Improved Approximation for Fair Correlation Clustering

Sara Ahmadian*
Google Research

Maryam Negahbani*
KatanaGraph

Abstract

Correlation clustering is a ubiquitous paradigm in unsupervised machine learning where addressing unfairness is a major challenge. Motivated by this, we study *fair correlation clustering* where the data points may belong to different protected groups and the goal is to ensure fair representation of all groups across clusters. Our paper significantly generalizes and improves on the quality guarantees of previous work (Ahmadi et al., 2020; Ahmadian et al., 2020) as follows.

- We allow the user to specify an arbitrary upper bound on the representation of each group in a cluster.
- Our algorithm allows individuals to have multiple protected features and ensure fairness simultaneously across them all.
- We prove guarantees for clustering quality and fairness in this general setting. Furthermore, this improves on the results for the special cases studied in previous work.

Our experiments on real-world data demonstrate that our clustering quality compared to the optimal solution is much better than what our theoretical result suggests.

1 Introduction

Machine learning algorithms are used in many sensitive applications such as awarding home loans (Khandani et al., 2010; Malhotra and Malhotra, 2003) and predicting recidivism (Angwin et al., 2016; Dressel and Farid, 2018; Chouldechova, 2017). Therefore, it is crucial to ensure these algorithms are *fair* and are not biased towards or against some specific groups in the population. Defining and practicing fairness in machine learning and optimization has been a major trend in recent years (Kamishima et al., 2011, 2012; Joseph et al., 2016; Celis et al., 2018b,a; Chierichetti et al., 2019; Yang and Stoyanovich, 2017). Clustering is one such learning paradigm with a line of work in fairness, starting with Chierichetti et al. (2017) and continuing with Bera et al. (2019), Ahmadian et al. (2019), Ahmadi et al. (2020), and Ahmadian et al. (2020), to name a few.

Correlation clustering is a popular unsupervised learning problem that has gained a lot of attention from both theory (Bansal et al., 2004; Ailon et al., 2008; Ji et al., 2020; Cohen-Addad et al., 2022) and applied communities (Van Gael and Zhu, 2007; Zheng et al., 2011; Cohen and Richman, 2002; McCallum and Wellner, 2003) (see survey by Wirth (2010) and references therein). In this problem, the input is a graph on a set of vertices (or nodes) corresponding to data entries along with similar($+$)/dissimilar($-$) labels on all pairs of nodes (i.e. labeled edges in a complete graph). The goal is to partition the nodes into so-called *clusters* in a way that respects the given similarities the best: minimizing the *clustering cost* defined as the total number of $+$ edges crossing clusters, in addition to $-$ edges inside clusters.

Ahmadi et al. (2020) and Ahmadian et al. (2020) studied a variant of fair correlation clustering where each node has a color, and each color encodes a value of a *protected feature*, e.g., red encodes woman for gender. In the case of ℓ colors and color 1 being the rarest one, p_i is defined as the ratio of nodes of color i to color 1. Then the goal is to ensure the color distribution in clusters is the same as the entire data, while minimizing the clustering cost. Ahmadi et al. (2020) and Ahmadian et al. (2020) design *approximation algorithms* for this problem where a β-approximate clustering is one with cost at most β times the cost of the optimal fair clustering. In particular, Ahmadi et al. (2020) presented an $O(\ell^2 \max_i p_i^2)$-approximation algorithm and Ahmadian et al. (2020) presented an $O(\ell^2)$-approximation when p_i's are 1. If the fairness constraint was instead, to ensure no cluster has a dominant color, one that takes over at least half of the cluster, Ahmadian et al. (2020) present a 256-approximation algorithm.

There are two main short-comings in both the previous work on Fair Correlation Clustering: **(1)** They do not cover the case where each node has *multiple* protected at-

*Authors contributed equally.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

tributes (e.g., gender, race, and age group), whereas, it is shown by Bera et al. (2019) that ensuring fairness with respect to only one attribute and oblivious to others, can produce clusters that are extremely unfair with respect to the rest of the protected features; similar to when a standard color-oblivious clustering algorithm can output extremely unfair clusters (Chierichetti et al., 2017; Ahmadian et al., 2020); **(2)** Ahmadi et al. (2020) and Ahmadian et al. (2020) do not cover many natural and important settings of fairness constraints such as allowing general thresholds that do not necessarily align with the distribution of the input data e.g., a uniform threshold of 80% in disparate impact doctrine (EEOC., 1979).

