

Online Resource

■ A Related Work

In the learning-augmented setting, Ergun *et al.* [1] gave a $(1 + 20\gamma)$ -approximation for the k -means problem assuming each auxiliary cluster P_i of size $\Omega(k/\gamma)$ with $\gamma \in [10 \log n / \sqrt{n}, 1/7]$. They also proposed a $(1 + \gamma')$ -approximation for the k -median problem, requiring each P_i to have size $\Omega(n/k)$ and $\gamma \leq O(\gamma'^4 / (k \log(k/\gamma')))$. Subsequently, Nguyen *et al.* [2] improved the approximation ratios to $1 + 7.7\gamma$ and $O(1 + O(\gamma))$ for the k -means and k -median problems, respectively, with $\gamma \in [0, 1/2]$.

Chierichetti *et al.* [3] introduced the definition of fairness with only two colors, requiring that the proportion of two colors has approximately equal representation in every cluster. Bercea *et al.* [4] proposed the notion of group fairness. They presented a 3-approximation with 1 violation (called essentially fair) for the group fairness constraints via linear programming (abbr. LP) and min-cost flow network for the GFkC problem, and developed a 5-approximation without fairness violation based on a tree-structure for the EGFkC problem. Note that the value of violation measures the extent to which the fairness constraints are violated. Informally, a cluster is essentially fair if there exists a fractional fair cluster, such that for each color h the number of color h points in this cluster differs by at most 1 from the mass of color h points in the fractional fair cluster (see Definition 1 with more details). Ahmadian *et al.* [5] studied the GFkC problem with only an upper bound constraint α , and gave a 3-approximation with an additive 2 violation (the definition of this violation is different from the one in [4]) using LP and min-cost flow network. For the GFkC problem under the condition that colors are allowed to overlap, a 4-approximation algorithm with $(4\Delta + 3)$ violation [6] and a 3-approximation algorithm with $(4\Delta + 3)$ violation [7] were presented, respectively, where Δ is the maximum number of colors a single point can belong to. Recently, several approximation algorithms were designed for the GFkC problem in the distributed model [8–10]. Regarding parameterized result, Bandyapadhyay *et al.* [11] employed the fair coresset technique, and proposed a $(3 + \epsilon)$ -approximation for the k -median problem under the group fairness constraints in parameterized time.

Kleindessner *et al.* [12] considered the data summarization fairness, and gave a constant-factor approximation algorithm in linear-time based on a swap technique for the k -center objective. This approximation was subsequently improved to 3 by [13] through the maximum matching method, matching the approximation ratio of the matroid center problem [14] that generalize the data summarization fairness for the k -center objective. For data summarization under the k -median objective, which can be generalized to the matroid median problem [15], the best known approximation ratio is $(7.081 + \epsilon)$ due to [16]. Thejaswi *et al.* [17] extended data summarization fairness by introducing an additional lower-bound constraint on the number of selected facilities, and provided a $(1 + 2/e + \epsilon)$ -approximation for the k -median objective, running in parameterized time. Furthermore, Zhang *et al.* [18] proposed $(1 + \epsilon)$ -approximation algorithm for the k -median objective in Euclidean metrics, operating in parameterized time.

■ B Missing Details of Section 2

We first describe the role of parameter τ_f in LA-SAMPLING. Note that for both the k -center problem and the GFkC problem, the cost of optimal solution is always the distance between two points. Therefore, we can guess the optimal solution value by considering a binary search over all the possible distances between points. More precisely, for a given instance and a parameter τ_f , by running LA-SAMPLING, we will try to obtain a feasible solution depending on τ_f . If such a solution does not exist, our guess value is too low, and we need to try larger values. Otherwise, we shall try smaller ones. For simplicity, we assume that the parameter τ_f in Algorithm LA-SAMPLING is the cost of optimal solution to given instance.

Lemma 1. Consider an instance (C, d, k) of the k -center related problem, an auxiliary clustering (P_1, \dots, P_k) of C , a label error rate γ and parameter τ_f . With probability at least $(1 - \gamma)^k$, LA-SAMPLING outputs a set $B = \{B_1, \dots, B_k\}$ of k local balls such that for any $v \in C$, the point v can be covered by the union $\bigcup_{i \in [k]} B_i$ of k local balls. Moreover, LA-SAMPLING runs in $O(nk)$ time.

