Approximation Algorithms for Fair Range Clustering

Sèdjro S. Hotegni [*1] **Sepideh Mahabadi** [*2] **Ali Vakilian** [*3]

Abstract

This paper studies the *fair range clustering* problem in which the data points are from different demographic groups and the goal is to pick k centers with the *minimum clustering cost* such that each group is at least *minimally represented* in the centers set and no group *dominates* the centers set. More precisely, given a set of n points in a metric space (P, d) where each point belongs to one of the ℓ different demographics (i.e., $P = P_1 \uplus P_2 \uplus \cdots \uplus P_\ell$) and a set of ℓ intervals $[\alpha_1, \beta_1], \cdots, [\alpha_\ell, \beta_\ell]$ on desired number of centers from each group, the goal is to pick a set of k centers C with minimum ℓ_p-clustering cost (i.e., $(\sum_{v \in P} d(v, C)^p)^{1/p}$) such that for each group $i \in \ell$, $|C \cap P_i| \in [\alpha_i, \beta_i]$. In particular, the fair range ℓ_p-clustering captures fair range k-center, k-median and k-means as its special cases. In this work, we provide an efficient constant factor approximation algorithm for the fair range ℓ_p-clustering for all values of $p \in [1, \infty)$.

1. Introduction

In recent years, the centroid-based clustering problem has been studied extensively from the fairness point of view. As in the human-centric applications, the input data usually comes from different demographics and the solution has supposedly some (in many cases, long-lasting) effects on the participants (e.g., college admissions, loan applications, and criminal justice), it is crucial to take into account the societal implications of the solution output by large scale automated processes in use. Specifically, since clustering is commonly used as a prepossessing step for more complicated ML pipelines, it is both easier and more effective to handle fairness consideration and bias reduction earlier in the pipeline rather than later.

Fair clustering was first studied by Chierichetti et al. (2017) and since then it has been studied with respect to various notions of fairness (Bera et al., 2019; Jung et al., 2020; Mahabadi & Vakilian, 2020; Ahmadi et al., 2022; Chen et al., 2019; Abbasi et al., 2021; Ghadiri et al., 2021; Brubach et al., 2020; 2021). Motivated by the application of centroid-based clustering as a means of data summarization (e.g., (Moens et al., 1999; Girdhar & Dudek, 2012)) for socioeconomic data, Kleindessner, Awasthi, and Morgenstern (2019) studied a notion of fair clustering in which the points belong to disjoint protected groups $P_1, \cdots, P_\ell$, and the goal is to pick exactly k_i centers from each population P_i, s.t. it minimizes the k-center clustering cost (i.e., maximum distance of any point to the selected centers), where $k = \sum_{i \in [\ell]} k_i$.

To give an example, consider an image search system. In practice, when a query is made, the user will only check the first few images output by the system. Those images (say, the first k ones) act as a summary or representative of the relevant images to the searched query. Notably, (Kay et al., 2015) observed that in a few jobs, including CEO, women are significantly underrepresented in Google image search results.[1] An approach to get around this disparity, as proposed by Kleindessner et al. (2019), is to force the solution to contain exactly k_i examples (or representative) from each protected group i, where k_is are input parameters. Although this *strict* center selection requirement seems to be a plausible fix for the unfairness issue, it may incur a significant loss in the quality of the output solution! See Figure 1.

In this paper, we study a relaxed requirement, called *fair range*, which only requires the number of selected centers from each protected group to be in a given interval specified by a lower bound and an upper bound. As we can easily tolerate slight deviations from the "expected" presence of each protected group, fair range is a more natural requirement and has a better alignment with practice, compared to the strict requirement. On the other hand, the larger the requirement interval becomes, the better the quality of the clustering becomes. While fair clustering with strict

[*]Equal contribution [1]African Institute for Mathematical Sciences–Rwanda [2]Microsoft Research–Redmond [3]Toyota Technological Institute at Chicago. Correspondence to: Sepideh Mahabadi <smahabadi@microsoft.com>, Ali Vakilian <vakilian@ttic.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]More recent studies also show that this phenomenon of unfairness in image search results is not fully resolved yet, e.g., (Feng & Shah, 2022).

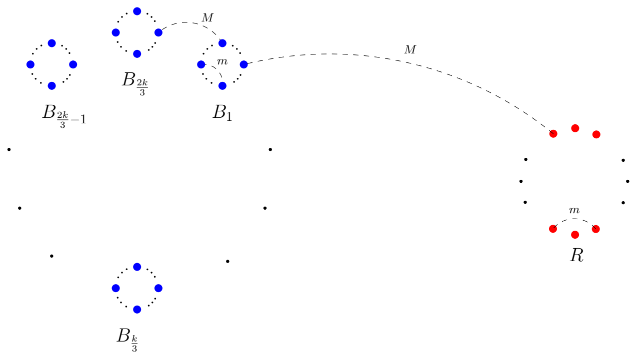

Figure 1. In this example, $\frac{n}{2}$ points with pairwise distance m belong to the red group on the right side of the figure. The other $\frac{n}{2}$ points belong to the blue group on the left side, partitioned into groups $B_1, \cdots, B_{\frac{2k}{3}}$ each of size $\frac{3n}{4k}$. For every $i \neq j \in [\frac{2k}{3}]$, the pairwise distance of $v, u \in B_i$ is m and the pairwise distance of $v \in B_i, u \in B_j$ is $M \gg m$. Moreover, the pairwise distance of any red and blue point is M. Note that the described distance function is metric. In the "strict" fair k-center, where $\frac{k}{2}$ centers should be picked from each of the red and blue group, the optimal solution has cost M. However, if we relax the requirement as in the fair range clustering and require $[\frac{k}{3}, \frac{2k}{3}]$ centers from each group, the k-center cost significantly reduces from M to m (the solution corresponds to picking one center from every B_i, for $i \in [\frac{2k}{3}]$ and arbitrary $\frac{k}{3}$ centers from R). While the latter solution is admissibly fair too, its quality is significantly better.

(i.e., exactly k_i from each group i) requirement—*which itself is a special case of clustering under partition matroid constraints*—is a well-studied problem in the domain of fair clustering (Hajiaghayi et al., 2010; Krishnaswamy et al., 2011; Charikar & Li, 2012; Swamy, 2016; Chen et al., 2016; Krishnaswamy et al., 2018; Jones et al., 2020; Chiplunkar et al., 2020), this natural generalization of the problem has not been studied in the context of fair clustering. We remark that, very recently, Nguyen et al. (2022) studied fair range k-center. They studied the problem both in the standard offline and the streaming models and provided constant factor approximations in both regimes. Moreover, Thejaswi et al. (2021) studied the problem when we are provided with lower bounds only and the objective is k-median cost. They showed fixed parameter algorithms for this problem which is referred to as *diversity-aware k-median*. However, the approximability of the general problem of fair range clustering with the ℓ_p-objective which includes the standard k-median and k-means still remains open.

Our contributions. In this paper, we study the fair range clustering with the ℓ_p-objective and provide a constant factor approximation algorithm for all values of $p \in [1, \infty)$. More precisely, our main result is as follows.

Theorem 1.1. *For all $p \in [1, \infty)$, there exists a constant factor approximation algorithm for fair range k-clustering with the ℓ_p-objective that runs in polynomial time.*

More generally, our algorithm works for fair range facility location with zero opening cost facilities which generalizes fair range clustering.

As fair range k-clustering with ℓ_p-objective has k-center as its special case (k-center is equivalent to fair range k-center where for every $i \in [\ell]$, $\alpha_i = 0$ and $\beta_i = k$), it is NP-hard to approximate it within a factor better than 2 (Gonzalez, 1985). So, our approximation factor for the problem is tight up to a constant factor.

To design our algorithm, we start with the framework of (Charikar et al., 2002; Krishnaswamy et al., 2011) to sparsify the input instance and find a feasible fractional solution of the standard LP-relaxation of the problem, called FAIRRANGELP, on the sparse instance. However, the main difficulty is within the rounding part. In our algorithm, we further exploit the combinatorial structures of an approximately optimal fractional solution of FAIRRANGELP. We consider another relaxation of the problem, STRUCTUREDLP, whose optimal solution is within an $e^{O(p)}$ factor of the optimal solution of FAIRRANGELP.[2] Then, using the techniques from Polyhedral Combinatorics, we show that STRUCTUREDLP has a half-integral optimal solution. Next, by a reduction to the network flow problem, we design a combinatorial algorithm that outputs an $e^{O(p)}$-approximation integral solution of STRUCTUREDLP, relying on the properties of a half-integral optimal solution of the LP. Finally, we can convert this feasible integral solution of STRUCTUREDLP to a feasible integral solution of FAIRRANGELP without losing more than a constant factor in the approximation ratio.

Refer to Section 2 for an overview of our algorithm and a summary of the first part of our algorithm which is standard and also used in some of prior works on (fair) clustering. Then, in Section 3, we describe our efficient rounding approach for fair range clustering.