We address these issues by designing an algorithm that allows the user to specify arbitrary upper bounds on the representation of each group in a cluster, where each individual node can have multiple protected features. We output clusters that are essentially fair across all these features simultaneously and bound the clustering cost of our algorithm with respect to the optimal solution, considerably improving the approximation ratios for the special cases studied in previous work.

1.1 Our Result

Our main contribution can be summarized as follows: We design an LP-rounding algorithm that given parameters $0 < \alpha_i \leq 1$ for colors $i \in \{1, \cdots, \ell\}$, and any small constant $\epsilon > 0$, returns a clustering with cost at most $O(\frac{1}{\epsilon \min_i \alpha_i})$ times that of the optimal fair solution. The clusters consist of singletons (that violate fairness by additive $+1$) and non-singleton clusters C, for which the number of points of color i in C is at most $(1+\epsilon)\alpha_i|C|$ for any i. To be more precise, we allow a $\max\{1, \epsilon|C| \max_i \alpha_i\}$ additive violation of the fairness constraints (See Theorem 1 for details).

Additive violation in group fairness is a recurring theme in *metric clustering* (Bera et al., 2019; Ahmadian et al., 2019; Bercea et al., 2019; Charikar et al., 2001). We suspect for achieving our low approximation ratio, fairness violation is necessary due to to NP-hardness of special cases, similar to what is proved by Bera et al. (2019). Comparing with Ahmadi et al. (2020) and Ahmadian et al. (2020) for the special cases addressed therein, in the case of $\alpha_i = 1/\ell$ our approximation ratio is better by a factor of ℓ (Corollary 1). For α's all equal to $1/2$ we get a $(4 + \frac{1}{\epsilon})$-approximation, a smaller constant compared to the 256-approximation of Ahmadian et al. (2020) for $\varepsilon > 1/252$ (Corollary 2).

Our empirical results in Section 5 demonstrate that our clustering cost is considerably better than the proven approximation ratio, namely, at most 15% more than the optimal cost even for $\epsilon = 0.01$. Furthermore, our approximation ratio captures the tension between getting a low-cost clustering and having strict bounds on representation of points, due to small ϵ and α's. This relation of clustering cost with ϵ and α's is also apparent from our experiments in Section 5.

Our algorithm is based on rounding a linear program (LP) formulation of Fair Correlation Clustering, similar to the work of (Ji et al., 2020) on approximating other variants of correlation clustering. Our technical contribution is for cases where carving out low-cost clusters is at odds with ensuring fairness. This happens when fairness constraints in LP are not effective due to integrality issues or when there are no clear cut fair and low-cost clusters in that region of the graph (See Section 4.2 for details).

1.2 Related Work

Fairness in machine learning and clustering. Fairness in machine learning has received a lot of attention and is a fast growing literature (survey by Caton and Haas (2020) and references therein). The efforts can be categorized into two main groups: (i) defining notions of fairness , and (ii) devising fair algorithms. Our work falls into the latter category and we concentrate on the notion of *disparate impact* which informally asks that the decisions made (by an algorithm) should not be disproportionately different for applicants in different protected classes. Under this notion, there are works spanning from fair classification (Feldman et al., 2015; Zafar et al., 2017), to fair ranking problems (Celis et al., 2018b), and to fair matroid optimization (Chierichetti et al., 2019). Chierichetti et al. (2017) introduced fair clustering problem based on this notion.

Chierichetti *et al*. Chierichetti et al. (2017) mainly defined fair clustering for two colors and later Rösner and Schmidt (2018) extended this definition to multiple colors. Both work required the distribution of colors in clusters to match the distribution of colors in the data and this definition was relaxed in Ahmadian et al. (2019) by allowing arbitrary distribution of colors in different clusters as long as presence of each color in each cluster was bounded. Bera et al. (2019) further generalized this notion by allowing lower and upper bounds per groups and also allowing overlapping groups. Other closely related problems to fair clustering are clustering with diversity constraints (Li et al., 2010), fair center selection (Chen et al., 2019), and clustering with proportionality constraints (Kleindessner et al., 2019).