Proof. For any $i \in [k]$, let $Q_i = P_i \cap P_i^*$. Thus, with probability $|Q_i|/|P_i| \geq (1 - \gamma) \max(|P_i|, |P_i^*|)/|P_i| \geq 1 - \gamma$, s_i falls in Q_i . By the definition of optimal clustering, the distance between each point in Q_i and s_i^* is at most τ_f . With probability at least $(1 - \gamma)^k$, $d(s_i, s_i^*) \leq \tau_f$ over all $i \in [k]$. By the process of getting local balls, we just prove that for any $v \in C$, $d(v, S) \leq 2\tau_f$. For any $v \in C$, assume that the point v is in optimal cluster P_i^* ($i \in [k]$). By the triangle inequality, we have $d(v, s_i) \leq d(v, s_i^*) + d(s_i^*, s_i) \leq \tau_f + \tau_f = 2\tau_f$, that is, the point v can be covered by the local ball B_i .

Finally, we bound the running time of LA-SAMPLING. In steps 2-5, there are k iterations, in each of which we sample the center s_i with $O(1)$ time. Thus, the total running time of steps 2-5 can be bounded by $O(k)$. Steps 6-9 can be done in $O(nk)$ time, since it involves k iterations, and in each iteration we consider all points in C to get local ball with time $O(n)$. Therefore, the overall running time of LA-SAMPLING can be bounded by $O(nk)$. \square

Lemma 2. Given an instance $\mathcal{I} = (C, d, k, \mathcal{G}, H, \alpha, \beta)$ of the GFkC problem, an auxiliary clustering (P_1, \dots, P_k) of C , a label error rate γ and parameter τ_f , let $B = \{B_1, \dots, B_k\}$ be the set of k local balls returned by LA-SAMPLING. Assume that τ_f is the cost of optimal solution of \mathcal{I} . Then, with probability at least $(1 - \gamma)^k$, there exists a mapping $\phi : C \rightarrow S$ satisfying the group fairness constraints such that for any $v \in C$, $d(v, \phi(v)) \leq 2\tau_f$.

Proof. Let (S^*, ϕ^*) be an optimal solution of \mathcal{I} , where $S^* = \{s_1^*, \dots, s_k^*\}$ is the set of k optimal centers. Let $P^* = \{P_1^*, \dots, P_k^*\}$ be the k optimal clusters with $\bigcup_{i \in [k]} P_i^* = C$ under mapping ϕ^* . For any $i \in [k]$ and $h \in H$, let $P_i^*(h)$ be the set of points in P_i^* with color h . For any $i \in [k]$, it is easy to get that $P_i^* = \bigcup_{h \in H} P_i^*(h)$. For any $i \in [k]$, we have that s_i is the closest center in S to s_i^* . Moreover, there is no such case that the centers in S^* have the same closest center in S . We use the mapping $\pi : S^* \rightarrow S$ to denote the one-to-one relation, and let $\pi(s_i^*) = \arg \min_{s \in S} d(s, s_i^*)$ denote the closest center in S to s_i^* for any $i \in [k]$. We now construct a mapping $\phi : C \rightarrow S$ as

Algorithm 1 LA-GROUP-FAIR

Input: An instance $\mathcal{I} = (C, d, k, \mathcal{G}, H, \alpha, \beta)$ of the GFkC problem, an auxiliary clustering (P_1, \dots, P_k) of C , a label error rate $\gamma < 1/2$ and parameter τ_f

Output: A solution (S, ϕ) of \mathcal{I}

- 1: $B \leftarrow \text{LA-SAMPLING}(\mathcal{I}, (P_1, \dots, P_k), \gamma, \tau_f)$;
- 2: Let $S = \{s_1, \dots, s_k\}$ be the corresponding k centers;
- 3: $\mathbf{x} \leftarrow$ obtain a fractional assignment from the points in C to the centers in S by solving LP-relaxation;
- 4: $\bar{\mathbf{x}} \leftarrow$ round the obtained fractional solution to an integral solution using min-cost flow network;
- 5: $\forall i \in S, j \in C : \phi(j) \leftarrow i$ if $\bar{x}_{ij} > 0$;
- 6: **return** (S, ϕ) .

