Fair Max-Min Diversification in Refined and Relaxed Metric Spaces

Ashley Qianxi Gao
Department of Computer Science,
University of Toronto, ON, Canada

Allan Borodin
Department of Computer Science,
University of Toronto, ON, Canada

Abstract

The fair Max-Min diversification problem for arbitrary metric spaces and Euclidean spaces has been previously studied in Moumoulidou et al. [ICDT 2021] and Addanki et al. [ICDT 2022]. The problem arises from the need to downsize a dataset while maintaining a diverse and representative (fair) subset. In this paper, we extend the problem to the setting of extended metric spaces. Given n points in a extended metric space $(\mathcal{U}, d)$ — where the triangle inequality is generalized to $c \cdot d(x, y) \leq d(x, z) + d(y, z)$ for $c \in (0, 2]$ — and each point belongs to a group $i \in [m]$. Our task is to select k_i points from each group i, with the goal of maximizing the minimum distance between the selected points. By ensuring k_i points selection for each group i, fairness is preserved, while maximizing the minimum distance between any two points guarantees diversity in the sample set.

In this paper, we present an algorithm for the refined metric space(i.e. extended metric space with distortion factor $c \in (1, 2]$), which achieves a $\frac{c^m+c-2}{(c-1)c^m}$-approximation, analogous to the $m + 1$ approximation algorithm for regular metric spaces(i.e. $c = 1$) introduced by Addanki et al. [ICDT 2022]. Furthermore, we examine the case of both refined($c \in (1, 2]$) and relaxed($c \in (0, 1)$) metric spaces with $m = 2$. By leveraging the 4-approximation algorithm for regular metric spaces proposed by Moumoulidou et al. [ICDT 2021], we demonstrate that a $\frac{4}{c^2}$-approximation can be achieved in the context of extended metrics where $c \in (0, 2]$.

A potential application of our work lies in its ability to handle high-dimensional datasets. For regular metric spaces, Moumoulidou et al. [ICDT 2021] demonstrated that achieving an approximation ratio better than 2 is infeasible unless $P = NP$. Notably, when $c > 1.5$, our algorithm achieves an approximation ratio better than 2 in such refined metric space.

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis

Keywords and phrases data selection, fairness constraints, diversity maximization, approximation algorithms, parameterized triangle inequality

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

Funding *Ashley Qianxi Gao*: University of Toronto Excellence Award (UTEA)

1 Introduction

In the contemporary era, data is used in diverse fields, including but not limited to machine learning, commerce, data mining, and healthcare. The datasets grow exponentially as the hardware evolves, and how we can utilize such datasets effectively becomes a challenge. This requires distilling the datasets into small, manageable, and high-quality datasets for practical applications. There are various ways to define high-quality subsets. In this paper, we adopt the setting from the work of [8, 2], define the high-quality sub-dataset in two aspects; in addition to maximizing the dissimilarity of the data that we choose, we may also need to ensure that each group in the original datasets is well represented. As [8] mentions, the concepts of $fair$ and $diverse$ are practical in the real-life scenario for the dataset selection.

For example, [6] mentions that the absence of fair representation in public media can lead to polarized opinions, while the loss of diversity in the media gives people homogeneous content.

As a simple example, similar to the Nobel laureates example demonstrated in [8], consider that the university plans to select a limited number of professors to give a series of public talks. Among hundreds of professors, only a few can be chosen. It is important to choose the set of professors carefully. To make the chosen set representative, the professors in the set should come from a broad range of faculties or disciplines, thereby ensuring diversity across academic domains; i.e., there should not be any two professors from the same discipline. Concurrently, it's important to avoid gender imbalance; the aim is to have almost the same number of representatives from different genders, which guarantees the fairness property of the chosen set.

The Fair Max-Min Diversification problem is defined as follows: Consider a metric space $(\mathcal{U}, d)$, where $\mathcal{U}$ is a universe of n elements divided into m non-overlapping groups, and d is a metric distance function. Given fairness constraints $k_1, k_2, \ldots, k_m$ for each group, we aim to select a total of $k = \sum_{i \in [m]} k_i$ points to form a subset $\mathcal{S}$ from the n points. Specifically, for each group i, we include k_i points in our solution set $\mathcal{S}$. The objective is to maximize the minimum distance between the points in $\mathcal{S}$. The problem for arbitrary metric spaces and particularly Euclidean space has been extensively studied by [8, 2, 7].

In this paper, we focus on the Fair Max-Min Diversification problem within refined and relaxed metric spaces, where the triangle inequality is modified to $c \cdot d(x, y) \leq d(x, z) + d(y, z)$ for $c \in (1, 2]$ and $c \in (0, 1)$ separately. To the best of our knowledge, this is the first work that discusses the Fair Max-Min Diversification problem in the context of refined metric or relaxed metric spaces. One of the applications for the refined framework is handling high-dimensional data, where the curse of dimensionality causes distances between points to become relatively uniform[3]. By assuming a refined triangle inequality for such high-dimensional data, our proposed Algorithm 1 provides an approximation ratio bounded by a constant, compared to existing $m + 1$-approximation algorithms when applied to these datasets.

We propose an improved version(Algorithm 1) of the FairGreedyFlow algorithm from [2], achieving a better $\frac{c^m+c-2}{(c-1)c^m}$-approximation[1] in refined metric spaces compared to the previous $m + 1$ approximation result in standard metric spaces. As long as $c > 1$, the approximation ratio $\frac{c^m+c-2}{(c-1)c^m}$ is bounded by a constant, and the bounded constant decreases(i.e. achieves a better approximation ratio) as c increases. This leads to significantly better results in the refined metric, reducing the bound from $m + 1$ to a constant. In addition to the previous algorithm and results for arbitrary m, we also explore the special case of $m = 2$. Specifically, we extend the FairSwap algorithm(Algorithm 3), originally introduced in [8] for $m = 2$, to extended metric spaces(i.e. refined and relaxed metrics, while $c \in (0, 2]$) and prove that the algorithm achieves a $\frac{4}{c^2}$-approximation in extended metric space, offering a better approximation ratio than the original 4 in refined metrics($c \in (1, 2]$), but a worse approximation ratio in relaxed metrics($c \in (0, 1)$).

[1] To be precise, here we are stating our approximation bound and that of [2] under the assumption that the value of the optimal diversity l^* under the fairness constraint is known. A standard divide-and-conquer approach can be used to achieve an approximation of $\left(\frac{c^m+c-2}{(c-1)c^m}\right)(1 + \epsilon)$ when l^* is not known. The result in [2] also needs the binary search idea to achieve their $(m + 1)(1 + \epsilon)$ approximation when l^* is not known.

2 Background and Preliminaries

2.1 Problem Definition

Approximation Ratio and Diversity

A standard measure of the quality of the approximation algorithm is the worst-case approximation ratio. In this framework, let us denote the value of an approximation algorithm as ALG and the value of an optimal solution as OPT. Typically, the approximation ratio is defined as $\alpha = ALG/OPT$. However, in the context of this paper, we focus on the max-min diversification problem, a maximization problem where OPT is always greater than or equal to ALG. To align better with the intuitive understanding obtained from other problems, we use a modified definition for the approximation ratio: $\alpha = OPT/ALG$. This adjustment ensures that the ratio is always greater than or equal to 1, providing a more standardized metric for comparison and analysis. A ratio greater than 1 indicates how close the approximation (ALG) is to the optimal (OPT) and a ratio of 1 means the solution is exactly optimal.

▶ **Definition 1** (Approximation Ratio). *For any possible input I, let $OPT(I)$ represent the optimal solution's value and $ALG(I)$ denote the value achieved by the approximation algorithm. The algorithm achieves an* ***approximation ratio*** *α if for all inputs I:*

$$\alpha \cdot ALG(I) \geq OPT(I).$$

▶ **Definition 2** (Diversity). *Throughout this paper, we refer to the* ***diversity*** *of a set S as*

$$div(S) = \min_{u,v \in S, u \neq v} d(u, v).$$

That is, diversity is defined as the minimum distance between two points in a given set. The objective is to maximize diversity while ensuring fairness.

Extended Metric Space

In regular metric spaces, the standard triangle inequality is used, which can be either relaxed or strengthened. We introduce a distortion factor c into the triangle inequality to adjust the metrics and name it extended metric space. This parameterized triangle inequality setting for a similar diversity problem, the so-called Max Sum Diversification problem, is discussed in [10]. According to [10], the distortion factor c ranges from $(0, 2]$ instead of all positive reals. We prove why the metric space does not make sense for $c > 2$ in Appendix B. Specifically, when $c = 1$, the metric space is the regular metric space with no distortion. When $c = 2$, the metric space becomes an identical metric space, meaning the distance between any two points is the same. In this paper, we refer to the extended metric space with a distortion factor $c \in (0, 1)$ as the *relaxed metric space*, and with $c \in (1, 2]$ as the *refined metric space*.

