ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Improved Fairlet Decomposition for Fair Correlation Clustering

Vacharapat Mettanant[1], Adisak Supeesun[2] and Jittat Fakcharoenphol[3]

ABSTRACT

We consider fairness in correlation clustering, a fundamental clustering problem when the similarity between data points is represented as graphs. The fairness conditions we focus on guarantee that the resulting clusters satisfy the required proportion of population groups, modeled as vertex colors. Recently, Ahmadian, Epasto, Kumar, and Mahdian considered fairness constraints parameterized by α, the maximum fraction of points in a cluster having the same color. They addressed fairness constraints in correlation clustering using the fairlet decomposition framework and solved the fairlet decomposition problem for many fairness constraints α using optimization problems on median costs. This paper gives algorithms for the median-cost problem with better approximation ratios for various cases, to obtain improvements on the approximation ratios for the correlation clustering problem from 256 to 174 when $\alpha = 1/2$ and from $16.48C^2$ to $40.96C+4.12$ when $\alpha = 1/C$, where C is the number of colors. We also consider the notion of proportional fairness, where there are only two colors with different population sizes. We give an approximation algorithm with provable approximation ratios, and provide connections to a star-covering problem.

Article information:

Keywords: Fairness, Correlation Clustering, Fairlet Decomposition

Article history:
Received: June 9, 2022
Revised: October 9, 2022
Accepted: December 23, 2022
Published: March 18, 2023 (Online)

DOI: 10.37936/ecti-cit.2023171.248695

1. INTRODUCTION

Data clustering is a fundamental unsupervised machine learning problem. The goal of data clustering is to assign data points into groups based on various distance measures between points. Typical clustering applications solely work under objective clustering costs. However, in many cases, social requirements must also be prioritized. Therefore, as in other areas of artificial intelligence, recent focus has been on ethical issues such as privacy [1-4], fairness [1, 5-7], and transparency [8-10]. When forming groups for democratic discussion physical meetings, one might want to include people who live close to each other in the same group. Using the standard clustering approach, we might end up with highly homogeneous groups with no interest in problems from other groups. To have meaningful discussions, each group should also have minority groups. This leads to the area of significantly interesting research called fair clustering [11-20].

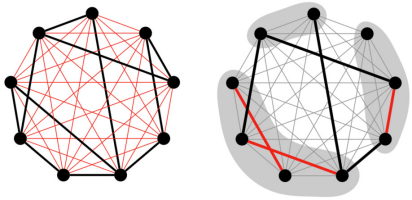


Fig. 1: *An example of correlation clustering: the graph instance (left) shows +1 edges (black) and -1 edges (red), and the clustering (right) shows disagreements.*

In this paper, we study a clustering problem with fairness requirements when relationships between data points are specified as a graph. In this section, we give a short overview of correlation clustering, a clustering problem on graphs, and fairness

[1,2] The authors are with Department of Computer Engineering, Kasetsart University, Sriracha Campus, Chonburi, Thailand., E-mail: vacharapat@eng.src.ku.ac.th and adisak@eng.src.ku.ac.th
[3] The author is with Department of Computer Engineering, Kasetsart University, Bangkok, Thailand., E-mail: jittat@gmail.com
[3] Corresponding author.

constraints in clustering. We summarize our contribution at the end of this section.

1.1 Correlation Clustering

Typical clustering problems work directly with data point representations (usually as a set of points in high dimensional spaces $\mathbb{R}^d$). However, situations where we only have access to information of the similarity between data points arise naturally when we consider network data such as social networks, interaction networks (such as academic citation networks), and financial transaction networks (see, e.g., [21-23]).

Even when data point representations are available, to partition the data points, we may choose to classify the relationship between data points first, using machine learning tools, then work with the resulting similarity relationships to cluster data points. This two-step process decouples the clustering problem from similarity classification and allows the usage of strong classification tools (such as deep neural networks) as a preprocessing step (see, e.g., [24]).

These two settings motivate correlation clustering, a well-known graph clustering problem. In this problem, first introduced by Bansal, Blum, and Chawla [25], a complete undirected graph $G = (V, E)$ is given with an edge label function $\sigma(e) = \sigma(u, v) \in \{+1, -1\}$ for each edge $e = (u, v) \in E$, suggesting that u and v should be in the same cluster if $\sigma(u, v) = +1$, and different clusters if $\sigma(u, v) = -1$. We want to partition all vertices in V into clusters, based on this information.

From the view of optimization, two objectives come out naturally: maximizing agreements and minimizing disagreements. We can measure agreements of a clustering as the number of positive edges whose endpoints are in the same cluster plus the number of negative edges whose endpoints are in different clusters. The other edges are counted as disagreements. Figure 1 shows an example of the correlation clustering problem and its feasible solution. In this figure, the left figure shows an instance as a graph, where the black lines and red lines refer to the edges with labels $+1$ and -1 respectively. The right figure shows a clustering, represented in grey. The thick lines represent all the edges that disagreed with the clustering. This clustering has seven disagreements.

In this work, we are interested in disagreement minimization. The reason for this choice is that when an instance contains many edges, although one can obtain a factor of 2 approximation for the agreements simply via randomization [25], one may still disagree about 50% of the edges; therefore, it is more meaningful to optimize for the minimum number of disagreements. We discuss more results in Section 1.4.

As noted in [25], one attractive advantage of correlation clustering over standard k-mean or k-median clustering is that there is no need to parameterize the problem with the number k of clusters, as the objective deals solely with agreements or disagreements.

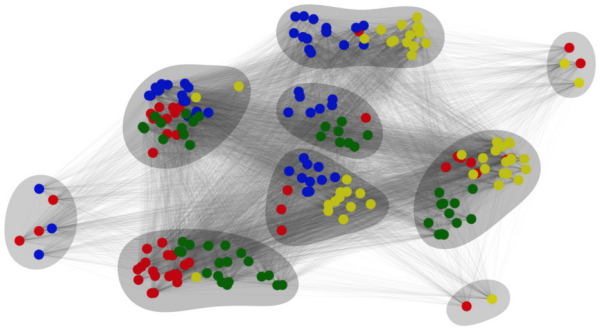


Fig.2: *An example of fair correlation clustering of* `reuters` *dataset from UCI when* α*=1/2. The graph shows only -1 edges.*

1.2 Fair Clustering and fairlet decomposition

The problem of fair clustering has been raised by Chierichetti, Kumar, Lattanzi, and Vassilvitskii [11], who formulated the k-center and k-median problems under the *disparate impact doctrine*, where each data point has a color, either red or blue, that represents its protected attribute, and the fairness criteria ensures that proportion of each color in each cluster is the same as in the entire set. We note that, unlike privacy where major definitions of private data analysis have been largely resolved, when considering problems related to fairness, it is crucial to state clearly the fairness doctrine used as there are many natural definitions, and a few are also in conflict with each other.

Chierichetti *et al.* [11] also introduced the powerful framework of *fairlet decomposition*, where input points to the original clustering problem are decomposed into smaller units that satisfy the fairness condition before passing that units into standard "unconstrained" clustering algorithms to produce required fair clustering.

For correlation clustering, Ahmadian, Epasto, Kumar, and Mahdian [13] defined the fair version of the unweighted correlation clustering and presented approximation algorithms for many interesting cases. They also generalized the fairness condition to include C colors with a parameter α, the maximum fraction of points in a cluster having the same color. This can be seen as a relaxation of the proportional constraint in [11], where no cluster is dominated by a single color.

Figure 2 shows an example of such a fair correlation clustering of samples from reuters dataset from the UCI Repository, when $\alpha = 1/2$. The clustering is constructed by our algorithm as one of the experiments (see details in Section 3.4). This clustering guarantees that each cluster is not dominated by any color, i.e., it contains no vertices of any particular

color for more than half of its size.

As in [11], the clustering is constructed based on the fairlet decomposition. However, since the problem is graph-based, the algorithm does not have explicit metric structures to work with. Ahmadian *et al.* [13] proceeded to show clustering cost reduction from the original instance to the clustering of fairlets (in the decomposition), i.e., they showed that given an η-approximation algorithm for the minimum cost fairlet decomposition and a β-approximation algorithm for the unconstrained correlation clustering algorithm, they can combine them to obtain a $(\beta(1+\eta)+\eta)$-approximation algorithm for fair correlation clustering problem.

To find a good fairlet decomposition, Ahmadian *et al.* [13] showed a reduction to the median-cost problem (defined in Section 2.3) which can be approximated using a variety of matching algorithms. We note that in this step, the reduction overhead in [13] depends on the maximum size f of the fairlets, which depends essentially on the fairness parameter α. For the fair correlation clustering, they presented a 256-approximation algorithm when $\alpha = 1/2$ for any $C \geq 2$, and a $(16.48C^2)$-approximation algorithm when $\alpha = 1/C$, i.e., when each cluster should have the same number of vertices from each color. For general parameter $\alpha = 1/t$, they only provided a reduction to show that if there is a γ-approximation algorithm for the fairlet decomposition, one would have an $O(t\gamma)$-approximation algorithm for the fair correlation clustering problem.

We note that very recently, Ahmadian and Negahbani [26] considered the same settings where small additive violations of fairness constraints are allowed; in that case, they gave improved approximation algorithms (see more discussion in Section 1.4).

The seminal fairness work of Chierichetti *et al.* [11] also considers the case where there are two colors, and the ratio between colors in the graph is 1:m. This problem is referred to as the *proportional fairness* problem. For correlation clustering, Ahmadi, Galhotra, Saha, and Schwartz [27] gave an $O(m^2)$ approximation for this case. Their algorithm can be extended to the case with more than two colors.

1.3 Overview of our contribution

In this paper, we focus on fairlet decomposition procedures. More specifically, we improve the approximation ratio for the fairlet decomposition median cost problem (defined in Section 2.3). Our contributions to the fair correlation clustering are the following.