Correlation clustering. Bansal et al. (2004) introduced and gave the first constant factor approximation for complete graphs with the current best being close to 1.994 by Cohen-Addad et al. (2022). Variants of the problem include complete signed graph (Bansal et al., 2004; Ailon et al., 2008), and weighted graphs (generalizing incomplete signed graph) (Charikar et al., 2005; Demaine et al., 2006). The problem is shown to be APX-hard in the former case (Demaine et al., 2006) and Unique-Games hard in the latter case (Chawla et al., 2006). The integrality gap[1] of the LP formulation of the problem for complete graphs is shown to

[1]The maximum ratio between the solution quality of the integer

be 2 (Charikar et al., 2005). Correlation clustering is also studied in the constraint setting. The most relevant is the upper-bounded correlation clustering, where each cluster is required to have size at most M for a given input parameter M. Ji et al. (2020) presents a bicriteria algorithm for this version.

2 Problem Definition and Preliminaries

The input of a Correlation Clustering problem is a complete undirected graph G = (V,E) with each edge uv labeled either $+$ or $-$ based on whether u and v are similar or dissimilar, respectively. Let E^+ denote the set of positive edges and E^- denote the set of negative edges, so $E = E^+ \cup E^-$. For subsets $S, T \subseteq V$, $E(S,T)$ denotes the set of edges between vertices in S and T, i.e., let $E(S,T) = E \cap (S \times T)$. We simplify this notation by defining $E(S) := E(S,S)$ and using u instead of $\{u\}$ when applicable, e.g., $E(u,v) = E(\{u\},\{v\})$. For a subset of edges $F \subseteq E$, let F^+ and F^- denote the intersection of edges in F with E^+ and E^- respectively, hence, $E^+(S,T) = E^+ \cap E(S,T)$, and $E^-(S,T) = E^- \cap E(S,T)$.

In a clustering problem, the goal is to find a partition $\mathcal{C}$ of points such that points inside each cluster are similar and points in different clusters are dissimilar. In the Correlation Clustering problem, this notation is naturally extended so the goal is to minimize the total number of disagreements where a disagreement happens when two similar vertices are separated, i.e, a positive inter-cluster edge, or when two dissimilar vertices are clustered together, i.e., a negative intra-cluster edge.

Definition 1. Given complete graph $G(V, E^+ \cup E^-)$, find a partition $\mathcal{C}$ of V that minimizes the *correlation cost* defined as follows

$$\mathsf{corr}(\mathcal{C}) = \left|\bigcup_{C\in\mathcal{C}} E^-(C)\right| + \left|\bigcup_{C,C'\in\mathcal{C}:C\neq C'} E^+(C,C')\right|.$$

For a subset F of edges and a given clustering, we define

$$\mathsf{corr}(F) = \left|\bigcup_{C\in\mathcal{C}} E^-(C)\cap F\right| + \left|\bigcup_{C,C'\in\mathcal{C}:C\neq C'} E^+(C,C')\cap F\right|.$$

The Fair Correlation Clustering is a generalization of Correlation Clustering problem where points may belong to groups corresponding to multiple protected features and there are constraints on representation of each group in each cluster. In this work, we consider the most general case with multiple features and different upper bound thresholds (suggested by Bera et al. (2019)) defined as follows.

Definition 2. In addition to the Correlation Clustering input, we are given a set of ℓ colors $V_1, V_2, \ldots, V_\ell \subseteq V$ that may overlap. Given *fairness parameters* $\alpha_1, \alpha_2, \ldots, \alpha_\ell \in [0,1]$, the goal is to find a clustering $\mathcal{C}$ minimizing the Correlation Clustering cost while satisfying the *fairness constraint* that for any $C \in \mathcal{C}$ and color $i \in \{0, \cdots, \ell\}$, $|V_i \cap C| \leq \alpha_i |C|$.

program and its relaxation is called integrality gap of an LP.