▶ **Definition 3** (Extended Metric Space). *$(\mathcal{U}, d)$ is an extended metric space with a factor $c \in (0, 2]$ if $\forall u, v, w \in \mathcal{U}$, the following are satisfied:*

1. *(identity)* $d(u, v) = 0 \iff u = v$
2. *(symmetry)* $d(u, v) = d(v, u)$
3. *(positive)* $d(u, v) \geq 0$
4. *(triangle)* $c \cdot d(u, w) \leq d(u, v) + d(v, w)$

With the necessary context established, we can now formally define the problem:

▶ **Definition 4** (Fair Max-Min Diversification Problem with Extended Metric Space)**.** *Let $(\mathcal{U}, d)$ be an **extended** metric space with $c \in (0, 2]$ where $\mathcal{U} = \bigcup_{i=1}^{m} \mathcal{U}_i$ is a universe of n elements partitioned into m non-overlapping groups and $d : \mathcal{U} \times \mathcal{U} \to \mathbb{R}_0^+$ is a metric distance function.*

Further, let $k_1, k_2, \cdots, k_m$ be non-negatice integers with $k_i \leq |\mathcal{U}_i|$, $\forall i \in [m]$, and $k = \sum_{i=1}^{m} k_i$.

The problem is defined as follows:

$$\max_{S \subseteq \mathcal{U}} \min_{u,v \in S, u \neq v} d(u, v),$$

or equivalently

$$\max_{S \subseteq \mathcal{U}} div(S).$$

subject to $|S \cap \mathcal{U}_i| = k_i$, $\forall i \in [m]$ (fairness constraints)

2.2 Related Work

The diversification problem has been widely studied with various objective functions, with distance-based objectives being one of the major focuses. For instance, the unconstrained Max-Sum Diversification problem is explored in [4], where the objective is to maximize the summation of pairwise distances. The unconstrained Max-Min Diversification problem is examined in [9], where the authors provide a 2-approximation algorithm. They also prove that for the unconstrained Max-Min Diversification problem, if the triangle inequality does not hold, no polynomial-time relative approximation algorithm exists unless $P = NP$. The Max-Average Diversification problem is also discussed in the same work [9], with the authors presenting a 4-approximation algorithm.

Constrained versions of the diversification problem, such as those involving matroid constraints, are introduced in [4, 5, 1]. The Fair Max-Min Diversification problem was first proposed by [8] and later improved upon by [2]. These works primarily focus on metric and Euclidean spaces. Specifically, [2] provides an $m+1$-approximation algorithm for the problem in metric spaces and introduces relaxed fairness algorithms, such as a 2-approximation algorithm for expected fairness and a 6-approximation algorithm for $(1 - \epsilon)$ fairness. [8] introduces a 4-approximation algorithm for the only 2 groups setting. In addition, the overlapping cases are studied in [8], with a 4-approximation ratio result for the 2 groups setting and a $\left(3\binom{m}{\lfloor m/2 \rfloor} - 1\right)$-approximation ratio for arbitrary m groups.

In the context of Euclidean spaces, [2] demonstrate that the problem can be solved exactly when dimension $D = 1$, and they propose a $(1 + \epsilon)$-approximation algorithm for constant dimensions and groups. Further improvements for constant-dimensional spaces are made in [7], where the authors present a constant-factor approximation algorithm that runs in near-linear time, contrasting with previous algorithms that required super-linear running time.

3 $\frac{c^m+c-2}{(c-1)c^m}$-Approximation Algorithm for Arbitrary m

▶ **Theorem 5.** *FairGreedyFlow for Refined Metric Space (Algorithm 1) using an approximation γ for the optimal value is a $\left(\frac{c^m+c-2}{(c-1)c^m}\right)(1 + \epsilon)$-approximation algorithm with perfect fairness for the Fair Max-Min Diversification problem with extended metric factor $c \in (1, 2]$ that runs in a time of $O(nkm^3\epsilon^{-1} \log n)$.*

Algorithm 1 FairGreedyFlow for Refined Metric Space

Input: $\mathcal{U}_1, \ldots, \mathcal{U}_m$: Universe of available elements
$k_1, \ldots, k_m \in \mathbb{Z}^+$
$\gamma \in \mathbb{R}^+$: A guess of the optimum fair diversity
Output: k_i points in $\mathcal{U}_i$ for $i \in [m]$

$\mathcal{R} \leftarrow \mathcal{U}$ ▷ Remaining elements
$\mathcal{C} \leftarrow \emptyset$ ▷ Subsets collection (also called clusters collection)
$\mathcal{P} \leftarrow \emptyset$ ▷ Points collection of the subsets in $\mathcal{C}$
$i \leftarrow 0$ ▷ Index of clusters
while $|\mathcal{R}| > 0$ **and** $|\mathcal{P}| \leq km$ **do**
  $i \leftarrow i + 1$
  $D_i \leftarrow \emptyset$ ▷ Current cluster
  $D_{i,\text{index}} \leftarrow \emptyset$ ▷ Index of groups that already have a point in current cluster D_i
  while an element $p \in \mathcal{R} \cap \mathcal{U}_j$ for some $j \in \{1, 2, \ldots, m\} \setminus D_{i,\text{index}}$ exists **do**
    if $|D_i| = 0$ **or** $d(p, x) < \frac{(c-1)c^m}{c^m+c-2}\gamma$ for some $x \in D_i$ **then**
      $D_i \leftarrow D_i \cup \{p\}$ ▷ Add point p to cluster D
      $D_{i,\text{index}} \leftarrow D_{i,\text{index}} \cup \{j\}$
    end if
  end while
  $\mathcal{R} \leftarrow \mathcal{R} \setminus \bigcup_{p \in D_i} \mathbf{B}(p, \frac{(c-1)c^m}{c^m+c-2}\gamma)$
  $\mathcal{P} \leftarrow \mathcal{P} \cup D_i$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{D_i\}$
  for all $j \in [m]$ **do**
    if $|\{D \mid D \in \mathcal{C} \text{ and } D \cap \mathcal{U}_j \neq \emptyset\}| \geq k$ **then**
      $\mathcal{R} \leftarrow \mathcal{R} \setminus \mathcal{U}_j$
    end if
  end for
end while

Let $t \leftarrow |\mathcal{C}|$ ▷ t is the number of the clusters
Construct directed graph $G = (V, E)$ where ▷ Construct flow graph
  $V = \{a, u_1, \ldots, u_m, v_1, \ldots, v_t, b\}$
  $E = \{(a, u_i)$ with capacity $k_i : i \in [m]\}$
    $\cup\{(v_j, b)$ with capacity $1 : j \in [t]\}$
    $\cup\{(u_i, v_j)$ with capacity $1 : |\mathcal{U}_i \cap D_j| \geq 1\}$
$\mathcal{S} \leftarrow \emptyset$ ▷ Initialization of the Solution Set $\mathcal{S}$
Compute maximum a-b flow in G
if flow size $< k = \sum k_i$ **then**
  return $\emptyset$ ▷ Abort
else ▷ Max flow is k
  for all (u_i, v_j) with flow equal to 1 **do**
    add the point in D_j with group i to $\mathcal{S}$
  end for
end if
return $\mathcal{S}$

In this section, we present the result that a slightly modified version of the *FairGreedyFlow* algorithm, as proposed in [2], can achieve a $\alpha(c,m) = \frac{c^m+c-2}{(c-1)c^m}$ result in the context of refined metrics. It is notable that in the context of refined metric spaces, the approximation ratio converges to a constant $\alpha(c,\infty)$ as m (i.e., the number of fairness groups) increases. Moreover, as the distortion factor c of the metric space approaches 2, the approximation ratio "constant" improves. For instance, if $c = 1.001$, the approximation ratio is bounded by 1000 as m tends to infinity. However, if $c = 1.5$, we can achieve an approximation bounded by 2 for even arbitrary large m.

The main idea of the algorithm is to form group points that are relatively close to each other into clusters, and select at most one point from each cluster to ensure the distances between the selected points are sufficiently large. To determine the "relative close" threshold for the given metrics, we give a guess γ of the optimal diversity l^* under the fairness constraints using binary search. We can find an $\epsilon \geq 0$ and define $\frac{l^*}{(1+\epsilon)} < \gamma \leq l^*$. If the distance between two points is less than $\frac{(c-1)c^m}{c^m+c-2}\gamma$, we consider them to be close and put them in the same group. Compared to the previous FairGreedyFlow algorithm in [2], the decision bound is changed from $\frac{1}{m+1}\gamma$ to $\frac{(c-1)c^m}{c^m+c-2}\gamma$ in the refined metric context. Note that by L'Hôpital's rule, $\lim_{c\to 1}\frac{(c-1)c^m}{c^m+c-2}\gamma = \frac{1}{m+1}\gamma$, which aligns with the approximation ratio for the standard metric.

