at most $\min\{n, \ell mk(32/\varepsilon)^D\}$ transmitted point records. For Euclidean space, we achieve a 5.84-approximation with at most $\min\{n, 122\ell mk\}$ transmitted point records.
- **Communication lower bound.** For one-way subset-summary protocols with only local distance access, we construct instances with $\Theta(n)$ workers on which achieving a factor strictly below 5 requires $\Omega(n)$ transmitted point records.
- **Empirical evaluation.** Experiments on Adult, SushiA, Celeb-A, and synthetic data evaluate clustering quality under varying budgets, quotas, and worker counts. A 100 GB instance further evaluates runtime under simulated parallel execution.

2 Preliminaries

For a positive integer q, let $[q] = \{1, \dots, q\}$. Let (P, d) be a finite metric space with $|P| = n$. For a point $p \in P$ and a set $A \subseteq P$, we define the distance between A and p as below:

$$d(p, A) = \begin{cases} \min_{a \in A} d(p, a), & A \neq \emptyset, \\ +\infty, & A = \emptyset. \end{cases}$$

Definition 1 (Data-Summarization Fair k-Center)**.** *Let $P = \dot{\bigcup}_{j=1}^{m} G_j$, where G_j is group j. Given group upper bounds $k_1, \dots, k_m \in \mathbb{Z}_{\geq 0}$ and $k = \sum_{j=1}^{m} k_j$, the goal is to find a set $C \subseteq P$ that minimizes* $\mathrm{cost}(C) = \max_{p \in P} d(p, C)$ *while satisfying*

$$|C \cap G_j| \leq k_j \qquad \textit{for every } j \in [m].$$

A center set satisfying the group upper bounds is called *feasible*. Since $k = \sum_{j=1}^{m} k_j$, every feasible center set also satisfies $|C| \leq k$. We study this upper-quota, at-most-k formulation throughout the paper.

Throughout the paper, C^* denotes an optimal feasible center set, and its optimal radius is

$$r^* = \mathrm{cost}(C^*) = \mathrm{OPT}.$$

We assume $r^* < +\infty$. We also assume $P \neq \emptyset$; together, these assumptions imply $k \geq 1$. The empty instance can be handled separately by returning the empty center set.

Algorithm 1: Local Summary Construction

Input: On worker i: the labeled local set P_i and the value k
Output: The local summaries $\Gamma_{i,1}, \ldots, \Gamma_{i,m}$

```
Bucket the points of P_i by group;
for j ← 1 to m do
    Γ_{i,j} ← GONZALEZ(P_i ∩ G_j, min{k, |P_i ∩ G_j|});
Send ⋃_{j=1}^{m} Γ_{i,j} to the coordinator;
```

Definition 2 (One-round coordinator MPC model)**.** *The input is partitioned as $P = \dot{\bigcup}_{i=1}^{\ell} P_i$, where worker i stores P_i and the group label of every local point. Each worker performs an independent local computation and sends one message containing a subset of P_i to a coordinator. The coordinator returns a solution using the union of the received subsets. We measure communication in transmitted point records and explicitly state the memory required to store the aggregate summary. A point record contains enough information to evaluate the distance between any two transmitted points; after a point is broadcast, a worker can also evaluate its distance to each local point.*

Definition 3 (Maximal τ-independent set)**.** *For $A \subseteq P$ and $\tau \geq 0$, let $\sigma(A, \tau)$ denote any inclusion-wise maximal subset of A such that*

$$d(p, q) > \tau \qquad \text{for all distinct } p, q \in \sigma(A, \tau).$$

Equivalently, maximality implies

$$d\big(p, \sigma(A, \tau)\big) \leq \tau \qquad \text{for every } p \in A.$$

3 A 7-Approximation in General Metrics

The algorithm has two stages. In the first stage (i.e., local computation stage), each worker first constructs one local representative set for every group and sends their union to the coordinator. In the second stage (i.e., coordinator computation stage), the coordinator then constructs a global independent set and uses a capacitated flow network to enforce the group upper bounds.