Other related work. Clustering has been an active area of research in the domain of fairness for algorithms and machine learning. In particular, it has been studied under various settings including group fairness notions such as *fair representation* (Chierichetti et al., 2017; Bera et al., 2019; Bercea et al., 2019; Backurs et al., 2019; Ahmadian et al., 2019; Dai et al., 2022), *social fairness* (Abbasi et al., 2021; Ghadiri et al., 2021; Makarychev & Vakilian, 2021; Chlamtáč et al., 2022; Ghadiri et al., 2022), *proportional fairness* (Chen et al., 2019; Micha & Shah, 2020) and *in-*

[2]In both relaxations, the objective is ℓ_p-objective raised to power of p. So, with respect to the original objective these relaxation are within a constant factor of each other.

dividual fairness (Jung et al., 2020; Mahabadi & Vakilian, 2020; Chakrabarty & Negahbani, 2021; Vakilian & Yalçıner, 2022; Ahmadi et al., 2022; Brubach et al., 2020; 2021).

A similar notion has been studied for the related problem of nearest neighbor search (Har-Peled & Mahabadi, 2019; Aumüller et al., 2020; 2021) and submodular maximization (El Halabi et al., 2020).

Remark 1.2. While the collection of solutions satisfying the given range constraints does not constitute a matroid (since these constraints do not satisfy the downward-closed property), (El Halabi et al., 2020) showed that the family of all "extendable" subsets is indeed a matroid.[3] In our context, a set of facilities $\tilde{F}$ is extendable if it is a subset of facilities $\overline{F}$ where $\overline{F}$ has size at most k and satisfies all given range constraints. More precisely, $\tilde{F}$ is extendable if and only if for all $i \in [\ell]$, $|\tilde{F} \cap P_i| \le \beta_i$ and $\sum_{i\in[\ell]} \max\{|\tilde{F} \cap P_i|, \alpha_i\} \le k$. Then, one can use an existing algorithm for the k-clustering under matroid constraint (e.g., (Krishnaswamy et al., 2011; Swamy, 2016; Krishnaswamy et al., 2018)) and find a constant factor approximation for fair range clustering. While the above algorithm only considers the k-median objective, it is known that this approach can be generalized to any value of $p \in [1, \infty)$, see (Vakilian & Yalçıner, 2022).

While working with the extendable sets reduces our problem to matroid k-clustering, the known algorithm will require solving LPs with exponentially many constraints. Although these LPs can be solved in polynomial time using the ellipsoid algorithm, the best known algorithm for solving such LPs, which is via cutting plane methods, runs in time $O(n^3 \log(1/\epsilon))$ and returns a $(1+\epsilon)$-approximate solution (Jiang et al., 2020). However, in our approach, we work with small size LPs of the fair range clustering that can be solved in time $O(n^{1.5}k^{1.5})$ via the interior point method (Van Den Brand et al., 2021). So, arguably our algorithm is more efficient (in particular, for the case $k = O(1)$, we obtain a quadratic improvement in the run-time).

2. Description of Our Algorithm

For the sake of clarity, similarly to (Swamy, 2016) on clustering under matroid constraints, we describe our algorithm for a more general problem of fair range facility location with the ℓ_p-objective in which the number of open facilities are required to be (at most) k. For our application of clustering, it suffices to consider the instances of fair range facility locations in which the set of *facilities* F is disjoint from the set of *clients*, where the clients are specified by a pair of *set of locations* D and a *demand function* $w : D \to \mathbb{Z}_{>0}$.Note that while F and D are disjoint, they may have points at the same location.

More specifically, let (P, d) be a metric space. In this problem, we are given a set of facilities $F := F_1 \uplus \cdots \uplus F_\ell$ where $F \subseteq P$, a set of integral range parameters $\{\alpha_i, \beta_i\}_{i\in[\ell]}$ where $\alpha_i, \beta_i \in \mathbb{Z}_{\ge 0}$, a set of locations $D \subseteq P$, and a demand function $w\colon D \to \mathbb{Z}_{>0}$ denoting the *total* number of clients in each location of D.[4] The goal is to open a subset $C \subseteq F$ of k facilities with minimum ℓ_p-cost assignment of clients to their closest facilities (i.e., $(\sum_{v\in D} w(v) \cdot d(v, C)^p)^{1/p}$), such that for every group $i \in [\ell]$, $|C \cap F_i| \in [\alpha_i, \beta_i]$.

Remark 2.1. In the description of our algorithm and in particular, in the objective of the LP-relaxations, we consider the ℓ_p-clustering raised to the power of p. Throughout the paper, we provide all approximation factors according to $\sum_{v\in D} w(v) \cdot d(v, C)^p$ denoted as cost. Then, at the end, to derive the result of the main theorem, we raise the approximation factor to $1/p$.

Our algorithm has two major components: (1) finding an approximate fractional solution (x, y) for an LP-relaxation of the problem with certain *well-separatedness* and *locality* properties, (2) rounding the fractional solution to an integral solution without losing more than a factor of $e^{O(p)}$ in the LP objective. The first part is essentially a standard approach in approximation algorithm for clustering, introduced in (Charikar et al., 2002), and is similar to the one used in (Krishnaswamy et al., 2011; Swamy, 2016). So, here we only describe the properties of the fractional solution at the end of the first part and for a detailed exposition, we refer to Appendix B.

The main technical contribution of our paper is an efficient rounding algorithm that preserves the clustering with ℓ_p-objective (for all $p \in [1, \infty)$) up to a constant factor, does not violate any of fairness constraints, and opens at most k centers.

2.1. Reducing the number of locations

Given a set of points P, we first run an existing efficient constant factor approximation algorithm for k-clustering with ℓ_p objective (ignoring all range constraints) to get a set of centers $C = (c_1, \cdots, c_k)$. Then, we construct the following instance of fair range clustering. We separate the set of clients (D) and facilities (F) as follows: Each point in P is moved to its closest center in C and the resulting set of points located at k locations $c_1, \cdots, c_k$ constitutes the set of clients. In other words, the set of clients can be described in terms of k locations plus the total number of clients in each location determined by $w' : D \to \mathbb{Z}_{>0}$. For

[3] The work of (El Halabi et al., 2020) used this idea to study the fairness of an objective different from clustering, namely submodular maximization.

[4] The integrality of α_i, β_i are without loss of generality as we can always replace them with $\lceil \alpha_i \rceil$ and $\lfloor \beta_i \rfloor$ and the solution space remains the same.

the facilities, we set $F = P$. Then, we solve the minimum cost fair-range clustering (or facility location w.r.t. D and F) that picks k centers (or facilities) from F and serves all clients.

Theorem 2.2. *Given an $O(\alpha)$-approximation algorithm of k-clustering with ℓ_p-objective, a β-approximate solution S of fair range clustering with ℓ_p-objective on the described instance (D, F) is an $O(\alpha\beta)$-approximation for fair range clustering with ℓ_p-objective on the original instance P.*

The proof of the theorem is in Appendix C.

In the rest of this paper, at the expense of losing a factor of $O(\alpha)$ in the approximation factor and running an α-approximation algorithm of k-clustering with ℓ_p-objective (with no range constraints), we assume that $|D| = k$ and $|F| = n$.

2.2. Constructing a structured fractional solution

In this section, we describe the sparsification approach of (Charikar et al., 2002) which outputs an instance with a subset of locations where pairwise distances of survived locations are "relatively large". First, we state a standard LP-relaxation of the problem as follows.

$$\textsc{FairRangeLP}(D, F, w, \{\alpha_i, \beta_i\}_{i\in[\ell]})$$

$$\text{minimize} \quad \textstyle\sum_{v\in D, u\in F} w(v) \cdot d(v,u)^p \cdot x_{vu}$$

$$\text{s.t.} \quad \sum_{u\in F} x_{vu} \ge 1 \quad \forall v \in D \tag{1}$$

$$\alpha_i \le \sum_{u\in F_i} y_u \le \beta_i \quad \forall i \in [\ell] \tag{2}$$

$$\sum_{u\in F} y_u \le k \tag{3}$$

$$0 \le x_{vu} \le y_u \quad \forall v \in D, u \in F \tag{4}$$

In our algorithm, we never modify the set F and the fairness constraints $\{\alpha_i, \beta_i\}_{i\in[\ell]}$. So, to specify an instance, we will provide the set of clients (D, w). Consider an optimal fractional solution (x^*, y^*) of $\textsc{FairRangeLP}(D, w)$. For any location $v \in D$, define $\mathcal{R}(v) := \left(\sum_{u\in F} x^*_{vu} \cdot d(v,u)^p\right)^{1/p}$ as the *fractional distance* of a unit of demand at location v w.r.t. the optimal solution (x^*, y^*). When y^* is integral, then $\mathcal{R}(v)$ is the distance of v to its closest open facility, specified by the vector y^*.

The sparsification approach of (Charikar et al., 2002) applied to our setting, described in Appendix B, outputs a sparse instance with the following properties. The proofs of theorems in this sections are deferred to Appendix C.

Theorem 2.3. *Given an instance (D, w) of fair range clustering with ℓ_p-cost and an optimal fractional solution (x, y) of $\textsc{FairRangeLP}(D, w)$ with cost OPT_D, there exists a polynomial time algorithm that returns a set of locations $D' \subseteq D$ and a demand function $w' : D' \to \mathbb{R}$ such that*

(Q1) For every pair of v_i, v_j in D', $d(v_i, v_j) \ge 2^{1+1/p} \max\{\mathcal{R}(v_i), \mathcal{R}(v_j)\}$.