- When $\alpha = 1/2$, we slightly improve the approximation ratio for the fairlet decomposition median-cost problem from 2 to 4/3, implying an improvement on the overall ratio from 256 to 174.
- When $\alpha = 1/C$ for C colors, we improve the approximation ratio for the median-cost problem from C to 2, resulting in the approximation ratio of $40.96C + 4.12$ instead of $16.48C^2$.
- We also consider the proportional fairness problem with two colors. We prove a connection between this problem and a capacitated star-cover problem. We also show a ξ-approximation algorithm where $\xi = O(\min\{m^2, m \log n\})$, slightly improving the bound of Ahmadi *et al.* [27] when $\log n < m$.

To obtain improvements in approximation ratios, we carefully analyze how Ahmadian *et al.* [13] construct the decompositions. When $\alpha = 1/2$, we show how to partition a cycle into three sets of short path covers satisfying the color constraint and the capacity constraint. When $\alpha = 1/C$, we use averaging argument to show the approximation ratio when choosing the color that minimizes the total cost as centers for computing the matchings. For proportional fairness, our reduction uses Hall's marriage theorem to establish the cost for the feasible solution of a matching.

Section 1.4 reviews the pioneering work of Chierichetti, Kumar, Lattanzi, and Vassilvitskii [11] that defines fairlet decomposition and related work. We give problem statements and definitions of fair correlation clustering and fairlet decomposition in Section 2. Section 3 describes our main contribution. We consider the case where $\alpha = 1/2$ in Section 3.1 and where $\alpha = 1/C$ in Section 3.2. We perform preliminary experiments to verify our improvements in Section 3.4. Finally, Section 4 focuses on proportional fairness with two colors with a population ratio of $1 : m$.

1.4 Related works

Correlation clustering [25] has been studied widely. When edges are unweighted, Chawla, Makarychev, and Schramm [28] presented the best-known 2.06-approximation algorithm that minimizes disagreement, improving on a 2.5-approximation algorithm of Ailon, Charikar, and Newman [29]. When edges have weights, Demaine, Emanuel, Fiat, and Immorlica [30] presented an $O(\log n)$-approximation algorithm.

Chierichetti, Kumar, Lattanzi, and Vassilvitskii [11] formulated the k-center and k-median problems under the *disparate impact doctrine*. Given a set X of points in some metric space, the goal is to partition X into k different clusters. For the fairness condition, each point has a *color*, either red or blue, that represents its protected attribute, and the fairness criteria ensures an appropriate proportion of each color in each cluster.

They introduced the powerful framework of fairlet decomposition that breaks up the clustering problem into two steps: the decomposition step that deals with fairness constraints and the second step that finds the unconstrained clustering. They showed that finding an optimum fairlet decomposition is NP-hard and presented a 4-approximation algorithm for the k-fair center problem and a $(t' + 1 + \sqrt{3} + \epsilon)$-approximation

algorithm for the k-median problem, where $t' \leq 1$ represents color balance constraint, i.e., the ratio between the number of points with minority color and the majority color in each group. The running time of the k-median fair clustering algorithm has been improved by Backurs, Indyk, Onak, Schieber, Vakilian, and Wagner [14] to run in near-linear time. Ahmadian, Epasto, Kumar, and Mahdian [12] considered the k-center clustering *without over-representation*. They parameterized the fairness requirement with α, the maximum fraction of points in a cluster having the same color, leading to the concept of α-fair, which can be seen as a relaxation of the fairness constraint in [11]. They gave a combinatorial approximation algorithm for $\alpha = 1/2$ and an LP-based approximation for general α.

Fairness in correlation clustering was first studied by Ahmadian, Epasto, Kumar, and Mahdian [13] and Ahmadi, Galhotra, Saha, and Schwartz [27]. With a small violation of the fairness constraints allowable, Ahmadian and Negahbani [26] recently improved the approximation guarantees from both [13] and [27]. Their result generalizes so that each color i can have a different fairness parameter α_i. For any $\epsilon > 0$, their algorithm, based on LP relaxation, gives a clustering such that each cluster is either a one-node cluster or a cluster with at most $(1+\epsilon)\alpha_i$ fraction for each color i with an approximation factor of $O(\frac{1}{\epsilon \min_i \alpha_i})$. The case where $\alpha_i = 1/2$ is the same as the first case we consider in Section 3.1 and [26] obtained a better bound *with* a possible small fairness violation. However, we note that the special case when all $\alpha_i = 1/C$ is the same as the second case that we consider in Section 3.2, and we obtain the same bound *without* the violation.

Instead of optimizing for clustering cost under the fairness condition, Esmaeili, Brubach, Srinivasan, and Dickerson [18] studied the problem under the condition that the fairness overhead is fixed and focused on optimizing various fairness violation costs across color groups. The objectives they considered are group-utilitarian (i.e., minimizing the total violation costs), group-egalitarian (i.e., minimizing the maximum violation costs), and group-leximin (where the worst violation cost is minimized, then the second-worst violation cost is minimized, and the third-worst violation cost is minimized and so on). They showed hardness results and gave bi-criteria approximation algorithms for the first two objectives.

We remark that there are other definitions of fair clustering described in the literature. We list a few of them here. Mahabadi and Vakilian [31] (and also [32, 33]) studied fair clustering from the perspective of individual fairness. Their constraint of fairness was developed upon the work of Jung, Kannan, and Lutz [34], which considered fairness in the problem of facility location. Abbasi, Bhaskara, and Venkatasubramanian [35] and, independently, Ghadiri, Samadi, and Vempala [36] focused mainly on how cluster centers represent other points in their clusters. (See also improvements by [37, 38].)

2. DEFINITIONS AND FRAMEWORK

This section describes the fair correlation clustering problem and the fairlet decomposition framework first used by Chierichetti, Kumar, Lattanzi, and Vassilvitskii [11] for the k-median and k-center problems. In section 2.2, we give an overview of how Ahmadian, Epasto, Kumar, and Mahdian [13] adapted this framework for the fair correlation clustering problem.

2.1 Fair correlation clustering

We consider fair correlation clustering. In the correlation clustering problem, we are given a complete graph $G = (V, E)$ with edge labeling $\sigma : E \to \mathcal{R}$, and we refer to set $E^+ = \{e : \sigma(e) > 0\}$ as positive edges and set $E^- = \{e : \sigma(e) < 0\}$ as negative edges. A clustering $\mathbf{C}=\{\mathcal{C}_1,\mathcal{C}_2,\ldots,\mathcal{C}_{|\mathbf{C}|}\}$ is a partition of V where $\mathcal{C}_i$'s are disjoint, i.e., each $\mathcal{C}_i$ is a non-empty subset of V, and each vertex in V belongs to exactly one $\mathcal{C}_i$. The goal is to exclude negative edges inside a cluster and to exclude positive edges between clusters; thus, the cost for clustering $\mathbf{C}=\{\mathcal{C}_1,\mathcal{C}_2,\ldots,\mathcal{C}_{|\mathbf{C}|}\}$ is defined as

$$\mathrm{COST}(G,\mathbf{C}) = \sum_{e \in \mathrm{intra}(\mathbf{C}) \bigcap \mathbf{E}^-} |\sigma(e)| + \sum_{e \in \mathrm{inter}(\mathbf{C}) \bigcap \mathbf{E}^+} |\sigma(e)|,$$

where $\mathrm{intra}(\mathbf{C})$ are edges inside the clustering, i.e., $\mathrm{intra}(\mathbf{C}) = \bigcup_{\mathcal{C} \in \mathbf{C}} \{(u,v) \in E : u, v \in \mathcal{C}\}$ and $\mathrm{inter}(\mathbf{C}) = E \backslash \mathrm{intra}(\mathbf{C})$. In this paper, we focus on the unweighted correlation clustering where $\sigma(e) \in \{+1, -1\}$ for each $e \in E$. We note that this is a version that minimizes disagreement. There is a version of correlation clustering that maximizes agreement and a simple randomized approach gives a factor of 2 approximation even with the fairness constraints [13].

To describe fairness constraints under the disparate impact doctrine as used in [11], each vertex $v \in V$ has a color $c(v) \in \{1, 2, \ldots, C\}$. For a given $\alpha \in (0, 1)$, the fairness condition given by Ahmadian *et al.* [13] states that the number of vertices of each color should not be greater than α-fraction of the total number of vertices in each cluster. They consider two cases:

- when $\alpha = 1/2$, where no color dominates any clusters, and
- when $\alpha = 1/C$, where every color shares the same fraction in each cluster.

In this paper, we also briefly consider the *proportional fairness* as in Chierichetti *et al.* [11], where we require that the number of vertices of each color

in each cluster should be proportional to the corresponding number in the graph.

2.2 Fairlet decomposition framework

Chierichetti, Kumar, Lattanzi, and Vassilvitskii [11] introduced the fairlet decomposition as a framework for fair clustering. Under this framework, the first step is to construct a fairlet decomposition of the problem instance where the fairness constraint is completely satisfied in each fairlet. Then one can use any clustering methods to find a good clustering of these fairlets to obtain the clustering of the original problem. When the cost of the fairlet decomposition is defined appropriately, a good decomposition implies a good fair clustering. Chierichetti *et al.* [11] defined the fairlet decomposition problem for the k-median and the k-center problems.

In the fairlet decomposition problem, given a metric space M with a distance function d, the input consists of a set of vertices $V \subseteq M$, where each vertex $u \in V$ has a color $c(u) \in \{1, \ldots, C\}$. We can consider the input as a complete graph $G = (V, E)$ where each edge $(u, v) \in E$ has a distance $d(u, v)$ that satisfies the metric constraint, and each vertex has a specific color.

Recall that a partition of V is a collection of disjoint subsets of V such that the union of all these sets is V. We call a partition $\mathcal{P}=\{P_1, \ldots, P_k\}$ of V a *fairlet decomposition* of V under some constraint when each fairlet $P_i \in \mathcal{P}$ satisfies the constraint, and no fairlet can be decomposed to smaller fairlets satisfying the constraint. The definition of [11] only works for clustering problems in metric spaces.