3.1 Local computation

Key idea. In this stage, each worker separately applies Gonzalez's farthest-first traversal algorithm to each group, retaining at most k representatives per group. Then, each worker sends the set of representatives to the coordinator. Note that the construction requires no knowledge of r^* and transmits at most km points per worker. Here, $\textsc{Gonzalez}(A, q)$ denotes the first q points selected by Gonzalez's farthest-first traversal. In particular, $\textsc{Gonzalez}(A, |A|) = A$ and $\textsc{Gonzalez}(\emptyset, 0) = \emptyset$. The detailed algorithm is as given in Alg. 1.

Then, the aggregate summary sent by the workers is $\Phi = \bigcup_{i \in [\ell],\, j \in [m]} \Gamma_{i,j}$. The following structural lemma shows that such Φ retains a feasible solution with radius at most $3r^*$.

Lemma 1. *Let C^* be an optimal feasible center set with radius r^*. Then there exists a center set $C \subseteq \Phi$ such that: (1) $|C| \leq k$; (2) $|C \cap G_j| \leq k_j$ for every $j \in [m]$; (3) $d(p, C) \leq 3r^*$ for every $p \in P$.*

The above lemma shows that our summary contains a feasible solution with radius at most $3r^*$, significantly improving on the $5r^*$ bound of Chiplunkar et al. (2020).

Moreover, for the error bound when computing summaries on the local machines, we have:

Lemma 2 (Local covering)**.** *For every $i \in [\ell]$, $j \in [m]$, and $p \in P_i \cap G_j$, $d(p, \Gamma_{i,j}) \leq 2r^*$.*

Lemma 3 (Local complexity)**.** *Assuming constant-time distance evaluations, worker i runs in $O\big(m + k|P_i|\big)$ time and sends at most*

$$\sum_{j=1}^{m} \min\{k, |P_i \cap G_j|\} \leq \min\{|P_i|, km\}$$

points. Hence, the total communication is at most $\min\{n, km\ell\}$ point records in one communication round.

Algorithm 2: Coordinator Computation

Input: The labeled summary Φ, bounds $k_1, \dots, k_m$, and r^*
Output: A feasible center set C, or FAIL

$\Psi \leftarrow \sigma(\Phi, 2r^*)$;
Create s, t, one node for each $u \in \Psi$, and group nodes $g_1, \dots, g_m$;
Add (s, u) with capacity 1 for every $u \in \Psi$;
Add (g_j, t) with capacity k_j for every $j \in [m]$;
foreach $u \in \Psi$ **do**
  for $j \leftarrow 1$ **to** m **do**
    if *there exists* $v \in \Phi \cap G_j$ *with* $d(u, v) \le 3r^*$ **then**
      Store one such point as $w(u, j)$ and add (u, g_j) with capacity 1;
Compute an integral maximum flow F;
if $\mathrm{val}(F) = |\Psi|$ **then**
  $C \leftarrow \{w(u, j) : F(u, g_j) = 1\}$;
  return C;
else
  return FAIL;

3.2 COORDINATOR COMPUTATION

Assuming r^* is known, the coordinator first computes a maximal $2r^*$-separated set $\Psi = \sigma(\Phi, 2r^*)$ that covers the summary within distance $2r^*$. Then, we construct an auxiliary graph: we create a source s, a sink t, one node for every $u \in \Psi$, and one group node g_j for every $j \in [m]$. An arc (u, g_j) stores a witness $w(u, j) \in \Phi \cap G_j$ within distance $3r^*$ of u. We consequently use an integral maximum flow to select a summarized center within $3r^*$ of each representative, with group capacities enforcing the quotas. The detailed algorithm is as depicted in Alg. 2.

Lemma 4 shows that the FAIL branch is never reached.

