# Parameterized Approximation Algorithm for Doubly Constrained Fair Clustering

**Xiaoliang Wu**[1,2], **Qilong Feng**[1*], **Junyu Huang**[1,2] and **Jianxin Wang**[1,2,3]
[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2]Xiangjiang Laboratory, Changsha 410205, China
[3]The Hunan Provincial Key Lab of Bioinformatics, Central South University, Changsha 410083, China
xiaoliangwu@csu.edu.cn, csufeng@mail.csu.edu.cn, junyuhuangcsu@foxmail.com, jxwang@mail.csu.edu.cn

## Abstract

Fair clustering has recently received considerable attention where numerous distinct fairness notions are developed. Despite being well-justified, these fairness notions are frequently studied in isolation, leaving the need to explore how they can be combined. Building on prior work, we focus on the doubly constrained fair clustering that incorporates two widely adopted demographic representation fairness notions in clustering: group fairness and data summarization fairness. Both fairness notions extend classical clustering formulation by associating each data point with a demographic label, where group fairness requires each cluster to proportionally reflect the population-level distribution of demographic groups, and data summarization fairness ensures the chosen facilities maintaining the population-level demographic representation of each group. In this paper, we study the Fixed-Parameter Tractable (FPT) approximation algorithms for doubly constrained fair clustering under the $k$-median objective, referred to DF-$k$-MED. The previous algorithms typically enumerate different demographic groups or construct fairness coreset, parameterized by both the number of opened facilities and demographic labels. By further leveraging the local fairness information, we propose a color-agnostic structural method that obtains the parameterized result independent of the number of demographic labels while effectively handling the combination of both fairness constraints. Specifically, we design a constant factor approximation for the DF-$k$-MED problem with fairness violation by one, which runs in FPT$(k)$-time, where $k$ is the number of opened facilities.

## 1 Introduction

Clustering is a fundamental problem in machine learning, and has numerous applications in data mining, image classification, and beyond. Given a set of points in a metric space, the goal is to partition these points into several disjoint clusters such that points within the same cluster are close to each other, while points in different clusters remain relatively far apart. Several classic clustering models have been extensively studied, such as $k$-center, $k$-means [Huang *et al.*, 2024], and $k$-median. In this paper, we focus on the $k$-median problem, in which we are given a set $\mathcal{C}$ of clients, a set $\mathcal{F}$ of facilities in a metric space $(\mathcal{X}, d)$, and a positive integer $k$. The goal of the metric $k$-median problem is to open a subset in $\mathcal{F}$ of $k$ facilities such that the total distance from each client to its closest opened facility is minimized. The metric $k$-median problem is NP-hard, leading to a rich line of research on obtaining efficient approximation algorithms. The first constant-factor approximation algorithm for the metric $k$-median problem was given by [Charikar *et al.*, 2002], which was improved to $(3+\epsilon)$ by [Arya *et al.*, 2004] using a local-search method [Huang *et al.*, 2023]. Currently, the best known approximation factor for the metric $k$-median problem is 2.675 [Byrka *et al.*, 2017] by the dependent rounding technique, and the problem is known to be NP-hard to approximate to a factor less than $1 + 2/e$ [Guha and Khuller, 1998]. Moreover, finding an optimal solution for the metric $k$-median problem is known to be W[2]-hard if parameterized by $k$ due to a reduction by [Guha and Khuller, 1998] ($n^{O(1)}g(k)$ time for an input size of $n$ and a positive function $g$, denoted by FPT$(k)$-time for brevity). Recently, [Cohen-Addad *et al.*, 2019] presented approximation algorithm in FPT$(k)$-time with ratio $(1+2/e+\epsilon)$ for the metric $k$-median problem, which is essentially tight assuming the Gap-ETH.

Recently, fair clustering has been extensively studied, and lots of definitions about fairness have been proposed, such as group fairness [Ahmadian *et al.*, 2019; Bera *et al.*, 2019; Bercea *et al.*, 2019; Harb and Lam, 2020], data summarization fairness [Chiplunkar *et al.*, 2020; Jones *et al.*, 2020; Kleindessner *et al.*, 2019; Angelidakis *et al.*, 2022; Zhang *et al.*, 2024b], proportional fairness [Chen *et al.*, 2019; Micha and Shah, 2020], individual fairness [Mahabadi and Vakilian, 2020; Negahbani and Chakrabarty, 2021], etc. Despite these notions being well-justified, they are often examined independently. Motivated by the fact that many clustering applications require a unified solution that simultaneously satisfies multiple fairness notions, [Dickerson *et al.*, 2023] introduced the doubly constrained fair clustering that integrates both group fairness and data summarization fairness. Both fairness notions are motivated by the principle of disparate

*Corresponding author

impact [Feldman *et al.*, 2015], which asserts that different groups should receive equitable treatment, making the intersection of both notions a natural consideration. As noted in [Dickerson *et al.*, 2023], combining group fairness and data summarization fairness is particularly relevant, since both focus on demographic fairness and the representation of groups. Furthermore, the incompatibility is studies between the above two fairness constraints and a family of other fairness constraints. In this paper, following the prior work, we focus on the doubly constrained fair clustering. More precisely, we consider the doubly constrained fair clustering under the $k$-median objective, referred to Doubly Constrained Fair $k$-Median (DF-$k$-MED). In the doubly constrained fair clustering instance, we are given a set of facilities and a set of clients in a metric space, where both facilities and clients are divided into several disjoint groups, and each facility or client is assigned a color to denote which group it is in. The goal is to form $k$ clusters such that the clustering objective is minimized, while ensuring that the proportion of clients of each color in every cluster is within a specified range, and that the number of opened facilities of each color is equal to a given value.

[Chierichetti *et al.*, 2017] introduced the definition of fairness involving only two colors, requiring that the proportion of both colors has approximately equal representation in every cluster. [Bercea *et al.*, 2019] proposed the notion of group fairness, and provided a 4.675-approximation for the $k$-median objective, with an additive 1 violation for the group fairness constraints using linear programming and min-cost flow network. The violation value represents the extent to which the fairness constraints are violated (see [Bercea *et al.*, 2019] with details). [Ahmadian *et al.*, 2019] studied the group fairness under the $k$-center objective with only an upper bound constraint, and presented a 3-approximation with an additive 2 violation, where the definition of violation differs from that in [Bercea *et al.*, 2019], using linear programming and min-cost flow network. For the group fairness under the condition that colors are allowed to overlap, [Bera *et al.*, 2019] developed a $(\rho + 2)$-approximation algorithm with $(4\Delta+3)$ violation, where $\rho$ is the factor given by any approximation algorithm for the $k$-median problem, and $\Delta$ is the maximum number of colors a single point can belong to, respectively. Regarding parameterized result, [Bandyapadhyay *et al.*, 2024] employed the fair coreset technique, and proposed a $(3 + \epsilon)$-approximation for the group $k$-median problem, respectively, in FPT$(k, m)$-time where $m$ is the number of client colors.

[Kleindessner *et al.*, 2019] 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 [Jones *et al.*, 2020] through the maximum matching method, matching the approximation ratio of the matroid center problem [Chen *et al.*, 2016] 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 [Krishnaswamy *et al.*, 2011], the best known approximation ratio is $(7.081 + \epsilon)$ due to [Krishnaswamy *et al.*, 2018]. [Thejaswi *et al.*, 2022] 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 FPT$(k, t)$-time, where $t$ denotes the number of facility colors. Furthermore, [Zhang *et al.*, 2024a] proposed $(1 + \epsilon)$-approximation algorithm for the $k$-median objective in Euclidean metrics, operating in FPT$(k, t)$-time.