follows. For any $v \in C$, let $\phi(v) = \pi(\phi^*(v))$. We first prove that for any $v \in C$, $d(v, \phi(v)) \leq 2\tau_f$. For any $v \in C$, assume that the point v is from optimal cluster P_i^* ($i \in [k]$). Then, we have $d(v, \phi(v)) = d(v, \pi(\phi^*(v))) \leq d(v, \phi^*(v)) + d(\phi^*(v), \pi(\phi^*(v))) \leq \tau_f + \tau_f = 2\tau_f$, where the first inequality holds due to the triangle inequality. We now show that the mapping ϕ satisfies the group fairness constraints. Note that the optimal solution (S^*, ϕ^*) is a feasible solution of \mathcal{I} . Thus, for any $i \in [k]$ and $h \in H$, we have $\beta_h \leq |P_i^*(h)|/|P_i^*| \leq \alpha_h$. By the process of constructing ϕ , it is easy to get that $\frac{|\{v \in C_h | \phi(v)=s\}|}{|\{v \in C | \phi(v)=s\}|} = \frac{|P_i^*(h)|}{|P_i^*|}$. Then, we have $\beta_h \leq \frac{|\{v \in C_h | \phi(v)=s\}|}{|\{v \in C | \phi(v)=s\}|} \leq \alpha_h$. Thus, ϕ satisfies the group fairness constraints. \square

■ C Missing Details of Subsection 3.1

Algorithm 1 details the process of LA-GROUP-FAIR. To proceed, we give the formal definition of fairness violation introduced by [4].

Definition 1. ([4]) Given an instance $\mathcal{I} = (C, d, k, \mathcal{G}, H, \alpha, \beta)$ of the GFkC problem, let \mathbf{x}^1 be an integral solution, but not necessarily feasible solution of \mathcal{I} . We call that \mathbf{x}^1 is a solution with 1 violation for the group fairness constraints if for any $i \in C$,

$$\begin{aligned} \lfloor \sum_{j \in C_h} x_{ij} \rfloor &\leq \sum_{j \in C_h} x_{ij}^1 \leq \lceil \sum_{j \in C_h} x_{ij} \rceil \quad \forall h \in H, \text{ and} \\ \lfloor \sum_{j \in C} x_{ij} \rfloor &\leq \sum_{j \in C} x_{ij}^1 \leq \lceil \sum_{j \in C} x_{ij} \rceil, \end{aligned}$$

where \mathbf{x} is a fractional feasible solution satisfying the group fairness constraints of \mathcal{I} .

Intuitively, a solution with 1 violation for the group fairness constraints is an integral assignment in which for every center, the number of points from each group and the total number differs from the fractional fair assignment by at most one point.

Lemma 3. Given an instance $\mathcal{I} = (C, d, k, \mathcal{G}, H, \alpha, \beta)$ of the GFkC problem in the learning-augmented setting, let \mathbf{x} be the solution returned by the LP-relaxation. Then, we can obtain a 2-approximate solution $\bar{\mathbf{x}}$ with 1 violation for group fairness constraints in polynomial time based on a min-cost flow network.

Proof. Let \mathbf{x} be the solution obtained by the LP-relaxation (step 3 of LA-GROUP-FAIR). For simplicity, for any $i \in S$ and $h \in H$, let

$T_i^h = \sum_{j \in C_h} x_{ij}$, $T_i = \sum_{j \in C} x_{ij}$, $B_i = \lfloor T_i \rfloor - \sum_{h \in H} \lfloor T_i^h \rfloor$, and $B = |C| - \sum_{i \in S} \lfloor T_i \rfloor$. We now begin to construct a min-cost flow network $(G = (V, E), b)$ with unit capacities and costs on the edges as well as b -values on the nodes as follows.

- $V = V_1 \cup V_2 \cup V_3 \cup V_4 \cup V_5$, where $V_1 = \{s\}$ with b -value $|C|$, $V_2 = \{v_j \mid j \in C\}$ with b -value 0, $V_3 = \{v_i^h \mid i \in S, h \in H\}$ with b -value $-\lfloor T_i^h \rfloor$, $V_4 = \{v_i \mid i \in S\}$ with b -value $-B_i$, and $V_5 = \{t\}$ with b -value $-B$.
- $E = E_1 \cup E_2 \cup E_3 \cup E_4$, where $E_1 = \{(s, v_j) \mid j \in C\}$ with cost 0, $E_2 = \{(v_j, v_i^h) \mid j \in C, i \in S, h \in H, j \in C_h, x_{ij} > 0\}$ with cost $d(i, j)$, $E_3 = \{(v_i^h, v_i) \mid i \in S, h \in H, T_i^h - \lfloor T_i^h \rfloor > 0\}$ with cost 0, and $E_4 = \{(v_i, t) \mid i \in S, T_i - \lfloor T_i \rfloor > 0\}$ with cost 0.