Lemma 4 (Completeness of the flow test)**.** *Let $\Psi = \sigma(\Phi, 2r^*)$. Then $|\Psi| \le k$, and the flow network admits a flow of value $|\Psi|$.*

Lemma 5 (Soundness of the flow test)**.** *Let $\Psi = \sigma(\Phi, 2r^*)$. If the flow network has a flow of value $|\Psi|$, the returned set C satisfies*

$$|C \cap G_j| \le k_j \quad \textit{for every } j \in [m], \qquad |C| \le k,$$

and

$$d(u, C) \le 3r^* \quad \textit{for every } u \in \Psi.$$

Theorem 1. *Assuming that the coordinator knows r^*, Algorithms 1 and 2 return a center set C satisfying*

$$|C \cap G_j| \le k_j \quad \textit{for every } j \in [m], \qquad |C| \le k,$$

and

$$\mathrm{cost}(C) \le 7r^*.$$

The algorithm uses one communication round. Worker i runs in $O(m + k|P_i|)$ time and sends at most $\min\{|P_i|, km\}$ points. Thus, the total communication is at most $\min\{n, km\ell\}$ point records. Writing $n_\Phi = |\Phi|$, the coordinator uses $O(n_\Phi + km)$ space and $O(kn_\Phi + k^2 m)$ time under constant-time distance evaluations.

Handling unknown r^*. We assume that the coordinator knows r^* to simplify the analysis. Since the local summaries are independent of r^*, this assumption can be removed by geometric guessing at the coordinator without additional communication Chiplunkar et al. (2020). Starting from a valid positive lower bound, and handling the zero-radius case separately, we increase the guess by a factor of $1+\varepsilon/5$ until the flow test succeeds. Since every guess $R \ge r^*$ succeeds, the returned guess satisfies $R \le (1+\varepsilon/5)r^*$, giving $\mathrm{cost}(C) \le 2r^* + 5R \le (7+\varepsilon)r^*$. If $L > 0$ is the starting lower bound and $U \ge r^*$ is a valid upper bound, the number of coordinator tests is $T = O(\log_{1+\varepsilon/5}(U/L))$. Thus, writing $n_\Phi = |\Phi|$, the coordinator time becomes $O(T(kn_\Phi + k^2 m))$, in addition to the one-time cost of obtaining L.

Extension to other metrics. Our framework extends to doubling metrics with fixed dimension D, achieving a $(3+\varepsilon)$-approximation with at most $\min\{n, \ell mk(32/\varepsilon)^D\}$ transmitted point records for any $\varepsilon \in (0,1]$. For Euclidean spaces of arbitrary dimension, it achieves a $\left(3+2\sqrt{2+2/h}\right)$-approximation for any integer $h \geq 2$, while transmitting at most $\min\{n, h\ell mk\}$ point records. The parameter h controls the trade-off between approximation quality and communication: $h = 122$ yields a 5.84-approximation, whereas $h = 8$ yields a 6-approximation. Details of the doubling-metric and Euclidean extensions appear in Appendices B and C, respectively.

4 A Lower Bound for One-Way Subset Summaries

Proposition 1 (Gap five for subset-summary protocols)**.** *Fix an integer $N \geq 2$ and $\delta \in [0,1]$. Let*

$$P = \{a_1, \ldots, a_N, b, q\},$$

where $a_1, \ldots, a_N$ belong to the red group and b, q belong to the blue group. Set

$$k_{\text{red}} = k_{\text{blue}} = 1.$$

For a hidden index $z \in [N]$, let d_z be the shortest-path metric induced by the following weighted tree:

$$d_z(b, a_z) = 5, \qquad d_z(b, a_i) = 1 \quad \textit{for } i \neq z, \qquad d_z(b, q) = 1.$$

Consequently, $d_z(a_z, a_i) = 6$ for $i \neq z$, $d_z(a_z, q) = 6$.