(Q2) (x, y) is a feasible solution of $\textsc{FairRangeLP}(D', w')$ of cost at most OPT_D,

(Q3) Any integral solution C of $\textsc{FairRangeLP}(D', w')$ of cost z, can be converted in polynomial time to a feasible solution of $\textsc{FairRangeLP}(D, w)$ of cost at most $4^p \cdot \mathrm{OPT}_D + 2^{p-1} \cdot z$.

Next, for every location $v \in D'$, we define the *ball* $\mathcal{B}(v) := \{u \in F | d(v,u) \le 2^{\frac{1}{p}} \cdot \mathcal{R}(v)\}$ to denote the set of facilities at distance at most $2^{\frac{1}{p}} \cdot \mathcal{R}(v)$ from v. Further, for every location $v \in D'$, we define $\mathcal{P}(v)$ as the super ball of v which consists of $\mathcal{B}(v)$ and a set of "private facilities" of v.

Observation 2.4. For any pair of $v, v' \in D'$ and $u' \in \mathcal{R}(v')$, $\frac{1}{2} d(v,v') \le d(v,u') \le \frac{3}{2} d(v,v')$.

Proof. Since d is a metric distance, by the triangle inequality, $d(v,u') + d(u',v') \ge d(v,v')$. Next, we bound $d(u',v')$ in terms of $d(v,v')$. Since $u' \in \mathcal{B}(v')$, $d(v',u') \le 2^{\frac{1}{p}} \mathcal{R}(v')$. On the other hand, by (Q1), $d(v,v') \ge 2^{1+\frac{1}{p}} \mathcal{R}(v')$. Hence, $d(v',u') \le \frac{1}{2} d(v,v')$. By another application of the triangle inequality, $d(v,u') \le d(v,v') + d(v',u') \le \frac{3}{2} d(v,v')$. ☐

Next, in polynomial time, we convert a fractional solution of $\textsc{FairRangeLP}(D, w)$ to a fractional solution (x, y) of $\textsc{FairRangeLP}(D', w')$ with the following further structural properties.

Theorem 2.5. *There exists a polynomial time algorithm that outputs a fractional solution (x, y) of $\textsc{FairRangeLP}(D', w')$ of cost $9^p \cdot \mathrm{OPT}_D$, where OPT_D is the cost of an optimal solution of $\textsc{FairRangeLP}(D, w)$, and a collection of super balls $\{\mathcal{P}(v)\}_{v\in D'}$ that satisfy the following properties:*

(P1) For every $v \in D'$, $\mathcal{B}(v) \subseteq \mathcal{P}(v)$,

(P2) For every $v \in D'$ and $u \in \mathcal{P}(v) \setminus \mathcal{B}(v)$, $x_{vu} > 0$ only if $\sum_{u\in\mathcal{B}(v)} y_u < 1$. Similarly, for every $v \in D'$ and $u \in F \setminus \mathcal{P}(v)$, $x_{vu} > 0$ only if $\sum_{u\in\mathcal{P}(v)} y_v < 1$.

(P3) For every $v \in D'$, if $x_{vu} > 0$, then either $u \in \mathcal{P}(v)$ or $u \in \mathcal{B}(v')$ where $v' = \mathsf{NN}_{D'}(v)$ denotes the nearest location in D' (other than v itself) to v,

(P4) For every $v \in D'$,

$$\sum_{u\in\mathcal{P}(v)} x_{vu} \ge \sum_{u\in\mathcal{B}(v)} x_{vu} \ge 1/2.$$

(P5) For every $v \in D'$, $u \in \mathcal{P}(v) \setminus \mathcal{B}(v)$, $d(v,u) \leq 2 \cdot d(v,v')$, where $v' = \mathsf{NN}_{D'}(v)$.

(P2) The set of super balls, $\{\mathcal{P}(v)\}_{v \in D'}$, are disjoint.

3. Rounding Algorithm

To do the rounding, first we write a new LP relaxation, called StructuredLP which is a simplification of FairRangeLP via the further structures of fractional solution (x,y) guaranteed by Theorem 2.5. In particular, StructuredLP is useful for our rounding algorithm because as we show in this section, the polyhedraon constructed by the constraints of StructuredLP is half-integral. A solution y to an LP relaxation is half-integral if every coordinates in y has value either $0, 1/2$ or 1.

In StructuredLP, we define $\Delta(v) := d(v,v')^p + \sum_{u \in \mathcal{P}(v)} \big(d(v,u)^p - d(v,v')^p\big) \cdot y_u$ to denote the minimum (fractional) distance of v to open facilities (i.e., the facilities vector y).

$$\textsc{StructuredLP}(D', w')$$

$$\begin{aligned}
\text{minimize } & \textstyle\sum_{v \in D'} w'(v) \cdot \Delta(v) & & \\
\text{s.t.} \quad & \alpha_i \leq \sum_{u \in F_i} y_u \leq \beta_i & \forall i \in [\ell] \quad & (5)\\
& \sum_{u \in F} y_u \leq k & & (6)\\
& \sum_{u \in \mathcal{B}(v)} y_u \geq 1/2 & \forall v \in D' \quad & (7)\\
& \sum_{u \in \mathcal{P}(v)} y_u \leq 1 & \forall v \in D' \quad & (8)\\
& y_u \geq 0 & \forall u \in F \quad & (9)
\end{aligned}$$

Lemma 3.1. *The optimal fractional solution of StructuredLP(D', w') is a valid solution for fair range clustering on (D', w') and has cost at most $e^{O(p)} \cdot \mathrm{OPT}_D$.*

Proof. By the cardinality constraint on the number of open facilities and the range constraints on the number of facilities from each group, a feasible solution of StructuredLP is a feasible solution of fair range clustering on (D', w').

Consider the solution (x,y) guaranteed by Theorem 2.5. It is straightforward to verify that y satisfies all constraints of StructuredLP(D', w') and is a feasible solution for the LP: (y) satisfies the first two sets of constraints in StructuredLP(D', w') because (x,y) is a feasible solution of FairRangeLP(D', w') and constraint 7 follows from (P4) in Theorem 2.5. Note that if y does not satisfy constraint (8), we can decrease the value of y accordingly so that (x,y) remains a feasible solution of FairRangeLP(D', w') with a lower cost.

Next, we bound the cost of y with respect to StructuredLP, $\mathrm{cost}_{\mathrm{STLP}}(y)$. We rewrite $\Delta(v)$ as follows:

$$\Delta(v) = \big(1 - \sum_{u \in \mathcal{P}(v)} y_u\big) d(v,v')^p + \sum_{u \in \mathcal{P}(v)} d(v,u)^p \cdot y_u$$

For every $v \in D'$,

$$\begin{aligned}
&\sum_{u \in F} d(v,u)^p x_{vu}\\
&= \sum_{u \in \mathcal{P}(v)} d(v,u)^p x_{vu} + \sum_{u' \in \mathcal{B}(v')} d(v,u')^p x_{vu}\\
&\geq \sum_{u \in \mathcal{P}(v)} d(v,u)^p x_{uv} + \frac{1}{2^p} \cdot d(v,v')^p \sum_{u \in \mathcal{B}(v')} x_{vu}\\
&= \sum_{u \in \mathcal{P}(v)} d(v,u)^p y_u + \frac{1}{2^p} \cdot d(v,v')^p (1 - \sum_{u \in \mathcal{P}(v)} y_u)\\
&\geq \frac{1}{2^p}\Big(\big(1 - \sum_{u \in \mathcal{P}(v)} y_u\big) d(v,v')^p + \sum_{u \in \mathcal{P}(v)} d(v,u)^p \cdot y_u\Big),
\end{aligned} \tag{10}$$

where the first equality follows from (P3), the first inequality is by Observation 2.4 and the second equality follows from (P2). Thus,

$$\begin{aligned}
\mathrm{cost}_{\mathrm{STLP}}(y) &= \sum_{v \in D'} w'(v) \Delta(v)\\
&\leq 2^p \sum_{v \in D'} w'(v) \sum_{u \in F} d(v,u)^p x_{vu}\\
&= 2^p \cdot \mathrm{cost}_{\mathrm{FLP}}(x,y)\\
&= e^{O(p)} \cdot \mathrm{OPT}_D
\end{aligned}$$

where OPT_D denotes the cost of an optimal solution of FairRangeLP(D, w). The last inequality bounding $\mathrm{cost}_{\mathrm{FLP}}(x,y)$ is by Theorem 2.5. ☐

Lemma 3.2. *Consider a half-integral solution $\tilde{y}$ of StructuredLP(D', w') of cost z. Then, $\tilde{y}$ is a feasible solution for FairRangeLP(D', w') with cost at most $\left(\frac{3}{2}\right)^p \cdot z$.*

Proof. Let $\tilde{x}$ be the assignment of clients to the opened facilities as follows: For every $v \in D'$,

- **Step 1.** For every $u \in \mathcal{P}(v)$, set $\tilde{x}_{vu} = \tilde{y}_u$. Let $\tilde{Y}(v) = \sum_{u \in \mathcal{P}(v)} y_u$ denote the total assignment of v at the end of this step.
- **Step 2.** While $\tilde{Y}(v) < 1$, iterate over $u \in \mathcal{B}(v')$ and at each iteration set $\tilde{x}_{vu} = \min\{1 - \tilde{Y}(v), \tilde{y}_u\}$ and $\tilde{Y}(v) = \tilde{Y}(v) + \tilde{x}_{vu}$.