Note that the b -value denotes the node's value in the min-cost flow network, where positive value indicates a supply node and negative value indicates a demand node. Observe that the sum of b -values of all nodes in V is equal to 0. Thus, we have flow conservation. Observe that all capacities, costs, and b -values of the min-cost flow network $(G = (V, E), b)$ are integral. Consequently, there always exists an integral optimal solution for $(G = (V, E), b)$. Let f be an optimal solution of the min-cost flow network. Then, it induces immediately a solution $\bar{\mathbf{x}}$ to instance \mathcal{I} by setting \bar{x}_{ij} equal to the flow value of edge $(v_j, v_i^h) \in E_2$ for any $j \in C$ and $i \in S$. Since the flow f is integral, this implies that $\bar{\mathbf{x}}$ is integral. Therefore, we have $\text{cost}(\bar{\mathbf{x}}) \leq 2\tau_f$ since $\bar{x}_{ij} > 0$ only if $d(i, j) \leq 2\tau_f$.

We now prove that the integral solution $\bar{\mathbf{x}}$ obtained above incurs 1 violation for the group fairness constraints. Observe that for each $i \in S$, there are at least $\lfloor T_i^h \rfloor$ points with color $h \in H$ and at least $\lfloor T_i \rfloor$ points in total assigned to i due to the b -values of nodes v_i^h and v_i . Since (v_i^h, v_i) and (v_i, t) have the unit capacity, and there is at most one outgoing arc for nodes v_i^h and v_i , we have at most $\lfloor T_i^h \rfloor$ points with color $h \in H$ and at most $\lfloor T_i \rfloor$ points in total assigned to i . Consequently, we have $\lfloor T_i^h \rfloor \leq \sum_{j \in C_h} \bar{x}_{ij} \leq \lfloor T_i^h \rfloor$ and $\lfloor T_i \rfloor \leq \sum_{j \in C} \bar{x}_{ij} \leq \lfloor T_i \rfloor$. By Definition 1, we get that $\bar{\mathbf{x}}$ is an integral solution with 1 violation for fairness constraints to \mathcal{I} .

We now discuss the running time of LA-GROUP-FAIR. The construction of the local balls, along with the use of linear programming and the min-cost flow network, can all be executed in polynomial time. Therefore, the overall running time of the algorithm is bounded by polynomial time. \square

■ D Missing Details of Subsection 3.2

Algorithm 2 details the process of LA-EXACT. Before describing the specific process of the function *Calculate*, we give some properties on which *Calculate* relies. For a given instance $\mathcal{I} = (C, d, k, \mathcal{G}, H, \alpha, \beta)$ of the EGFkC problem with $\beta_h = \alpha_h = |C_h|/|C|$ for any $h \in H$, assume that (S, ϕ) is a feasible solution of \mathcal{I} , and let $O = \{O_1, \dots, O_k\}$ be the corresponding clusters induced by ϕ . Let $g = \gcd(|C_1|, \dots, |C_m|)$. Then for each cluster O_i ($i \in [k]$), there exists a positive integer $r \in \mathbb{N}_{\geq 1}$ such that O_i contains exactly $r \cdot \frac{|C_h|}{g}$ points with color h for each $h \in H$ and $r \cdot \sum_{h \in H} \frac{|C_h|}{g}$ points in total. Thus, each cluster O_i must contain at least $\frac{|C_h|}{g}$ points with color h for each $h \in H$. We call a set of points that contains exactly $\frac{|C_h|}{g}$ points

Algorithm 2 LA-EXACT

Input: An instance $\mathcal{I} = (C, d, k, \mathcal{G}, H, \alpha, \beta)$ of the EGfK problem with $\beta_h = \alpha_h = |C_h|/|C|$ ($h \in H$), an auxiliary clustering (P_1, \dots, P_k) of C , a label error rate $\gamma < 1/2$ and parameter τ_f

Output: A solution (S, ϕ) of \mathcal{I}

- 1: $B \leftarrow \text{LA-SAMPLING}(\mathcal{I}, (P_1, \dots, P_k), \gamma, \tau_f)$;
 - 2: Let $S = \{s_1, \dots, s_k\}$ be the corresponding k centers;
 - 3: Construct $U(S')$ for each non-empty subset $S' \subseteq S$;
 - 4: $(R_i^1, R_i^2, \dots, R_i^m) \leftarrow \text{Calculate}(U_i)$, for each $i \in S$;
 - 5: Construct a min-cost flow network based on the returned values;
 - 6: $\mathbf{x} \leftarrow$ find a min-cost flow on the constructed network;
 - 7: $\forall i \in S, j \in C : \phi(j) \leftarrow i$ if $x_{ij} > 0$;
 - 8: **return** (S, ϕ) .
-

with color h for each color $h \in H$ a fair-set. Therefore, a fair-set is a minimal set satisfying fairness constraints. For example, consider an instance consisting of a set of 50 points and $k = 2$, where 10 points with the color red, 15 points with the color yellow, and 25 points with the color green. Then, a fair-set of this instance is a set of 10 points, where 2 points with the color red, 3 points with the color yellow, and 5 points with the color green. Therefore, for each cluster O_i of the feasible solution, it contains one or more fair-sets.