3 The Fair Correlation Clustering Algorithm

In this section, we present our algorithm for solving Fair Correlation Clustering. The main idea is to first solve a linear program (LP) relaxation of the problem to obtain a fractional solution and then use this fractional solution to form as many "almost fair clusters" as possible without sacrificing approximation factor by too much. Our aim is to get a *bicriterion* approximation factor, i.e., solution which will violate the fairness constraint mildly (say, $(1+\epsilon)$ factor) with cost at most β times the optimal fractional solution which itself is at most the optimal cost. More precisely, for any input $\epsilon > 0$, we can show that each cluster C of our algorithm is either of size 1 or is *ϵ-fair*, meaning $|V_i \cap C| \leq (1+\epsilon)\alpha_i |C|$ for any color class V_i, with approximation factor $\beta = O(\frac{1}{\epsilon \min_i \alpha_i})$.

3.1 An LP Formulation

The standard LP for Correlation Clustering has been studied extensively and heuristic approaches have been developed for solving it Downing et al. (2010). Our linear programming relaxation is just the extension of this LP with fairness constraints and is the LP relaxation (denoted by FCC-LP) of the following integer program (IP)

$$\min \sum_{uv\in E^+} x_{uv} + \sum_{uv\in E^-} (1-x_{uv}) \qquad \text{(FCC-IP)}$$

$$\sum_{v\in V_i}(1-x_{uv}) \leq \alpha_i \sum_{v\in V}(1-x_{uv}) \quad \forall i \in [\ell], \forall u \in V, \qquad \text{(LP-fair)}$$

$$x_{uv} + x_{vw} \geq x_{uw} \qquad \forall u,v,w \in V, \quad (\triangle\text{-ineq})$$

$$x_{uv} = x_{vu} \qquad \forall u,v \in V,$$

$$x_{uu} = 0, \qquad \forall u \in V,$$

$$x_{uv} \in \{0,1\} \qquad \forall u,v \in V,$$

Here, the indicator variable x_{uv} denotes whether vertices u and v are assigned to different clusters or not; 0 and 1 values indicate same and different cluster respectively. Constraint (LP-fair) captures the fairness requirement as $\sum_{v\in V}(1-x_{uv})$ is the size of cluster containing vertex u and $\sum_{u\in V_i}(1-x_{uv})$ is the number of vertices of color i in this cluster. The rest of the constraints ensure that x defines a distance metric (encoding three axioms for defining a metric). In particular, constraint ($\triangle$-ineq) also known as *triangle inequality* captures that if vertex v and w are

assigned to different clusters then u cannot be in the same cluster as both v and w at the same time.

As mentioned, we solve the LP relaxation of this IP, called FCC-LP, where for all $u, v \in V$ we allow x_{uv} to take any (possibly fractional) value from 0 to 1. Observe that for any clustering, we can define a feasible x that satisfies the constraints and the objective will be equal to the clustering cost, i.e., number of disagreements. Hence the optimal LP cost lower bounds the optimal IP cost. In fact, we get the claimed approximation by bounding the cost of our solution in terms of the optimal LP cost.

For a subset of the edges $F \subseteq E$, we use the notation $\mathsf{LP}(F)$ to denote the FCC-LP *cost-share* of F. That is

$$\mathsf{LP}(F) := \sum_{uv \in F^+} x_{uv} + \sum_{uv \in F^-} (1 - x_{uv}),$$

which we simplify to $\mathsf{LP}(uv)$ when $F = \{uv\}$.

3.2 The Algorithm

Our algorithm is based on rounding an optimal solution x of the Fair Correlation Clustering LP (FCC-LP) which defines a metric on the vertices. It takes as input three parameters: $\epsilon > 0$, the allowed degree of violation in the fairness constraint, and parameters $0 < \rho \leq 1/2$ and $0 < \sigma \leq \rho/2$ which will be fixed later. The high-level idea of the algorithm is to carve out ϵ-fair clusters as much as possible and then return all the remaining uncovered (unclustered) vertices as singleton clusters. We use the term *degenerate* to refer to such singleton clusters which will be collected in the set $\mathcal{C}^1$ and we use the term *non-degenerate* to refer to non-singleton clusters, i.e., all ϵ-fair clusters.