For the fair correlation clustering, which works on graphs, Ahmadian *et al.* [13] showed how this framework can be applied and gives an approximation algorithm for the fairlet decomposition via a reduction to a median cost problem (defined below). Considering an unweighted correlation clustering instance $\mathcal{G}$ with a particular fairness constraint, we let $\mathcal{F}$ be a family of all possible clusters satisfying this constraint, referred to later as feasible clusters. We assume, as in [13], that $\mathcal{F}$ has composability property, i.e., if F_1,$F_2 \in \mathcal{F}$, $F_1 \cup F_2$ is also in $\mathcal{F}$. Given $\mathcal{F}$, a *fairlet decomposition* for a correlation clustering is a partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of V into subsets in $\mathcal{F}$. The cost of the decomposition should capture the clustering cost incurred later. Ahmadian *et al.* [13] defined the decomposition cost scheme FCOST below and proved the following main theorem.

For each fairlet P_i, let

$$\text{FCOST}^{in}(P_i) = |E^- \cap \text{intra}(P_i)|,$$

and for fairlets P_i and P_j, let

$$\text{FCOST}^{out}(P_i, P_j) = \min(|E^-(P_i, P_j)|, |E^+(P_i, P_j)|).$$

I.e., $\text{FCOST}^{in}(P_i)$ is the number of negative edges inside P_i, and $\text{FCOST}^{out}(P_i, P_j)$ is the number of edges between them with the minority sign. For a decomposition $\mathcal{P}$, we let $\text{FCOST}^{in}(\mathcal{P})=\sum_{P_i \in P} \text{FCOST}^{in}(P_i)$ and $\text{FCOST}^{out}(\mathcal{P})=\sum_{P_i \neq P_j} \text{FCOST}^{out}(P_i, P_j)$, and

$$\text{FCOST}(\mathcal{P}) = \text{FCOST}^{in}(\mathcal{P}) + \text{FCOST}^{out}(\mathcal{P})$$

Given an instance G and a fairlet decomposition $\mathcal{P} = \{P_1, \ldots, P_k\}$ for G, a *reduced correlation clustering instance* $G^{\mathcal{P}}$ is a complete graph on vertices $\{p_1, p_2, \ldots, p_k\}$, where each vertex p_i corresponds to a fairlet $P_i \in \mathcal{P}$, and the label $\sigma(p_i, p_j)$ of the edge between p_i and p_j is the majority sign of edges in $E(P_i, P_j)$ multiplied by the number of edges in $E(P_i, P_j)$ with the majority sign. We note that while G is an unweighted instance, the reduced instance $G^{\mathcal{P}}$ is weighted.

The following theorem by [13] shows that the fairlet framework applies to the correlation clustering problem under this costing scheme.

Theorem 1: (Theorem 3.4 in [13]) Assume there is an η-approximation algorithm A_1 for the minimum cost fairlet decomposition $\mathcal{P}$ and a β-approximation algorithm A_2 that solves the unconstrained correlation clustering instance $G^{\mathcal{P}}$. Using A_2 to solve the reduced correlation clustering instance $G^{\mathcal{P}}$ using the fairlet decomposition $\mathcal{P}$ from algorithm A_1 returns a $(\beta(1+\eta)+\eta)$-approximate constrained correlation clustering for G.

One can solve the weighted problem $G^{\mathcal{P}}$ directly using an $O(\log n)$-approximation algorithm by Demain, Emanuel, Fiat, and Immorlica [30]. However, since the weighted instance is from the reduction, [13] showed an improved approximation when the ratio between the maximum and minimum fairlet sizes is small. The following lemma states this fact.

Lemma 1: (Lemma 3.5 from [13]) There exists an approximation algorithm for $G^{\mathcal{P}}$ with approximation ratio $\beta = \min(\log n, 2\rho r^2)$ where $r = \frac{\max_{P \in \mathcal{P}} |P|}{\min_{P \in \mathcal{P}} |P|}$ and ρ is the approximation factor for unweighted correlation clustering.

We note that our work improves the approximation ratios η for the fairlet decomposition problem under two cases of α; thus, in turn, improving the approximation ratio for the fair correlation clustering algorithm.

2.3 The median cost problem

To solve the fairlet decomposition problem, Ahmadian *et al.* [13] gave a reduction to the following median-cost problem. Given a distance function d defined on a metric space M that contains all points for all vertices V, for a fairlet decomposition $\mathcal{P}=\{P_1, \ldots, P_k\}$, we define its me-

dian cost MCOST($\mathcal{P}$)=$\sum_{P_i \in \mathcal{P}}$ MCOST(P_i), where MCOST(A) for any subset $A \subseteq V$ is defined as followed.

$$\text{MCOST}(A) = \min_{\mu \in M} \sum_{v \in A} d(\mu, v),$$

where metric space M defines a set of all possible centers. For simplicity, if $S \in E$ represents a set of edges, we also use MCOST(S) as the median cost of the set of vertices that appear in S. Moreover, in our analysis, we let $W(S)$ be the total edge cost of S, i.e., $W(S) = \sum_{(u,v) \in S} d(u, v)$. Given the distance function d on a metric space M, the *Fairlet Decomposition with Median Cost* problem is to find a fairlet decomposition $\mathcal{P}$ under the considered constraint such that MCOST($\mathcal{P}$) is minimized.

Given an instance $G = (V = \{v_1, v_2, \ldots, v_n\}, E)$ with edge sets E^+ and E^- of the correlation clustering problem, Ahmadian et al. [13] defined a mapping $\phi : V \to [0, 1]^n$ to be such that

$$\phi(u) = [\phi_{v_1}(u), \phi_{v_2}(u), \ldots, \phi_{v_n}(u)],$$

and

$$\phi_v(u) = \begin{cases} 1 & \text{if } u = v \text{ or } (u, v) \in E^+, \\ 0 & \text{if } (u, v) \in E^- \end{cases}$$

and let metric space (M, d) be such that $M = [0, 1]^n$ and d is the Hamming distance. Given ϕ, the distance $d(u, v)$ is the Hamming distance between $\phi(u)$ and $\phi(v)$, i.e., $d(u, v) = |\phi(u) - \phi(v)|$.

The median-cost problem relates to the original fairlet decomposition via the following lemma. We note that there is a dependency on the maximum size f of the fairlets in the reduction.

Theorem 2: (Lemmas 4.1 and 4.2 and Theorem 4.3 from [13]) Given metric space (M, d) as defined above, for any fairlet decomposition $\mathcal{P}$, we have

$$\frac{1}{2} \cdot MCOST(\mathcal{P}) \leq FCOST(\mathcal{P}) \leq 2f \cdot MCOST(\mathcal{P}),$$

where $f = \max_{P \in \mathcal{P}} |P|$. Furthermore, if there is a γ-approximation algorithm for the fairlet decomposition with median cost that returns the decomposition consisting of fairlets of size at most f, then the decomposition produced by this algorithm is a $(4f\gamma)$-approximation to the minimum cost fairlet decomposition.

3. FAIRLET DECOMPOSITION WITH α-FAIR

This section describes new approximation algorithms for the Fairlet Decomposition with Median Cost defined in Section 2.2 under the constraint of α-fair. We note that while in the context of fairlet decomposition, the metric (M, d) used to compute the edge cost is derived from the input graph, our algorithms in this section do not use additional structures from that reduction, and they work for any metric.

We consider the cases that $\alpha = 1/2$ and $\alpha = 1/C$ where C is the number of distinct colors. Ahmadian et al. [13] proposed algorithms that give the ratio of 2 when $\alpha = 1/2$ and C when $\alpha = 1/C$. Our algorithms improve these factors to 4/3 and 2 respectively.

3.1 When α=1/2

In this case, the constraint states that no cluster should have more than half of its vertices with the same color. Therefore, no cluster is dominated by a single color. In this case, each fairlet in the fairlet decomposition consists of 2 or 3 vertices with distinct colors.

The key ingredient is a 2-*factor*, a subgraph in which all vertices have degree two (where edges are allowed to have multiplicity). The algorithm of [13] constructs subgraph $H = (V, E_H)$ where $(u, v) \in E_H$ iff u and v have different colors and finds a minimum-cost 2-factor F of H, under the cost d, using a polynomial-time algorithm (see, e.g., Chapter 21 from [39]). Let $W(F)$ be the cost of F. Since F is a collection of cycles, the algorithm then breaks each cycle in F into a set of paths of length 2 and possibly one path of length 3; these paths become fairlets in the decomposition.

Let $\mathcal{P}^*$ be the optimum fairlet decomposition for $\alpha = 1/2$. Ahmadian *et al.* [13] showed that the cost of the minimum 2-factor can be bounded using MCOST($\mathcal{P}^*$).

Lemma 2: (Lemma 4.4 in [13]) Let F be the minimum-cost 2-factor in H and $W(F)$ be its cost. $W(F) \leq 2\text{MCOST}(\mathcal{P}^*)$.

They also showed that the median cost of the fairlet decomposition is at most $W(F)$. With Lemma 2, this gives the approximation ratio of 2. We obtain an improvement to 4/3 by more careful analysis. We start with the following lemma related to a process for finding three consecutive vertices in an odd-length cycle with different colors.

Lemma 3: Consider an odd-length cycle $\mathcal{C}$ of length l with vertices $v_1, v_2, \ldots, v_l$. Assume that the color of vertex v_1 is a. There exists vertex v_i such that i is odd, vertex v_i is of color $c \neq a$, and vertices

$$v_1, v_3, v_5, \ldots, v_{i-2}$$

are of color a. Moreover, vertices v_{i-2}, v_{i-1}, and v_i have distinct colors.