Finding an appropriate γ is not difficult, as there are at most $\binom{n}{2}$ possible values for γ. A suitable γ can be efficiently determined using binary search. If $\gamma < l^*$, the algorithm will fail and return the empty set. Therefore, we can adjust the value of γ until an optimal point is reached.

When forming a cluster D, we impose the condition that $|D \cap \mathcal{U}_i| \leq 1$ for each $i \in [m]$. This means that each cluster can contain at most one point from each fairness group i. If there are additional points from $\mathcal{U}_i$ near cluster D, but we have already selected a point $p \in (D \cap \mathcal{U}_i)$, the remaining points from $\mathcal{U}_i$ nearby will be excluded from further consideration when forming clusters.

Once a sufficient number of clusters is obtained using this approach, we utilize the resulting clusters to construct a network flow graph. Consider the fairness groups denoted as $u_1, u_2, \ldots, u_m$ and clusters represented as $v_1, v_2, \ldots v_t$. We designate a as the source and b as the sink. Initially, we establish connections from the source a to each group u_i, assigning a weight k_i to each edge, where $i \in [m]$. Subsequently, we create edges between each corresponding pair of u_i and v_j with a weight of 1. Similarly, we connect each cluster v_j to the sink b with edges also weighted at 1. We give a concrete example in Figure 1.

A natural question regarding this algorithm is how we ensure that there are enough points from each fairness group within the clusters to form a valid solution set, by choosing at most one point from each cluster. In the proof(Appendix A), we demonstrate that for every point in the optimal solution set, either the point itself or its substitute will be contained within a cluster. Additionally, each cluster can contain at most one such point. Consequently, we can find a valid solution set by applying the network flow algorithm to the algorithm-constructed graph.

In the proof, we first utilize γ to establish the approximation bound and then incorporate the $(1+\epsilon)$ factor into our analysis. The proof involves some technical details, which are elaborated in Appendix A, where Claim 8, Claim 11, and Lemma 12 are proven thoroughly. We will provide a high-level overview of the proof as follows.

To establish the approximation bound, denoted as $\frac{c^m+c-2}{(c-1)c^m}$, we first prove that for all

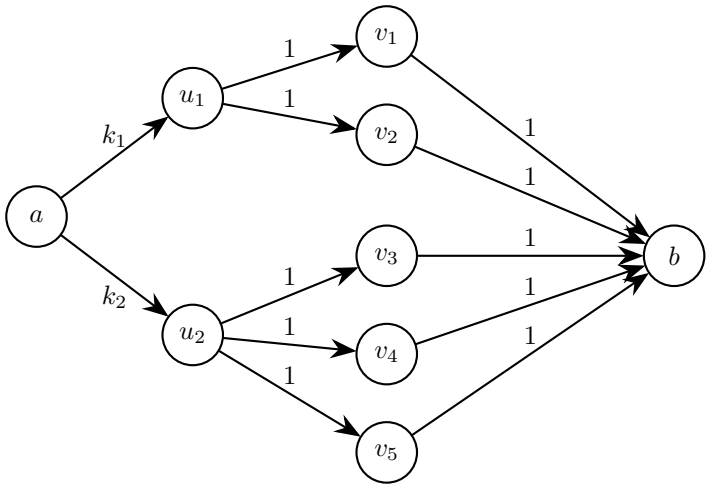


Figure 1 Network Flow Graph Example

p_1, p_2 in a cluster D formed by the algorithm, the following inequality holds:

$$d(p_1, p_2) < \frac{c^2(c^{m-2} + c - 2)}{c^m + c - 2}\gamma,$$

i.e., for any set of points in the refined metric, the upper bound of the distance between any two points is $\frac{c^2(c^{m-2}+c-2)}{c^m+c-2}\gamma$. The detailed proof of Claim 8 appears in Appendix A. Next, we prove Claim 11, which describes the trapezoid relationship for any four points p_1, p_2, p_3, p_4 in the refined metric space:

$$d(p_1, p_2) \leq \frac{1}{c}d(p_1, p_3) + \frac{1}{c^2}d(p_2, p_4) + \frac{1}{c^2}d(p_3, p_4).$$

Using these two claims, we conclude Lemma 12, which states that

$$\left| \bigcup_{p \in D} B(p, \frac{(c-1)c^m}{c^m + c - 2}\gamma) \cap f(\mathcal{S}^*) \leq 1 \right|$$

for every cluster $D \in \mathcal{C}$. The function $f(\mathcal{S}^*)$ is a representative set for the points in the optimal solution set $\mathcal{S}^*$. This lemma essentially means that for all points in a cluster D formed by our algorithm, at most one point from $f(\mathcal{S}^*)$ can exist near the cluster. In other words, integrating the previous two claims ensures that no more than one point from $f(\mathcal{S}^*)$ is close to the same cluster. This allows us to construct our solution set $\mathcal{S}$ to be $f(\mathcal{S}^*)$ by running the network flow algorithm on the flow graph, thereby achieving the desired approximation ratio.

With this lemma, we prove that FairGreedyFlow for Refined Metric Space is a $\left(\frac{c^m+c-2}{(c-1)c^m}\right)(1+\epsilon)$-approximation algorithm in the refined metric space setting. The $(1+\epsilon)$ factor comes from the assumption $\frac{l^*}{1+\epsilon} < \gamma \leq l^*$.

4 $\frac{4}{c^2}$-Approximation Algorithm for $m = 2$

▶ **Theorem 6.** *FairSwap is a $\frac{4}{c^2}$-approximation algorithm for the Fair Max-Min Diversification problem with extended metric factor c when $m = 2$ that runs in time $O(kn)$*

FairSwap, proposed in [8], is a 4-approximation algorithm for metric space that focuses on groups of size two, with the time complexity $O(kn)$. In this section, we show that the FairSwap algorithm achieves a $\frac{4}{c^2}$-approximation in the extended metric space with $c \in (0, 2]$

The algorithm is fundamentally greedy. It begins by applying the GMM algorithm [9] to find an initial solution, treating all groups as identical. Let the intermediate set be $\mathcal{S}$. The solution is then adjusted for fairness by modifying group assignments as needed. Specifically, one of the two groups contains fewer points than required, i.e., there exists an $i \in \{1, 2\}$ such that $|\mathcal{S} \cap \mathcal{U}_i| \leq k_i$. We refer to this as the under-satisfied group, denoted by S_U, where $U \in \{1, 2\}$. The next step is to run the GMM algorithm again, but this time exclusively on group U, selecting a set of points to add to the final solution to ensure that the fairness constraint for U is satisfied. Finally, the over-satisfied group, O, is adjusted by removing the points from $\mathcal{S}_O$ that are closest to the points just added. The pseudo-code Algorithm 2 [9] and Algorithm 3 [8] is provided to help understand the algorithm.

Algorithm 2 GMM Algorithm

Input: $\mathcal{U}$: Universe of available elements
$k \in \mathbb{Z}^+$
I: An initial set of elements
Output: $\mathcal{S} \subseteq \mathcal{U}$ of size k

$\mathcal{S} \leftarrow \emptyset$
if $I = \emptyset$ **then**
  $\mathcal{S} \leftarrow$ a randomly chosen point in $\mathcal{U}$
end if
while $|\mathcal{S}| < k$ **do**
  $x \leftarrow \arg\max_{u \in \mathcal{U}} \min_{s \in \mathcal{S} \cup I} d(u, s)$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{x\}$
end while
return $\mathcal{S}$

Algorithm 3 Fair-Swap

Input: $\mathcal{U}_1$,$\mathcal{U}_2$: Set of points of group 1 and 2
$k_1, k_2 \in \mathbb{Z}^+$
Output: k_i points in $\mathcal{U}_i$ for $i \in \{1, 2\}$

Color-Blind Phase:
$\mathcal{S} \leftarrow \text{GMM}(\mathcal{U}, k, \emptyset)$
$\mathcal{S}_i \leftarrow \mathcal{S} \cap \mathcal{U}_i$ for $i \in \{1, 2\}$
Balancing Phase:
$U \leftarrow \arg\min_i(|\mathcal{S}_i| - k_i)$ ▷ Under-satisfied set
$O \leftarrow 3 - U$ ▷ Over-satisfied set
Compute:
$E \leftarrow \text{GMM}(\mathcal{U}_U, \mathcal{S}_U, k_U - |\mathcal{S}_U|)$
$R \leftarrow \left\{ \arg\min_{e \in S_O} d(x, e) : e \in E \right\}$
return $(S_U \cup E) \cup (S_O \setminus R)$

Running Time Analysis.

The analysis of the running time follows the proof in [8], resulting in a time complexity of $O(kn)$.