Suppose that every red point a_i is stored on a separate worker, while b and q are stored together on one additional worker and are always communicated. Each worker can access distances only among its locally stored points, sends a subset of its local input in one communication round, and the coordinator must select centers from the communicated points.

Then every deterministic protocol guaranteeing a ratio strictly below 5 communicates all N red points in the worst case. Moreover, a randomized protocol succeeding with probability at least $1-\delta$ communicates at least $(1-\delta)N$ red points in expectation on some input.

5 Experiments

In this section, we empirically evaluate our algorithms to compare with 6 baselines on 3 real-world datasets and 2 synthetic datasets. On real-world datasets, we compare clustering quality under different center-budget ratios k/n, group m, and numbers of machines ℓ. On synthetic datasets, we compare clustering costs against a constructed optimal solution and evaluate running time on a 100 GB dataset.

5.1 Experimental setup

Datasets. Following Chiplunkar et al. (2020), we use three real-world datasets: Celeb-A, SushiA and Adult. For the synthetic dataset, we set $k = 25$ and place the remaining points within distance 0.5 of the grid points. We vary the number of groups over $\{2, 4, 6, 8, 10\}$ and set each group's quota to its number of grid points. Thus, the grid points form a quota-feasible solution with cost 0.5. We also use a separate 100 GB synthetic dataset to evaluate scalability. Table 1 summarizes the datasets. More details can be found in the Appendix.

Dataset	#Record	#Dim.	Feature	#Group
SushiA	5,000	45	gender / age / gender×age	2 / 6 / 12
Celeb-A	5,000	15,360	gender / gender×age	2 / 4
Adult	32,561	6	gender / race / gender×race	2 / 5 / 10
SynA	2K, 4K, 6K, 8K, 10K	2	-	2 / 4 / 6 / 8 / 10
Syn100G	4,294,623	1000	-	4

Table 1: Datasets Summary

Baselines We compare our algorithms for general metrics, Euclidean spaces and doubling metrics (named Ours, Ours (Euclidean), and Ours (Doubling)) with six baselines for the fair k-center problem:

- Chiplunkar et al. (2020) (D): a distributed $(17+\epsilon)$-approximation algorithm.
- Chiplunkar et al. (2020) (S): a two-pass streaming $(3+\epsilon)$-approximation algorithm.
- Guo et al. (2026a): a one-pass streaming $(1+2\sqrt{3}+\epsilon)$-approximation algorithm.
- Wu et al. (2024): a distributed $(4+\epsilon)$-approximation algorithm for data-summarization fair k-center in bounded-doubling metrics.
- Kleindessner et al. (2019): a centralized $(3\cdot 2^{m-1}-1)$-approximation algorithm.
- Jones et al. (2020): a centralized 3-approximation algorithm.

Experimental setting. Unless otherwise stated, we use $\ell = 8$ machines and set $\epsilon = 0.1$ throughout the evaluation. In experiments, radius searches start from a Gonzalez lower bound computed on the full dataset using the corresponding metric. For the center-budget experiments, we vary the target ratio k/n from 0.2% to 1.0% in increments of 0.2%. Integer group quotas are allocated approximately in proportion to group sizes and adjusted to sum to k. We vary the number of machines over $\ell \in \{1, 2, 4, 8, 16, 32\}$ while fixing k at approximately 1% of the dataset size. On real-world datasets, Ours (Euclidean) selects centers using ℓ_2 distances, while all returned center sets are evaluated using a common ℓ_1 cost. On SynA, all methods use Euclidean distances for both center selection and cost evaluation. Each experiment is repeated five runs, except for the 100 GB experiment, which is run once due to its computational cost. We report mean results over the five runs, while the box plots summarize the individual runs.

All algorithms are implemented in Python and evaluated on a Linux server equipped with a 32-core processor and 256 GB of RAM. We simulate the ℓ machines through parallel execution on this multicore server.