Proof: The lemma is clearly true for $l = 3$ since every vertex has a different color. Consider the case where $l > 3$. Assume that vertices $v_3, v_5, \ldots, v_{l-2}$,

and v_l are of color a as well. This implies that v_l and v_1 have the same color and we obtain a contradiction. We let i be the first vertex on the list whose color is not a. ■

The following lemma is our key analysis. It shows that we can construct a set of short path covers with good properties.

Lemma 4: Given a 2-factor F, there is a set S of paths such that

1. *Each path in S consists of 2 or 3 vertices with distinct colors,*
2. *S covers every vertex in G,*
3. $W(S) \leq \frac{2}{3}W(F)$.

Proof: We consider each cycle $\mathcal{C}$ in F and break it to obtain a set of paths whose total cost is at most $2W(\mathcal{C})/3$. Taking the union of these paths yields the lemma.

We first deal with simple cases. If the length of $\mathcal{C}$ is even, $\mathcal{C}$ decomposes into two edge-disjoint matchings. We pick the minimum cost one whose weight is at most $W(\mathcal{C})/2$. Clearly, each edge joins vertices of different colors. Also, if the length of $\mathcal{C}$ is 3, we remove the edge with maximum cost to get a path of length 2 that covers all three vertices. Since the maximum cost is at least $W(\mathcal{C})/3$, the remaining path weighs at most $\frac{2}{3}W(\mathcal{C})$. Since previously $\mathcal{C}$ is a triangle, every vertex has a distinct color.

Consider the case where the length of $\mathcal{C}$ is an odd integer $l > 3$. We will show that there are three sets of paths S_1, S_2, S_3 such that each S_i is a set of paths with a length of at most 3 that covers all vertices in $\mathcal{C}$, and the total weight of these three sets is at most $2W(\mathcal{C})$.

From Lemma 3, we know that there exist three consecutive vertices in $\mathcal{C}$ with distinct colors. Let P be a path containing these 3 vertices. Let u_1 be the unique vertex of degree 2 in path P. Denote vertices of $\mathcal{C}$ as $u_1, u_2, \ldots,$ and u_l in cyclic order. We let S_1 consists of path P along with other alternating paths of length 2, i.e., we let

$$\begin{aligned} S_1 =&\{(u_l, u_1, u_2)\} \\ &\cup \{(u_3, u_4), (u_5, u_6), \ldots, (u_{l-2}, u_{l-1})\}. \end{aligned}$$

Assume that the color of u_1 is c. To construct S_2, we apply Lemma 3 again, starting at u_1 to find a vertex u_i such that i is odd, u_i has color $a \neq c$, and all vertices $u_1, u_3, u_5, \ldots, u_{i-2}$ have color c. We let

$$\begin{aligned} S_2 =&\{(u_1, u_2), (u_3, u_4), \ldots, (u_{i-4}, u_{i-3})\} \\ &\cup \{(u_{i-2}, u_{i-2}, u_i)\} \\ &\cup \{(u_{i+1}, u_{i+2}), \ldots, (u_{l-1}, u_l)\}. \end{aligned}$$

To construct S_3, we find the maximum index j such that j is even and u_j does not have color c. We know that $j \geq i-1$ since u_{i-1} is adjacent to vertex u_{i-2} of color c. We let

$$\begin{aligned} S_3 =&\{(u_2, u_3), \ldots, (u_{j-2}, u_{j-1})\} \\ &\cup \{(u_j, u_{j+1}, u_{j+2})\} \\ &\cup \{(u_{j+3}, u_{j+4}), \ldots, (u_{l-2}, u_l), (u_l, u_1)\}. \end{aligned}$$

It is clear from our construction that each of these sets S_1,S_2, and S_3 covers cycle $\mathcal{C}$. Each set contains exactly one path of length 3, and as we argue above, these paths have proper coloring. To see that each edge of $\mathcal{C}$ is used at most twice, we partition the edges of $\mathcal{C}$ into two sets

$$A = \{(u_1, u_2), \ldots, (u_{i-2}, u_{i-1})\}$$

and

$$B = \{(u_{i-1}, u_i), \ldots, (u_{l-1}, u_l), (u_l, u_1)\}$$

We note that S_1 and S_2 may share edges in A, but S_3 does not use these shared edges. Also, S_1 and S_3 may share edges in B, but S_2 does not use these shared edges. (See Figure 3, for example.)

We choose one of S_1,S_2, or S_3 with the minimum cost to cover $\mathcal{C}$ with the required set of paths whose cost is at most $2W(\mathcal{C})/3$. ■

From the above Lemma 4, if we break each cycle of the 2-factor into a set of paths of length 2 (with one path of length 3 if necessary) with the minimum overall cost, and let S be the union of these paths from every cycle, representing a fairlet decomposition, we will have a decomposition such that every fairlet consists of 2 or 3 vertices with different colors, satisfying the fair constraint when $\alpha = 1/2$. The lemma also ensures that the total median cost of the fairlets is at most $W(S) \leq 2/3W(F)$. Using Lemmas 2 and 4, we have the following theorem.

Theorem 3: There is a $\frac{4}{3}$-approximation algorithm for the Fairlet Decomposition with Median Cost when $\alpha = 1/2$.

The running time of the decomposition algorithm clearly is dominated by the running time for finding the minimum cost 2-factor. To find the 2-factor, one can use an algorithm for minimum bipartite matching between two copies of the vertex set V, which runs in time $O(|V|^2 \log |V| + |V||E|)$ (see, e.g., [39]). Since $|V| = n$ and $|E|$ is bounded by $|V|^2$, the overall running time is $O(n^3)$.

3.2 When α=1/C

In this section, we assume that the input graph has C colors, each with the same number of vertices.

For $\alpha = 1/C$, we want each fairlet to have exactly one vertex from each color.

Our algorithm essentially chooses the best color c^* as the starting color for building fairlets. For each color i, we find $C-1$ minimum matchings between vertices with color i and those in other colors. Formally, for each $j \neq i$, let $M_{i,j}$ be the minimum bipartite matching between vertices of colors i and vertices of color j and let $S_i = \bigcup_{j \neq i} M_{i,j}$. The algorithm then chooses the result from the best i, i.e., let $S = \text{argmin}_{S_i} W(S_i)$ and uses each of its connected components as a fairlet, i.e., the decomposition $\mathcal{P}$ consists of connected components of S. We note that this algorithm runs in polynomial time. The following theorem uses an averaging argument to prove the bound.

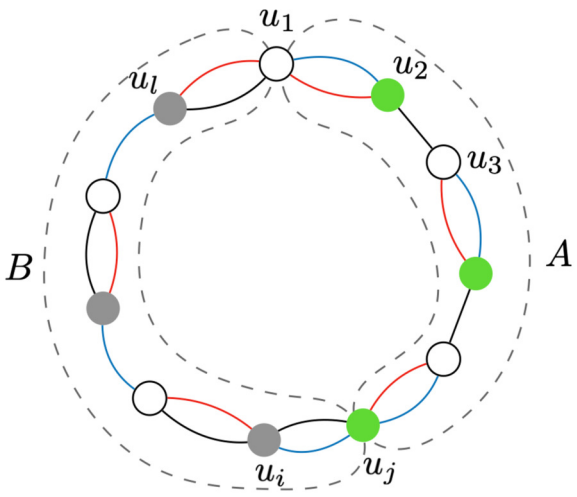


Fig.3: *Example of three sets of paths corresponding to Lemma 4., where S_1, S_2, S_3 are represented in red, blue, and black respectively.*

Theorem 4: The algorithm gives 2-approximation for the Fairlet Decomposition with Median Cost when $\alpha = 1/C$.

Proof: Let $\mathcal{P}$ be the fairlet decomposition whose fairlets are connected components of S, and let $\mathcal{P}^*$ be the optimal one. Let c be the color of the center vertex in each component of S. Since one can view S as a collection of fairlets with vertices of color c as centers, this gives one possible center choice for each fairlet; thus, $\text{MCOST}(\mathcal{P}) \leq W(S)$.

We consider $\mathcal{P}^*$. Let μ_f be the median of each fairlet $f \in \mathcal{P}^*$. Since our solution places the median vertices at vertices of a particular color, we would transform the optimal solution to be under that condition as well. Let c^* be the color such that the summation of distances from the vertex of color c^* to the median of each fairlet is minimum. More formally, let $v_f(i)$ is the vertex of color i in the fairlet f. We have that

$$c^* = \text{argmin}_i \sum_{f \in \mathcal{P}^*} d(v_f(i), \mu_f).$$

If we move the median of each fairlet f in $\mathcal{P}^*$ from μ_f to $v_f(c^*)$, the new total median cost Z becomes

$$\begin{aligned}
Z &= \sum_{f \in \mathcal{P}^*} \sum_{i \neq c^*} d\left(v_f(i), v_f(c^*)\right) \\
&= \sum_{i \neq c^*} \sum_{f \in \mathcal{P}^*} d\left(v_f(i), v_f(c^*)\right) \\
&\leq \sum_{i \neq c^*} \sum_{f \in \mathcal{P}^*} d(v_f(i), \mu_f) + d(\mu_f, v_f(c^*)) \qquad (*) \\
&= \sum_{i \neq c^*} \left(\sum_{f \in \mathcal{P}^*} d(v_f(i), \mu_f) + \sum_{f \in \mathcal{P}^*} d(\mu_f, v_f(c^*)) \right) \\
&\leq \sum_{i \neq c^*} \left(2 \sum_{f \in \mathcal{P}^*} d(v_f(i), \mu_f) \right) \qquad (**) \\
&= 2 \sum_{i \neq c^*} \sum_{f \in \mathcal{P}^*} d(v_f(i), \mu_f) \\
&\leq 2 \cdot \text{MCOST}(\mathcal{P}^*)
\end{aligned}$$

Note that step (*) uses the triangle inequality and in step (**), we use the fact that c^* minimizes $\sum_f d(v_f(c^*), \mu_f)$; thus, $\sum_f d(v_f(c^*), \mu_f) \leq \sum_f d(v_f(i), \mu_f)$, for any i.