Approximation-Factor Analysis.

The notations used in this section slightly differ from those of the previous section. In the previous section, we used $\mathcal{S}^*$ to represent the optimal solution under fairness constraints, whereas in this section, we use $\mathcal{F}^*$ for the same purpose. Here, $\mathcal{S}^*$ denotes the optimal solution **without** any fairness constraints (i.e., the groupless optimal solution). Similarly, while l^* was previously used to denote the optimal diversity with fairness, in this section, l^* refers to the optimal diversity in the groupless setting (i.e., $l^* = \text{div}(\mathcal{S}^*)$), and l^*_{fair} represents the optimal diversity under fairness constraints. For clarity, the definitions of the variables used in this proof are provided below.

- $\mathcal{S}^*$: The set of k points in $\mathcal{U}$ that maximize the diversity when there are no fairness constraints, which means that we assume all the points in the same group and don't care fairness anymore.
- l^*: The diversity of the set $\mathcal{S}^*$, which is: $l^* = div(\mathcal{S}^*)$
- $\mathcal{F}^* = \mathcal{F}^*_1 \cup \mathcal{F}^*_2$: Where $\mathcal{F}^*$ is the optimal solution for the Fair Max-Min diversification, and
$$\mathcal{F}^*_1 = \mathcal{F}^* \cap \mathcal{U}_1, \qquad \mathcal{F}^*_2 = \mathcal{F}^* \cap \mathcal{U}_2$$
- l^*_{fair}: The diversity of the set $\mathcal{F}^*$, which is: $l^*_{\text{fair}} = div(\mathcal{F}^*)$
- $\mathcal{S}, \mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_U, \mathcal{S}_O, E, R, U, O$: Adopt the same meaning that defined in the Algorithm 3.
  - In detail, $\mathcal{S}$ is the intermediate output of the algorithm with no fairness constraints(output for the Color-Blind Phase).
  - $\mathcal{S}_1$ is the points that both in $\mathcal{S}$ and group 1, while $\mathcal{S}_2$ is the points that both in $\mathcal{S}$ and group 2.
  - $\mathcal{S}_U$ is one of the $\mathcal{S}_1$ or $\mathcal{S}_2$ which under satisfy the constraint for the group (which means that we need to add more points from $\mathcal{S}_U$ to satisfy the fairness constraint), while $\mathcal{S}_O$ is the other set which over-satisfied.
  - E is the intermediate output of the GMM algorithm. This algorithm randomly and greedily chooses the points that could be added to $\mathcal{S}_U$ and returns the adding set. R is the set of the points that would be removed from $\mathcal{S}_O$, where the points are greedily selected by traversing all the possibilities.
  - U is the index number for the unsatisfied group, and O is the index for the satisfied group.

We claim that the output set $(\mathcal{S}_U \cup E) \cup (\mathcal{S}_O \setminus R)$ is a $\frac{c^2}{4}$-approximation solution for the Max-Min Fair Diversification.

Proof. First, note that
$$l^* \geq l^*_{\text{fair}}.$$
Since the l^* is the *diversity* of the solution in the groupless setting, then by applying more constraints, the div of the solution $\mathcal{F}^*$ would be worse, in other words, l^*_{fair} is smaller or equal than l^*.

Then note that
$$\forall i \in \mathcal{S}, \quad \left|\{p \in \mathcal{S}^* \mid d(p,i) < \frac{cl^*}{2}\}\right| \leq 1.$$

This means that **for each** point i in $\mathcal{S}$, there's **at most one point** p in $\mathcal{S}^*$ that satisfies: $d(p,i) < \frac{cl^*}{2}$, in other word, there's only one point in $\mathcal{S}^*$, such that distance $< \frac{cl^*}{2}$ from each point of $\mathcal{S}$. We can prove this by contradiction. Assume that there exist at least two points satisfy (2), then we have $\exists i \in \mathcal{S}$, $\exists p, q \in \mathcal{S}^*$, s.t. $d(p,i) < \frac{cl^*}{2}$ and $d(q,i) < \frac{cl^*}{2}$.

Then from triangle inequality and the previous two inequalities, which are

$$c \cdot d(p,q) \leq d(p,i) + d(q,i),$$

$$d(p,i) < \frac{cl^*}{2}, \quad d(q,i) < \frac{cl^*}{2}.$$

We have

$$c \cdot d(p,q) \leq d(p,i) + d(q,i) < \frac{cl^*}{2} + \frac{cl^*}{2} = cl^* \implies d(p,q) < l^*.$$

Since $l^* = div(\mathcal{S}^*)$, which is the Min distance between points in $\mathcal{S}^*$, but now we have points p, q with $d(p,q) < l^*$, which contradicts to the Minimum distance l^*. Then we've proved that for each point in $\mathcal{S}$, there's at most one point in $\mathcal{S}^*$ such that $d(p,i) < \frac{cl^*}{2}$.

Therefore, while GMM has picked less than k elements, in the worst case, each point in $\mathcal{S}$ has 1 corresponding point in $\mathcal{S}^*$ such that the distance between them $< \frac{cl^*}{2}$ (and different points in $\mathcal{S}$ may correspond to the same point in $\mathcal{S}^*$ as well), then we can create a bijective mapping between the points in $\mathcal{S}$ and the points in $\mathcal{S}^*$. By the Pigeon Hole Principle, since $|\mathcal{S}| < |\mathcal{S}^*| = k$, there must exist a good point (in $\mathcal{S}^*$, but we can also choose the point that not in $\mathcal{S}^*$ if it is considered as a better option to current setting) that can be greedily selected and added to $\mathcal{S}$, with distance $\geq \frac{cl^*}{2}$ from all the points already selected. Also with the fact that the algorithm greedily picks the next point, which is the point farthest away, we guarantee that the good point that we mentioned would be chosen if there's no other better option. Then we naturally have

$$div(\mathcal{S}) \geq \frac{cl^*}{2} \geq \frac{cl^*_{\text{fair}}}{2}.$$

Since $\mathcal{S}_U$ is a subset of $\mathcal{S}$, which has less point. We observe that

$$div(\mathcal{S}_U) \geq div(\mathcal{S}) \geq \frac{cl^*_{\text{fair}}}{2}.$$

Then we look at set E, which is the set that includes the points that would be added to $\mathcal{S}_U$ later. Similar to previous reasoning, when we are running GMM with parameter $(\mathcal{U}_U, \mathcal{S}_U, k_U - |\mathcal{S}_U|)$, there is at most one point in $\mathcal{F}^*_U$ that is distance $< \frac{cl^*_{\text{fair}}}{2}$ from each point in $\mathcal{S}_U \cup E$. Formally,

$$\forall i \in \mathcal{S}_U \cup E, \quad \left| \{p \in \mathcal{F}^*_U \mid d(p,i) < \frac{cl^*}{2}\} \right| \leq 1.$$

Similarly, when GMM has picked less than $k - |\mathcal{S}_U|$ elements, there exists at least one element that can be selected with a distance greater or equal to $\frac{cl^*_{\text{fair}}}{2}$ from the points already selected. Since the algorithm picks the next point farthest away from the points already chosen, the next point is at least $\frac{cl^*_{\text{fair}}}{2}$ from the existing points. We naturally have

$$div(\mathcal{S}_U \cup E) \geq \frac{cl^*_{\text{fair}}}{2}.$$

Our output is $(\mathcal{S}_U \cup E) \cup (\mathcal{S}_O \setminus R)$, from the $div(S) \geq \frac{cl^*}{2} \geq \frac{cl^*_{\text{fair}}}{2}$ and $div(S_U \cup E) \geq \frac{cl^*_{\text{fair}}}{2}$, we already proved that $div(\mathcal{S})$ and $div(\mathcal{S}_U \cup E)$ are greater or equal to $\frac{cl^*_{\text{fair}}}{2}$, the only thing

that left for our proof is $div(E \cup \mathcal{S}_O)$. For this case, by the algorithm's logic, we remove the closest point in $\mathcal{S}_O$ from E. Note that by application of the triangle inequality and the fact that $div(\mathcal{S}_O) \geq \frac{cl^*_{\text{fair}}}{2}$, for each $x \in E$ there can be at most one point $y \in \mathcal{S}_O$ such that $d(x, y) < \frac{c^2 l^*_{\text{fair}}}{4}$, and it is obvious that the size of E and R are the same. Hence, after the removal of the closest points the distance between all pairs is as required and we have the Theorem 6. ◀

5 Conclusion

In this paper, we introduced approaches to solving the Fair Max-Min Diversification problem in refined and relaxed metric spaces, extending the existing work on traditional metric spaces. Our primary contribution is a $\frac{c^m+c-2}{(c-1)c^m}$-approximation algorithm for refined metrics, which improves on the standard $m + 1$ approximation by leveraging the strengthened triangle inequality. Additionally, for the special case of $m = 2$, we presented a $\frac{4}{c^2}$-approximation algorithm in extended metric spaces. These results demonstrate that refining the metric space leads to better approximation ratios.