Evaluation metrics. We measure clustering quality by $\mathrm{cost}_q(C) = \max_{p\in P} \min_{c\in C} \|p-c\|_q$, using $q = 1$ for the real-world experiments and $q = 2$ for SynA. Since SynA has optimal cost $\mathrm{OPT} = 0.5$, we report the empirical approximation factor $\mathrm{cost}_2(C)/\mathrm{OPT}$. We also report algorithm running time in seconds.

5.2 Experimental results

5.2.1 Clustering results on synthetic datasets

Figure 1 reports the empirical approximation factors on SynA for $n = 10{,}000$, $d = 2$, and $k = 25$. Across the five group settings, the median factor of Ours lies between 1.50 and 1.62, far below its

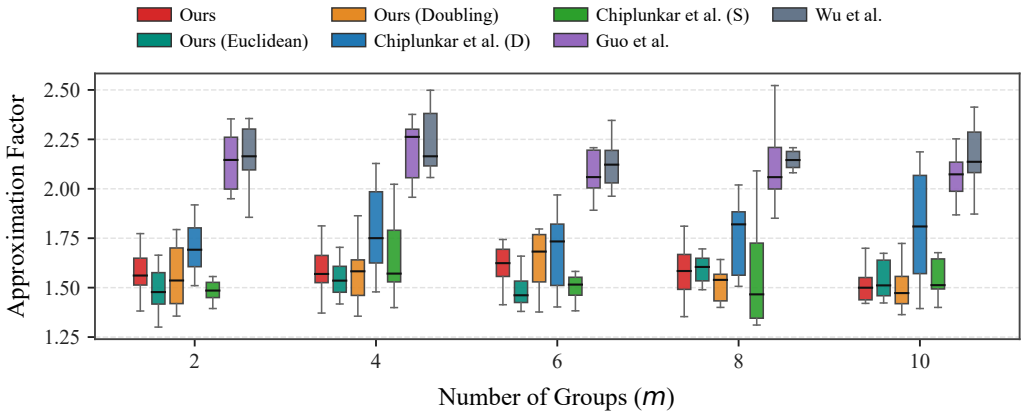

Figure 1: Normalized clustering costs on SynA as the number of groups varies, with $n = 10{,}000$.

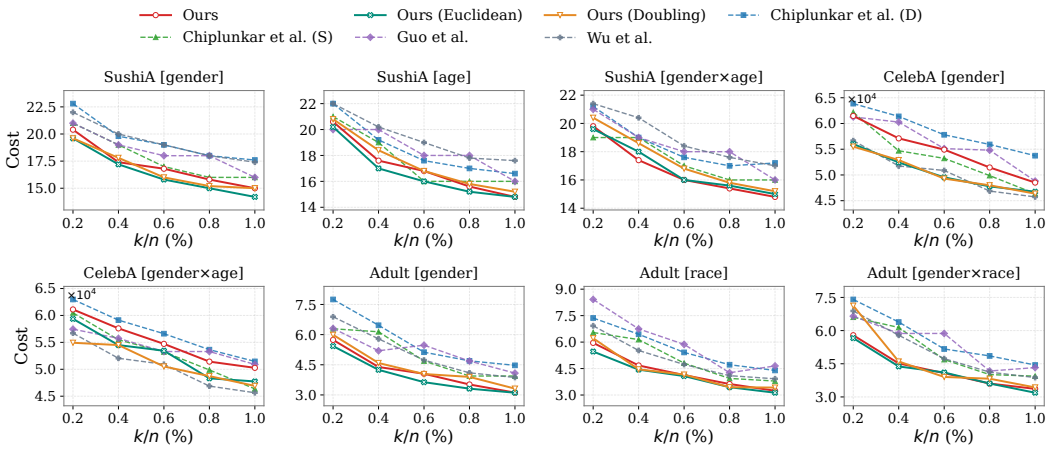

Figure 2: Clustering cost on real-world datasets across varying center ratios k/n.

worst-case guarantee of $7 + \varepsilon$. The Euclidean variant has medians between 1.46 and 1.61, also well below its theoretical bound of $5.84 + \varepsilon$. Increasing m from 2 to 10 does not produce a consistent increase in either method's median cost. On these instances, the worst-case bounds are therefore conservative; these grid-based instances appear easier to cover than the configurations considered in the analysis.