We now give the specific process of the function *Calculate* and explain the returned values. Recall that a fair-set is a minimal set of points satisfying fairness constraints. Hence, for each single-combiner U_i ($i \in S$), by examining the number of points with each color in U_i , it is easy to get that the minimum number of points with each color added to U_i will make U_i satisfy fairness constraints. More formally, we use R_i^h to denote the minimum number of points with each color h for each single-combiner U_i . Thus, for each single-combiner U_i , if we add R_i^h points with color h for each $h \in H$ and $\sum_{h \in H} R_i^h$ points in total to U_i , then U_i will satisfy fairness constraints. Note that for any $h \in H$, it is possible that $R_i^h \geq \frac{|C_h|}{g}$ since there is more than one fair-set of the optimal cluster. For such case, U_i will contain exactly $r \in \mathbb{N}_{\geq 1}$ fair-sets.

Lemma 4. Given an instance $\mathcal{I} = (C, d, k, \mathcal{G}, H, \alpha, \beta)$ of the EGfK problem with $\beta_h = \alpha_h = |C_h|/|C|$ for any $h \in H$, we can obtain a 2-approximate solution without fairness violation of \mathcal{I} in polynomial time with the learning-augmented setting.

Proof. Let $C' = C \setminus \cup_{i \in S} J_{\{i\}}$ be the union of all multi-combiners. For any $i \in S$ and $h \in H$, let R_i^h be the value returned by *Calculate* (step 4 of LA-EXACT). We now begin to construct a min-cost flow network $(G = (V, E), b)$ with capacities and costs on the edges as well as b -values on the nodes as follows.

- $V = V_1 \cup V_2 \cup V_3 \cup V_4 \cup V_5$, where $V_1 = \{s\}$ with b -value $|C'|$, $V_2 = \{v_j \mid j \in C'\}$ with b -value 0, $V_3 = V_3^1 \cup V_3^2$, where $V_3^1 = \{v_i^h \mid i \in S, h \in H\}$ with b -value $-R_i^h$, and $V_3^2 = \{a^h \mid h \in H\}$ with b -value $-(|\{v \in C' \mid v \in C_h\}| - \sum_{i \in S} R_i^h)$, $V_4 = \{v_i \mid i \in S\} \cup \{a\}$ with b -value 0, and $V_5 = \{t\}$ with b -value 0.

- $E = E_1 \cup E_2 \cup E_3 \cup E_4$, where $E_1 = \{(s, v_j) \mid j \in C'\}$ with cost 0 and capacity 1, $E_2 = E_2^1 \cup E_2^2$, where $E_2^1 = \{(v_j, v_i^h) \mid j \in C', i \in S, h \in H, j \in C_h, j \in B_i\}$ with cost $d(i, j)$ and capacity 1, $E_2^2 = \{(v_j, a^h) \mid j \in C', h \in H, j \in C_h\}$ with cost 0 and capacity 1, $E_3 = \{(v_i^h, v_i) \mid i \in S, h \in H\} \cup \{(a^h, a) \mid h \in H\}$ with cost 0 and capacity 0, and $E_4 = \{(v_i, t) \mid i \in S\} \cup \{(a, t)\}$ with cost 0 and capacity 0.

Compared to the min-cost flow network constructed for the GfK problem, here we introduce $m + 1$ nodes additionally, as some points may remain unassigned to any center in S . Indeed, these points are composed of multiple fair-sets. More precisely, if there exists a multi-combiner $U(S')$ that contains an unassigned fair-set, then we can assign the points in the fair-set to any center in S' . Therefore, for each color $h \in H$, we construct an additional node a^h to receive the unassigned nodes in V_2 .