[Dickerson *et al.*, 2023] considered the doubly constrained fair clustering for the $k$-center objective, and the method is not workable for the DF-$k$-MED problem due to different optimization objective. The previous parameterized algorithms in [Bandyapadhyay *et al.*, 2024; Zhang *et al.*, 2024a; Thejaswi *et al.*, 2022] mainly enumerate feasible facilities based on their colors or construct coreset satisfying fairness constraints. However, these results are parameterized by both the number of opened facilities and the number of colors. The reason behind is that the related color information of fairness notions is essential for satisfying the group fairness or data summarization fairness. Therefore, in doubly constrained fair clustering, it seems inevitable that the approximation results have parameterized dependency on the number of colors. Moreover, due to the existence of two fairness constraints, the problem is considerably more challenging than its non-fair or one-fair counterpart. Since even if a set of $k$ facilities is provided, it remains NP-hard to find a solution satisfying the group fairness constraints [Esmaeili *et al.*, 2021]. Therefore, we must strategically ensure that the resulting solution satisfies both requirements while maintaining good approximation guarantees. Naturally, for the DF-$k$-MED problem, we then ask whether effective approximation algorithm can be developed in FPT$(k)$-time.

### 1.1 Our Contributions

In this paper, we propose a color-agnostic structural method to overcome the aforementioned obstacles. Our main contribution provides a positive answer to the stated question, summarized as follows.

- By exploring the local fairness information, we propose a color-agnostic structural method, which directly avoids the enumeration of clients or facilities with different colors, and exploits the inherent structures associated with fairness notions itself. Based on the proposed method, we obtain constant-factor approximation result for the DF-$k$-MED problem in FPT$(k)$-time, parameterized solely by the number of opened facilities.
- The proposed color-agnostic structural method simultaneously handles both fairness constraints while maintaining good approximation guarantees. Specifically, we theoretically prove that there must exist a $(4 + \epsilon)$-approximation for the DF-$k$-MED problem, running in FPT time.

We summarize the results in the literature and ours in Table 1. Formally, we have following result for the DF-$k$-MED problem.

**Theorem 1.** *For any $\epsilon > 0$, there exists a randomized $(4+\epsilon)$-approximation algorithm of fairness violation by one for the*

| Fairness Constraints | Approximation | Time | Reference |
|---|---|---|---|
| Group Fair | $3+\epsilon$ | FPT$(k,m)$ | [Bandyapadhyay *et al.*, 2024] |
| Data Summarization | $1+2/e+\epsilon$ | FPT$(k,t)$ | [Thejaswi *et al.*, 2022] |
| Doubly Fair | $4+\epsilon$ | FPT$(k)$ | Theorem 1 |

Table 1: Parameterized approximation results for the group fairness and data summarization fairness in metric space.

DF-$k$-MED *problem with running time* $f(k,\epsilon)n^{O(1)}$*, where* $f(k,\epsilon)=(O(\epsilon^{-2}k\log k))^k$.

## 2 Preliminaries

For any $m\in\mathbb{Z}^+$, let $[m]=\{1,\dots,m\}$. Given a metric space $(\mathcal{X},d)$, let $\mathcal{F}\subseteq\mathcal{X}$ and $\mathcal{C}\subseteq\mathcal{X}$ denote the set of facilities and clients considered in this paper, respectively. Let $|\mathcal{F}\cup\mathcal{C}|=n$. Given a client $c\in\mathcal{C}$ and a subset $\mathcal{H}\subseteq\mathcal{F}$, let $d(c,\mathcal{H})=\min_{f\in\mathcal{H}}d(c,f)$ be the distance from $c$ to its nearest facility in $\mathcal{H}$.

**Definition 2** (the metric $k$-median problem)**.** *An instance of the $k$-median problem is denoted by $((\mathcal{X},d),\mathcal{F},\mathcal{C},k)$, where $(\mathcal{X},d)$ is a metric space, $\mathcal{F}\subseteq\mathcal{X}$ is a set of facilities, $\mathcal{C}\subseteq\mathcal{X}$ is a set of clients, and $k$ is a positive integer, respectively. The goal is to find a subset $\mathcal{H}\subseteq\mathcal{F}$ of $k$ facilities such that the cost $\sum_{v\in\mathcal{C}}d(c,\mathcal{H})$ is minimized.*

The coreset is a commonly used tool in clustering algorithms [Har-Peled and Mazumdar, 2004], defined formally as follows.

**Definition 3** (coreset)**.** *Given an instance $((\mathcal{X},d),\mathcal{F},\mathcal{C},k)$ of the metric $k$-median problem and a parameter $\eta>0$, a coreset is a subset of clients $\mathcal{C}^\dagger\subseteq\mathcal{C}$ with associated weights $\{w(c):c\in\mathcal{C}^\dagger\}$ such that for any subset of facilities $\mathcal{H}\subseteq\mathcal{F}$ of size $k$, $\sum_{c\in\mathcal{C}^\dagger}w(c)d(c,\mathcal{H})\in[1-\eta,1+\eta]\cdot\sum_{c\in\mathcal{C}}d(c,\mathcal{H})$.*

The metric $k$-means problem has a similar definition of the coreset. [Chen, 2009] introduced the first coreset for the metric $k$-median problem, and the following result is the best known construction.

**Theorem 4** ([Feldman and Langberg, 2011])**.** *Given an instance $((\mathcal{X},d),\mathcal{F},\mathcal{C},k)$ of the metric $k$-median problem and parameters $\eta>0$, $\gamma<1/2$, there exists a randomized algorithm that, with probability at least $(1-\gamma)$, computes a coreset $\mathcal{C}^\dagger\subseteq\mathcal{C}$ of size $|\mathcal{C}^\dagger|=O(\eta^{-2}(k\log n+\log\frac{1}{\gamma}))$ in time $O(k(n+k)+\log^2\frac{1}{\gamma}\log^2\frac{1}{n})$.*

Theorem 4 implies that the original client set $\mathcal{C}$ can be reduced such that the total distance to any chosen facility set is distorted by at most a factor of $(1+\eta)$.

**Definition 5** (the DF-$k$-MED problem)**.** *An instance of the doubly constrained fair $k$-median problem is denoted by $((\mathcal{X},d),\mathcal{F},\mathcal{P}_1,\mathcal{C},\mathcal{P}_2,\vec{\theta},\vec{\alpha},\vec{\beta},t,m,k)$, where $(\mathcal{X},d)$ is a metric space, $\mathcal{F}\subseteq\mathcal{X}$ is a set of facilities with a partition $\mathcal{P}_1=\{\mathcal{F}_1,\dots,\mathcal{F}_t\}$ satisfying $\cup_{h=1}^t\mathcal{F}_h=\mathcal{F}$, $\mathcal{C}\subseteq\mathcal{X}$ is a set of clients with a partition $\mathcal{P}_2=\{\mathcal{C}_1,\dots,\mathcal{C}_m\}$ satisfying $\cup_{h=1}^m\mathcal{C}_h=\mathcal{C}$, $\vec{\theta}=(k_1,\dots,k_t)$ satisfying $\sum_{h=1}^t k_h=k$, $\vec{\alpha}=(\alpha_1,\dots,\alpha_m),\vec{\beta}=(\beta_1,\dots,\beta_m)$ are three fairness vectors, and $t,m,k$ are three positive integers, respectively. The goal is to find a subset $\mathcal{H}\subseteq\mathcal{F}$ of $k$ facilities and a mapping $\phi:\mathcal{C}\to\mathcal{H}$ such that the cost $\sum_{v\in\mathcal{C}}d(c,\phi(v))$ is minimized, and the following conditions hold: (1) The set $\mathcal{H}$ satisfies the data summarization fairness constraints, i.e., for any $h\in[t]$, $|\mathcal{H}\cap\mathcal{F}_h|\le k_h$; (2) The mapping $\phi$ satisfies the group fairness constraints, i.e., for any $f\in\mathcal{H},h\in[m]$, $\beta_h\le\frac{|\{c\in\mathcal{C}_h|\phi(c)=f\}|}{|\{c\in\mathcal{C}|\phi(c)=f\}|}\le\alpha_h$.*

Given an instance $((\mathcal{X},d),\mathcal{F},\mathcal{P}_1,\mathcal{C},\mathcal{P}_2,\vec{\theta},\vec{\alpha},\vec{\beta},t,m,k)$ the DF-$k$-MED problem, a pair $(\mathcal{H},\phi)$ is called a feasible solution of this instance if $\mathcal{H}\subseteq\mathcal{F}$ is a set with size $k$ satisfying data summarization fairness constraints, and $\phi:\mathcal{C}\to\mathcal{H}$ is a mapping satisfying group fairness constraints. Let $(\mathcal{H}^*,\phi^*)$ be an optimal solution with cost $opt=\sum_{c\in\mathcal{C}}d(c,\phi^*(c))$, and let $\mathcal{O}_1^*,\dots,\mathcal{O}_k^*$ be the corresponding $k$ optimal clusters under mapping $\phi^*$ with $\mathcal{C}=\cup_{i\in[k]}\mathcal{O}_i^*$. For any $i\in[k]$ and $h\in[m]$, let $\mathcal{O}_i^*(h)$ be the set of clients in $\mathcal{O}_i^*$ with color $h$. Then, we have $\mathcal{O}_i^*=\cup_{h\in[m]}\mathcal{O}_i^*(h)$.