Across the displayed group settings, Ours and its Euclidean variant have lower median factors than Chiplunkar et al. (2020) (D), Guo et al. (2026a), and Wu et al. (2024). The Euclidean variant also has a narrower interquartile range than Chiplunkar et al. (2020) (D) at every displayed group count, indicating that its lower median is accompanied by less variation across runs. The two-pass streaming method of Chiplunkar et al. (2020) (S) achieves comparable medians and is lower than Ours for some group counts, including $m = 8$. The box plots also indicate that these results are not driven by a few favorable runs: none of our three variants exceeds a factor of 2 in the displayed trials. Ours (Doubling) achieves the lowest median among our variants at $m = 8$ and $m = 10$, showing that the doubling-metric construction can yield a lower clustering cost on these instances.

5.2.2 Clustering results on real-world datasets

Figure 2 compares clustering costs as k/n increases from 0.2% to 1% across the real-world datasets. Costs generally decrease with larger budgets k. This trend is expected for the k-center objective: a larger center budget permits more representatives and can reduce the maximum distance from a point to its nearest center. Our general-metric algorithm achieves a lower cost than the distributed algorithm of Chiplunkar et al. (D) on all datasets, with particularly clear gains on Adult and SushiA. Our algorithm retains well-separated candidate centers, while the coordinator selects centers subject to the group fairness requirement. The observed improvements are consistent with our theoretical approximation guarantee and show that the approach also performs well on real-world data.

We also compare against methods with smaller worst-case approximation factors under their respective assumptions, including the MPC algorithm of Wu et al. and the streaming algorithms of Guo et al. and Chiplunkar et al. (S). Our algorithm achieves comparable or lower costs in many configurations, indicating that our algorithm design remains competitive in clustering quality. Moreover, our Euclidean result typically matches or improves upon the general-metric one, even when the real-world clustering costs are evaluated using ℓ_1 distance. The explanation is that its larger group-wise summaries allow workers to transmit more candidate centers, giving the coordinator greater flexibility to select quota-feasible centers with better coverage.

5.2.3 Running time

Figure 3 shows that the effect of increasing the number of machines on running time varies across algorithms and datasets. Some configurations benefit from smaller local workloads, while others exhibit limited improvement or increased runtime. The observed trends reflect the combined cost