The cost Z is the cost of the star cover, a collection of stars that cover all vertices, that is a union of matching between the vertices of color c^* and other colors. Since the algorithm chooses a star cover S that has the minimum cost, we have that $W(S) \leq Z \leq 2\ \text{MCOST}(\mathcal{P}^*)$. Therefore, $\text{MCOST}(\mathcal{P}) \leq 2\text{MCOST}(\mathcal{P}^*)$, as required. Note that we do not need to find the color c^*, as it is used only in the theoretical analysis. The only thing we need is that there exists such a color c^*. ■

To analyze the running time for this case, first note that the number of vertices for each color is $|V|/C$. We iterate over all colors to find c; for each color, we find $C-1$ matchings, each in time $O((|V|/C)^2 \cdot \log(|V|/C) + (|V|/C)|E|)$, where E is the set of edges connecting this color to one another. With $|E| = (|V|/C)^2$ and $|V| = n$, each matchings can be computed in time $O((n/C)^3)$. Thus, combining all steps, the total running time is $O(n^3)$.

3.3 Application to the fair correlation clustering

This section shows how results from the previous two sections apply to the fair correlation clustering problem. Recall the unweighted version of correlation clustering where we are given a complete graph $G = (V, E)$ with edge label $\sigma(e) = \sigma(u, v) \in \{+1, -1\}$ for each edge $e = (u, v) \in E$.

From Theorem 2, our results can be used to approximate the minimum-cost fairlet decomposition problem, which is used to approximate the fair correlation clustering by Theorem 1 and Lemma 1. Recall that the approximation factor in Theorem 2 depends on the approximation factor of the median-cost problem γ and the maximum fairlet size f.

When $\alpha = 1/2$, using our algorithm, each fairlet has size 2 or 3, so $f = 3$ and $\gamma = 4/3$. Using Theorem 2, we have a 16-approximation algorithm for the fairlet decomposition problem. Applying Theorem 1 and Lemma 1, using the best known $\rho = 2.06$ from [28], we can conclude as follows.

Corollary 1: There is a 173.59-approximation algorithm for the Fair Correlation Clustering when $\alpha = 1/2$.

This gives an improvement over the 256-approximation algorithm presented in [13].

Consider the case that $\alpha = 1/C$. In this case, each fairlet has the same size of C. From Theorem 4 and 2, using $f = C$ and $\gamma = 2$, we get a $8C$-approximation algorithm for the fairlet decomposition problem. Applying Theorem 1 from [13], we have the following corollary that gives an improvement over the ratio of $\Theta(C^2)$.

Corollary 2: There is an approximation algorithm with a ratio of $40.96C + 4.12$ for the Fair Correlation Clustering when $\alpha = 1/C$.

The running time of the clustering algorithm depends on the decomposition step and the final clustering algorithm over the fairlets. The correlation clustering algorithm by Chawla *et al.* [28] is linear-programming based; thus it runs in polynomial time.

3.4 Experimental results

We demonstrate our algorithms from Section 3 with `reuters` dataset from the UCI Repository[1]. The same dataset has been used for the fair correlation clustering problem in [13] and fair clustering in [11]. We note that Ahmadian *et al.* [13] also used `victorian` and `amazon`.

The `reuters` dataset includes text data from 50 authors; for each author, the dataset contains between 50 and 100 English texts by that author. In our settings, the graph includes each text as a vertex whose color represents the author. The edges in the graph are calculated from the cosine similarity between the semantic embeddings of pairs of the texts. We assign the label +1 to the top θ fractions of the edges and -1 to the rest, for $\theta \in \{0.25, 0.50, 0.75\}$.

We perform two sets of experiments, where (1) $\alpha = 1/2$ and (2) $\alpha = 1/C$. For each set, for each parameter θ and for each C=2,3,4,7,8,15, and 16, we construct five random graphs from a randomly selected group of C authors. The graphs contain 50 vertices for each color when C=2,4,8,16, and 49 vertices for each color when C=3,7,15. We note that for odd C, we intentionally let the number of vertices per color be odd to test the effects of parity in matching.

We remark that the dataset `reuters` provides us a way to generate graph instances based on semantic similarity between texts. Our use of authors as colors in the problem formulation could be interpreted as a fairness constraint, requiring that no document clusters should be dominated by a single author. Typically, the `reuters` dataset is used to evaluate text classification algorithms (using provided document categories as the ground truth). However, we mostly use it to generate realistic graph instances. As for the baseline comparison, we note that once we consider fairness conditions to be additional constraints for clustering, we may view the goal of the clustering as to obtain a good trade-off between fairness and accuracy, i.e., one would compare the performance (the number of disagreements) between the clustering with fairness constraints and without the constraints.

In each graph instance, the algorithms we consider are the following:

1. **Unconstrained correlation clustering algorithms** as baselines for accuracy when no fairness conditions are considered.
   For the unconstrained correlation clustering, we use the local search algorithm (LS) implemented by [13], which performs several iterations to improve the clustering by moving each vertex locally to the better cluster, and the standard pivot algorithm (PV) presented in [29]. Note that we run the local search algorithm and the pivot algorithm ten times and use the best result.
2. **Two baseline fair algorithms** that report a single cluster that contains every vertex (SINGLE) and random fair clusters (RF).
   For baseline fair algorithms, the random fair clustering (RF) works by finding a random matching between vertices with different colors (for $\alpha = 1/2$) and finding a random clustering where each cluster contains one vertex of each color (for $\alpha = 1/C$).
3. **Fair correlation clustering algorithms** that include ones from [13] and our improved algorithms.
   For the fair clustering algorithms, when $\alpha = 1/2$, we use the algorithm presented in the experiment of [13] (MATCH+LS), which decomposes fairlets by computing a minimum-cost perfect matching (1-factor), the algorithm presented in the analysis of [13] (R_FACTOR+LS) and the proposed algorithm (O_FACTOR+LS) described in Section 3.1. We note that the MATCH+LS used in the experiments by [13] does not work when the number of vertices per color is odd. The O_FACTOR+LS algorithm computes the minimum-cost fairlet decomposition from 2-factors, whereas the R_FACTOR+LS algorithm

[1] http://archive.ics.uci.edu/ml/datasets/Reuter_50_50

Table 1: *Results for $\alpha = 1/C$. Each cell reports the ERROR and the IMBALANCE (shown in parenthesis) of the algorithm.*

C	θ	Algorithms					
		Unconstrained		Baseline fair		Fair	
		LS	PV	SINGLE	RF	PATH+LS	STAR+LS
2	0.25	0.100 (0.465)	0.117 (0.448)	0.750 (0)	0.258 (0)	**0.236 (0)**	0.237 (0)
	0.5	0.161 (0.398)	0.178 (0.379)	0.500 (0)	0.504 (0)	0.359 (0)	**0.358 (0)**
	0.75	0.197 (0.134)	0.205 (0.137)	**0.250 (0)**	0.748 (0)	**0.250 (0)**	**0.250 (0)**
3	0.25	0.094 (0.547)	0.113 (0.524)	0.750 (0)	0.260 (0)	**0.243 (0)**	0.244 (0)
	0.5	0.149 (0.426)	0.166 (0.430)	0.500 (0)	0.504 (0)	**0.389 (0)**	0.391 (0)
	0.75	0.180 (0.299)	0.191 (0.263)	0.250 (0)	0.746 (0)	**0.246 (0)**	**0.246 (0)**
4	0.25	0.107 (0.58)	0.126 (0.548)	0.750 (0)	0.261 (0)	0.242 (0)	**0.240 (0)**
	0.5	0.173 (0.44)	0.195 (0.426)	0.500 (0)	0.503 (0)	0.383 (0)	**0.382 (0)**
	0.75	0.224 (0.157)	0.235 (0.15)	0.250 (0)	0.745 (0)	**0.248 (0)**	0.250 (0)
7	0.25	0.127 (0.572)	0.155 (0.540)	0.750 (0)	0.261 (0)	0.252 (0)	**0.247 (0)**
	0.5	0.238 (0.425)	0.272 (0.403)	0.500 (0)	0.502 (0)	0.413 (0)	**0.405 (0)**
	0.75	0.246 (0.052)	0.268 (0.096)	**0.250 (0)**	0.742 (0)	**0.250 (0)**	**0.250 (0)**
8	0.25	0.123 (0.618)	0.152 (0.580)	0.750 (0)	0.261 (0)	0.257 (0)	**0.253 (0)**
	0.5	0.207 (0.468)	0.241 (0.441)	0.500 (0)	0.502 (0)	0.441 (0)	**0.431 (0)**
	0.75	0.248 (0.077)	0.265 (0.110)	**0.250 (0)**	0.742 (0)	**0.250 (0)**	**0.250 (0)**
15	0.25	0.135 (0.567)	0.169 (0.529)	0.750 (0)	0.261 (0)	0.257 (0)	**0.250 (0)**
	0.5	0.244 (0.410)	0.278 (0.385)	0.500 (0)	0.501 (0)	0.450 (0)	**0.420 (0)**
	0.75	0.249 (0.058)	0.275 (0.111)	**0.250 (0)**	0.741 (0)	**0.250 (0)**	**0.250 (0)**
16	0.25	0.140 (0.592)	0.172 (0.534)	0.750 (0)	0.260 (0)	0.257 (0)	**0.252 (0)**
	0.5	0.248 (0.425)	0.285 (0.386)	0.500 (0)	0.501 (0)	0.457 (0)	**0.426 (0)**
	0.75	0.250 (0.008)	0.284 (0.147)	**0.250 (0)**	0.741 (0)	**0.250 (0)**	**0.250 (0)**

selects an arbitrary one.

When $\alpha = 1/C$, we experiment with the algorithm PATH+LS from [13] and the algorithm STAR+LS based on the star cover described in Section 3.2. We note that after these fair algorithms decompose the fairlets, they use the LS algorithm to solve correlation clustering.