## 3 An Overview of Our Algorithms

In this paper, we propose a color-agnostic structural method for the DF-$k$-MED problem based on the parameterized approximation framework developed by [Cohen-Addad *et al.*, 2019]. This framework first identifies a set of clients, called leaders, which are close to the facilities opened by an optimal solution. It then selects the facilities to open by searching within annular regions centered on these leaders, where the radius of each region approximates the distance between the leader and its associated facility. Over the past few years, this framework has yielded various approximation algorithms in FPT time for clustering problem [Cohen-Addad and Li, 2019; Bandyapadhyay *et al.*, 2024; Zhang *et al.*, 2024a; Thejaswi *et al.*, 2022], including capacitated clustering, fair clustering, etc. In the context of fair clustering, due to the existence of fairness constraints, it seems unavoidable that the parameterized approximation results include the parameter with respect to the number of colors. For instance, [Thejaswi *et al.*, 2022] enumerated the set of facilities with different colors to satisfy the data summarization fairness constraints, while [Bandyapadhyay *et al.*, 2024] constructed a composable coreset for group fairness through separately sampling on clients with different colors, resulting in the parameterized dependency on the number of colors to facilities or clients. Instead of enumerating feasible facilities or applying fair coreset techniques, both of which depend on the number of colors, our improved strategy avoids color-based parameters by adding an additional partition matroid constraint on the selected facilities and enumerating clients without considering their colors, based on the inherent structural properties exploited from fairness constraints itself.

We now provide an intuitive overview of our algorithm for the DF-$k$-MED problem. The algorithm begins by utilizing a coreset technique for the $k$-median problem to reduce the number of clients to a weighted set, without considering fairness constraints, as incorporating fairness into the coreset will incur parameter dependence on the number of colors [Bandyapadhyay *et al.*, 2024]. In the resulting weighted set, our algorithm then identifies a set of leaders closed to the facilities in an optimal solution. By guessing the leaders and their distances to these facilities, the algorithm enumerates potential configurations of facilities within suitable distance ranges. Next, it constructs a monotone submodular function by introducing fictitious facilities, and solves a monotone submodular maximization problem under two partition matroid constraints, thus ensuring data summarization fairness. Finally, a weighted fair assignment problem is defined to assign the weighted clients to the chosen facilities satisfying the group fairness constraints, where a widely used linear programming procedure guarantees the desired approximation ratio. By the above process, we develop an FPT($k$)-time approximation algorithm with approximation ratio $(4+\epsilon)$ of 1 fairness violation for the DF-$k$-MED problem, where $\epsilon > 0$ is a given parameter.

## 4 Parameterized Algorithm for Doubly Constrained Fair Clustering

We now present our color-agnostic structural algorithm for the DF-$k$-MED problem. The high-level idea of our algorithm is as follows. Consider an instance $((\mathcal{X}, d), \mathcal{F}, \mathcal{P}_1, \mathcal{C}, \mathcal{P}_2, \vec{\theta}, \vec{\alpha}, \vec{\beta}, t, m, k)$ of the DF-$k$-MED problem and parameters $\eta, \gamma, \delta > 0$. Let $(\mathcal{H}^*, \phi^*)$ be an optimal solution, where $\mathcal{H}^* = \{f_1^*, \ldots, f_k^*\}$ is the set of $k$ facilities opened by the optimal solution. Our algorithm consists of two phases. The first phase (steps 1-15 of Algorithm 1) starts with the coreset technique on the instance with parameter $\eta$, yielding a weighted set $\mathcal{C}^\dagger$ of size $O(\eta^{-2}k\log n)$ and an associated weight function $w$. Now our goal is to identify a subset $\mathcal{H} \subseteq \mathcal{F}$ satisfying the data summarization fairness and a weighted assignment $\psi$ from the clients in $\mathcal{C}^\dagger$ to $\mathcal{H}$, minimizing the corresponding cost. Note that the first phase only focuses on the finding of $\mathcal{H}$. Next we guess a set of leaders and distances between the leaders and the corresponding facilities. For any $i \in [k]$, let $\ell_i$ be the closest client in $\mathcal{C}^\dagger$ to $f_i^* \in \mathcal{H}^*$. We call $\ell_i$ the leader of $f_i^*$. Let $\lambda_i$ be the distance $d(\ell_i, f_i^*)$, rounded down to the closest integer power of $(1+\delta)$ with parameter $\delta$. The algorithm then enters an enumeration phase, with $|\mathcal{C}^\dagger|^k$ choices for $\{\ell_1, \ldots, \ell_k\}$ and $O(\delta^{-1}\log n)^k$ choices for $\{\lambda_1, \ldots, \lambda_k\}$ (see Subsection 4.1 with more details). By enumerating over all $|\mathcal{C}^\dagger|^k O(\delta^{-1}\log n)^k$ combinations, we can assume that we have identified the correct leaders and distances. For each leader $\ell_i$ ($i \in [k]$), let $\mathcal{N}_i$ denote the set of facilities $f \in \mathcal{F}$ satisfying $\lambda_i \le d(f, f_i^*) < (1+\delta)\lambda_i$. To construct a monotone submodular function, we add a fictitious facility $f_i'$ to each $\mathcal{N}_i$ ($i \in [k]$). For any $\mathcal{H} \subseteq \mathcal{F}$, we then formulate a monotone submodular maximization problem under two partition matroid constraints to achieve the desired set $\mathcal{H}$ that satisfies data summarization constraints. The second phase (steps 16-17 of Algorithm 1) aims to find an assignment of clients from $\mathcal{C}^\dagger$ to the facilities in $\mathcal{H}$ satisfying the group fairness constraints. This phase avoids the use of fair coreset. To this end, we define a weighted assignment problem, and prove theoretical that there must exist a $(4+\epsilon)$-approximate solution for the DF-$k$-MED problem. For obtaining such solutions, we model the defined problem to the linear programming procedure, and utilize the min-cost flow network to solve it. Algorithm 1 details the specific process of our parameterized algorithm for the DF-$k$-MED problem.

**Algorithm 1:** An algorithm for the DF-$k$-MED problem

**Input**: An instance $((\mathcal{X}, d), \mathcal{F}, \mathcal{P}_1, \mathcal{C}, \mathcal{P}_2, \vec{\theta}, \vec{\alpha}, \vec{\beta}, t, m, k)$ of the DF-$k$-MED problem, parameters $\eta, \gamma, \delta > 0$
**Output**: A pair $(\mathcal{H}, \phi)$