Since for every $v \in D'$, $\frac{1}{2} \leq \sum_{u \in \mathcal{B}(v)} \tilde{y}_u \leq 1$, this procedure terminates with a feasible fractional assignment of clients to facilities, $\tilde{x}$: for every $v \in D', u \in F$, $\tilde{x}_{vu} \leq \tilde{y}_u$ and $\sum_{u \in F} \tilde{x}_{vu} = 1$. Moreover, if $\tilde{y}$ is half-integral, clearly $\tilde{x}$ is half-integral too.

For every $v \in D'$,

$$\begin{aligned}
&\sum_{u\in F} d(v,u)^p \cdot \tilde{x}_{vu} \\
&= \sum_{u\in\mathcal{P}(v)} d(v,u)^p \cdot \tilde{x}_{vu} + \sum_{u'\in\mathcal{B}(v')} d(v,u')^p \cdot \tilde{x}_{vu'} \\
&\leq \sum_{u\in\mathcal{P}(v)} d(v,u)^p \cdot \tilde{x}_{vu} + \big(\frac{3}{2}\big)^p \sum_{u'\in\mathcal{B}(v')} \tilde{x}_{vu'} \quad \triangleright \text{ by Obs. 2.4} \\
&= \sum_{u\in\mathcal{P}(v)} d(v,u)^p \cdot \tilde{y}_u + \big(\frac{3}{2}\big)^p \big(1 - \sum_{u\in\mathcal{P}(v)} \tilde{y}_u\big) \cdot d(v,v')^p \\
&\leq \big(\frac{3}{2}\big)^p \Big(\sum_{u\in\mathcal{P}(v)} d(v,u)^p \cdot \tilde{y}_u + \big(1 - \sum_{u\in\mathcal{P}(v)} \tilde{y}_u\big) \cdot d(v,v')^p \Big) \\
&= \big(\frac{3}{2}\big)^p \cdot \Delta(v)
\end{aligned} \tag{11}$$

Thus,

$$\begin{aligned}
\mathrm{cost}_{\mathrm{FLP}}(\tilde{x},\tilde{y}) &= \sum_{v\in D'} w'(v) \sum_{u\in F} d(v,u)^p \tilde{x}_{vu} \\
&\leq \sum_{v\in D'} w'(v) \cdot \big(\frac{3}{2}\big)^p \cdot \Delta(v) \quad \triangleright \text{ by Eq. (11)} \\
&\leq \big(\frac{3}{2}\big)^p \cdot \mathrm{cost}_{\mathrm{STLP}}(\tilde{y})
\end{aligned}$$

□

3.1. Constructing a half-integral solution of StructuredLP

Next, we show that the constraints of StructuredLP forms a totally unimodular (TU) matrix. Then, by scaling the constraint properly, we can find a half-intergal optimal solution of StructuredLP in polynomial time. Recall that a matrix is TU if every square submatrix has determinant $-1, 0$ or 1. Totally unimodular matrices are extremely important in polyhedral combinatorics and combinatorial optimization since if A is TU and b is integral, then the linear programs of the form $\{\min cx \mid Ax \geq b, x \geq 0\}$ has integral optima, for any cost vector c.

Lemma 3.3. *The matrix corresponding to the constraints of* StructuredLP *is a TU matrix.*

Proof. We consider StructuredLP in the form $\{\min cy \mid Ay \geq b, y \geq 0\}$. Note that values of all entries in A belong to $\{0, \pm 1\}$. Moreover, in each column A_i, corresponding to each variable y_i, there are at most five non-zero entries: one with value -1 corresponding to "picking at most k centers" constraint, one with value -1 to upper bound the contribution of the facilities of the color class to which facility i belongs, one with value 1 to lower bound the contribution of the facilities of the color class to which facility i belongs, one with value 1 to open at least $1/2$ facilities from the super ball containing facility i and finally one with value -1 to open at most 1 from the super ball containing facility i.

Now, we show that A is a TU matrix by the result of Ghouila-Houri (1962); see Theorem A.4. They showed that if for every subset R of the rows there is an assignment $s_R : R \to \{-1, 1\}$ of signs to rows so that $\sum_{r\in R} s_R(r) A_r$ has all entries in $\{0, \pm 1\}$, then A is a TU matrix. To show that, for any subset of rows R we write $R = R_{\text{center}} \cup R_{\text{upper}} \cup R_{\text{lower}} \cup R_{\text{superball}}$. Next, we consider the following two cases:

- $R_{\text{center}} \neq \emptyset$. In this case, we assign -1 to the row corresponding to the total facility budget constraint, (6). Next, we consider the color classes C_{both} for which both rows corresponding to the upper and lower bounds constraints on the number of centers from the color class exist in R, s_R assigns 1 to both such rows. Note that as the rows corresponding to the lower and upper bound constraints of a specific color class sum up to zero vector, such assignment makes the contribution of such rows in the sum of rows of A in R zero. Otherwise, as color classes are disjoint, if exactly one of the rows corresponding to the upper and lower bound constraints belong to R, we can set s_R so that the contribution of entries in such rows become exactly 1. Hence, so far (by summing up terms $\sum_{r\in R_{\text{center}}} s_R(r) A_r$, $\sum_{r\in R_{\text{upper}}} s_R(r) A_r$ and $\sum_{r\in R_{\text{lower}}} s_R(r) A_r$), each entry has value either 0 or -1. Next, we consider the set of rows in $R_{\text{superball}}$. If both rows corresponding to the upper and lower bounds constraints on the contributions of facilities in a super ball are in R, then s_R assigns 1 to both rows and their total contributions in the final sum will become zero. For the remaining rows in $R_{\text{superball}}$, we set the sign so that the contribution of each row to the final sum becomes exactly 1 (i.e., if the row corresponds to constraint (7), s_R assigns 1, and assigns -1 otherwise). Since the super balls are disjoint, in the weighted sum of rows with s_R, each entry is either 0 or ± 1.
- $R_{\text{center}} = \emptyset$. Our approach is similar to the previous case. However, here first we set the assignment of rows corresponding to $R_{\text{lower}}, R_{\text{upper}}$ and makes sure that by summing up $\sum_{r\in R_{\text{upper}}} c(r) A_r, \sum_{r\in R_{\text{lower}}} c(r) A_r$), each entry has value either 0 or 1. Then, we follow a similar approach to the previous case and exploiting the fact super balls are disjoint, we show that it is possible to set the sign of rows in $R_{\text{superball}}$ so that in the overall signed sum of the rows, each entry is either 0 or ± 1.

□

Theorem 3.4. STRUCTUREDLP *has an optimal half-integral solution.*

Proof. To see this, let vector $\hat{y} = y/2$ and then note that by Lemma 3.3, the matrix corresponding to the constraints of STRUCTUREDLP is a TU matrix. Hence, the polytope specified by $A\hat{y} \leq 2b$ is integral which implies that STRUCTUREDLP admits a half integral optimal solution. □

We remark that such half-integral optimal solution can be found in via an efficient polynomial time algorithm.

3.2. Constructing an Integral Solution of STRUCTUREDLP

In this section, exploring the structure imposed by a half-integral optimal solution of STRUCTUREDLP, we construct an integral $e^{O(p)}$-approximate solution of STRUCTUREDLP.

Step 1. Partitioning facilities. First, we show that we can partition facilities into $L \leq k$ disjoint sets $S_1, \cdots, S_L$ such that any solution that opens at least one facility from each S_i for all $i \in L$ is an $e^{O(p)}$-approximate solution of STRUCTUREDLP. Here is the description of the partitioning procedure given a half-integral optimal solution of STRUCTUREDLP, y:

Algorithm 1 Partitioning Facilities.

```
Input: A set of locations D', half-integral vector y
for all location v_i ∈ D' do
    R_i ← the minimum assignment cost of a unit of demand at v_i w.r.t. y: i.e., R_i = 1/2 (d(v_i, u_{i_1})^p + d(v_i, u_{i_1})^p) where u_{i_1}, u_{i_2} are respectively the primary and secondary facilities serving v_i
    S_i ← {u_{i_1}} ∪ {u_{i_2}}
end for
D'' ← D', D̄ ← ∅
while D'' is nonempty do
    let v_i ← argmin_{v_j ∈ D''} R_j
    add v_i to D̄
    remove all locations v_j ∈ D'' such that S_j ∩ S_i ≠ ∅
end while
```

Next, we show the following useful property of the $\{S_i\}_{\{i \mid v_i \in \overline{D}\}}$.