As in [13], for each setting, we compare various algorithms in two quality measures: (1) the clustering ERROR, which is the cost of the correlation clustering, i.e., the fraction of "misclustered" edges, and (2) the IMBALANCE, which is the fraction of vertices violating the fairness constraints. We note that fairness is a "hard" constraint; thus, we would like our fair algorithms to have zero IMBALANCE. However, we would like also to have low ERROR; ideally, the ERROR should be close to the baseline unconstrained clustering.

In our experiment, there are five graph instances for each combination of C and θ. We repeat all algorithms ten times for each instance and report the average results.

When $\alpha = 1/2$, our algorithm O_FACTOR+LS although has a better worst-case theoretical guarantee, the experimental results in Table 3 in Appendix A.1 show that it is not significantly different from R_FACTOR+LS and MATCH+LS. See the discussion in Appendix A.1. Figure 2 shows one of these experiments when $\theta = 0.25$ and $C = 4$, where each color in the figure represents texts by an author. The algorithm gives us a clustering of nine clusters with an ERROR of approximately 0.2.

We turn our attention to results where $\alpha = 1/C$, shown in Table 1, where we highlight fair results with smallest errors. We first note that when θ is 0.75, the numbers of positive edges are large, and it is best to have a single cluster; thus, both PATH+LS and STAR+LS performed well similarly. When θ gets smaller, we see that our proposed algorithm performs consistently better as the number C of colors increases. Theoretically, in the worse case, our algorithm should perform better by a factor of $C/2$; but the actual cost of the fairlet decomposition is not as bad as the worst-case analysis for PATH.

For comparison with the unconstrained versions and the baseline versions, we refer to the discussion in [13] as our algorithm performs fairly similarly to theirs.

Cluster validation. We perform brief experiments to confirm the cluster validity. In this section, we only show results for $\alpha = 1/C$; for the cases where $\alpha = 1/2$, we refer to Appendix A.1. As correlation clustering works directly on graphs, the key measure should be related to the number of misclustered edges (shown as ERROR in Table 1 and Table 3).

In Table 2 below, we break down the errors into inter-cluster errors (i.e., the fractions of +1 edges between clusters) and intra-cluster errors (i.e., the fractions of -1 edges inside all clusters). The low error rates imply that the clustering satisfies the +1/-1 re-

Table 2: *Error breakdown and average distance result for $\alpha = 1/C$. On the left columns, each cell reports the inter-cluster ERROR and the intra-cluster ERROR (shown in parenthesis) of the algorithms. On the right columns, each cell reports the intra-cluster average distance and the average inter-cluster distance (displayed in parenthesis) of the algorithm.*

C	θ	Algorithms (ERROR Breakdown)				Algorithms (Average distance)			
		Unconstrained		Fair		Unconstrained		Fair	
		LS	PV	PATH+LS	STAR+LS	LS	PV	PATH+LS	STAR+LS
2	0.25	0.08 (0.02)	0.08 (0.03)	0.13 (0.11)	0.13 (0.11)	0.04 (0.20)	0.04 (0.20)	0.11 (0.18)	0.11 (0.18)
	0.5	0.12 (0.04)	0.13 (0.05)	0.08 (0.28)	0.08 (0.28)	0.07 (0.24)	0.07 (0.24)	0.14 (0.22)	0.14 (0.22)
	0.75	0.10 (0.10)	0.10 (0.11)	0 (0.25)	0 (0.25)	0.13 (0.26)	0.14 (0.26)	0.17 (-)	0.17 (-)
4	0.25	0.08 (0.03)	0.09 (0.04)	0.21 (0.03)	0.21 (0.03)	0.07 (0.26)	0.08 (0.26)	0.18 (0.23)	0.18 (0.23)
	0.5	0.14 (0.04)	0.15 (0.05)	0.23 (0.15)	0.21 (0.17)	0.11 (0.30)	0.12 (0.29)	0.19 (0.25)	0.19 (0.25)
	0.75	0.09 (0.13)	0.09 (0.15)	0.01 (0.24)	0 (0.25)	0.19 (0.31)	0.20 (0.31)	0.22 (0.28)	0.22 (0.27)

quirements of the graphs. Note that we do not perform cluster validation for fair baseline algorithms as these algorithms output either trivial or random clustering.

We also look at traditional distance-based measures for cluster validation as we have raw data for the input graphs. On the rightmost columns of Table 2, we show the average distances based on cosine dissimilarity between vertices inside clusters (intra-cluster distances) and between pairs of vertices in different clusters (inter-cluster distances). The results show clear differences between vertices inside clusters and vertices outside clusters. The table contains missing values when $C = 2$ and $\theta = 0.75$ because every algorithm output a single cluster.

4. PROPORTIONAL FAIRNESS

In this section, we consider another notion of fairness called proportional fairness studied in [11]. We assume that there are two vertex colors, 1 and 2, with a ratio of $1 : m$, i.e.,

$$\frac{n_1}{n_2} = \frac{1}{m},$$

where n_i is the number of vertices with color i. Let V_1 and V_2 be the set of vertices of color 1 and 2 respectively. In this case, we want each fairlet to have the same population ratio of the two colors. Therefore, each fairlet must contain exactly one vertex from V_1 and m vertices from V_2.

Ahmadi *et al.* [27] considered this problem, proved that it is NP-hard, and gave an $O(m^2)$-approximation algorithm that can be extended to deal with more than two colors. Here, we prove a connection between this problem and a capacitated star cover problem. For the two-color case, we also present a slightly improved algorithm with an approximation ratio of $O(\min\{m^2, m \log n\})$. Note that this is better than $O(m^2)$ when $\log n < m$.

As in the case that $\alpha = 1/C$, our algorithm finds a low-cost star cover such that each star represents a fairlet. Under the condition of proportional fairness, each star must have exactly one vertex of color 1 and m vertices of color 2. We find the star cover such that the center of each star is of color 1, and each star contains exactly $m+1$ vertices, using a min-cost flow algorithm [40]. Using the triangle inequality as in the proof of Theorem 4, we have the following theorem.

Theorem 5: There is an m-approximation algorithm for the Fairlet Decomposition with Median Cost under the proportional fairness constraint.

We cannot hope for a better approximation guarantee for this approach. Figure 4 shows an example where all vertices in $f \cap V_2$ in a fairlet f of $\mathcal{P}^*$ are very close to each other, but the vertex u of color 1 is far away. When we move the center of this fairlet from the best median to u, the cost grows approximately m times.

There seems to be a close connection to the star cover problem from the previous approach. Another strategy is to directly find the star cover, where each star is of size m, for vertices of color 2. Let $G_2 = (V_2, E_2)$ be a complete subgraph consisting of vertices of color 2. We would like to find a subgraph of G_2 such that every connected component is a star of size m and every vertex is a member of some star; we refer to this problem as the *Minimum m-size Star Cover problem*. The theorem below shows that if one can find a τ-approximation to this problem, one can obtain a good approximation to the fairlet decomposition as well.

Theorem 6: Given a τ-approximation algorithm for the Minimum m-size Star Cover problem, there exists an approximation algorithm for the Fairlet Decomposition with Median Cost under the proportional fairness constraint with ratio $4\tau + 2$.

Proof: We first apply the τ-approximation algorithm for the Minimum m-size Star Cover problem in G_2 to obtain S as the resulting star cover.

For each star s in S, let v_s be the vertex at the center of s. We need to connect each star containing entirely color 2 with a vertex of color 1. To do so, we find the minimum bipartite matching M between the vertices in V_1 and $U = \{v_s : s \in S\}$. The union of M and S will be a star cover of G where each star consists of exactly m vertices of color 2 and a vertex of color 1. Each star becomes a fairlet in the resulting

fairlet decomposition $\mathcal{P}$.

The median cost of our decomposition $\mathcal{P}$ is at most $W(S) + W(M)$. Let $\mathcal{P}^*$ be the optimal fairlet decomposition. Next, we show that $W(S) + W(M) \leq (4\tau + 2)\text{MCOST}(\mathcal{P}^*)$.

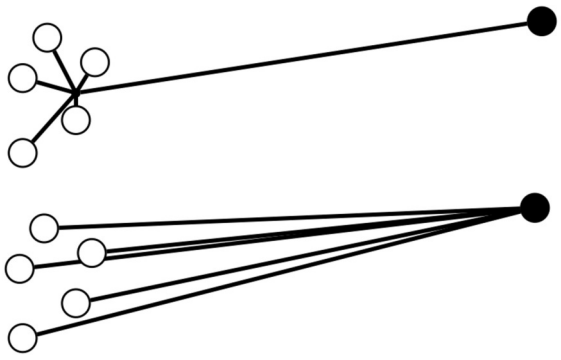


Fig.4: *An example of a fairlet whose cost grows approximately m times when we move the center to vertex u of color 1.*

We start with $W(S)$. Consider each fairlet f in $\mathcal{P}^*$. Let μ_f be its median center and let $v_f \in f \cap V_2$ be the closest vertex of color 2 to f's median μ_f. Consider a star cover S^* such that each star corresponds to a fairlet f with v_f as its center. For each new center v_f, its distance to the center decreases by $d(v_f, \mu_f)$, and for the vertex in V_1, its distance to the center increases by at most $d(v_f, \mu_f)$. For each other vertex $v \in f \cap V_2 \backslash \{v_f\}$, its new distance to the center is $d(v, v_f) \leq d(v, \mu_f) + d(v_f, \mu_f) \leq 2d(v, \mu_f)$. Therefore, the cost of the star is at most twice the median cost of $\mathcal{P}^*$. We partition the star cover S^* into a star cover S' consisting of only vertices from V_2 and a matching M' from vertices in V_1 and the star centers. From the previous discussion, we have

$$W(M') + W(S') \leq 2 \cdot \text{MCOST}(\mathcal{P}^*). \tag{1}$$

Since S' is an m-size star cover over V_2 and S is an m-size star cover with τ-approximation guarantee, we get that

$$W(S) \leq \tau W(S'). \tag{2}$$

To bound $W(M)$, we will show that there is a matching M'' between V_1 and U, whose cost can be bounded in terms of $W(M')$ and $W(S')$. Since M is a matching with minimum cost, this will give us a bound of $W(M)$. The main issue we deal with when bounding $W(M)$ is that in the matching M' obtained from the optimal solution $\mathcal{P}^*$ vertices from V_1 are matched with centers from the near-optimal star cover S', while in M they are matched, instead, with centers from S. To construct the matching, we would find "routes" from centers of stars in S' to centers of stars in S.