1 Let $\mathcal{C}^\dagger$ be the weighted set of clients constructed by Theorem 4 with $((\mathcal{X}, d), \mathcal{F}, \mathcal{C}, k, \eta, \gamma)$ as the input, and let $w : \mathcal{C}^\dagger \to \mathbb{Z}^+$ be the corresponding weighted function;
2 Let $d_{\min}$ and $d_{\max}$ be the maximum and minimum distances between any two points in $\mathcal{F} \cup \mathcal{C}$, respectively;
3 $\Lambda \leftarrow \{d_{\min}, (1+\delta)d_{\min}, (1+\delta)^2 d_{\min}, \ldots, d_{\max}\}$;
4 **for** *each multi-set* $\{\ell_1, \ldots, \ell_k\} \subseteq \mathcal{C}^\dagger$ **do**
5  **for** *each multi-set* $\{\lambda_1, \ldots, \lambda_k\} \subseteq \Lambda$ **do**
6  **for** $i = 1$ **to** $k$ **do**
7  $\mathcal{N}_i \leftarrow \{f \in \mathcal{F} \mid \lambda_i \le d(f, \ell_i) < (1+\delta)\lambda_i\}$;
8  Construct a new facility $f_i'$, and add it to $\mathcal{N}_i$;
 **for** *each* $f \in \mathcal{N}_i$ **do**
9  $d(f, f_i') \leftarrow 2\lambda_i$;
10  **for** *each* $c \notin \mathcal{N}_i$ **do**
11  $d(c, f_i') \leftarrow \min_{f \in \mathcal{N}_i}(d(c, f) + d(f, f_i'))$;
12  $\mathcal{F}' \leftarrow \{f_1', \ldots, f_k'\}$;
13  Define $\Delta(\mathcal{H}) = \Phi(\mathcal{C}^\dagger, \mathcal{F}') - \Phi(\mathcal{C}^\dagger, \mathcal{H} \cup \mathcal{F}')$ for any $\mathcal{H} \subseteq \mathcal{F}$;
14  Find $\mathcal{H} \subseteq \mathcal{F}$ maximizing function $\Delta(\mathcal{H})$ such that $|\mathcal{H} \cap \mathcal{N}_i| = 1$ for any $i \in [k]$, and $|\mathcal{H} \cap \mathcal{F}_i| = k_h$ for any $h \in [t]$;
15 Output $\mathcal{H}$ such that $\Phi(\mathcal{C}^\dagger, \mathcal{H})$ is minimized among all $\mathcal{H}$ computed in Line 14;
16 $\psi \leftarrow$ solve the Weighted Assignment problem on $\mathcal{C}^\dagger$ and $\mathcal{H}$;
17 $\phi \leftarrow$ obtain the assignment from the original client set $\mathcal{C}$ to $\mathcal{H}$ based on $\psi$;
18 **return** $(\mathcal{H}, \phi)$.

### 4.1 Finding Feasible Facilities

In this section, we describe how to select a set $\mathcal{H}$ of $k$ facilities that satisfies the data summarization fairness constraints in the first phase. The previous method in [Thejaswi *et al.*, 2022] enumerated the feasible facilities with different colors, leading to the result parameterized by the number of colors to facilities. To avoid the enumeration, we explore the inherent structure related to the data summarization fairness constraints, and model it to a partition matroid constraint.

Consider an instance $((\mathcal{X}, d), \mathcal{F}, \mathcal{P}_1, \mathcal{C}, \mathcal{P}_2, \vec{\theta}, \vec{\alpha}, \vec{\beta}, t, m, k)$ of the DF-$k$-MED problem and a parameter $\eta > 0$. By applying a coreset construction with parameter $\eta$, we reduce the original client set $\mathcal{C}$ to a weighted set $\mathcal{C}^\dagger$ of size $O(\eta^{-2}k \log n)$, where each client in $\mathcal{C}^\dagger$ is weighted by an integer indicating how many original clients it represents. However, the above process does not consider the inherent structural information of the clients, such as their colors, leading to a loss of the crucial attribute. Therefore, to address the later group fairness constraints, we record the number of points with each color assigned to each weighted client in $\mathcal{C}^\dagger$. Specifically, consider a client $c \in \mathcal{C}^\dagger$ with weight $w(c)$, and for any $h \in [m]$, let $n_h^c$ denote the number of clients with color $h$ assigned to $c$. Consequently, $\sum_{h=1}^{m} n_h^c = w(c)$. Moreover, we introduce a function $\rho : \mathcal{C} \to \mathcal{C}^\dagger$ denoting the mapping relation. By applying Theorem 4, we can obtain a weighted instance $((\mathcal{X}, d), \mathcal{F}, \mathcal{P}_1, \mathcal{C}^\dagger, w, \mathcal{P}_2, \vec{\theta}, \vec{\alpha}, \vec{\beta}, t, m, k)$. For simplicity, we do not explicitly track the number of clients in the original set with different color assigned to the client in this weighted instance.

Recall that the goal of this section is to find a set $\mathcal{H}$ of facilities satisfying the data summarization fairness constraints. Let $\mathcal{H}^* = \{f_1^*, \ldots, f_k^*\}$ denote the set of $k$ facilities opened by an optimal solution. For each $f_i^* \in \mathcal{H}^*$ ($i \in [k]$), let $\ell_i$ be the closest client in $\mathcal{C}^\dagger$ from $f_i^*$. Assume that we have correctly guessed the set of $k$ leaders (i.e., $\{\ell_1, \ldots, \ell_k\}$) with respect to the $k$ optimal facilities. We now discuss how to guess the distance between $f_i^*$ and $\ell_i$ for any $i \in [k]$. Although the optimal facility $f_i^*$ is unclear, it is known that $d(f_i^*, \ell_i)$ must be the distance between some facility and client in $\mathcal{C} \cup \mathcal{F}$. Thus, we can use the following discretization trick. Let $d_{\min}$ and $d_{\max}$ be the minimum distance and the maximum distance of any two points in $\mathcal{X}$, respectively. Then, the aspect ratio of this metric is $\frac{d_{\max}}{d_{\min}}$, capturing the ratio of the largest to smallest pairwise distance among points in $\mathcal{C} \cup \mathcal{F}$. It is well-known that we can assume that the aspect ratio can be bounded by polynomial in $n$ [Cohen-Addad *et al.*, 2019]. For a small parameter $\delta > 0$, we can guess the distance $d(f_i^*, \ell_i)$ from the set $\{d_{\min}, (1+\delta)d_{\min}, (1+\delta)^2 d_{\min}, \ldots, d_{\max}\}$, which has at most $\log_{1+\delta} \frac{d_{\max}}{d_{\min}} = O(\log n)$ possible values. Assume that we have guess $d(f_i^*, \ell_i) \in [\lambda_i, (1+\delta)\lambda_i)$ where $\lambda_i = (1+\delta)^j d_{\min}$ for some $j \in [O(\log n)]$. Similarly, we can guess the discretized distances, denoted by $\lambda_1, \ldots, \lambda_k$, from each leader to its corresponding facility. Therefore, by enumerating over $|\mathcal{C}^\dagger|^k (O(\log n))^k$ choices, we can guess the right $\ell_i$ and $\lambda_i$ for all $i \in [k]$.

For each leader $\ell_i$ ($i \in [k]$), we construct a facility set $\mathcal{N}_i = \{f \in \mathcal{F} \mid \lambda_i \le d(f, \ell_i) < (1+\delta)\lambda_i\}$. Then, if we pick one arbitrary facility from each $\mathcal{F}_i$, a good distance-based property will be obtained with respect to an optimal pick. However, the resulting set by this way fails to satisfy the data summarization fairness constraints, and incurs a large loss in approximation guarantee. To achieve a better approximation while satisfying fairness constraints, we first construct a monotone submodular function, and then invoke the monotone submodular maximization problem under two partition matroid constraints, capturing the data summarization fairness constraints. We now show how to construct a monotone submodular function. For each set $\mathcal{N}_i$ ($i \in [k]$), we introduce a new facility $f_i'$, and add it to $\mathcal{N}_i$. Further, for each $f \in \mathcal{N}_i$, we set $d(f, f_i') = 2\lambda_i$. For each client $c \in \mathcal{C}^\dagger$, we define $d(c, f_i') = \min_{f \in \mathcal{N}_i}(d(c, f) + d(f, f_i'))$. Let $\mathcal{F}' = \{f_1', \ldots, f_k'\}$ be the set of $k$ new facilities. For any $\mathcal{A} \subseteq \mathcal{X}$, let $\Phi(\mathcal{C}^\dagger, \mathcal{A}) = \sum_{c \in \mathcal{C}^\dagger} w(c) d(c, \mathcal{A})$. For any set $\mathcal{H} \subseteq \mathcal{F}$, we define the improvement function $\Delta(\mathcal{H}) = \Phi(\mathcal{C}^\dagger, \mathcal{F}') - \Phi(\mathcal{C}^\dagger, \mathcal{H} \cup \mathcal{F}')$, and prove that it is monotone and submodular.