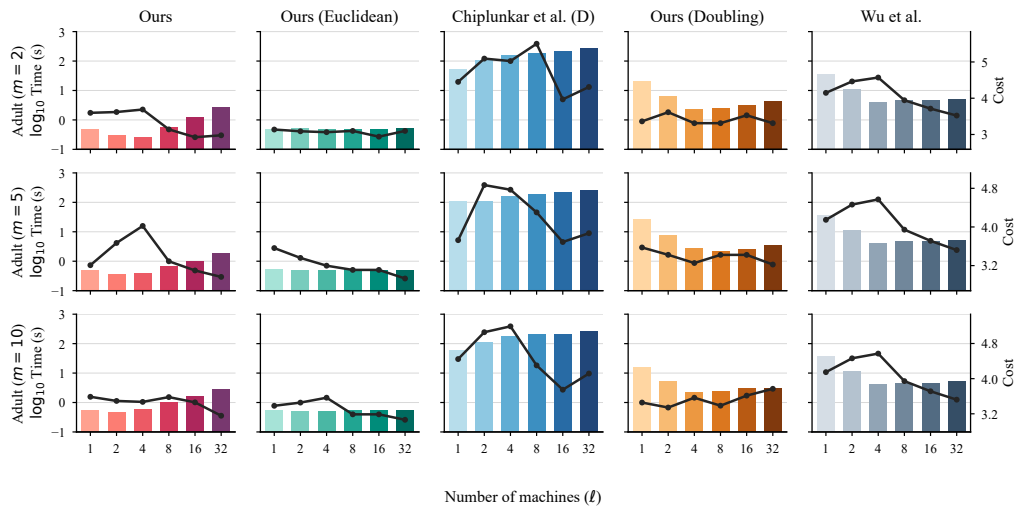

Figure 3: Running time and clustering cost on Adult as the number of machines varies. Rows correspond to $m = 2, 5, 10$; the time axis uses a logarithmic scale.

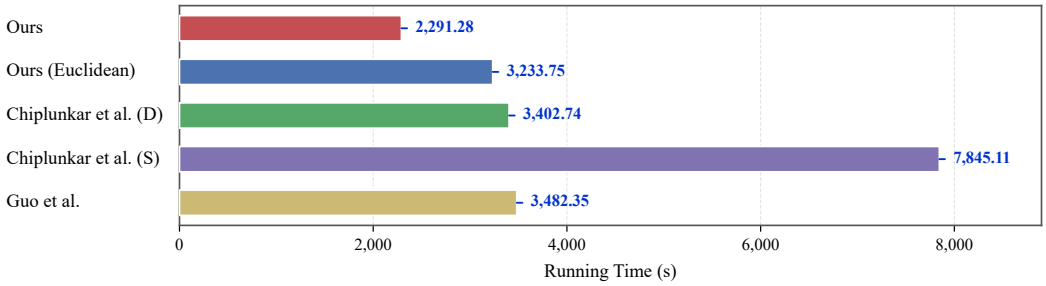

Figure 4: Running time on the 100 GB synthetic dataset with $\ell = 8$ machines.

of local processing, parallel execution overhead, and coordinator computation. Figure 4 reports a single-run timing comparison on the 100 GB synthetic dataset containing 4,294,623 points in 1,000 dimensions. We use four groups, quotas $\mathcal{K} = (2, 2, 2, 2)$, $\ell = 8$ machines, and local blocks of 10,000 points. Among the methods shown, our general-metric algorithm has the shortest measured runtime, whereas the two-pass streaming algorithm of Chiplunkar et al. (2020) has the longest. The comparison reports the observed runtime under this configuration.

6 Conclusion

We presented a group-aware summary framework for fair k-center in the coordinator-based MPC model using a single worker-to-coordinator communication round. In general metrics, it achieves a $(7 + \varepsilon)$-approximation, improving the previous factor of 17 while maintaining $O(km)$ transmitted point records per worker. Refined summaries yield a $(3 + \varepsilon)$-approximation in bounded-doubling metrics and a $(5.84 + \varepsilon)$-approximation in Euclidean spaces of arbitrary dimension. The analysis rests on a common principle: local summaries preserve same-group replacements for optimal centers, while geometric separation ensures the feasibility of the coordinator's quota-constrained flow. For one-way subset-summary protocols with only local distance access, we showed that achieving an approximation factor strictly below 5 can require $\Omega(n)$ transmitted point records on instances with $\Theta(n)$ workers. Experiments on synthetic and real-world datasets demonstrate competitive clustering quality and practical efficiency under simulated parallel execution.

AI USE STATEMENT

The authors used an AI language model to assist with language editing, proofreading, and identifying relevant literature. The authors take full responsibility for the final manuscript, including its theoretical claims, proofs, code, and reported experimental results.

REFERENCES

Haris Angelidakis, Adam Kurpisz, Leon Sering, and Rico Zenklusen. Fair and fast k-center clustering for data summarization. In *International Conference on Machine Learning*, pp. 669–702. PMLR, 2022.