Let $V_{S'}$ and V_S be the sets of stars from S' and S, respectively. Consider a bipartite graph $H = (V'_S \cup V_S, E_H)$ where edge $(s', s) \in E_H$, for $s' \in V_{S'}$ and $s \in V_S$, exists iff stars s' and s intersect. Since every star in S' and S has exactly m vertices, we can use Hall's marriage theorem [41] to conclude that there is a perfect matching D in H. Consider each edge $(s', s) \in D$. Since $(s', s) \in E_H$, stars s' and s intersect. We let $R^s_{s'}$ be the path of length at most 2 from the center of s' to the center of s containing only edges of stars s and s'. We also note that if two edges (s', s) and (t', t) in H share no endpoints, paths $R^s_{s'}$ and $R^t_{t'}$ are also disjoint.

Given the matching D in H and a matching M' obtained from $\mathcal{P}^*$, we construct a new matching M'' and bound its cost as follows. For each edge $(u, v_{s'})$ in M', where $u \in V_1$ and $v_{s'}$ is the center of s' in S', there is an edge (s', s) in the matching D. We include (u, v_s) in M'' where v_s is the center of s in S. Note that the cost of (u, v_s) is at most the cost of edge $(u, v_{s'})$ and the cost of $R^s_{s'}$. We note that since M' and D are matchings, the set of edges in M'' forms a matching from V_1 to U as well. The cost of M'' satisfies

$$\begin{aligned} W(M'') &= \sum_{(u,v_s)\in M''} d(u, v_s) \\ &\leq \sum_{\substack{u,\, s',\, s;\\ (u,v_{s'})\in M',\,(s',s)\in D}} d(u, v_{s'}) + W(R^s_{s'}) \\ &= \sum_{(u,v_{s'})\in M'} d(u, v_{s'} + \sum_{(s',s)\in D} W(R^s_{s'}) \\ &\leq W(M') + W(S') + W(S). \end{aligned}$$

Since M'' is a matching between V_1 and U and M is the minimum cost matching between V_1 and U, we have $W(M) \leq W(M'')$. Therefore, using (2) the median cost of the fairlet decomposition from this algorithm is at most

$$\begin{aligned} W(S) + W(M) &\leq W(M') + W(S') + 2W(S) \\ &\leq (2\tau + 1)W(S') + W(M') \\ &\leq (2\tau + 1)(W(S') + W(M')). \end{aligned}$$

Using (1), we conclude that the fairlet decomposition $\mathcal{P}$ has the median cost at most $(4\tau + 2)\text{MCOST}(\mathcal{P}^*)$. ■

The m-size star cover is a variation of the k-median problem where the set of facility locations equals the set of clients, and every facility has a uniform capacity of m. The capacitated k-median problem has been studied in many works, and it is one of the fundamental problems that still does not have a constant approximation factor or proof that it does not exist. When parameterized by k, there are constant-factor FPT approximation algorithms [42, 43]. However, in this case, the number of facilities k

can be significant (i.e., $k = n/(1+m)$). The best approximation algorithm applicable to this case is an $O(\log k)$-approximation algorithm by Adamczyk, Byrka, Marcinkowski, Meesum, and Wlodarczyk [42] based on a tree embedding [44]. Using this result together with Theorem 6, we have the following corollary.

Corollary 3: There exists an $O(\log n)$-approximation algorithm for the Fairlet Decomposition with Median Cost under the proportional fairness constraint.

To obtain an algorithm for fair correlation clustering, we note that each fairlet has the exact size of $m+1$. Using Theorem 5 and 6, we have the following result.

Corollary 4: The Fair Correlation Clustering under the proportional fairness constraint can be approximated within the factor of $O(\min\{m^2, m\log n\})$.

5. CONCLUSIONS

In this paper, we give approximation algorithms for the fair correlation clustering problem in different fairness constraints by proposing new approximation algorithms for the fairlet decomposition with median cost. Using the reduction in [13], our results can be applied to improve the approximation ratios of the fair correlation clustering. Under the α-fair condition, we give a ratio of 173.59 when $\alpha = 1/2$, and 40.96C+4.12 when $\alpha = 1/C$ where C is the number of distinct colors. We also consider the proportional fairness condition where there are two colors with a ratio of $1 : m$. For this constraint, we give an approximation algorithm with the factor of $O(\min\{m^2, m\log n\})$.

ACKNOWLEDGMENT

We would like to thank anonymous referees for valuable comments that help to improve our experiments and our presentation significantly.

Funding: Vacharapat Mettanant is supported by the Faculty of Engineering at Sriracha Graduate Scholarship, Kasetsart University. Jittat Fakcharoenphol is supported by the Thailand Research Fund, Grant RSA-6180074.

Conflicts of interest: No conflicts of interest.

References

[1] M. Kearns and A. Roth, "Ethical Algorithm Design," *ACM SIGecom Exchanges*, vol. 18, pp. 31–36, Jul. 2020.

[2] C. Dwork, N. Kohli and D. Mulligan, "Differential Privacy in Practice: Expose your Epsilons!," *Journal of Privacy and Confidentiality*, vol. 9, no. 2, pp. 1-22, Oct. 2019.

[3] C. Dwork, "Differential Privacy in Distributed Environments: An Overview and Open Questions," in *Proceedings of the 2021 ACM Symposium on Principles of Distributed Computing*, pp. 5, Jul. 2021.

[4] C. Rösner and M. Schmidt, "Privacy Preserving Clustering with Constraints," in *The 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, vol. 107, pp. 96:1-96:14, 2018.

[5] M. Hardt, E. Price and N. Srebro, "Equality of Opportunity in Supervised Learning," in *The 30th International Conference on Neural Information Processing Systems*, vol. 29, pp. 1-9, 2016.

[6] H. Elzayn, et al., "Fair Algorithms for Learning in Allocation Problems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp.170-179, Jan. 2019.

[7] K. Donahue and J. Kleinberg, "Fairness and Utilization in Allocating Resources with Uncertain Demand," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp.658-668, Jan. 2020.

[8] A. Weller, "Transparency: Motivations and Challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp.23-40, 2019.

[9] B. Wagner, et al., "Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 261-271, Jan. 2020.

[10] A. Masoomi, et al., "Explanations of Black-Box Models based on Directional Feature Interactions," in *International Conference on Learning Representations*, pp. 1-31, 2022.

[11] F. Chierichetti, et al. "Fair Clustering through Fairlets," in *The 31th International Conference on Neural Information Processing Systems*, pp. 1-9, 2017.

[12] S. Ahmadian, et al., "Clustering without Over-Representation," in *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.267-275, Jul. 2019.

[13] S. Ahmadian, et al., "Fair Correlation Clustering," in *Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020*, 2020.

[14] A. Backurs, et al., "Scalable fair clustering," in *Proceedings of the 36 th International Conference on Machine Learning*, 2019.

[15] S. K. Bera, et al., "Fair Algorithms for Clustering," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, no. 446, pp. 4954-4965, 2019.

[16] I.O. Bercea, et al., "On the Cost of Essentially Fair Clusterings, in *Approximation, Randomization, and Combinatorial Optimization. Al-*

gorithms and Techniques, APPROX/RANDOM 2019, no.18, pp.18:1-18:22, 2019.

[17] S. A. Esmaeili, et al., "Probabilistic Fair Clustering," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, pp. 1-13, 2020.

[18] S. A. Esmaeili, et al., "Fair Clustering Under a Bounded Cost," in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, pp. 1-13, 2021.

[19] M. Kleindessner, et al., "Guarantees for Spectral Clustering with Fairness Constraints," in *Proceedings of the 36 th International Conference on Machine Learning*, pp.1-10, 2019.

[20] M. Schmidt, C. Schwiegelshohn, and C. Sohler, "Fair Coresets and Streaming Algorithms for Fair k-means," in *Approximation and Online Algorithms*, pp. 232-251, 2020.

[21] J. Tang, et al., "A Survey of Signed Network Mining in Social Media," *ACM Computing Surveys*, vol. 49, no. 3, pp. 1-37, 2016.

[22] P. Li, H. Dau, G. Puleo and O. Milenkovic, "Motif clustering and overlapping clustering for social network analysis," *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pp. 1-9, 2017.

[23] Y. Yan, L. Chen and W.-C. Tjhi, "Semi-supervised fuzzy co-clustering algorithm for document categorization," *Knowledge and information systems*, vol. 34, pp. 55-74, 2013.

[24] W.W. Cohen and J. Richman, "Learning to Match and Cluster Large High-Dimensional Data Sets for Data Integration," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 475-480, Jul. 2002.

[25] N. Bansal, A. Blum and S. Chawla, "Correlation Clustering," *Machine Learning*, vol. 56, pp. 89-113, 2004.

[26] S. Ahmadian and M. Negahbani, "Improved Approximation for Fair Correlation Clustering," *arXiv preprint arXiv:2206.05050*, 2022.

[27] S. Ahmadi, et al., "Fair Correlation Clustering," *arXiv preprint arXiv:2002.03508*, 2020.

[28] S. Chawla, et al., "Near Optimal LP Rounding Algorithm for CorrelationClustering on Complete and Complete k-Partite Graphs," in *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pp. 219-228, Jun. 2015.

[29] N. Ailon, M. Charikar and A. Newman, "Aggregating Inconsistent Information: Ranking and Clustering," *Journal of the ACM*, vol. 55, no. 5, pp. 1-27, 2008.