Future Work

In this paper, we employ the Max-Min function to measure diversity, but exploring alternative diversity definitions(e.g. Max-Sum) in the extended metric space could also be an interesting direction. Another promising direction involves enhancing results for relaxed metrics (i.e., when $c < 1$). This could be achieved by adopting existing algorithms for regular metrics or inventing novel ones. For both refined and relaxed metrics, improving bounds where specific algorithm results are not yet tight is also a possible avenue for future work. Furthermore, while [8, 2] demonstrated a negative result for regular Fair Max-Min Diversification, showing that no approximation algorithm can achieve better than a factor of 2 unless P=NP, a possible future direction would be to investigate whether refined metrics lead to weaker negative results due to more restricted choices, or if relaxed metrics result in stronger negative results as the freedom is given. Motivated by our results that relaxing metric space leads to worse ratios and c could be arbitrary small, we conjecture that, without the metric assumption, there's no constant approximation ratio for the Fair Max-Min Diversification problem unless $P = NP$. Finally, while our current work focuses on disjoint groups, exploring extended metrics when groups overlap as in [8] opens another area for future research.

References

1 Zeinab Abbassi, Vahab S. Mirrokni, and Mayur Thakur. Diversity maximization under matroid constraints. In *KDD*, pages 32–40. ACM, 2013.

2 Raghavendra Addanki, Andrew McGregor, Alexandra Meliou, and Zafeiria Moumoulidou. Improved approximation and scalability for fair max-min diversification. In Dan Olteanu and Nils Vortmeier, editors, *25th International Conference on Database Theory, ICDT 2022, March 29 to April 1, 2022, Edinburgh, UK (Virtual Conference)*, volume 220 of *LIPIcs*, pages 7:1–7:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. URL: `https://doi.org/10.4230/LIPIcs.ICDT.2022.7`, `doi:10.4230/LIPICS.ICDT.2022.7`.

3 Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In Jan Van den Bussche and Victor Vianu, editors, *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer, 2001. `doi:10.1007/3-540-44503-X\_27`.

4 Allan Borodin, Aadhar Jain, Hyun Chul Lee, and Yuli Ye. Max-sum diversification, monotone submodular functions, and dynamic updates. *ACM Trans. Algorithms*, 13(3):41:1–41:25, 2017. `doi:10.1145/3086464`.

5 Matteo Ceccarello, Andrea Pietracaprina, and Geppino Pucci. A general coreset-based approach to diversity maximization under matroid constraints. *ACM Trans. Knowl. Discov. Data*, 14(5):60:1–60:27, 2020. `doi:10.1145/3402448`.

6 Marina Drosou, H. V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in big data: A review. *Big Data*, 5(2):73–84, 2017. URL: `https://doi.org/10.1089/big.2016.0054`, `doi:10.1089/BIG.2016.0054`.

7 Yash Kurkure, Miles Shamo, Joseph Wiseman, Sainyam Galhotra, and Stavros Sintos. Faster algorithms for fair max-min diversification in r^d. *Proc. ACM Manag. Data*, 2(3):137, 2024. `doi:10.1145/3654940`.

8 Zafeiria Moumoulidou, Andrew McGregor, and Alexandra Meliou. Diverse data selection under fairness constraints. In Ke Yi and Zhewei Wei, editors, *24th International Conference on Database Theory, ICDT 2021, March 23-26, 2021, Nicosia, Cyprus*, volume 186 of *LIPIcs*, pages 13:1–13:25. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. URL: `https://doi.org/10.4230/LIPIcs.ICDT.2021.13`, `doi:10.4230/LIPICS.ICDT.2021.13`.

9 S. S. Ravi, Daniel J. Rosenkrantz, and Giri Kumar Tayi. Heuristic and special case algorithms for dispersion problems. *Oper. Res.*, 42(2):299–310, 1994. URL: `https://doi.org/10.1287/opre.42.2.299`, `doi:10.1287/OPRE.42.2.299`.

10 Marcin Sydow. Improved approximation guarantee for max sum diversification with parameterised triangle inequality. In Troels Andreasen, Henning Christiansen, Juan Carlos Cubero Talavera, and Zbigniew W. Ras, editors, *Foundations of Intelligent Systems - 21st International Symposium, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014. Proceedings*, volume 8502 of *Lecture Notes in Computer Science*, pages 554–559. Springer, 2014. `doi:10.1007/978-3-319-08326-1\_60`.

A Proof for $\frac{c^m+c-2}{(c-1)c^m}$-Approximation Algorithm

▶ **Theorem 5.** *FairGreedyFlow for Refined Metric Space (Algorithm 1) using an approximation γ for the optimal value is a $\left(\frac{c^m+c-2}{(c-1)c^m}\right)(1+\epsilon)$-approximation algorithm with perfect fairness for the Fair Max-Min Diversification problem with extended metric factor $c \in (1,2]$ that runs in a time of $O(nkm^3\epsilon^{-1}\log n)$.*

In the context of refined metric space($c \in (1,2]$, we first demonstrate that the algorithm produces a valid solution with perfect fairness and that this solution is $\left(\frac{c^m+c-2}{(c-1)c^m}\right)(1+\epsilon)$-

approximation, with the assumption that our guess of $\gamma \leq \frac{l^*}{1+\epsilon}$. Then we will show that the running time of the algorithm is $O(nkm^3\epsilon^{-1}\log n)$.

Approximation-Factor Analysis.

The input, output and variables for the network flow are as defined in Algorithm 1.

Definition of $f(y_i)$: Let $\mathcal{S}^* = \{y_1, \ldots, y_k\}$ be the optimal solution set for the given input. For $i \in [k]$, if y_i is "removed" by the algorithm after forming the cluster D, define $f(y_i) \in D$ to be a point in the cluster D that is in the same group as y_i. If y_i is not "removed", but in the cluster D, then $f(y_i) = y_i$. In addition, We will prove in Lemma 12 that it is not possible to have $f(y_i), f(y_j)$ $(i \neq j)$ in the same cluster D, then for all $i \in [m]$, we can pick corresponding $f(y_i)$ from different clusters to form a solution set with perfect fairness.

We will utilize Lemma 12 to establish Theorem 5. To prove Lemma 12, it is necessary to demonstrate Claim 8 and Claim 11 first.

The assumption on γ is $\frac{l^*}{1+\epsilon} < \gamma \leq l^*$. In the following, we will disregard the $(1+\epsilon)$ factor for l^* and focus on γ to demonstrate the approximation factor, and the $(1+\epsilon)$ term will be reintroduced at the end of the analysis.

▶ Remark 7. For the cluster $D \in \{D_1, \ldots, D_t\}$ formed by the algorithm, it has the following properties:

- D contains at most m points, m is the number of group given in input.
- For any points p in the cluster D, let p' denotes the closest point from p in D other than p itself, $d(p, p') < \frac{(c-1)c^m}{c^m+c-2}\gamma$.

  For simplicity, let us use $x = \frac{(c-1)c^m}{c^m+c-2}\gamma$ in the rest of the section.

▷ Claim 8. For such cluster D with the properties in Remark 7, the distance between any two points in D is less than

$$\left(\sum_{i=1}^{m-2}\frac{1}{c^i} + \frac{1}{c^{m-2}}\right)x = \frac{c^2(c^{m-2}+c-2)}{c^m+c-2}\gamma$$

in the extended metric space with extended factor $c \in (1, 2]$.

Proof. First, we need to figure out that given a set of points, what the shape of the points that the theoretical upper bound for the max distance between the farthest two points reaches is. From the Definition 3, for any three point x, y, z in the metric space with factor c,

$$d(x,z) \leq \frac{1}{c}\left(d(x,y) + d(y,z)\right).$$

To determine the theoretical upper bound for $d(x, z)$, we observe that the maximum value is achieved when the inequality becomes equality. When $c = 1$, it means that the given extended metric space is a regular metric space, and the way points are arranged to achieve this maximum distance is a chain. We can consider the points in the extended metric space to be in a "chain" shape as well. The critical feature that the points reach the theoretical maximum distance bound is the equality $d(x,z) = \frac{1}{c}\left(d(x,y) + d(y,z)\right)$ is satisfied.