Lemma 3.5. *The clustering cost of any set of k facilities C that opens at least one facility from each $\{S_i\}_{\{i \mid v_i \in \overline{D}\}}$ is at most $\left(\frac{9}{2}\right)^p$ times the cost of an optimal solution of* STRUCTUREDLP(D', w').

Proof. We compute the cost of a unit of demand in each location of $v_i \in D'$ when it is assigned to its closest facility in $\mathcal{S} := \bigcup_{i: v_i \in \overline{D}} S_i$. In particular, we compare this cost to the fractional cost imposed by the half-integral solution y of STRUCTUREDLP. We consider the following two cases:

- $v_i \in \overline{D}$**.** It is straightforward to check that in this case the assignment cost of v_i to a facility in $\mathcal{S}$ is at most twice the its fractional assignment cost with respect to y.
- $v_i \notin \overline{D}$ Let $v_j \in C$ be the client who removed v_i from D'', i.e., v_j is the minimum cost location whose opened facilities, S_j, has non-empty intersection with S_i. Let $u = S_j \cap C$. Then, we bound $d(v_i, u)^p$ in terms of R_i and R_j. Note that in this case $S_i \cap S_j \neq \emptyset$, however, their intersection might be a facility u' different from u. So, by the approximate triangle inequality (see Corollary A.2)

$$\begin{aligned} d(v_i, u)^p &\leq 3^{p-1}(d(v_i, u')^p + d(u', v_j)^p + d(v_j, u)^p) \\ &\leq 3^{p-1}(R_i + 2R_j) \\ &\leq 3^p \cdot R_i \end{aligned}$$

Hence, the total cost of such clustering is at most

$$\begin{aligned} \sum_{v_i \in D'} w'(v) \cdot d(v_i, C)^p &\leq 3^p \cdot \sum_{v_i \in D'} R_i \\ &\leq \left(\frac{9}{2}\right)^p \cdot \mathrm{cost}_{\mathrm{STLP}}(y), \end{aligned}$$

where the last inequality follows from Lemma 3.2. □

Step 2. Constructing an integral solution. Finally, we show that we can always find an integral solution of STRUCTUREDLP that picks at least one center from every set $S_1, \cdots, S_L$. Note that by showing the existence of such a solution, automatically (via Lemma 3.5) we have the guarantee that it is a 3^p-approximation solution of STRUCTUREDLP. We show the existence of such an integral solution via an application of max-flow problem.

Lemma 3.6. *Given a collection of disjoint sets $S_1, \cdots, S_L$ where $L \leq k$, there exists an integral solution that picks a set of k centers C with the following extra properties:*

- *For every $j \in [L]$, $C \cap S_j \neq \emptyset$, and*
- *For every group $i \in [\ell]$, $\alpha_i \leq |C \cap P_i| \leq \beta_i$.*

Proof. To show the existence of such a solution, we construct the following instance of network-flow. As in Figure 2, we create a network with 6 layers. Layer 0 consists of a single source vertex s and layer 1 consists of $L+1$ vertices corresponding to sets $S_1, \cdots, S_l$ and a dummy set $\overline{S}$. Moreover, for every $i \in [L]$, the source vertex s is connected to S_i with an edge of capacity 1. There is also an edge from s to $\overline{S}$ with capacity $k - L$. In layer 2, there are $|F|$ vertices corresponding to each facility in F. Moreover, for every

$i \in [L]$, S_i is connected to the facilities u_{i_1} and u_{i_2} where $S_i = \{u_{i_1}, u_{i_2}\}$ with capacity 1. Moreover, $\overline{S}$ has an edge to every facility with capacity 1. In layer 3, we have a vertex g_i for every group $i \in [\ell]$. Then, the edges between layer 2 and layer 3 specify the membership of the facilities to the groups. Note that as groups are disjoint, each facility is connected to exactly one group. Finally, the edges between the vertices in layer 4, representing the groups, and the vertex t_1 in layer 5 assures that the number of facilities selected from each group i is in the given integral $[\alpha_i, \beta_i]$: there is an edge with respectively lower and upper capacities of α_i, β_i connecting g_i to t_1. Finally, there is an edge from t_1 to t_2 with capacity k to bound the total number of selected centers.

It is straightforward to check that if there exists an integral flow of value k from s to t_2, then the set of facilities whose corresponding vertices in layer 2 receive a unit of flow is a set of centers that satisfies all requirements of fair range clustering. So, it remains to show that this network has a flow of value k. Since all capacities are integral, by the integrality theorem of max-flow, this implies that an integral flow of value k exists too. Hence, it suffices to show that the network has a (possibly fractional) flow of value k. Consider the half-integral optimal solution of StructuredLP, y, from which we constructed the sets $S_1, \cdots, S_L$. By the definition of $S_1, \cdots, S_L$, we can send one unit of flow from s to each S_i (corresponding to $y_{u_{i_1}}$ and $\min\{y_{u_{i_2}}, 1/2\}$) then $1/2$ flow is sent from S_i to each of u_{i_1} and u_{i_2} (which can be the same vertex too). Note that S_i are all disjoint and no facility receives more than one unit of flow at the end of this step. Then, we send $k - L$ flow from s to $\overline{S}$ and from $\overline{S}$ to facilities so that the vertex corresponding to each facility u receives exactly y_u units of flow overall. Next, each facility vertex u send y_u units of flow to the color class i that contains u. Note that by feasibility of y for StructuredLP, no edge capacity will be violated. Lastly, as $\sum_{u \in F} y_u = k$, the amount of flow from t_1 to t_2 constructed from solution y is exactly k. $\square$

Now, we are ready to prove the main theorem of this paper.

Proof of Theorem 1.1. Given an instance of the problem $\{(D, w), F\}$ with the specified set of fairness constraints, first we compute the sparse instance (D', w') guaranteed in Theorem 2.5. Next, we find a half-integral optimal solution y of StructuredLP(D', w') (Theorem 3.4). Then, by Lemma 3.5 and 3.6, we can find a set of k centers C, such that the clustering (D', w') using the set of centers C has cost at most $\left(\frac{9}{2}\right)^p \cdot \mathrm{cost}_{\mathrm{STLP}}(y)$. Next, by the optimality of y for StructuredLP and by Lemma 3.1, we show that the clustering cost of the instance with the set of centers C is at most $\left(\frac{9}{2}\right)^p \cdot \mathrm{cost}_{\mathrm{STLP}}(y) = e^{O(p)} \cdot \mathrm{OPT}_D$ where OPT_D is the cost of an optimal solution of FairRangeLP(D, w).

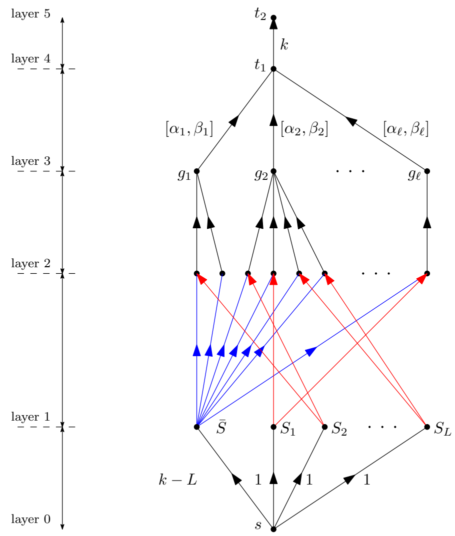

Figure 2. An example of network flow instance corresponding to a fair range clustering instance.

Finally, by Theorem 2.3 and raising the approximation factor to the power of $1/p$, the clustering using the center set C is an $O(1)$-approximate solution on instance (D, w) with the given fairness constraints $\{\alpha_i, \beta_i\}_{i \in [\ell]}$. $\square$

Conclusion

In this paper, we study the fair range clustering problem which is a generalization of several well-studied problems including fair k-center (Kleindessner et al., 2019) and clustering under partition matroid. We designed efficient constant-factor approximation algorithms. Our result is the first pure multiplicative approximation algorithm for fair range clustering with general ℓ_p-objective.

References

Abbasi, M., Bhaskara, A., and Venkatasubramanian, S. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 504–514, 2021.

Ahmadi, S., Awasthi, P., Khuller, S., Kleindessner, M., Morgenstern, J., Sukprasert, P., and Vakilian, A. Individual preference stability for clustering. In *International Conference on Machine Learning (ICML)*, 2022.

Ahmadian, S., Epasto, A., Kumar, R., and Mahdian, M. Clustering without over-representation. In *Proceedings of the SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 267–275, 2019.

Aumüller, M., Pagh, R., and Silvestri, F. Fair near neighbor search: Independent range sampling in high dimensions.

In *Proceedings of the SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 191–204, 2020.

Aumüller, M., Har-Peled, S., Mahabadi, S., Pagh, R., and Silvestri, F. Sampling a near neighbor in high dimensions–who is the fairest of them all? *arXiv preprint arXiv:2101.10905*, 2021.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. Scalable fair clustering. In *Proceedings of the International Conference on Machine Learning*, pp. 405–413, 2019.

Bera, S., Chakrabarty, D., Flores, N., and Negahbani, M. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems*, pp. 4955–4966, 2019.