[30] E.D. Demaine, et al., "Correlation Clustering in General Weighted Graphs," *Theoretical Computer Science*, vol. 361, no. 2-3, op. 172-187, Sep. 2006.

[31] S. Mahabadi and A. Vakilian. "Individual Fairness for k-Clustering," in *Proceedings of the 37th International Conference on Machine Learning*, no. 611, pp. 6586-6596, 2020.

[32] A. Vakilian and M. Yalçıner, "Improved Approximation Algorithms for Individually Fair Clustering," *arXiv preprint arXiv:2106.14043*, 2021.

[33] D. Chakrabarty and M. Negahbani, "Better Algorithms for Individually Fair k-Clustering," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[34] C. Jung, S. Kannan, and N. Lutz, "A Center in Your Neighborhood: Fairness in Facility Location," *ArXiv:abs/1908.09041*, 2019.

[35] M. Abbasi, A. Bhaskara, and S. Venkatasubramanian. "Fair Clustering via Equitable Group Representations," in *The 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 504-514, Mar .2021.

[36] M. Ghadiri,, S. Samadi, and S. Vempala. "Socially Fair K-Means Clustering," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 438-448, Mar. 2021.

[37] Y. Makarychev and A. Vakilian. "Approximation Algorithms for Socially Fair Clustering," in *Proceedings of Thirty Fourth Conference on Learning Theory*, vol. 134, pp. 1-19, 2021.

[38] D. Goyal and R. Jaiswal, "FPT Approximation for Socially Fair Clustering," *ArXiv:abs/2106.06755*, 2021.

[39] A. Schrijver, '*Combinatorial Optimization - Polyhedra and Efficiency*, Springer, 2003.

[40] R.K. Ahuja, T.L. Magnanti and J.B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall, 1993.

[41] P. Hall, "On Representatives of Subsets," *Journal of the London Mathematical Society*, vol. s1-10, no. 1, pp. 26-30, 1935.

[42] M. Adamczyk,, et al. "Constant-Factor FPT Approximation for Capacitated k-Median," in *27th Annual European Symposium on Algorithms (ESA 2019)*. 2019.

[43] D. Goyal, R. Jaiswal and A. Kumar. "FPT Approximation for Constrained Metric k-Median/Means," in *15th International Symposium on Parameterized and Exact Computation (IPEC 2020)*, pp. 14:1-14:19, 2020.

[44] J. Fakcharoenphol, S. Rao and K. Talwar, "A Tight Bound on Approximating Arbitrary Metrics by Tree Metrics," *Journal of Computer and System Sciences*, vol. 69, no. 3, pp. 485-497, Nov. 2004.

Table 3: *Results for $\alpha = 1/2$. Each cell reports the ERROR and the IMBALANCE (shown in parenthesis) of the algorithm.*

C	θ	Algorithms)						
		Unconstrained		Baseline fair		Fair		
		LS	PV	SINGLE	RF	MATCH+LS	R_FACTOR+LS	O_FACTOR+LS
2	0.25	0.100 (0.465)	0.117 (0.448)	0.750 (0)	0.258 (0)	**0.236 (0)**	0.239 (0)	0.238 (0)
	0.5	0.161 (0.398)	0.178 (0.379)	0.500 (0)	0.504 (0)	**0.359 (0)**	**0.359 (0)**	0.360 (0)
	0.75	0.197 (0.134)	0.205 (0.137)	**0.250 (0)**	0.748 (0)	**0.250 (0)**	**0.250 (0)**	**0.250 (0)**
3	0.25	0.094 (0.357)	0.113 (0.326)	0.750 (0)	N/A	N/A	**0.202 (0)**	**0.202 (0)**
	0.5	0.149 (0.205)	0.166 (0.212)	0.500 (0)	N/A	N/A	0.365 (0)	**0.363 (0)**
	0.75	0.180 (0.111)	0.191 (0.097)	0.250 (0)	N/A	N/A	**0.245 (0)**	**0.245 (0)**
4	0.25	0.107 (0.297)	0.126 (0.264)	0.750 (0)	0.253 (0)	0.178 (0)	**0.177 (0)**	0.178 (0)
	0.5	0.173 (0.15)	0.195 (0.143)	0.500 (0)	0.501 (0)	**0.231 (0)**	0.232 (0)	0.232 (0)
	0.75	0.224 (0.046)	0.235 (0.043)	0.250 (0)	0.749 (0)	**0.240 (0)**	**0.240 (0)**	**0.240 (0)**
7	0.25	0.127 (0.110)	0.155 (0.085)	0.750 (0)	N/A	N/A	**0.153 (0)**	0.154 (0)
	0.5	0.238 (0.045)	0.272 (0.031)	0.500 (0)	N/A	N/A	**0.259 (0)**	**0.259 (0)**
	0.75	0.246 (0.004)	0.268 (0.009)	**0.250 (0)**	N/A	N/A	**0.250 (0)**	**0.250 (0)**
8	0.25	0.123 (0.138)	0.152 (0.090)	0.750 (0)	0.252 (0)	**0.147** (0)	0.148 (0)	0.148 (0)
	0.5	0.207 (0.012)	0.241 (0.019)	0.500 (0)	0.5 (0)	**0.220 (0)**	**0.220 (0)**	**0.220 (0)**
	0.75	0.248 (0.011)	0.265 (0.013)	**0.250 (0)**	0.749 (0)	**0.250 (0)**	**0.250 (0)**	**0.250 (0)**
15	0.25	0.135 (0.022)	0.169 (0.014)	0.750 (0)	N/A	N/A	**0.143 (0)**	**0.143 (0)**
	0.5	0.244 (0.005)	0.278 (0.005)	0.500 (0)	N/A	N/A	**0.254 (0)**	**0.254 (0)**
	0.75	0.249 (0.001)	0.275 (0.001)	**0.250 (0)**	N/A	N/A	**0.250 (0)**	**0.250 (0)**
16	0.25	0.140 (0.021)	0.172 (0.013)	0.750 (0)	0.251 (0)	**0.147 (0)**	**0.147 (0)**	**0.147 (0)**
	0.5	0.248 (0.006)	0.285 (0.004)	0.500 (0)	0.500 (0)	**0.255 (0)**	**0.255 (0)**	**0.255 (0)**
	0.75	**0.250 (0)**	0.284 (0.001)	**0.250 (0)**	0.749 (0)	**0.250 (0)**	**0.250 (0)**	**0.250 (0)**

Table 4: *Error breakdown and average distance result for $\alpha = 1/2$. On the left columns, each cell reports the inter-cluster ERROR and the intra-cluster ERROR (shown in parenthesis) of the algorithms. On the right columns, each cell reports the intra-cluster average distance and the average inter-cluster distance (displayed in parenthesis) of the algorithm.*

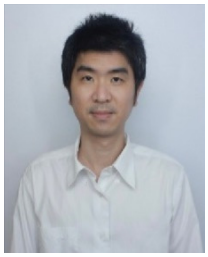
C	θ	Algorithms (ERROR Breakdown)				Algorithms (Average distances)			
			Fair				Fair		
		PV	M+LS	R_F+LS	O_F+LS	PV	M+LS	R_F+LS	O_F+LS
2	0.25	0.08 (0.03)	0.13 (0.11)	0.12 (0.12)	0.12 (0.12)	0.04 (0.20)	0.11 (0.18)	0.11 (0.18)	0.11 (0.18)
	0.5	0.13 (0.05)	0.08 (0.28)	0.08 (0.28)	0.07 (0.29)	0.07 (0.24)	0.14 (0.22)	0.14 (0.22)	0.14 (0.22)
	0.75	0.10 (0.11)	0 (0.25)	0 (0.25)	0 (0.25)	0.14 (0.26)	0.17 (-)	0.17 (-)	0.17 (-)
4	0.25	0.09 (0.04)	0.09 (0.09)	0.09 (0.09)	0.09 (0.09)	0.08 (0.26)	0.11 (0.26)	0.11 (0.26)	0.11 (0.26)
	0.5	0.15 (0.05)	0.12 (0.12)	0.12 (0.12)	0.12 (0.11)	0.12 (0.29)	0.15 (0.30)	0.15 (0.30)	0.15 (0.30)
	0.75	0.09 (0.15)	0.01 (0.23)	0.01 (0.23)	0.01 (0.23)	0.20 (0.31)	0.22 (0.29)	0.22 (0.29)	0.22 (0.29)

APPENDIX

A. MORE EXPERIMENTAL RESULTS

A.1 Results for α=1/2

For each parameter θ and each $C = 2, 3, 4, 7, 8, 15$, and 16, we construct five random graphs from a randomly selected set of C authors. The graphs contain 50 vertices for each color when $C = 2, 4, 8, 16$, and 49 vertices for each color when $C = 3, 7, 15$. We perform various fairlet decomposition algorithms together with the local search algorithm (LS) as described in Section 3.4. Note that MATCH+LS and RF do not work when the numbers of vertices per color are odd as they have to find perfect matchings; therefore, they are not evaluated in cases $C = 3, 7, 15$. Table 3 shows the results and Table 4 shows cluster validation as in the case where $\alpha = 1/C$. For cluster validation, we leave the columns for the local search algorithm (LS) out as they are very close to the pivot algorithm (PV).



Vacharapat Mettanant received his M.Eng. degree in computer engineering from Kasetsart University, Thailand. Currently, he is an assistant professor at the Department of Computer Engineering, Kasetsart University, Sriracha campus. His research interests include approximation algorithms, theoretical machine learning, and ethical issues in algorithms.

Adisak Supeesun received his Ph.D. degree in computer engineering from Kasetsart University, Thailand. Currently, he is a lecturer at the Department of Computer Engineering, Kasetsart University, Sriracha campus. His research interest lies in theoretical machine learning.

Jittat Fakcharoenphol works as an assistant professor at the Department of Computer Engineering, Kasetsart University. His research area focuses on various aspects of algorithms.