Second, because we are working on the extended metric space, the upper bound can be different for the same points setting(shape), we need to figure out what is the tightest upper bound for the max distance between the points in the set. For example, consider a set of 5 points in the extended metric space,

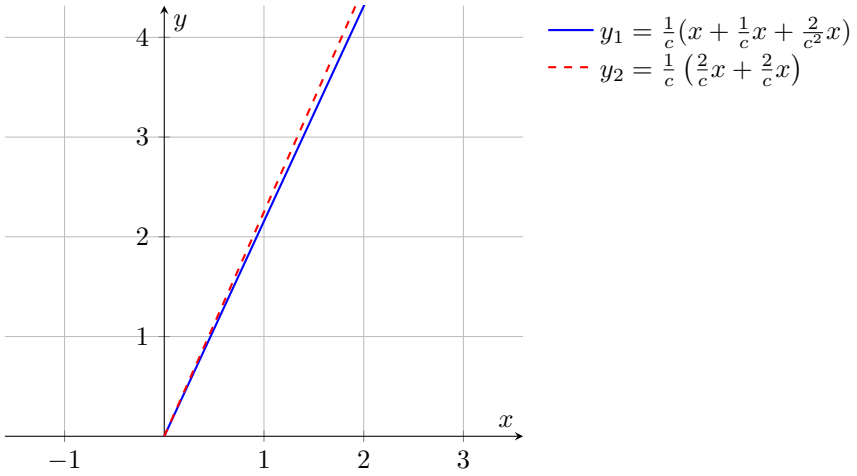
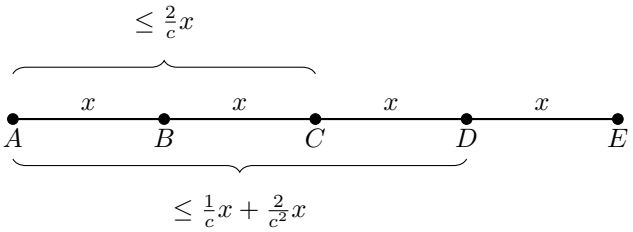


Figure 2 Plot of the two upper bound functions for $d(A, E)$ with $c \in (1, 2)$



where $d(A, B), d(B, C), d(C, D), d(D, E)$ are x. By applying the triangle inequality to points A, B and C, $d(A, C) \leq \frac{1}{c}(d(A, B)+d(B, C)) = \frac{2}{c}x$. Similarly, $d(A, D) \leq \frac{1}{c}\left(d(A, C) + d(C, D)\right) = \frac{1}{c}(x + \frac{2}{c}x) = \frac{1}{c}x + \frac{2}{c^2}x$. However, if we would like to find the maximum length for AE, there are two upper bounds.

$$d(A, E) \leq \frac{1}{c}(d(A, D) + d(D, E)) \quad \text{and} \quad d(A, E) \leq \frac{1}{c}(d(A, C) + d(C, E)).$$

For simplicity, we ignore the case where $d(A, E) \leq \frac{1}{c}(d(A, B) + d(B, E))$, as it is merely a symmetric case of $d(A, E) \leq \frac{1}{c}(d(A, D) + d(D, E))$.

We would like to find the tightest upper bound for $d(A, E)$, we want

$$d(A, E) \leq \min\left\{\frac{1}{c}(d(A, D) + d(D, E)), \frac{1}{c}(d(A, C) + d(C, E))\right\},$$

which are

$$d(A, E) \leq \min\left\{\frac{1}{c}\left(x + \frac{1}{c}x + \frac{2}{c^2}x\right), \frac{1}{c}\left(\frac{2}{c}x + \frac{2}{c}x\right)\right\}.$$

In Figure 2, we plot $y_1 = \frac{1}{c}\left(x + \frac{1}{c}x + \frac{2}{c^2}x\right)$ and $y_2 = \frac{1}{c}\left(\frac{2}{c}x + \frac{2}{c}x\right)$ for some $c \in (1, 2)$. From the figure, it is evident that $y_1 < y_2$, and thus the inequality can be simplified as $d(A, E) \leq \frac{1}{c}\left(x + \frac{1}{c}x + \frac{2}{c^2}x\right)$. Note that in the special case when $c = 2$, $y_1 = y_2$; however, this does not affect the validity of the result above.

▶ **Definition 9.** *Define $\tilde{d}(q)$ to be the tightest bound of the maximum distance for the given set of $q + 1$ points.*

The $\tilde{d}(q)$ function is defined recursively. For $q \geq 2$,

$$\tilde{d}(q) = \begin{cases} x & \text{if } q = 1,\ x = \frac{(c-1)c^m}{c^m+c-2}\gamma \\ \min\left\{\forall r \in \left[\left\lfloor \frac{q}{2} \right\rfloor\right] \mid \frac{1}{c}(\tilde{d}(r) + \tilde{d}(q-r))\right\} & \text{if } q \geq 2 \end{cases}$$

Note that we traverse all the r from 1 to $\left\lfloor \frac{q}{2} \right\rfloor$ rather than 1 to $q-1$, because $\tilde{d}(r)+\tilde{d}(q-r) = \tilde{d}(q-r)+\tilde{d}(r)$.

For example, if we apply $\tilde{d}$ notation onto the set $\{A, B, C, D, E\}$ as the previous example,

Function	Expression
$\tilde{d}(1)$	x
$\tilde{d}(2)$	$\frac{1}{c}(\tilde{d}(1) + \tilde{d}(1)) = \frac{2}{c}x$
$\tilde{d}(3)$	$\frac{1}{c}(\tilde{d}(1) + \tilde{d}(2)) = \frac{1}{c}x + \frac{2}{c^2}x$
$\tilde{d}(4)$	$\min\left\{\frac{1}{c}(\tilde{d}(1) + \tilde{d}(3)), \frac{1}{c}(\tilde{d}(2) + \tilde{d}(2))\right\} = \frac{1}{c}\left(x + \frac{1}{c}x + \frac{2}{c^2}x\right)$

▶ **Definition 10.** *Define $\tilde{d}'(q)$ for cluster D with the properties in Remark 7 recursively:*

$$\tilde{d}'(q) = \begin{cases} x & \text{if } q = 1,\ x = \frac{(c-1)c^m}{c^m+c-2}\gamma \\ \frac{1}{c}(\tilde{d}'(1) + \tilde{d}'(q-1)) & \text{if } q \geq 2 \end{cases}$$

We have introduced all definitions that are needed for the proof. In the following proof, we will find the closed form for $\tilde{d}'(q)$ by induction. Then prove that $\tilde{d}(q) = \tilde{d}'(q)$ to get the closed form for $\tilde{d}(q)$. For the set D that satisfies the requirements in the Claim 8, the distance between the farthest two points in D must be smaller to $\tilde{d}(m-1)$, because D has m points. Since we already know the closed form of $\tilde{d}(q)$, we can plug $q = m-1$ into the closed form, then we will get $\tilde{d}(m-1) = \left(\sum_{i=1}^{m-2} \frac{1}{c^i} + \frac{1}{c^{m-2}}\right) x = \frac{c^2(c^{m-2}+c-2)}{c^m+c-2}\gamma$. Consequently, our Claim 8 follows directly.

We will find the closed form for $\tilde{d}'(q)$ by induction.

Statement:

$$\tilde{d}'(q) = \begin{cases} x & \text{if } q = 1,\ x = \frac{(c-1)c^m}{c^m+c-2}\gamma \\ \left(\sum_{i=1}^{q-1} \frac{1}{c^i} + \frac{1}{c^{q-1}}\right) x & \text{if } q \geq 2 \end{cases}$$

Base Case:

$\tilde{d}'(1) = x.$

$$\tilde{d}'(2) = \left(\sum_{i=1}^{2-1} \frac{1}{c^i} + \frac{1}{c^{2-1}}\right) x = \frac{2}{c}x.$$

The statement holds for $q = 1$ and $q = 2$.

Inductive Step:

Inductive Hypothesis: Assume the statement holds for some arbitrary positive integer $q \geq 2$, i.e.,

$$\tilde{d}'(q) = \left(\sum_{i=1}^{q-1} \frac{1}{c^i} + \frac{1}{c^{q-1}} \right) x.$$

Next, prove that the statement holds for $q+1$:

$$\tilde{d}'(q+1) = \left(\sum_{i=1}^{q} \frac{1}{c^i} + \frac{1}{c^q} \right) x.$$

Using the inductive hypothesis, we have:

$$\begin{aligned}
\tilde{d}'(q+1) &= \frac{1}{c} \left(\tilde{d}'(1) + \tilde{d}'(q) \right) \\
&= \frac{1}{c} \left(x + \left(\sum_{i=1}^{q-1} \frac{1}{c^i} + \frac{1}{c^{q-1}} \right) x \right) \\
&= \frac{1}{c} x + \left(\sum_{i=2}^{q} \frac{1}{c^i} + \frac{1}{c^q} \right) x \\
&= \left(\sum_{i=1}^{q} \frac{1}{c^i} + \frac{1}{c^q} \right) x.
\end{aligned}$$

Therefore,

$$\tilde{d}'(q+1) = \left(\sum_{i=1}^{q} \frac{1}{c^i} + \frac{1}{c^q} \right) x.$$

By the principle of mathematical induction, the statement holds for all $q \geq 1$.