**Lemma 6.** *For any $\mathcal{H} \subseteq \mathcal{F}$, the above defined function $\Delta(\mathcal{H})$ is monotone and submodular with $\Delta(\emptyset) = 0$.*

*Proof.* It is easy to see that $\Delta(\emptyset) = 0$ by definition. We first show the function $\Delta(\mathcal{H})$ is monotone. Consider subsets $\mathcal{H} \subseteq \mathcal{H}' \subseteq \mathcal{F}$. We need to prove that $\Delta(\mathcal{H}) \le \Delta(\mathcal{H}')$. Then, we have $\Phi(\mathcal{C}^\dagger, \mathcal{F}' \cup \mathcal{H}) = \sum_{c \in \mathcal{C}^\dagger} d(c, \mathcal{F}' \cup \mathcal{H}) \ge \sum_{c \in \mathcal{C}^\dagger} d(c, \mathcal{F}' \cup \mathcal{H}') = \Phi(\mathcal{C}^\dagger, \mathcal{F}' \cup \mathcal{H}')$. Thus, we have $\Delta(\mathcal{H}) \le \Delta(\mathcal{H}')$.

We now need to prove that the function $\Delta(\mathcal{H})$ is monotone and submodular. Consider subsets $\mathcal{H} \subseteq \mathcal{H}' \subseteq \mathcal{F}$ and facility $f \in \mathcal{F}$. For each client $c \in \mathcal{C}^\dagger$, we have $x - \min(x, y) = \max(0, x - y)$, for any real numbers $x$ and $y$. Then, we have

$$\begin{aligned}
&d(c, \mathcal{F}' \cup \mathcal{H}) - d(c, \mathcal{F}' \cup (\mathcal{H} \cup \{f\})) \\
=&d(c, \mathcal{F}' \cup \mathcal{H}) - \min(d(c, \mathcal{F}' \cup \mathcal{H}), d(c, \{f\})) \\
=&\max(0, d(c, \mathcal{F}' \cup \mathcal{H}) - d(c, \{f\})) \\
\ge&\max(0, d(c, \mathcal{F}' \cup \mathcal{H}') - d(c, \{f\})) \\
=&d(c, \mathcal{F}' \cup \mathcal{H}') - \min(d(c, \mathcal{F}' \cup \mathcal{H}'), d(c, \{f\})) \\
=&d(c, \mathcal{F}' \cup \mathcal{H}') - d(c, \mathcal{F}' \cup (\mathcal{H}' \cup \{f\})).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\Delta(\mathcal{H} \cup \{f\}) - \Delta(\mathcal{H}) \\
=&\Phi(\mathcal{C}^\dagger, \mathcal{F}' \cup \mathcal{H}) - \Phi(\mathcal{C}^\dagger, \mathcal{F}' \cup \mathcal{H} \cup \{f\}) \\
=&\sum_{c \in \mathcal{C}^\dagger} w(c) d(c, \mathcal{F}' \cup \mathcal{H}) - \sum_{c \in \mathcal{C}^\dagger} w(c) d(c, \mathcal{F}' \cup \mathcal{H} \cup \{f\}) \\
\ge&\sum_{c \in \mathcal{C}^\dagger} w(c) d(c, \mathcal{F}' \cup \mathcal{H}') - \sum_{c \in \mathcal{C}^\dagger} w(c) d(c, \mathcal{F}' \cup \mathcal{H}' \cup \{f\}) \\
=&\Phi(\mathcal{C}^\dagger, \mathcal{F}' \cup \mathcal{H}') - \Phi(\mathcal{C}^\dagger, \mathcal{F}' \cup \mathcal{H}' \cup \{f\}) \\
=&\Delta(\mathcal{H}' \cup \{f\}) - \Delta(\mathcal{H}'),
\end{aligned}$$

which proves that the function is submodular. □

We apply the approximation algorithm of [Lee *et al.*, 2009] for monotone submodular maximization under multiple matroid constraints. We are interested in the set $\mathcal{H}$ that consists of one center from each $\mathcal{F}_i$ while satisfying the data summarization constraints, since one such set is the desired $\mathcal{H}^*$. Here, we aims to maximize $\Delta(\mathcal{H})$ under two partition matroid constraints, where the first one requires containing exactly one facility from each set $\mathcal{N}_i$, and the second one asks for including at most $k_h$ facility from each set $\mathcal{F}_h$. Note that the latter one captures the data summarization fairness constraints. Let $\mathcal{H}$ be the output in step 15 of Algorithm 1. By the result in [Lee *et al.*, 2009], we have $\Delta(\mathcal{H}) \le 1/2 \Delta(\mathcal{H}^*)$.

**Lemma 7.** *Consider an instance $\mathcal{I}$ of the* DF-$k$-MED *problem with parameters $\eta, \gamma, \delta > 0$. Let $\mathcal{C}^\dagger$ and $\mathcal{H} = \{f_1, \ldots, f_k\}$ be the output in step 1 and step 15 of Algorithm 1, respectively. Then, for any parameter $\epsilon_1 > 0$, we have $\Phi(\mathcal{C}^\dagger, \mathcal{H}) \leq (2+\epsilon_1)opt$, where opt is the optimal cost of $\mathcal{I}$.*

*Proof.* Let $\mathcal{F}' = \{f'_1, \ldots, f'_k\}$. We first bound the cost induced by the set that opens facilities in $\mathcal{F}'$. Consider a client $c \in \mathcal{C}^\dagger$ with $d(c, \mathcal{H}) = d(c, f_i)$. Observe that since $\ell_i$ is the closest client from $f_i^*$, we have $d(c, f_i^*) \geq d(\ell_i, f_i^*) \geq \lambda_i/(1+\delta)$. Thus, by the triangle inequality and the definition of facility $f'_i$, we have $d(c, f'_i) \leq d(c, f_i^*) + d(f_i^*, f'_i) = d(c, f_i^*) + 2\lambda_i \leq d(c, f_i^*) + 2(1+\delta)d(c, f_i^*) \leq (3+2\delta)d(c, f_i^*)$. Further, $d(c, \mathcal{F}') \leq (3+2\delta)d(c, \mathcal{H}^*)$. Combing over all client $c \in \mathcal{C}^\dagger$, $\sum_{c \in \mathcal{C}^\dagger} w(c)d(c, \mathcal{F}') = (3+2\delta)\sum_{c \in \mathcal{C}^\dagger} w(c)d(c, \mathcal{H}^*)$, and thus we have $\Phi(\mathcal{C}^\dagger, \mathcal{F}') \leq (3+2\delta)\Phi(\mathcal{C}^\dagger, \mathcal{H}^*)$.

We next show that the set $\mathcal{F}'$ will not decrease the cost. Consider a client $c \in \mathcal{C}^\dagger$ with $d(c, \mathcal{H}) = d(c, f_i)$, i.e., $f_i$ is the closest facility in $\mathcal{H}$ to $c$. Let $\sigma(c) \in \mathcal{N}_i$ be the closest facility to $c$ in $\mathcal{N}_i$. We claim that client $c$ is closer to $f_i \in \mathcal{H}$ than to $f'_i \in \mathcal{F}'$. Then, we have $d(c, f_i) \leq d(c, \sigma(c)) + d(\sigma(c), \ell_i) + d(\ell_i, f_i) \leq d(c, \sigma(c)) + \lambda_i + \lambda_i = d(c, \sigma(c)) + d(\sigma(c), f'_i) = d(c, f'_i)$, where the first inequality follows from the triangle inequality, and the second inequality and the second equality use the definition of leader $\ell_i$ and facility $f'_i$, respectively. Thus, $d(c, \mathcal{H}) \leq d(c, \mathcal{F}')$. Summing over all clients in $\mathcal{C}^\dagger$, we obtain $\sum_{c \in \mathcal{C}^\dagger} w(c)d(c, \mathcal{H}) = \sum_{c \in \mathcal{C}^\dagger} w(c)d(c, \mathcal{H} \cup \mathcal{F}')$. Thus, we have

$$\begin{aligned}
\Phi(\mathcal{C}^\dagger, \mathcal{H}) &= \Phi(\mathcal{C}^\dagger, \mathcal{H} \cup \mathcal{F}') = \Phi(\mathcal{C}^\dagger, \mathcal{F}') - \Delta(\mathcal{H}) \\
&\leq \Phi(\mathcal{C}^\dagger, \mathcal{F}') - 1/2\Delta(\mathcal{H}^*) \\
&\leq \Phi(\mathcal{C}^\dagger, \mathcal{F}') - 1/2(\Phi(\mathcal{C}^\dagger, \mathcal{F}') - \Phi(\mathcal{C}^\dagger, \mathcal{H}^*)) \\
&= 1/2\Phi(\mathcal{C}^\dagger, \mathcal{F}') + 1/2\Phi(\mathcal{C}^\dagger, \mathcal{H}^*) \\
&\leq 1/2 \cdot (3+2\delta)\Phi(\mathcal{C}^\dagger, \mathcal{H}^*) + 1/2\Phi(\mathcal{C}^\dagger, \mathcal{H}^*) \\
&= (2+\delta)\Phi(\mathcal{C}^\dagger, \mathcal{H}^*).
\end{aligned}$$