Observe that the sum of b -values of all nodes in V is equal to 0. Thus, we have flow conservation. Observe that all capacities, costs, and b -values of the min-cost flow network $(G = (V, E), b)$ are integral. Consequently, there always exists an integral optimal solution for $(G = (V, E), b)$. Let f be an optimal solution of the min-cost flow network. Since the flow f is integral, this implies an integral assignment of all points in C' to centers in S . The integral flow f induces an integral solution \mathbf{x} to instance \mathcal{I} as follows. Firstly, for any $j \in C'$ and $i \in S$, we set x_{ij} equal to the flow value of edge $(v_j, v_i^h) \in E_2^1$. Secondly, for any $i \in S$, let $x_{ij} = 1$ if $j \in U_i$. For any $j \in C'$ and $h \in H$, if the flow value of edge $(v_j, a^h) \in E_2^2$ is non-zero, then each node a^h will receive $r \cdot \frac{|C_h|}{g}$ ($r \in \mathbb{N}_{\geq 1}$) flows from nodes with color h , which implies that C' contains r fair-sets unassigned to any center. More precisely, consider a multi-combiner $U(S')$ containing some unassigned points by the above operation, and denote by U the set of unassigned points in $U(S')$. It must be the case that U is composed of one or more fair-sets. Recall that a fair-set is a minimal set satisfying fairness constraints. Finally, we assign all points in U as a whole to any center in S' , i.e., for a fixed center $i \in S$, let $x_{ij} = 1$ for any $j \in U$. By the above construction of \mathbf{x} , we have $\text{cost}(\mathbf{x}) \leq 2\tau_f$ due to the distance-bounded property (a point can only be assigned to a center within a distance $2\tau_f$).

We now prove that \mathbf{x} is an integral solution without any violation for the fairness constraints. By the function *Calculate*, we get that for each $i \in S$, a single-combiner U_i needs exactly R_i^h points with color h for each $h \in H$ and $\sum_{h \in H} R_i^h$ points in total such that U_i satisfies fairness constraints. Observe that for each $i \in S$, there are at least R_i^h points with color h assigned to i due to the b -values of node v_i^h . Since there is at most one outgoing arc (v_i^h, v_i) of capacity 0 for node v_i^h , we have exactly R_i^h points with color h and $\sum_{h \in H} R_i^h$ points in total assigned to i . Consequently, we obtain an integral solution \mathbf{x} without fairness violation.

The analysis of the running time is similar to Lemma 3, and we can obtain that LA-EXACT is polynomial. \square

■ E A Deterministic Heuristic Algorithm

In this section, we propose a deterministic heuristic algorithm,

Algorithm 3 LA-GREEDY

Input: An instance (C, d, k) of the k -center related problem, an auxiliary clustering (P_1, \dots, P_k) of C , a label error rate γ and parameters τ_f, l

Output: A set B of k local balls

```

1:  $S \leftarrow \emptyset, B \leftarrow \emptyset;$ 
2: for  $i = 1$  to  $k$  do
3:    $G_i \leftarrow$  pick  $l$  points from  $P_i$  by the greedy strategy;
4:    $s_i \leftarrow$  a point with minimum cost in  $G_i$  of  $P_i$ ;
5:    $S \leftarrow S \cup \{s_i\};$ 
6: end for
7: for  $i = 1$  to  $k$  do
8:    $B_i \leftarrow \{v \in C \mid d(v, s_i) \leq 2\tau_f\};$ 
9:    $B \leftarrow B \cup \{B_i\};$ 
10: end for
11: return  $B$ .
```

Table 1 Datasets.

Dataset	Reuters	Victorian	4area
Size	2,500	4,500	35,385
Dimension	10	10	8
Number of Groups	50	45	4
Sensitive	author	author	area

called LA-GREEDY (see Algorithm 3), that uses the greedy strategy in [19] for the k -center problem to select points. Here we briefly review how the algorithm in [19] works. Given an instance (C, d, k) of the k -center problem, the algorithm first selects an arbitrary point from C as center. Then, it iteratively selects the next center that is the farthest point from all chosen centers until k centers are chosen. In algorithm LA-GREEDY, we select l points by the above strategy, where $l \in \mathbb{N}^+$ is a given parameter. Note that compared with LA-SAMPLING, the difference is that in LA-GREEDY, we select a subset G_i of l points from each auxiliary cluster P_i ($i \in [k]$), and obtain a point $s_i \in G_i$ with the minimum cost of P_i . Since LA-GREEDY is deterministic to obtain S , it is hard to get the relation between the points in S and the optimal solution with an approximation bound, compared with LA-SAMPLING using probabilistic method to bound the approximation loss. Therefore, we can obtain two deterministic algorithms, called LAG-GROUP-FAIR and LAG-EXACT, by using LA-GREEDY to replace the step 1 of algorithms LA-GROUP-FAIR and LA-EXACT for the GFkC problem and its special case, respectively.