Bercea, I. O., Groß, M., Khuller, S., Kumar, A., Rösner, C., Schmidt, D. R., and Schmidt, M. On the cost of essentially fair clusterings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, 2019.

Brubach, B., Chakrabarti, D., Dickerson, J., Khuller, S., Srinivasan, A., and Tsepenekas, L. A pairwise fair and community-preserving approach to k-center clustering. In *International Conference on Machine Learning*, pp. 1178–1189, 2020.

Brubach, B., Chakrabarti, D., Dickerson, J., Srinivasan, A., and Tsepenekas, L. Fairness, semi-supervised learning, and more: A general framework for clustering with stochastic pairwise constraints. In *Proc. Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Chakrabarty, D. and Negahbani, M. Better algorithms for individually fair k-clustering. *arXiv preprint arXiv:2106.12150*, 2021.

Charikar, M. and Li, S. A dependent lp-rounding approach for the k-median problem. In *International Colloquium on Automata, Languages, and Programming*, pp. 194–205, 2012.

Charikar, M., Guha, S., Tardos, É., and Shmoys, D. B. A constant-factor approximation algorithm for the k-median problem. *Journal of Computer and System Sciences*, 65 (1):129–149, 2002.

Chen, D. Z., Li, J., Liang, H., and Wang, H. Matroid and knapsack center problems. *Algorithmica*, 75(1):27–52, 2016.

Chen, X., Fain, B., Lyu, L., and Munagala, K. Proportionally fair clustering. In *International Conference on Machine Learning*, pp. 1032–1041, 2019.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pp. 5036–5044, 2017.

Chiplunkar, A., Kale, S., and Ramamoorthy, S. N. How to solve fair k-center in massive data models. In *Proceedings of the International Conference on Machine Learning*, pp. 1877–1886, 2020.

Chlamtáč, E., Makarychev, Y., and Vakilian, A. Approximating fair clustering with cascaded norm objectives. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 2664–2683, 2022.

Dai, Z., Makarychev, Y., and Vakilian, A. Fair representation clustering with several protected classes. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 814–823, 2022.

El Halabi, M., Mitrović, S., Norouzi-Fard, A., Tardos, J., and Tarnawski, J. M. Fairness in streaming submodular maximization: Algorithms and hardness. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 13609–13622, 2020.

Feng, Y. and Shah, C. Has CEO gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11882–11890, 2022.

Ghadiri, M., Samadi, S., and Vempala, S. Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 438–448, 2021.

Ghadiri, M., Singh, M., and Vempala, S. S. Constant-factor approximation algorithms for socially fair k-clustering. *arXiv preprint arXiv:2206.11210*, 2022.

Ghouila-Houri, A. Caractérisation des matrices totalement unimodulaires. *Comptes Redus Hebdomadaires des Séances de l'Académie des Sciences (Paris)*, 254: 1192–1194, 1962.

Girdhar, Y. and Dudek, G. Efficient on-line data summarization using extremum summaries. In *International Conference on Robotics and Automation*, pp. 3490–3496, 2012.

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

Hajiaghayi, M., Khandekar, R., and Kortsarz, G. Budgeted red-blue median and its generalizations. In *Proceedings of the European Symposium on Algorithms*, pp. 314–325, 2010.

Har-Peled, S. and Mahabadi, S. Near neighbor: Who is the fairest of them all? *Advances in Neural Information Processing Systems*, 32, 2019.

Jiang, H., Lee, Y. T., Song, Z., and Wong, S. C.-w. An improved cutting plane method for convex optimization, convex-concave games, and its applications. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 944–953, 2020.

Jones, M., Nguyen, H., and Nguyen, T. Fair k-centers via maximum matching. In *Proceedings of the International Conference on Machine Learning*, pp. 4940–4949, 2020.

Jung, C., Kannan, S., and Lutz, N. A center in your neighborhood: Fairness in facility location. In *Proceedings of the Symposium on Foundations of Responsible Computing*, pp. 5:1–5:15, 2020.

Kay, M., Matuszek, C., and Munson, S. A. Unequal representation and gender stereotypes in image search results for occupations. In *Conference on Human Factors in Computing Systems (CHI)*, pp. 3819–3828, 2015.

Kleindessner, M., Awasthi, P., and Morgenstern, J. Fair k-center clustering for data summarization. In *Proceedings of the International Conference on Machine Learning*, pp. 3448–3457, 2019.

Krishnaswamy, R., Kumar, A., Nagarajan, V., Sabharwal, Y., and Saha, B. The matroid median problem. In *Proceedings of the Symposium on Discrete Algorithms*, pp. 1117–1130, 2011.

Krishnaswamy, R., Li, S., and Sandeep, S. Constant approximation for k-median and k-means with outliers via iterative rounding. In *Proceedings of the Symposium on Theory of Computing*, pp. 646–659, 2018.

Mahabadi, S. and Vakilian, A. Individual fairness for k-clustering. In *Proceedings of the International Conference on Machine Learning*, pp. 6586–6596, 2020.

Makarychev, K., Makarychev, Y., and Razenshteyn, I. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the Symposium on Theory of Computing*, pp. 1027–1038, 2019.

Makarychev, Y. and Vakilian, A. Approximation algorithms for socially fair clustering. In *Conference on Learning Theory*, pp. 3246–3264. PMLR, 2021.

Micha, E. and Shah, N. Proportionally fair clustering revisited. In *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*, 2020.

Moens, M.-F., Uyttendaele, C., and Dumortier, J. Abstracting of legal cases: the potential of clustering based on the selection of representative objects. *Journal of the American Society for Information Science*, 50(2):151–161, 1999.

Nguyen, H. L., Nguyen, T., and Jones, M. Fair range k-center. *arXiv preprint arXiv:2207.11337*, 2022.

Swamy, C. Improved approximation algorithms for matroid and knapsack median problems and applications. *ACM Transactions on Algorithms (TALG)*, 12(4):1–22, 2016.

Thejaswi, S., Ordozgoiti, B., and Gionis, A. Diversity-aware k-median: Clustering with fair center representation. In *Machine Learning and Knowledge Discovery in Databases*, pp. 765–780, 2021.

Vakilian, A. and Yalçıner, M. Improved approximation algorithms for individually fair clustering. In *International Conference on Artificial Intelligence and Statistics*, pp. 8758–8779. PMLR, 2022.

Van Den Brand, J., Lee, Y. T., Liu, Y. P., Saranurak, T., Sidford, A., Song, Z., and Wang, D. Minimum cost flows, MDPs, and ℓ_1-regression in nearly linear time for dense instances. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 859–869, 2021.

A. Preliminaries

A.1. Approximate Triangle Inequality

Lemma A.1 (Lemma A.1 (Makarychev et al., 2019))**.** *Let $x, y_1, \cdots, y_n$ be non-negative real numbers and $\lambda > 0, p \geq 1$. Then,*

$$(x + \sum_{i=1}^{n} y_i)^p \leq (1+\lambda)^{p-1} x^p + \Big(\frac{(1+\lambda)n}{\lambda}\Big)^{p-1} \sum_{i=1}^{n} y_i^p.$$

The following approximate variants of triangle inequalities are direct corollaries of the above lemma.

Corollary A.2. *Let (P, d) be a metric space. Consider distance function $d(u, v)^p$. Then, $\forall u_0, \cdots, u_r \in P, d(u_0, u_r)^p \leq r^{p-1} \cdot \sum_{i=0}^{r-1} d(u_i, u_{i+1})^p$.*

Proof. Follows from the triangle inequality for the distance function d and application of Lemma A.1 with $\lambda = r - 1$. □

Corollary A.3. *Let (P, d) be a metric space. Consider distance function $d(u, v)^p$. Then, $\forall u, v, w \in P, d(u, w)^p \leq 2^{p-1} \cdot (d(u, v)^p + d(v, w)^p)$.*

A.2. Totally Unimodular Matrices

A matrix is called *totally unimodular* if every square submatrix of the given matrix has a determinant that is either 0, 1, or -1. This property has significant implications in linear programming (LP). One of the key benefits of totally unimodular matrices is their close relationship with integer programming and LP. When a LP has a constraint matrix that is totally unimodular, it guarantees that the LP has an integral solution. Moreover, the integral solution can be found in polynomial time using various algorithms such as the ellipsoid method or the interior point method.