Sayan Bandyapadhyay, Tanmay Inamdar, Shreyas Pai, and Kasturi Varadarajan. A constant approximation for colorful k-center. *arXiv preprint arXiv:1907.08906*, 2019.

Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

Suman K Bera, Syamantak Das, Sainyam Galhotra, and Sagar Sudhir Kale. Fair k-center clustering in mapreduce and streaming settings. In *Proceedings of the ACM Web Conference 2022*, pp. 1414–1422, 2022.

Matteo Ceccarello, Andrea Pietracaprina, and Geppino Pucci. Solving k-center clustering (with outliers) in mapreduce and streaming, almost as accurately as sequentially. *Proceedings of the VLDB Endowment*, 12, 2019.

Matteo Ceccarello, Andrea Pietracaprina, and Geppino Pucci. Fast and accurate fair k-center clustering in doubling metrics. In *Proceedings of the ACM Web Conference 2024*, pp. 756–767, 2024.

Danny Z Chen, Jian Li, Hongyu Liang, and Haitao Wang. Matroid and knapsack center problems. *Algorithmica*, 75(1):27–52, 2016.

Xianrun Chen, Sai Ji, Chenchen Wu, Yicheng Xu, and Yang Yang. An approximation algorithm for diversity-aware fair k-supplier problem. *Theoretical Computer Science*, 983:114305, 2024.

Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *International conference on machine learning*, pp. 1032–1041. PMLR, 2019.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

Ashish Chiplunkar, Sagar Kale, and Sivaramakrishnan Natarajan Ramamoorthy. How to solve fair k-center in massive data models. In *International Conference on Machine Learning*, pp. 1877–1886. PMLR, 2020.

Ameet Gadekar and Suhas Thejaswi. Capacitated fair-range clustering: Hardness and approximation algorithms. *arXiv preprint arXiv:2505.15905*, 2025.

Ameet Gadekar, Aristides Gionis, and Suhas Thejaswi. Fair clustering for data summarization: Improved approximation algorithms and complexity insights. In *Proceedings of the ACM on Web Conference 2025*, pp. 4458–4469, 2025.

Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical computer science*, 38:293–306, 1985.

Branko Grünbaum, Victor Klee, Micha A Perles, and Geoffrey Colin Shephard. *Convex Polytopes*, volume 16. Springer, 1967.

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.

Alireza Haqi and Hamid Zarrabi-Zadeh. Almost optimal massively parallel algorithms for k-center clustering and diversity maximization. In *Proceedings of the 35th ACM Symposium on Parallelism in Algorithms and Architectures*, pp. 239–247, 2023.

Sungjin Im and Benjamin Moseley. Fast and better distributed mapreduce algorithms for k-center clustering. In *Proceedings of the 27th ACM symposium on Parallelism in Algorithms and Architectures*, pp. 65–67, 2015.

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. *arXiv preprint arXiv:1810.06267*, 2018.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pp. 3819–3828, 2015.

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.

Ting Liang, Xiaoliang Wu, Jinhui Xu, and Qilong Feng. The distributed algorithms for the lower-bounded k-center clustering in metric space. *Theoretical Computer Science*, 1027:114975, 2025.

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.

Sepideh Mahabadi and Stojan Trajanovski. Core-sets for fair and diverse data summarization. *Advances in Neural Information Processing Systems*, 36:78987–79011, 2023.

Sepideh Mahabadi and Ali Vakilian. Individual fairness for k-clustering. In *International conference on machine learning*, pp. 6586–6596. PMLR, 2020.

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. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pp. 165–178. Springer, 2008.

Maryam Negahbani and Deeparnab Chakrabarty. Better algorithms for individually fair k-clustering. *Advances in Neural Information Processing Systems*, 34:13340–13351, 2021.

Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pp. 232–251. Springer, 2019.

Suhas Thejaswi. On the hardness of approximation of the fair k-center problem. *arXiv preprint arXiv:2602.16688*, 2026.

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.