Then we will prove $\tilde{d}(q) = \tilde{d}'(q)$.

Statement:

$$\begin{aligned}
\tilde{d}(q) &= \min \left\{ \forall r \in \left[\left\lfloor \frac{q}{2} \right\rfloor \right] \;\middle|\; \frac{1}{c} (\tilde{d}(r) + \tilde{d}(q-r)) \right\} \\
&= \tilde{d}'(q).
\end{aligned}$$

Base Case:

$$\tilde{d}(1) = \tilde{d}'(1).$$
$$\tilde{d}(2) = \tilde{d}'(2) = \frac{2}{c} x.$$

The statement holds for $q = 1$ and $q = 2$.

Inductive Step:

Inductive Hypothesis 1: Assume the statement holds for all positive integer k such that $2 \leq k \leq q$, i.e.,

$$\begin{aligned}\tilde{d}(k) &= \min\left\{\forall r \in \left[\left\lfloor \frac{k}{2} \right\rfloor\right] \mid \frac{1}{c}(\tilde{d}(r) + \tilde{d}(k-r))\right\} \\ &= \tilde{d}'(k).\end{aligned}$$

Next, prove that the statement holds for $q+1$:

$$\begin{aligned}\tilde{d}(q+1) &= \min\left\{\forall r \in \left[\left\lfloor \frac{q+1}{2} \right\rfloor\right] \mid \frac{1}{c}(\tilde{d}(r) + \tilde{d}(q+1-r))\right\} \\ &= \tilde{d}'(q+1).\end{aligned}$$

To demonstrate that $\tilde{d}(q+1) = \tilde{d}'(q+1)$, it is necessary to establish the following sequence of inequalities:

$$\begin{aligned}\tilde{d}'(q+1) = \frac{1}{c}(\tilde{d}(1) + \tilde{d}(q)) &\leq \frac{1}{c}(\tilde{d}(2) + \tilde{d}(q-1)) \\ &\leq \cdots \leq \frac{1}{c}(\tilde{d}(r) + \tilde{d}(q+1-r)) \leq \frac{1}{c}(\tilde{d}(r+1) + \tilde{d}(q-r)) \\ &\leq \cdots \leq \frac{1}{c}\left(\tilde{d}\left(\left\lfloor \frac{q+1}{2} \right\rfloor\right) + \tilde{d}\left(q - \left\lfloor \frac{q+1}{2} \right\rfloor\right)\right).\end{aligned}$$

To prove this chain of inequalities, another induction must be employed.

The statement for this inner induction is that

$$\frac{1}{c}(\tilde{d}(r) + \tilde{d}(q+1-r)) \leq \frac{1}{c}(\tilde{d}(r+1) + \tilde{d}(q-r))$$

for $r \in \left[\left\lfloor \frac{q+1}{2} \right\rfloor - 1\right]$.

The base case for this inner induction is when $r = 1$. i.e.

$$\frac{1}{c}(\tilde{d}(1) + \tilde{d}(q)) \leq \frac{1}{c}(\tilde{d}(2) + \tilde{d}(q-1)).$$

From the Inductive Hypothesis 1, we know that it is true for all $k \leq q$ that $\tilde{d}(k) = \tilde{d}'(k)$. Therefore, $\tilde{d}(q) = \tilde{d}'(q)$ and $\tilde{d}(q-1) = \tilde{d}'(q-1)$. We can plug in the closed form of $\tilde{d}'(q)$ and $\tilde{d}'(q-1)$ to prove the base case.

$$\left(\sum_{i=1}^{q-1} \frac{1}{c^i} + \frac{1}{c^{q-1}}\right) x.$$

Rearrange the base case inequality, we want to prove $\tilde{d}(2) + \tilde{d}(q-1) - \tilde{d}(1) - \tilde{d}(q) \geq 0$

$$\begin{aligned}
\tilde{d}(2) + \tilde{d}(q-1) - \tilde{d}(1) - \tilde{d}(q) &= \tilde{d}'(2) + \tilde{d}'(q-1) - \tilde{d}'(1) - \tilde{d}'(q) \\
&= \frac{2}{c}x + \left(\sum_{i=1}^{q-2}\frac{1}{c^i} + \frac{1}{c^{q-2}}\right)x - x - \left(\sum_{i=1}^{q-1}\frac{1}{c^i} + \frac{1}{c^{q-1}}\right)x \\
&= \left(\frac{2}{c} - 1 + \sum_{i=1}^{q-2}\frac{1}{c^i} + \frac{1}{c^{q-2}} - \sum_{i=1}^{q-1}\frac{1}{c^i} - \frac{1}{c^{q-1}}\right)x \\
&= \left(\frac{2}{c} - 1 + \frac{1}{c^{q-2}} - \frac{2}{c^{q-1}}\right)x \\
&= \frac{2c^{q-2} - c^{q-1} + c - 2}{c^{q-1}}x \\
&= \frac{c^{q-2}(2-c) - (2-c)}{c^{q-1}}x \\
&= \frac{(c^{q-2}-1)(2-c)}{c^{q-1}}x \geq 0 \quad \text{, for } c \in (1,2].
\end{aligned}$$

Therefore, the base case for the inner induction is true.

Assume $\tilde{d}(r) + \tilde{d}(q+1-r) \leq \tilde{d}(r+1) + \tilde{d}(q-r)$ is true for some $r \in \left[\left\lfloor \frac{q+1}{2} \right\rfloor - 2\right]$, then we will show $\tilde{d}(r+1) + \tilde{d}(q-r) \leq \tilde{d}(r+2) + \tilde{d}(q-r-1)$

Rearrange the inequality, we want to show:

$\tilde{d}(r+2) + \tilde{d}(q-r-1) - \tilde{d}(r+1) - \tilde{d}(q-r) \geq 0$

Prove the above inequality:

$$\begin{aligned}
\text{Left Side} &= \tilde{d}(r+2) + \tilde{d}(q-r-1) - \tilde{d}(r+1) - \tilde{d}(q-r) \\
&= \left(\sum_{i=1}^{r+1}\frac{1}{c^i} + \frac{1}{c^{r+1}}\right)x + \left(\sum_{i=1}^{q-r-2}\frac{1}{c^i} + \frac{1}{c^{q-r-2}}\right)x - \left(\sum_{i=1}^{r}\frac{1}{c^i} + \frac{1}{c^r}\right)x - \left(\sum_{i=1}^{q-r-1}\frac{1}{c^i} + \frac{1}{c^{q-r-1}}\right)x \\
&= \left(\frac{2}{c^{r+1}} - \frac{1}{c^r} - \frac{2}{c^{q-r-1}} + \frac{1}{c^{q-r-2}}\right)x \\
&= \left(\frac{2-c}{c^{r+1}} - \frac{2-c}{c^{q-r-1}}\right)x \\
&= (2-c)\left(\frac{1}{c^{r+1}} - \frac{1}{c^{q-r-1}}\right)x \\
&= \frac{(2-c)(c^{q-2r-2}-1)}{c^{q-r-1}}x. \qquad \text{(Because } r+1 < r+2 \leq q-r-1\text{)}
\end{aligned}$$

Since $c \in (1,2]$, $\frac{(2-c)(c^{q-2r-2}-1)}{c^{q-r-1}}x \geq 0$. Then we have proved that $\tilde{d}(q) = \tilde{d}'(q)$. ◀

From the definition of D, we know that the distance between the farthest two points in D is smaller than $\tilde{d}(m-1)$ as D has at most m points.

$$\tilde{d}(m-1) = \left(\sum_{i=1}^{m-2}\frac{1}{c^i} + \frac{1}{c^{m-2}}\right)x \qquad (x = \tfrac{(c-1)c^m}{c^m+c-2}\gamma)$$

▷ Claim 11. Given 4 points(A, B, C, D) in the extended metric space with factor $c \in (1,2]$. Without loss of the generality,

$$d(A,B) \leq \frac{1}{c}d(A,C) + \frac{1}{c^2}d(B,D) + \frac{1}{c^2}d(C,D).$$

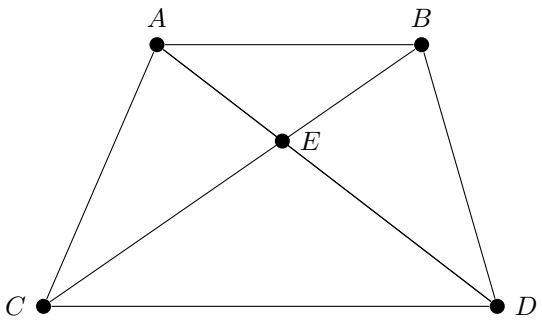
Proof.

$$c \cdot d(A,B) \le d(A,C) + d(B,C) \implies d(A,B) \le \frac{1}{c}(d(A,C) + d(B,C))$$

(Divide c in the both side)

$$\implies d(A,B) \le \frac{1}{c}(d(A,C) + \frac{1}{c}(d(B,D) + d(C,D))).$$

(Replace $d(B,C)$ by $\frac{1}{c}(d(B,D) + d(C,D))$)



Note: The figure above illustrates one particular configuration of the four points discussed. Please note that while this example helps in visualizing the proof, the arguments presented are applicable to any generic arrangement of these points. Thus, the generality of the theorem is not restricted to the scenario depicted here. ◀

▶ **Lemma 12.** $\left|\bigcup_{p\in D} B(p, \frac{(c-1)c^m}{c^m+c-2}\gamma) \cap f(\mathcal{S}^*) \le 1\right|$ *for every cluster* $D \in \mathcal{C}$.