Theorem A.4 (Ghouila-Houri (1962))**.** *A matrix $A \in \mathbb{R}^{m \times n}$ is totally unimodular if and only if for every subset of the rows $R \subseteq [m]$, there is a partition of $R = R_- \uplus R_+$ such that for every $j \in [n]$,*

$$\sum_{i \in R_+} A_{ij} - \sum_{i \in R_-} A_{ij} \in \{-1, 0, 1\}.$$

B. Constructing well-separated locations

In this section, we follow the *location consolidation* approach of (Charikar et al., 2002) and output an instance with a sparsified set of locations where pairwise distance of survived locations are "relatively large". To describe this step, we work with the relaxation FAIRRANGELP. As in our algorithm we never modify the set F and the range constraints $\{\alpha_i, \beta_i\}_{i \in [\ell]}$, to specify an instance, from now on we will only specify the set of clients (D, w). Consider an optimal fractional solution (x, y) of FAIRRANGELP(D, w)—*throughout the paper we refer to the cost of an optimal solution of this LP as* OPT_D. For any location $v \in D$, we define $\mathcal{R}(v) := \big(\sum_{u \in F} x_{vu} \cdot d(v, u)^p\big)^{1/p}$ as the *fractional distance* of a unit of demand at location v w.r.t. the optimal solution (x, y). Note that if (x, y) is an integral solution, then $\mathcal{R}(v)$ is simply the distance from v to its closest open facility, which is specified by y. Next, we process the locations as follows. We sort locations in a non-decreasing order of their fractional distance; we index the locations in D as $v_1, \cdots, v_n$ such that $\mathcal{R}(v_1) \leq \mathcal{R}(v_2) \leq \cdots \leq \mathcal{R}(v_n)$. We iterate over the locations in this order and for every $i \in [n]$, when we are processing v_i, we check for all locations v_j s.t. $j > i$ and $d(v_i, v_j) \leq 2^{1+1/p} \cdot \mathcal{R}(v_j)$. For each such location, we move the demand at location v_j to v_i and set the demand at location v_j to zero. At the end of this step, the algorithm returns a new demand function w' supported on $D' \subseteq D$.

For every location $v \in D'$, we define the *ball* $\mathcal{B}(v) := \{u \in F | d(v, u) \leq 2^{1/p}\mathcal{R}(v)\}$ to denote the set of facilities at distance at most $2^{1/p}\mathcal{R}(v)$ from v. The above claim implies the following lemma.

Lemma B.1. *For every $v \in D'$, $\sum_{u \in \mathcal{B}(v)} x_{vu} \geq 1/2$.*

Proof. Note that

$$\mathcal{R}(v)^p = \sum_{u \in \mathcal{B}(v)} x_{vu} d(u, v)^p + \sum_{u \in F \setminus \mathcal{B}(v)} x_{vu} d(u, v)^p \geq \sum_{u \in \mathcal{B}(v)} x_{vu} d(u, v)^p + (1 - \sum_{u \in \mathcal{B}(v)} x_{vu}) \cdot 2\mathcal{R}(v)^p$$

which proves that $\sum_{u \in \mathcal{B}(v)} x_{vu} \geq 1/2$. □

Algorithm 2 Consolidating Locations.

Input: (x, y) is an optimal solution of FAIRRANGELP(D, w)
$\mathcal{R}(v) \leftarrow (\sum_{u \in F} d(v, u)^p \cdot x_{vu})^{1/p}$ **for all** $v \in D$
$w'(v) \leftarrow w(v)$ **for all** $v \in D$
sort locations in D so that $\mathcal{R}(v_1) \leq \cdots \leq \mathcal{R}(v_n)$
for $i = 1$ to $n - 1$ **do**
  if $w'(v_i) > 0$ **then**
    for $j = i + 1$ to n **do**
      if $w'(v_j) > 0$ **and** $d(v_i, v_j) \leq 2^{1+1/p} \cdot \mathcal{R}(v_j)$ **then**
        $w'(v_i) \leftarrow w'(v_i) + w(v_j)$ and $w'(v_j) \leftarrow 0$
      end if
    end for
  end if
end for

Next, we follow the reduction steps in (Krishnaswamy et al., 2011; Swamy, 2016) to construct a fractional solution of FAIRRANGELP with further structures. We construct a fractional solution (x', y) in which for each location v, its demands are served by the facilities in a *super ball* $\mathcal{P}(v)$ with fraction $\gamma \geq 1/2$ and by the facilities in $\mathcal{P}(v')$ with fraction $1-\gamma_u$ where $v' := \mathsf{NN}_{D'}(v)$ denotes the nearest location in D' (other than v itself) to v and the collection of super balls $\{\mathcal{P}(v)\}_{v \in D'}$ are disjoint.

We start with the feasible solution (x^1, y) of FAIRRANGELP(D', w') where $x^1_{vu} = x_{vu}$ for all $v \in D'$ and $u \in F$ and (x, y) is the optimal solution of FAIRRANGELP(D, w) from which the sparsified set of locations (D', w') is constructed. Note that by Theorem 2.3-(Q2), (x^1, y) is a feasible solution of FAIRRANGELP(D', w') with cost at most OPT_D.

Private facilities. In Theorem 2.3-(Q1), we showed the collection of balls $\{\mathcal{B}(v)\}_{v \in D'}$ are disjoint. Here, we construct a solution (x^2, y) of FAIRRANGELP(D', w') with cost $e^{O(p)} \cdot \text{OPT}_D$ such that each facility $u \in F \setminus \bigcup_{v \in D'} \mathcal{B}(v)$, serves the clients of at most one location. In other words, each (fractionally) opened facility outside the balls *only* serves one location. For each $v \in D'$, the super ball $\mathcal{P}(v)$ consists of $\mathcal{B}(v)$ and the private facilities of v.

Consider a facility u that does not belong to any ball in $\{\mathcal{B}(v)\}_{v \in D'}$ and serves clients from more than one locations $v_1, \cdots, v_r$. Lets assume locations are ordered according their distance to u; $d(v_1, u) \leq \cdots \leq d(v_r, u)$.

Claim B.2. *For every $j \in \{2, \cdots, r\}$ and every facility $u' \in \mathcal{B}(v_1)$, $d(v_j, u') \leq 3d(v_j, u)$.*

Proof. Since for every $j \in \{2, \cdots, r\}$, $d(v_j, u) \geq d(v_1, u)$,

$$d(v_j, v_1) \leq d(v_j, u) + d(v_1, u) \leq 2d(v_j, u). \tag{12}$$

Moreover, for a facility $u' \in \mathcal{B}(v_1)$,

$$\begin{aligned} d(v_j, u') &\leq d(v_j, v_1) + d(v_1, u') \\ &\leq \frac{3}{2} \cdot d(v_j, v_1) && \triangleright \text{ Theorem 2.3-(Q1)} \\ &\leq 3d(v_j, u) && \triangleright \text{ Eq. (12)} \end{aligned}$$

□

Now, in (x^2, y), for every $j \in \{2, \cdots, r\}$, we set the assignment of v_j to u to zero; $x^2_{v_j u} = 0$ and instead increase the assignment of v_j to facilities $u \in \mathcal{B}(v_1)$ with total increase of $x^1_{v_j u}$. A formal procedure of this reassignment step is described in Algorithm 3.

Lemma B.3. *(x^2, y) is a feasible solution of* FAIRRANGELP(D', w') *with cost at most* $3^p \cdot \text{OPT}_D$.

Algorithm 3 Assignments to Private Facilities.

Input: (x^1, y) is a feasible solution of $\textsc{FairRangeLP}(D', w')$
$x^2_{vu} \leftarrow x^1_{vu}$ **for all** $v \in D', u \in F$
for all $u \in F \setminus \bigcup_{v \in D'} \mathcal{B}(v)$ with $y(u) > 0$ **do**
  sort locations $\{v \in D' \mid x^1_{vu} > 0\}$ according to their distance to u so that $d(v_1, u) \leq \cdots \leq d(v_r, u)$
  for $j = 2$ to r **do**
    $x^2_{v_j u} \leftarrow 0$ and $b = x^1_{v_j u}$
    for all $u' \in \mathcal{B}(v_1)$ **do**
      $x^2_{v_j u'} \leftarrow \min(y_{u'}, b + x^1_{v_j u'})$ and $b \leftarrow b - \min(y_{u'} - x^1_{v_j u'}, b)$
    end for
  end for
end for

Proof. First we prove the feasibility of (x^2, y).

$$\begin{aligned} \sum_{u_1 \in \mathcal{B}(v_1)} y_{u_1} &\geq \sum_{u_1 \in \mathcal{B}(v_1)} x^1_{v_1 u_1} && \triangleright \text{ by feasibility of } (x^1, y) \\ &\geq 1/2 && \triangleright \text{ Lemma B.1} \end{aligned}$$

By another application of Lemma B.1, for every $j \in \{2, \cdots, r\}$,

$$x^1_{v_j u} + \sum_{u'_j \in F \setminus \mathcal{B}(v_j)} x^1_{v_j u'_j} \leq 1 - \sum_{u'_j \in \mathcal{B}(v_j)} x_{v_j u'_j} \leq 1/2$$

Hence, the facilities in $\mathcal{B}(v_1)$ have enough slack to accommodate extra $x^1_{v_j u}$ in their assignment from v_j. This implies that (x^2, y) as constructed in Algorithm 3 is a feasible solution of $\textsc{FairRangeLP}(D', w')$.

Moreover, since by Claim B.2 for every $j \in \{2, \cdots, r\}$ and $u' \in \mathcal{B}(v_1)$, $d(v_j, u') \leq 3d(v_j, u)$, $\mathrm{cost}(x^2, y) \leq 3^p \cdot \mathrm{cost}(x^1, y) = 3^p \cdot \mathrm{OPT}_D$. $\square$

C. Omitted Proofs of Section 2

Proof of Theorem 2.2. As the set of facilities $F = P$, the set S is a feasible solution for fair range clustering on the original instance P too.