F Experiments

In this section, we consider the GFkC problem, and compare our proposed algorithms with the state-of-the-art algorithms. All experiments are conducted on 12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz. Following the prior settings [1, 2], we run each algorithm for 10 times, and report the average clustering cost and running time.

Datasets. We conduct experiments on the following 3 real datasets frequently used in fair clustering, including Reuters, Victorian and

Table 2 Comparison of average costs returned by algorithms for the GFkC problem on real datasets.

Datasets	β	α	HARB	SAMPLING	GREEDY
Reuters	0	0.05	2.325	2.583	2.306
	0	0.20	2.325	2.583	2.305
	0	0.40	2.326	2.581	2.305
Victorian	0	0.10	5.630	4.963	4.706
	0	0.30	5.047	4.984	4.589
	0	0.50	4.974	4.983	4.602
4area	0	0.45	12.584	13.089	12.505
	0	0.60	12.584	13.089	12.505
	0	0.80	12.584	13.089	12.505

4area from [7]. Reuters dataset contains 50 English language texts from each of 50 authors, where author is considered as the sensitive attribute to generate 50 groups. Victorian dataset contains the texts from 45 English language authors, in which each text consists of 1000-word sequences obtained from a book of the author. We use author as sensitive attribute to get 45 groups. 4area dataset contains researchers from four areas of computer science, where the main area of research is considered as the sensitive attribute to generate 4 groups. We normalize all used datasets to have zero mean and standard deviation of one. The datasets used in our experiments are summarized in Table 1.

Learning-Augmented Setting. As pointed out in [1], the predictor used typically derives labels via heuristic algorithms or trained neural networks. However, obtaining optimal clustering is computationally challenging in practice. Thus, heuristic algorithms are commonly employed to approximate optimal clustering, and error rate is estimated based on the guess of different values with respect to the clustering cost. Consequently, even with classical clustering problems, it is unclear whether the learning-augmented condition is satisfied. In our setting, incorporating fairness constraints further complicates the task of achieving an optimal clustering. Hence, we consider the heuristic algorithm to approximate the optimal solution as the works in [1, 2]. In particular, we use the greedy algorithm of [19], which is a 2-approximation algorithm for the k -center problem, to generate a feasible clustering in replacement of the optimal clustering (P_1^*, \dots, P_k^*) for given instance. We take (P_1^*, \dots, P_k^*) with the minimum clustering cost after executing the greedy algorithm on 10 independently trails. While previous work [1, 2] estimates the error rate through heuristic guesswork related to clustering costs, our theoretical results are independent of the error rate. Therefore, in our experiments, we directly use the computed optimal clustering as the predictor.

Baselines Algorithms. We consider the baseline algorithm to be the one in [7], which solves the GFkC problem and achieves a 3-approximation with additive violation 7. Note that the algorithm in [7] is practical, and has better performance than previous algorithm for the GFkC problem. We denote the baselines algorithm as HARB, and denote LA-GROUP-FAIR and LAG-GROUP-FAIR as SAMPLING and GREEDY, respectively.

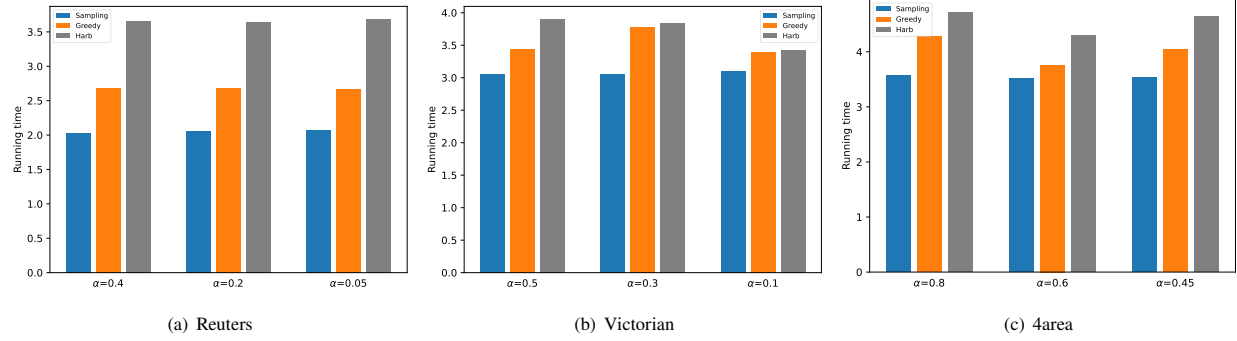


Fig. 1 Comparison of average time in seconds returned by algorithms for the GFkC problem on real datasets.