By Definition 3 and the fact that the cost induced by $\mathcal{H}^*$ is no more than that of $(\mathcal{H}^*, \phi^*)$, we get $\Phi(\mathcal{C}^\dagger, \mathcal{H}^*) \leq (1+\eta)\Phi(\mathcal{C}, \mathcal{H}^*) \leq (1+\eta)opt$. Thus, $\Phi(\mathcal{C}^\dagger, \mathcal{H}) \leq (2+\epsilon_1)opt$ with $\epsilon_1 = O(\eta\delta)$. □

### 4.2 Solving the Weighted Assignment Problem

Recall that a set $\mathcal{H}$ satisfying the data summarization fairness constraints is obtained in previous section. The goal of this section is to find an assignment from the weighted client set $\mathcal{C}^\dagger$ to the facilities in $\mathcal{H}$, where the assignment satisfies the group fairness. To proceed, we modify the weighted instance $((\mathcal{X}, d), \mathcal{F}, \mathcal{P}_1, \mathcal{C}^\dagger, w, \mathcal{P}_2^\dagger, \vec{\theta}, \vec{\alpha}, \vec{\beta}, t, m, k)$, since some structural information is not considered in the process of constructing coreset. Recall that for each client $c \in \mathcal{C}^\dagger$, we record how many points with each color assigned to $c$, denoted as $n_1^c, \ldots, n_m^c$. Then, we have $\sum_{h=1}^m n_h^c = w(c)$. For each client $c \in \mathcal{C}^\dagger$, we divide $c$ into $m$ clients $c^1, \ldots, c^m$, where each client $c^h$ ($h \in [m]$) with the same position as $c$ is assigned a weight $n_h^c$ with color $h$. Let $\mathcal{C}^\ddagger$ denote the new constructed set with the corresponding weight function $w^\dagger : \mathcal{C}^\ddagger \to \mathbb{Z}^+ \cup \{0\}$. Therefore, the above problem can be defined as the following weighted assignment problem formally.

**Definition 8** (the Weighted Assignment problem)**.** *Given a weighted set $\mathcal{C}^\ddagger$ of clients in a metric space $(\mathcal{X}, d)$ associated with weight function $w^\dagger : \mathcal{C}^\ddagger \to \mathbb{Z}^+ \cup \{0\}$, a set $\mathcal{P}_2^\ddagger = \{\mathcal{C}_1^\ddagger, \ldots, \mathcal{C}_m^\ddagger\}$ of $m$ disjoint groups with $\cup_{h=1}^m \mathcal{C}_h^\ddagger = \mathcal{C}^\ddagger$, two fairness vectors $\vec{\alpha} = (\alpha_1, \ldots, \alpha_m), \vec{\beta} = (\beta_1, \ldots, \beta_m)$, and a set $\mathcal{H}$ of $k$ facilities, the goal is to find a mapping $\psi : (\mathcal{C}^\ddagger \times \mathcal{H}) \to \mathbb{Z}^+ \cup \{0\}$ satisfying $\sum_{f \in \mathcal{H}} \psi(c, f) = w^\dagger(c)$ for any $c \in \mathcal{C}^\ddagger$, and the weighted group fairness constraints: for any $f \in \mathcal{H}, h \in [m]$, $\beta_h \leq \frac{\sum_{c \in \mathcal{C}_h^\ddagger} \psi(c,f)}{\sum_{c \in \mathcal{C}^\ddagger} \psi(c,f)} \leq \alpha_h$, such that the cost $\sum_{c \in C^\ddagger} \sum_{f \in \mathcal{H}} \psi(c, f)d(c, f)$ is minimized.*

Given an instance $\mathcal{J} = ((\mathcal{X}, d), \mathcal{C}^\ddagger, w^\dagger, \mathcal{P}_2^\ddagger, \vec{\alpha}, \vec{\beta}, \mathcal{H})$ of the Weighted Assignment problem, we call $\psi$ a feasible solution of $\mathcal{J}$ if $\psi$ satisfies weighted group fairness constraints and $\sum_{f \in \mathcal{H}} \psi(c, f) = w^\dagger(c)$ for any $c \in \mathcal{C}^\ddagger$. We define the cost of $\psi$ as $\text{cost}(\psi) = \sum_{c \in C^\ddagger} \sum_{f \in \mathcal{H}} \psi(c, f)d(c, f)$, which is the sum of the product of $\psi(c, f)$ and distance between a client $c \in C^\ddagger$ and a facility $f \in \mathcal{H}$. Note that for any $c \in C^\ddagger, f \in \mathcal{H}$, if $\psi(c, f) > 0$, the $\psi(c, f)$ unit of weight of the client $c$ is assigned to $f$. Let $\psi^* : (\mathcal{C}^\ddagger \times \mathcal{H}) \to \mathbb{Z}^+ \cup \{0\}$ denote the optimal solution of $\mathcal{J}$ such that the cost $\sum_{c \in C^\ddagger} \sum_{f \in \mathcal{H}} \psi^*(c, f)d(c, f)$ is minimized.

**Lemma 9.** *Given an instance $\mathcal{I}$ of the* DF-$k$-MED *problem and parameters $\eta, \gamma, \delta > 0$, let $\mathcal{J}$ be the obtained weighted assignment problem instance. Then, there must exist a solution $\psi$ satisfying weighted group fairness constraints such that the cost of $\psi$ is at most $(4+\epsilon)$ times the optimal cost of $\mathcal{I}$, where $\epsilon = O(\eta\delta)$.*

*Proof.* Our proof strategy begins by constructing a mapping $\phi$ from the original client set $\mathcal{C}$ to the facility set $\mathcal{H}$, which is then converted into the weighted mapping $\psi : \mathcal{C}^\ddagger \to \mathcal{H}$. In this process, each client is assigned to the facility in $\mathcal{H}$ closest to its corresponding optimal facility. Since the optimal solution is feasible, we can ensure that the constructed mapping satisfies the weighted group fairness constraints. Moreover, by using the triangle inequality, we can bound the cost of the resulting assignment, thus obtaining the stated approximation guarantees.

We first show how to construct the mapping $\psi$. For any $f_i^* \in \mathcal{H}^*$ ($i \in [k]$), let $\pi(f_i^*) = \arg\min_{f \in \mathcal{H}} d(f, f_i^*)$ denote the closest facility in $\mathcal{H}$ to $f_i^*$. For any $c \in \mathcal{C}$, let $\phi(c) = \pi(\phi^*(c))$. Note that the mapping $\phi$ is for the original client set $\mathcal{C}$, and we need convert it to the weighted client set $\mathcal{C}^\ddagger$ to obtain $\psi$. For any client $c \in \mathcal{C}$, assume that $c$ is assigned to $c^h$ in $\mathcal{C}^\ddagger$. By the definition of mapping $\rho$, we get that $\rho(c)$ is the closest client in $\mathcal{C}^\dagger$ to $c$. Recall that for each color $h \in [m]$, the weighted set $\mathcal{C}^\ddagger$ contains a client $c^h$ with the same position as $\rho(c) \in \mathcal{C}^\dagger$. Thus, we have $d(c^h, c) = d(\rho(c), c)$ For any $c \in \mathcal{C}$, if client $c$ is assigned to the facility $\phi(c) \in \mathcal{H}$ by $\phi$, we assign 1 unit of weight of the client $c^h$ to $\phi(c)$, i.e., set $\psi(c^h, \phi(c)) = 1$. By the triangle inequality, we have $d(c^h, \phi(c)) \leq d(c^h, c) + d(c, \phi(c)) = d(\rho(c), c) + d(c, \phi(c))$. Consider a client $c \in \mathcal{O}_i^*$ ($i \in [k]$),