Proof. Let's prove it by contradiction. Without loss of generality, we can assume that there exists $f(y_1)$ and $f(y_2)$, where $y_1, y_2 \in S^*$ and $y_1 \neq y_2$, such that

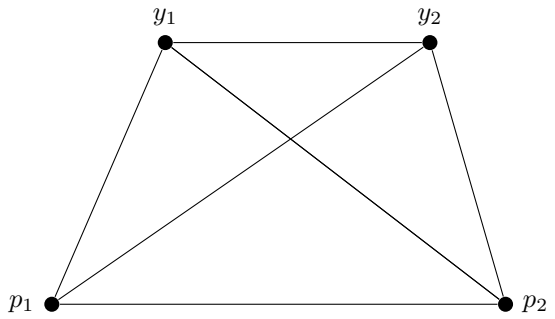
$$f(y_1), f(y_2) \in \bigcup_{p\in D} B\left(p, \frac{(c-1)c^m}{c^m+c-2}\gamma\right).$$

From the construction, we know that there exists at least one corresponding point for $f(y_1)$ and $f(y_2)$ respectively in D, call them p_1 and p_2, such that $d(f(y_1), p_1) < \frac{(c-1)c^m}{c^m+c-2}\gamma$ and $d(f(y_2), p_2) < \frac{(c-1)c^m}{c^m+c-2}\gamma$.

Since two points from the same group are not allowed to be in the same cluster, and we have m groups in total, from the algorithm design, we know that there are at most m points in the cluster D. As we already get the upper bound for the maximum distance in D in the

Claim 8, combined with the Claim 11, we can infer

$$
\begin{aligned}
d(y_1, y_2) &\leq \frac{1}{c} d(y_1, p_1) + \frac{1}{c^2} d(y_2, p_2) + \frac{1}{c^2} d(p_1, p_2) \\
&\quad \text{(By definition, } d(y_1, p_1), d(y_2, p_2) < x. \text{ By Claim 8, } d(p_1, p_2) < \left(\textstyle\sum_{i=1}^{m-2} \frac{1}{c^i} + \frac{1}{c^{m-2}} \right) x) \\
&< \frac{1}{c} x + \frac{1}{c^2} x + \frac{1}{c^2} \left(\sum_{i=1}^{m-2} \frac{1}{c^i} + \frac{1}{c^{m-2}} \right) x \\
&\leq (\sum_{i=1}^{m} \frac{1}{c^i} + \frac{1}{c^m}) x \\
&\leq (\frac{\frac{1}{c}(1 - \frac{1}{c^m})}{1 - \frac{1}{c}} + \frac{1}{c^m}) x \\
&\leq (\frac{\frac{1}{c}(1 - \frac{1}{c^m})}{1 - \frac{1}{c}} + \frac{1}{c^m}) \cdot \frac{(c-1)c^m}{c^m + c - 2} \gamma \\
&\leq \gamma.
\end{aligned}
$$



Then $d(y_1, y_2) < \gamma$. By definition, $y_1, y_2 \in \mathcal{S}^*$, then $d(y_1, y_2) \geq \gamma$, which leads to the contradiction. We've proved that $\left| \bigcup_{p \in D} B(p, \frac{(c-1)c^m}{c^m + c - 2} \gamma) \cap f(\mathcal{S}^*) \leq 1 \right|$ for every cluster $D \in \mathcal{C}$. ◀

Then we can use Lemma 12 to establish the Theorem 5.

It is quite obvious that the $\frac{c^m + c - 2}{(c-1)c^m}$-approximation solution exists if and only if we can find a valid flow with size k in the graph that we construct by Algorithm 1 (Figure 1). Because the distance between clusters $D_1, D_2, \cdots, D_t$ are at least $\frac{(c-1)c^m}{c^m + c - 2} \gamma$, the solution $\mathcal{S}$ given by the algorithm is a $\frac{c^m + c - 2}{(c-1)c^m}$-approximation solution set. Then, the next step is to show that there must exist a solution set $\mathcal{S}$ s.t.

$$
|\mathcal{S} \cap D_i| \leq 1 \text{ for every } D_i \in D_1, \cdots, D_t \text{ and } |\mathcal{S} \cap \mathcal{U}_j| = k_j \text{ for } j \in [m].
$$

In other words, we need to show that it is possible to form a $\mathcal{S}$ that satisfies the fairness constraints by selecting at most one point from each cluster D_i.

Note that we have all points in $f(\mathcal{S}^*)$ in $D_1, D_2, \cdots, D_t$ formed by Algorithm 1. We can justify this by considering two cases. First case, if y_i in $\mathcal{R}$ (Remaining points set) and we select it into the cluster D_j, then it is good. Otherwise, if y_i is not in the $\mathcal{R}$, it must be removed by some $f(y_i)$ that would be in the same cluster as y_i, which means that $f(y_i)$ exists in a formed cluster. It is not possible that the point y_i is removed by some point y_j or $f(y_j)$ or any points in the cluster of $f(y_j)$, because of Lemma 12. Therefore, any points

in $f(\mathcal{S}^*)$ each belongs to a distinct cluster, which means that we can form a solution set $\mathcal{S} = f(\mathcal{S}^*)$ by selecting at most one point from each cluster.

By our assumption that $l^* \leq \gamma \leq \frac{l^*}{1+\epsilon}$, we know that $\forall x, y \in \mathcal{S}$, $d(x,y) \geq \frac{(c-1)c^m}{c^m+c-2}\gamma = \frac{(c-1)c^m}{(c^m+c-2)}\frac{l^*}{(1+\epsilon)}$.

Then we've proved FairGreedyFlow for Refined Metric Space(Algorithm 1) is a $\frac{c^m+c-2}{(c-1)c^m}(1+\epsilon)$-approximation algorithm with perfect fairness for the Fair Max-Min Diversification problem with extended metric factor $c \in (1,2]$.

Running Time Analysis.

The running time for this extended metric space version algorithm doesn't change the time complexity from the original version of the algorithm proposed by [2] in metric space, then the running time for *FairGreedyFlow for Refined Metric Space* algorithm is $O(nkm^3\epsilon^{-1}\log n)$.

B The Range of Distortion Factor c

We cannot have $c > 2$ for the factor of extended metric space. In addition, if $c = 2$, it enforces that all non-zero distances in such metric to become identical. Notice that the problem of interest would be trivial if the distances are identical. We will give proof for c's range first and then show that the distances are identical when $c = 2$.

The range of c is $(0, 2]$.

Proof. Let $c \in \mathbb{R}^+$

From the definition, we have

$$\begin{aligned} & x + y \geq c \cdot z, \;\; y + z \geq c \cdot x, \;\; x + z \geq c \cdot y \\ \implies & x + y \geq c \cdot (c \cdot x - y) \\ \implies & x + y \geq c^2 x - c \cdot y \\ \implies & (1 + c)y \geq (c^2 - 1)x. \end{aligned}$$

By symmetry, $(1 + c)x \geq (c^2 - 1)y$.

Rearrange the above inequalities:

$$\left(\frac{1+c}{c^2-1}x\right) \geq y,$$

$$\left(\frac{1+c}{c^2-1}y\right) \geq x.$$

We need to have $\frac{1+c}{c^2-1} \geq 1$ to make both possible, regardless of the value of x, y. Then

$$1 + c \geq c^2 - 1 \implies c \leq 2.$$

Then we reach our conclusion where 2 is the maximum value for the c. ◀

Identical distance metric sapce when $c = 2$.

Proof. Assume that $c = 2$, the distances between three points are x, y, z.

From this setup, we have $x + y \geq 2z, \;\; x + z \geq 2y, \;\; y + z \geq 2x$.

Then $x + y \geq 2z \geq 2(2y - x) \implies x + y \geq (4y - 2x) \implies 3x \geq 3y$.

Similarly, we get $3y \geq 3x$, then we have $x = y$. By symmetry, $x = y = z$. The distances in this metric are identical. ◀