For carving out a non-degenerate cluster, the algorithm relies on the distance metric defined by x and only takes points that are in close proximity of each other. We basically look for a central point for which all the unclustered points at maximum distance ρ from it, form an ϵ-fair cluster and have average distance of at most σ. We refer to the latter condition as *density* condition and use the term *sparse* when this condition is not satisfied for a (prospective) cluster. This concludes the high-level idea of our algorithm presented in Algorithm 1.

3.3 The Main Result

To prove Algorithm 1 produces an approximately optimal solution, we bound the cost of edges by the LP cost. Our main result is the following:

Theorem 1. *There is an LP rounding algorithm that given an instance of Correlation Clustering and an $\epsilon > 0$, returns clustering $\mathcal{C}$ such that for each $C \in \mathcal{C}$ either $|C| = 1$ or $|V_i \cap C| \leq (1+\epsilon)\alpha_i |C|$ for all $i \in [\ell]$. This algorithm produces a β-approximation for*

$$\beta := \max\left\{\frac{1}{\epsilon\alpha^*}, 4 + \frac{1}{\epsilon}\right\},$$

where $\alpha^* := \min_{i\in[\ell]} \alpha_i/(1-\alpha_i)$.

Algorithm 1 Fair-CC algorithm

Input: $G = (V, E^+ \cup E^-)$, parameters $\epsilon, \sigma, \rho \in \mathbf{R}^+$ s.t. $2\sigma \leq \rho \leq .5$, FCC-LP solution $\{x_{uv} : u, v \in V\}$
Output: Clustering $\mathcal{C}$ including singletons in $\mathcal{C}^1$
$\mathcal{C} \leftarrow \emptyset$
$U \leftarrow V$
while $U \neq \emptyset$ **do**
  $T_u \leftarrow \{v \in U : x_{uv} \leq \rho\}, \forall u \in U$
  if $\exists u \in U : (\sum_{v\in T_u} x_{uv})/|T_u| \leq \sigma$ and $(|V_i \cap T_u|)/|T_u| \leq (1+\epsilon)\alpha_i, \forall i$ **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup T_u$
    $U \leftarrow U \backslash T_u$
  else
    $\mathcal{C}^1 \leftarrow U$
    $\mathcal{C} \leftarrow \mathcal{C} \cup \bigcup_{u\in U}\{u\}$
    $U \leftarrow \emptyset$
  end if
end while

Note that for special cases studied previously (Ahmadi et al., 2020; Ahmadian et al., 2020), we get the following improvements on approximation ratio modulo the fairness violation.

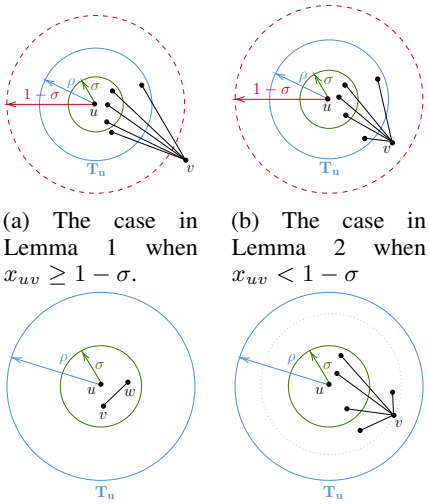
Corollary 1. For the special case of $\alpha_i = \frac{p_i}{\sum_i p_i}$ for color classes $V_1, \cdots, V_l$ with $p_1 = 1$ and arbitrary choice of p_i for $i \geq 2$, our algorithm gets an ϵ-fair solution within $O(\epsilon^{-1}\sum_i p_i)$ times optimum which is bounded by $O(\epsilon^{-1}\ell \max_i p_i)$ factor of the optimal cost.

Corollary 2. For the special case of $\alpha_i = \frac{1}{2}$ for color classes $V_1, \cdots, V_l$, our algorithm gets an ϵ-fair solution within $(4 + \frac{1}{\epsilon})$-factor of the optimal cost.

In the next section, we prove Theorem 1, a direct followup of Theorems 2 and 3.