and let $f$ be the closest facility in $\mathcal{H}$ to the client $\rho(c)$ in $\mathcal{C}^\ddagger$ that $c$ is assigned. By the triangle inequality, we have $d(c,\phi(c)) \le d(c,f_i^*)+d(f_i^*,\phi(c)) \le d(c,f_i^*)+d(f_i^*,f) \le d(c,f_i^*)+d(c,f_i^*)+d(c,f) \le 2d(c,f_i^*)+d(c,\rho(c))+d(\rho(c),f)$, where the second inequality follows from that $\phi(c)$ is the closest facility in $\mathcal{H}$ to $f_i^*$. Thus, we have that $d(c^h,\phi(c)) \le 2d(c,f_i^*)+2d(c,\rho(c))+d(\rho(c),f)$. Combing over all client $c \in \mathcal{C}$, we can get that $\mathrm{cost}(\psi) \le 2opt + 2\eta opt + (2+\epsilon_1)opt = (4+\epsilon_1+2\eta)opt$. Hence, the cost of $\psi$ is at most $(4+\epsilon)opt$, where $\epsilon = O(\eta\delta)$.

The remaining task is to prove that $\psi$ satisfies the weighted group fairness constraints. We first prove that the mapping $\phi$ on the original client set $\mathcal{C}$ satisfies the group fairness constraints. Since $(\mathcal{H}^*,\phi^*)$ is a feasible solution of $\mathcal{I}$, for any $i \in [k]$ and $h \in [m]$, we have $\beta_h \le \frac{|\mathcal{O}_i^*(h)|}{|\mathcal{O}_i^*|} \le \alpha_h$. For any $f \in \mathcal{H}$, let $N(f) = \{f_i^* \in \mathcal{H}^* \mid \pi(f_i^*) = f\}$ denote all facilities in $\mathcal{H}^*$ such that $f$ is the closest center. Note that $\{c \in \mathcal{C} \mid \phi(c) = f\} = \cup_{f_i^* \in N(f)} \mathcal{O}_i^*$. Similarly, for any $h \in [m]$, we have $\{c \in \mathcal{C}_h \mid \phi(c) = f\} = \cup_{f_i^* \in N(f)} \mathcal{O}_i^*(h)$. Consequently, for any $f \in \mathcal{H}$ and $h \in [m]$, we get that $\frac{|\{c\in\mathcal{C}_h \mid \phi(c)=f\}|}{|\{c\in\mathcal{C} \mid \phi(c)=f\}|} = \frac{\sum_{f_i^*\in N(f)}|\mathcal{O}_i^*(h)|}{\sum_{f_i^*\in N(f)}|\mathcal{O}_i^*|}$. By using the scaling technique, we have that $\min_{f_i^*\in N(f)} \frac{|\mathcal{O}_i^*(h)|}{|\mathcal{O}_i^*|} \le \frac{\sum_{f_i^*\in N(f)}|\mathcal{O}_i^*(h)|}{\sum_{f_i^*\in N(f)}|\mathcal{O}_i^*|} \le \max_{f_i^*\in N(f)} \frac{|\mathcal{O}_i^*(h)|}{|\mathcal{O}_i^*|}$. Then, we get that $\beta_h \le \frac{\sum_{f_i^*\in N(f)}|\mathcal{O}_i^*(h)|}{\sum_{f_i^*\in N(f)}|\mathcal{O}_i^*|} \le \alpha_h$. Thus, $\phi$ satisfies the group fairness constraints. We now prove that the mapping $\psi$ satisfies the weighted group fairness constraints. By the above process, we get that the total weight of clients with color $h \in [m]$ in $\mathcal{C}^\ddagger$ assigned to a facility $f \in \mathcal{H}$ is exactly equal to the number of clients of this color assigned to $f$ in the solution $(\mathcal{H},\phi)$, i.e., for any $f \in \mathcal{H}$ and $h \in [m]$, we have $\sum_{c^\ddagger\in\mathcal{C}_h^\ddagger}\psi(c^\ddagger,f) = |\{c \in \mathcal{C}_h \mid \phi(c) = f\}|$. Then, for any $f \in \mathcal{H}$, we have $\sum_{c^\ddagger\in\mathcal{C}^\ddagger}\psi(c^\ddagger,f) = |\{c \in \mathcal{C} \mid \phi(c) = f\}|$. Thus, for any $f \in \mathcal{H}$ and $h \in [m]$, $\frac{\sum_{c^\ddagger\in\mathcal{C}_h^\ddagger}\psi(c^\ddagger,f)}{\sum_{c^\ddagger\in\mathcal{C}^\ddagger}\psi(c^\ddagger,f)} = \frac{|\{c\in\mathcal{C}_h \mid \phi(c)=f\}|}{|\{c\in\mathcal{C} \mid \phi(c)=f\}|}$ holds. Since the mapping $\phi$ satisfies the group fairness constraints, the mapping $\psi$ satisfies the weighted group fairness constraints. □

Lemma 9 implies that there must exist a $(4+\epsilon)$-approximate solution $\psi$ satisfying the weighted group fairness constraints. To obtain such a solution, the general idea is to reduce the assignment problem to a linear programming problem. The unknown optimal assignment can be naturally expressed in terms of linear inequalities, along with the condition that the assignment is fair. However, the issue is that in general the optimal fractional solution to this linear programming problem is not integral, and we must covert it to integral solution. Naturally, we start with the following linear programming. For any client $c_j \in \mathcal{C}^\ddagger$ and facility $f_i \in \mathcal{H}$, we introduce a variable $x_{ij}$ denoting how much weight from the client $c_j$ is assigned to the facility $f_i$. Note that the value of $x_{ij}$ represents the assignment $\psi$. The goal is to minimize the cost $\sum_{c_j\in\mathcal{C}^\ddagger}\sum_{f_i\in\mathcal{H}} d(c_j,f_i)x_{ij}$ satisfying the following conditions.

$$\sum_{f_i\in\mathcal{H}} x_{ij} = w^\dagger(c_j) \qquad \forall c_j \in \mathcal{C}^\ddagger, \tag{1}$$

$$\sum_{c_j\in\mathcal{C}_h^\ddagger} x_{ij} \le \alpha_h \sum_{c_j\in\mathcal{C}^\ddagger} x_{ij} \qquad \forall i \in [k], h \in [m], \tag{2}$$

$$\sum_{c_j\in\mathcal{C}_h^\ddagger} x_{ij} \ge \beta_h \sum_{c_j\in\mathcal{C}^\ddagger} x_{ij} \qquad \forall i \in [k], h \in [m], \tag{3}$$

$$x_{ij} \ge 0 \qquad \forall i \in [k], c_j \in \mathcal{C}^\ddagger. \tag{4}$$

Constraint (1) ensures that each client $c_j \in \mathcal{C}^\ddagger$ is assigned to facilities with weight $w^\dagger(c_j)$. Constraints (2) and (3) capture the group fairness constraints. The above linear programming can be solved in polynomial time, yielding a fractional assignment satisfying the group fairness. To obtain an integral assignment, by using a deterministic rounding method [Bercea *et al.*, 2019] that rounds the feasible fractional assignment obtained to an integral assignment, we have the following result. Note that we now only need to satisfy the group fairness constraints.