Let $D = \{c_1, \cdots, c_k\}$ denote an α-approximate solution of clustering with ℓ_p-objective on P. Let $\mathrm{OPT}_{\mathrm{org}}$ and $\mathrm{OPT}_{\mathrm{rdc}}$ respectively denote optimal solutions of fair range clustering with ℓ_p-objective on the original instance P and the reduced instance (D, F). First we bound the cost of $\mathrm{OPT}_{\mathrm{org}}$ on the reduced instance (D, F):

$$\begin{aligned} \mathrm{cost}_{\mathrm{rdc}}(\mathrm{OPT}_{\mathrm{org}}) &= \sum_{v \in D} w'(v) \cdot d(v, \mathrm{OPT}_{\mathrm{org}})^p \\ &\leq \sum_{v \in P} w(v) \cdot 2^{p-1} \cdot \big(d(v, D)^p + d(v, \mathrm{OPT}_{\mathrm{org}})^p\big) && \triangleright \text{ approximate triangle inequality} \\ &\leq 2^{p-1} \big(\sum_{v \in P} w(v) d(v, D)^p + \sum_{v \in P} w(v) d(v, \mathrm{OPT}_{\mathrm{org}})^p \big) \\ &\leq 2^{p-1} \cdot (\alpha^p \cdot \mathrm{cost}(\mathrm{OPT}_{\mathrm{org}}) + \mathrm{cost}(\mathrm{OPT}_{\mathrm{org}})) && \triangleright\ D \text{ is an } \alpha\text{-approximate solution} \\ &= 2^{p-1} (\alpha^p + 1) \cdot \mathrm{cost}(\mathrm{OPT}_{\mathrm{org}}) && (13) \end{aligned}$$

Next, we bound the cost of the solution S on the original instance P,

$$\begin{aligned}
\mathrm{cost}(S) &= \sum_{v\in P} w(v)\cdot d(v,S)^p \\
&= \sum_{v\in P} w(v)\cdot 2^{p-1}\cdot \big(d(v,D)^p + d(\mathsf{NN}_D(v),S)^p\big) && \triangleright \text{ approximate triangle inequlity} \\
&\le 2^{p-1}\big(\sum_{v\in P} w(v)d(v,D)^p + \sum_{v\in D} w'(v)d(v,S)^p\big) \\
&\le 2^{p-1}\cdot(\alpha^p\cdot \mathrm{cost}(\mathrm{OPT}_{\mathrm{org}}) + \beta^p\cdot \mathrm{cost}_{\mathrm{rdc}}(\mathrm{OPT}_{\mathrm{rdc}})) \\
&\le 2^{p-1}\cdot(\alpha^p + \beta^p\cdot 2^{p-1}(\alpha^p+1))\cdot \mathrm{cost}(\mathrm{OPT}_{\mathrm{org}}) && \triangleright \text{ by Eq (13)}
\end{aligned}$$

Thus, S is an $O(\alpha\beta)$-approximation for fair range clustering with ℓ_p-objective on the original instance P. □

Proof of Theorem 2.3. (Q1): Wlog, lets assume that v_1 located before v_2 in the ordering considered by Algorithm 2; hence, $\mathcal{R}(v_2)\ge\mathcal{R}(v_1)$. However, if $d(v_1,v_2)\ge 2^{1+1/p}\mathcal{R}(v_2)$, then Algorithm 2 moves the demand of v_2 to v_1 and v_2 will not survive the process which is a contradiction. Thus, $d(v_1,v_2)\le 2^{1+1/p}\mathcal{R}(v_2) = 2^{1+1/p}\cdot\max\{\mathcal{R}(v_2),\mathcal{R}(v_1)\}$.

(Q2): In FAIRRANGELP, no constraint depends on demands and thus the constraints of FAIRRANGELP(D',w') are subsets of the ones in FAIRRANGELP(D,w). Thus, (x,y) is a feasible solution of FAIRRANGELP(D',w') too.

Next, we show that the cost of (x,y) w.r.t. FAIRRANGELP(D',w') is not more that the cost (x,y) w.r.t. FAIRRANGELP(D,w). For each location v, let $\overline{v}$ denote the location in D' that receives the demand of v by the end of Algorithm 2.

$$\begin{aligned}
\mathrm{cost}(x,y;D',w') := \sum_{q\in D'} w'(q)\sum_{u\in F} x_{qu} d(q,u)^p &= \sum_{q\in D'} w'(q)\mathcal{R}(q)^p \\
&= \sum_{q\in D'}\sum_{v\in D:\overline{v}=q} w(v)\mathcal{R}(q)^p \\
&= \sum_{v\in D} w(v)\mathcal{R}(\overline{v})^p \\
&\le \sum_{v\in D} w(v)\mathcal{R}(v)^p = \mathrm{OPT}_D
\end{aligned}$$

(Q3): Algorithm 2 may either move the demand of a location v to some other location $\overline{v}$ or keep it at v. Let $v'=\overline{v}$ in the former case and $v'=v$ in the latter case. In either case $d(v,v')\le 2^{1+1/p}\cdot\mathcal{R}(v)$. Therefore, by the approximate triangle inequality (Corollary A.3),

$$d(v,C)^p \le 2^{p-1}(d(v,v')^p + d(v',C)^p) \le 2^{2p}\mathcal{R}(v)^p + 2^{p-1}d(v',C)^p$$

Summing over all locations,

$$\sum_{v\in D} w(v)d(v,C)^p \le \sum_{v\in D} w(v)\big(4^p\mathcal{R}(v)^p + 2^{p-1}d(v',C)^p\big) \le 4^p\cdot\mathrm{OPT}_D + 2^{p-1}\cdot z$$

□

Proof of Theorem 2.5. We initialize $(\overline{x},\overline{y})$ to the solution (x^2,y) as constructed in Algorithm 3; $\overline{x}=x^2$ and $\overline{y}=y$. We keep $\overline{y}=y$ throughout the process but modify $\overline{x}$ to satisfy the desired properties.

By Claim B.2, we can partition open facilities, $\{u\in F \mid y_u>0\}$ into disjoint super balls $\{\mathcal{P}(v)\}_{v\in D'}$. So, (P1) will be satisfied by $\overline{y}$.

Next, we modify the assignment vector x_2 so that for every $v\in D'$ and $u\in\mathcal{P}(v)\setminus\mathcal{B}(v)$, $\overline{x}_{vu}>0$ only if $\sum_{u\in\mathcal{B}(v)}\overline{y}<1$. Similarly, for every $v\in D'$ and $u\in F\setminus\mathcal{P}(v)$, $\overline{x}_{vu}>0$ iff $\sum_{u\in\mathcal{P}(v)}\overline{y}_v<1$. So, by the construction of $(\overline{x},\overline{y})$, (P2) is satisfied.

Consider a location $v \in D'$. Since the total $\overline{y}$ contributions in each of $\mathcal{B}(v)$ and $\mathcal{B}(v')$ is at least half, in $\overline{x}$ we assign the remaining demand of each location v, which are not satisfied by the facilities in $\mathcal{P}(v)$, to the facilities in $\mathcal{B}(v')$. So, $(\overline{x}, \overline{y})$ satisfies (P3) and (P4).

Without loss of generality, we can assume that (P5) holds, otherwise, we could transfer the $\overline{x}_{vu}$ fraction of assignment of v from u to a facility in $\mathcal{B}(v')$, set $\overline{y}_u = 0$, and reduce the total cost. Again, this reassignment is always possible because for every $v \in D'$,

$$\overline{x}_{vu} + \sum_{u' \in F \setminus \mathcal{P}(v)} \overline{x}_{vu'} \leq 1/2 \leq \sum_{u' \in \mathcal{B}(v')} \overline{y}_{u'}.$$

Finally, the last property follows from Claim B.2 and the following bound. For every $u'' \in \mathcal{B}(v'')$ and $u' \in \mathcal{B}(v')$ where $w \notin \{v, v'\}$,

$$\begin{aligned}
d(v, u') &\leq d(v, v') + d(v', u') && \triangleright \text{ triangle inequality} \\
&\leq (1 + \frac{1}{2}) \cdot d(v, v') \\
&\leq (1 + \frac{1}{2}) \cdot d(v, v'') && \triangleright\ d(v, v') \leq d(v, v'') \\
&\leq 3 \cdot d(v, u''),
\end{aligned}$$

where the second inequality holds since $d(v, v') \geq 2^{1+\frac{1}{p}} \mathcal{R}(v')$ and $d(v', u') \leq 2^{\frac{1}{p}} \mathcal{R}(v')$. Similarly, the fourth inequality follows from $d(v, v'') \geq 2^{1+\frac{1}{p}} \mathcal{R}(v'')$ and $d(v'', u'') \leq 2^{\frac{1}{p}} \mathcal{R}(v'')$ (which implies that $d(v, u'') \geq \frac{1}{2} d(v, v'')$). Thus, $\mathrm{cost}(\overline{x}, \overline{y}) \leq 3^p \cdot \mathrm{cost}(x^2, y) \leq 9^p \cdot \mathrm{OPT}_D$. □