4 Analysis

In this section, we present the required ingredients for proving Theorem 2 and Theorem 3 for bounding the cost of non-degenerate and degenerate clusters in $\mathcal{C}$. Roughly speaking non-degenerate clusters are more well-behaved and using the density property, i.e., bounded average distance, we can charge the correlation cost to the optimal LP cost comfortably. The degenerate clusters require more building arguments based on various insights such as using the density of cluster around a point at the time of removal, and charging to points of the color class that violates the fairness constraint. The main idea is to charge the cost of a disagreement edge, a negative intra-cluster edge or a positive inter-cluster edge



(a) The case in Lemma 1 when $x_{uv} \geq 1-\sigma$.

(b) The case in Lemma 2 when $x_{uv} < 1-\sigma$

(c) The case in Lemma 3 when $x_{uv}, x_{uw} \leq \sigma$.

(d) The case in Lemma 4 when $x_{uv} > \sigma$.

Figure 1: Analysis of inter-cluster edges incident with vertices covered by non-degenerate clusters. First two figures, positive edges and last two figures, negative edges.

in $\mathcal{C}$, to the LP cost of a set of edges. As long as we can ensure that no edge gets charged too many times, we can bound the total cost of $\mathcal{C}$ by the maximum factor an edge is charged.

4.1 Non-degenerate Clusters

In this section, we look at non-degenerate clusters in $\mathcal{C}$ and prove the following theorem which can be found in Ji et al. (2020) with minor changes in notation. But for the sake of completeness and making a gentler introduction to our contributions, we bring an outline of the proofs here and include full proofs in Appendix A.

Theorem 2. *For $\mathcal{C}$ output of Algorithm 1, the correlation cost of the set of edges incident to non-degenerate clusters, i.e., $F = \bigcup_{C\in\mathcal{C}} E(C) \cup E(C, V\backslash C)$ is at most* $\max\{\frac{1}{\rho-\sigma}, \frac{1}{1-\rho-\sigma}\}\mathsf{LP}(F)$.

We fix a non-degenerate cluster C and use U to denote the set of uncovered vertices at the time C is formed. Let T_u be the set corresponding to C which includes all vertices in U with maximum distance ρ from u. Refer to Figure 1 for an accompanying diagram. We start with positive inter-cluster edges and do case analysis based on whether their endpoint outside of T_u is close to the central vertex u or not (see Figure 1). Note that for any $v \in U\backslash T_u$, we have $x_{uv} > \rho$, but this is not enough for bounding the length of crossing edge between vertices in T_u to v. We first look at the case that the endpoint is "far enough" from u.

Lemma 1. For any vertex $v \in U\backslash T_u$ with $x_{uv} \geq (1-\sigma)$, $|E^+(v, T_u)| \leq \frac{1}{1-\rho-\sigma}\mathsf{LP}(E^+(v, T_u))$.

The remaining set of positive inter-cluster edges correspond to vertices that are not "far enough" from u. Here the length of a crossing edge may be short and so we rely on the fact that the cluster T_u is dense, i.e., $\sum_{w\in T_u} x_{uw} \leq \sigma|T_u|$, and so on average the length of crossing edges are long. In the following lemma, the number of these "short" positive edges from a v to T_u are bounded by the LP cost of *all* edges from v to T_u, including the negative ones.

Lemma 2. For any vertex $v \in U\backslash T_u$ with $x_{uv} < (1-\sigma)$, $|E^+(v, T_u)| \leq \frac{1}{\rho-\sigma}\mathsf{LP}(E(v, T_u))$.

For negative edges inside the cluster T_u, we do a similar case analysis based on whether the endpoints are close or far from u (see accompanying diagram in Figure 1). Let us start with the case where endpoints are close and hence the length of the negative edge is short by $\triangle$-ineq.

Lemma 3. Let $N_s := \{vw \in E^-(T_u) : x_{uv}, x_{uw} \leq \frac{\rho}{2}\}$, then $|N_s| \leq \frac{1}{1-\rho}\mathsf{LP}(N_s)$.

To complete the cost analysis of the edges incident to T_u, it remains to bound the cost of negative intra-cluster edges where at least one endpoint is far from u. In order to not overcharge any edge, we need to fix an ordering on the vertices of T_u. For $v, w \in T_u$, define $v < w$ if $x_{uv} \leq x_{uw}$ (break ties consistently) and $T_u^<(v) = \{w \in T_u : w < v\}$. Again the proof relies on the density of T_u and a counting argument.