Parameters and Results. For the GFkC problem, the parameters are fixed as $k = 5$, $\beta = 0$, and α is selected with feasible value following in [7]. For Reuters, Victorian and 4area datasets, l is set to be 120, 50 and 100, respectively. We use CPLEX to solve the LP. Table 2 and Figure 1 compare the average clustering costs and running time returned by algorithms for the GFkC problem on real datasets. For clustering cost and running time, our proposed algorithm has better performance on all datasets. Theoretically, as the value of α increases, the clustering cost is expected to decrease, since the fairness constraints become more relaxed, allowing points to be assigned closer to their nearest centers. Conversely, as α decreases, points may be forced to deviate from their nearest centers to satisfy tighter fairness constraints, causing clustering cost to increase. In the particular case of the 4area dataset, the value of α is sufficiently large to permit most points to remain assigned to their nearest centers, thus the clustering cost remains stable across tested values of α .

References

- [1] Ergun J C, Feng Z, Silwal S, Woodruff D, Zhou S. Learning-augmented k -means clustering. In: Proceedings of the 10th International Conference on Learning Representations. 2022
- [2] Nguyen T D, Chaturvedi A, Nguyen H L. Improved learning-augmented algorithms for k -means and k -medians clustering. In: Proceedings of the 11th International Conference on Learning Representations. 2023
- [3] Chierichetti F, Kumar R, Lattanzi S, Vassilvitskii S. Fair clustering through fairlets. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 5029–5037
- [4] Bercea I O, Groß M, Khuller S, Kumar A, Rösner C, Schmidt D R, Schmidt M. On the cost of essentially fair clusterings. In: Proceedings of the 22nd International Conference on Approximation Algorithms for Combinatorial Optimization Problems and 23rd International Conference on Randomization and Computation. 2019, 18:1–18:22
- [5] Ahmadian S, Epasto A, Kumar R, Mahdian M. Clustering without over-representation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019, 267–275
- [6] Bera S K, Chakrabarty D, Flores N, Negahbani M. Fair algorithms for clustering. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019, 4955–4966
- [7] Harb E, Lam H S. KFC: A scalable approximation algorithm for k -center fair clustering. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020, 14509–14519
- [8] Bera S K, Das S, Galhotra S, Kale S S. Fair k -center clustering in mapreduce and streaming settings. In: Proceedings of the ACM Web Conference 2022. 2022, 1414–1422
- [9] Wu X, Feng Q, Huang Z, Xu J, Wang J. 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. 2024, 1938–1946
- [10] Ceccarello M, Pietracaprina A, Pucci G. Fast and accurate fair k -center clustering in doubling metrics. In: Proceedings of the ACM on Web Conference 2024. 2024, 756–767
- [11] Bandyapadhyay S, Fomin F V, Simonov K. On coresets for fair clustering in metric and euclidean spaces and their applications. Journal of Computer and System Sciences, 2024, 142: 103506
- [12] Kleindessner M, Awasthi P, Morgenstern J. Fair k -center clustering for data summarization. In: Proceeding of the 36th International Conference on Machine Learning. 2019, 3448–3457
- [13] Jones M, Lê Nguyễn H, Nguyen T. Fair k -centers via maximum matching. In: Proceedings of the 37th International Conference on Machine Learning. 2020, 4940–4949
- [14] Chen D Z, Li J, Liang H, Wang H. Matroid and knapsack center problems. Algorithmica, 2016, 75(1): 27–52
- [15] Krishnaswamy R, Kumar A, Nagarajan V, Sabharwal Y, Saha B. The matroid median problem. In: Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms. 2011, 1117–1130
- [16] Krishnaswamy R, Li S, Sandeep S. Constant approximation for k -median and k -means with outliers via iterative rounding. In: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing. 2018, 646–659
- [17] Thejaswi S, Gadekar A, Ordozgoiti B, Osadnik M. Clustering with fair-center representation: Parameterized approximation algorithms and heuristics. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022, 1749–

- [18] Zhang Z, Chen X, Liu L, Chen J, Huang J, Feng Q. Parameterized approximation schemes for fair-range clustering. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. 2024, 60192–60211
- [19] Gonzalez T F. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 1985, 38: 293–306