**Lemma 10.** *For the* DF-$k$-MED *problem, there is a deterministic rounding algorithm that returns an integral assignment $\psi$ with cost at most $(4+\epsilon)$opt with an additive 1 violation for group fairness constraints.*

### 4.3 The Analysis of Running Time

We now bound the running time of Algorithm 1. Indeed, we only need to discuss the iteration of steps 4-5, since all other things can be executed in polynomial time, including constructing coreset, optimizing improvement function, and solving the Weighted Assignment problem. Since the size of the weighted set is $|\mathcal{C}^\dagger| = O(\eta^{-2}k\log n)$, there are at most $|\mathcal{C}^\dagger|^k = (O(\eta^{-2}k\log n))^k$ different multi-sets of size $k$. Additionally, there are $(\log_{1+\delta}\frac{d_{\max}}{d_{\min}})^k = (O(\log n))^k$ choices for $\lambda_i$ for each $i \in [k]$, since the aspect ratio $\frac{d_{\max}}{d_{\min}}$ can be assumed to polynomially bounded in $n$. Therefore, the number of iterations in steps 4-5 of the algorithm can be bounded by $(O(\eta^{-2}k\log n))^k \cdot (O(\log n))^k \le (O(\eta^{-2}\log n))^k$. If $k < \log n/\log\log n$, we can obtain $(O(\eta^{-2}\log n))^k \le (O(\eta^{-2}))^k \cdot (\log n)^{\log n/\log\log n} = (O(\eta^{-2}))^k \cdot n$. Otherwise, $\log n \le O(k\log k)$, implying $(O(\eta^{-2}\log n))^k = (O(\eta^{-2}k\log k))^k$. Therefore, the running time of Algorithm 1 can be bounded by $(O(\eta^{-2}k\log k))^k \cdot n^{O(1)}$. By Lemma 10 and the above discussion, Theorem 1 can be proved.

## 5 Conclusions

In this paper, we study the doubly constrained fair clustering. Due to the doubly fairness constraints, it is thus a non-trivial task to obtain a solution that satisfies the fairness constraints and meanwhile achieves a small approximation ratio. The main contribution of this paper is a $(4+\epsilon)$-approximation for the DF-$k$-MED problem, in FPT$(k)$-time parameterized only by the number of opened facilities. Considering the considerable attention received by the theoretical aspect of the fair clustering problem, we believe that gaining these new parameterized results are of independent interest.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (62432016, 62172446) and Central South University Research Program of Advanced Interdisciplinary Studies (2023QYJC023).

## References

[Ahmadian *et al.*, 2019] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 267–275, 2019.

[Angelidakis *et al.*, 2022] Haris Angelidakis, Adam Kurpisz, Leon Sering, and Rico Zenklusen. Fair and fast $k$-center clustering for data summarization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 669–702, 2022.

[Arya *et al.*, 2004] Vijay Arya, Naveen Garg, Rohit Khandekar, Adam Meyerson, Kamesh Munagala, and Vinayaka Pandit. Local search heuristics for $k$-median and facility location problems. *SIAM Journal on Computing*, 33(3):544–562, 2004.

[Bandyapadhyay *et al.*, 2024] Sayan Bandyapadhyay, Fedor V Fomin, and Kirill Simonov. On coresets for fair clustering in metric and euclidean spaces and their applications. *Journal of Computer and System Sciences*, 142:103506, 2024.

[Bera *et al.*, 2019] Suman Kalyan Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 4955–4966, 2019.

[Bercea *et al.*, 2019] Ioana Oriana Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R. Schmidt, and Melanie Schmidt. 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*, pages 18:1–18:22, 2019.

[Byrka *et al.*, 2017] Jarosław Byrka, Thomas Pensyl, Bartosz Rybicki, Aravind Srinivasan, and Khoa Trinh. An improved approximation for $k$-median and positive correlation in budgeted optimization. *ACM Transactions on Algorithms*, 13(2):1–31, 2017.

[Charikar *et al.*, 2002] Moses Charikar, Sudipto Guha, Éva Tardos, and David B Shmoys. A constant-factor approximation algorithm for the $k$-median problem. *Journal of Computer and System Sciences*, 65(1):129–149, 2002.

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

[Chen *et al.*, 2019] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1032–1041, 2019.

[Chen, 2009] Ke Chen. On coresets for $k$-median and $k$-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[Chierichetti *et al.*, 2017] 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*, pages 5029–5037, 2017.

[Chiplunkar *et al.*, 2020] Ashish Chiplunkar, Sagar Kale, and Sivaramakrishnan Natarajan Ramamoorthy. How to solve fair $k$-center in massive data models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1877–1886, 2020.

[Cohen-Addad and Li, 2019] Vincent Cohen-Addad and Jason Li. On the fixed-parameter tractability of capacitated clustering. In *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming*, pages 41–1, 2019.

[Cohen-Addad *et al.*, 2019] Vincent Cohen-Addad, Anupam Gupta, Amit Kumar, Euiwoong Lee, and Jason Li. Tight FPT approximations for $k$-median and $k$-means. In *Proceedings of the 46th International Colloquium on Automata, Languages, and Programming*, pages 42–1, 2019.

[Dickerson *et al.*, 2023] John Dickerson, Seyed A Esmaeili, Jamie Morgenstern, and Claire Jie Zhang. Doubly constrained fair clustering. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 13267–13293, 2023.

[Esmaeili *et al.*, 2021] Seyed A Esmaeili, Brian Brubach, Aravind Srinivasan, and John P Dickerson. Fair clustering under a bounded cost. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 14345–14357, 2021.

[Feldman and Langberg, 2011] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 569–578, 2011.

[Feldman *et al.*, 2015] 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*, pages 259–268, 2015.

[Guha and Khuller, 1998] Sudipto Guha and Samir Khuller. Greedy strikes back: improved facility location algorithms. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 649–657, 1998.

[Har-Peled and Mazumdar, 2004] Sariel Har-Peled and Soham Mazumdar. On coresets for $k$-means and $k$-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 291–300, 2004.

[Harb and Lam, 2020] Elfarouk Harb and Ho Shan Lam. K-FC: A scalable approximation algorithm for $k$-center fair clustering. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 14509–14519, 2020.

[Huang *et al.*, 2023] Junyu Huang, Qilong Feng, Ziyun Huang, Jinhui Xu, and Jianxin Wang. Linear time algorithms for $k$-means with multi-swap local search. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 45651–45680, 2023.

[Huang *et al.*, 2024] Junyu Huang, Qilong Feng, Ziyun Huang, Jinhui Xu, and Jianxin Wang. Near-linear time approximation algorithms for $k$-means with outliers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 19723–19756, 2024.

[Jones *et al.*, 2020] Matthew Jones, Huy Lê Nguyên, and Thy Nguyen. Fair $k$-centers via maximum matching. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4940–4949, 2020.

[Kleindessner *et al.*, 2019] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair $k$-center clustering for data summarization. In *Proceeding of the 36th International Conference on Machine Learning*, pages 3448–3457, 2019.

[Krishnaswamy *et al.*, 2011] Ravishankar Krishnaswamy, Amit Kumar, Viswanath Nagarajan, Yogish Sabharwal, and Barna Saha. The matroid median problem. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1117–1130, 2011.

[Krishnaswamy *et al.*, 2018] Ravishankar Krishnaswamy, Shi Li, and Sai Sandeep. 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*, pages 646–659, 2018.

[Lee *et al.*, 2009] Jon Lee, Vahab S Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, pages 323–332, 2009.

[Mahabadi and Vakilian, 2020] Sepideh Mahabadi and Ali Vakilian. Individual fairness for $k$-clustering. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6586–6596, 2020.

[Micha and Shah, 2020] Evi Micha and Nisarg Shah. Proportionally fair clustering revisited. In *Proceedings of the 47th International Colloquium on Automata, Languages, and Programming*, pages 85:1–85:16, 2020.

[Negahbani and Chakrabarty, 2021] Maryam Negahbani and Deeparnab Chakrabarty. Better algorithms for individually fair $k$-clustering. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 13340–13351, 2021.

[Thejaswi *et al.*, 2022] Suhas Thejaswi, Ameet Gadekar, Bruno Ordozgoiti, and Michal Osadnik. Clustering with fair-center representation: Parameterized approximation algorithms and heuristics. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1749–1759, 2022.

[Zhang *et al.*, 2024a] Zhen Zhang, Xiaohong Chen, Limei Liu, Jie Chen, Junyu Huang, and Qilong Feng. Parameterized approximation schemes for fair-range clustering. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024.

[Zhang *et al.*, 2024b] Zhen Zhang, Junfeng Yang, Limei Liu, Xuesong Xu, Guozhen Rong, and Qilong Feng. Towards a theoretical understanding of why local search works for clustering with fair-center representation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, number 15, pages 16953–16960, 2024.