Lemma 4. For any $v \in T_u$ with $\rho/2 < x_{uv}$, $|E^-(v, T_u^<(v))| \leq \frac{1}{1-\rho-\sigma}\mathsf{LP}(E(v, T_u^<(v)))$.

4.2 Degenerate Clusters

In this section, we focus on proving the following theorem for the cost of degenerate clusters. Here, the only disagreement edges are positive edges to other clusters and since we already counted the inter-cluster edges to non-degenerate clusters (Lemma 1 and Lemma 4), we just need to focus on bounding the cost of positive edges between degenerate clusters with respect to the cost that LP pays.

Theorem 3. *For $\mathcal{C}$ output of Algorithm 1, the correlation cost of the set of edges between degenerate clusters, i.e., $F = \bigcup_{u,u'\in\mathcal{C}^1}\{uu'\}$ is at most*

$$\Big[\max\{\frac{1}{\rho} + \frac{1-\rho}{\epsilon\rho}, \frac{1}{\sigma}, \frac{1}{2\sigma} + \frac{1}{2\epsilon\alpha^*}, \frac{1}{\epsilon\alpha^*}\}\Big]\mathsf{LP}(F) + \frac{1}{2\epsilon}\cdot\frac{1-\rho}{\rho}\mathsf{LP}(E(\mathcal{C}^1, V\backslash\mathcal{C}^1)),$$

where $\alpha^* := \min_{i\in[\ell]} \alpha_i/(1-\alpha_i)$.

Recall in Algorithm 1, degenerate cluster $\{u\}$ is added to $\mathcal{C}$ when at least one of the conditions in the if statement in Line 7 is not satisfied. At this time, all degenerate clusters, i.e., $\bigcup_{u\in\mathcal{C}^1}\{u\}$ are added to $\mathcal{C}$. Note that, unlike non-degenerate clusters, here T_u and $T_{u'}$ for degenerate clusters

$\{u\}, \{u'\} \in \mathcal{C}$ may overlap. Positive edges crossing T_us are easy to charge using the following Fact.

Fact 1. The correlation clustering cost of any $(u, v) \in E^+$, for any $\rho > 0$, is at most $\mathsf{LP}(uv)/\rho$ if $x_{uv} > \rho$.

So our focus throughout this section is on edges inside T_u's, starting with sparse T_u's (Lemma 5 and then moving to dense T_u's. We conclude by using these lemmas to prove Theorem 2.

Lemma 5. For a degenerate cluster $\{u\} \in \mathcal{C}$ with a sparse T_u, i.e., $(\sum_{v\in T_u} x_{uv})/|T_u| > \sigma$,

$$|T_u| \leq \frac{1}{\sigma}\mathsf{LP}(E(v, T_u)). \tag{1}$$

Proof. This just follows from rearranging density definition to $|T_u| \leq \sigma \sum_{uv\in E(u,T_u)} x_u v$ and applying Fact 2. □

Fact 2. For any $F \subseteq E(u, T_u)$, $\sum_{uv\in F} x_{uv} \leq \mathsf{LP}(F)$.

Proof. Take any $uv \in F$. If $uv \in E^+$ then $x_{uv} \leq \mathsf{LP}(uv)$. Now if $uv \in E^-$, since $x_{uv} \leq \rho \leq 0.5$, $x_{uv} \leq 1 - x_{uv} = \mathsf{LP}(uv)$. □

It remains to bound the cost of edges in $E(u, T_u)$ for a degenerate cluster u with dense T_u.

Lemma 6. For a degenerate cluster $\{u\} \in \mathcal{C}$ with a dense T_u, i.e., $(\sum_{v\in T_u} x_{uv})/|T_u| \leq \sigma$,

$$|T_u| \leq \frac{1}{\epsilon}[\frac{1}{\alpha^*}\mathsf{LP}(F_1) + \frac{1-\rho}{\rho}\mathsf{LP}(F_2) + (1 + \frac{1}{\rho-\sigma})\mathsf{LP}(F_3)], \tag{2}$$

where $F_1 = E(u, T_u)$, $F_2 = E(u, \mathcal{C}^1\backslash T_u)$, $F_3 = E(u, V\backslash\mathcal{C}^1)$, and $\alpha^* := \min_{i\in[\ell]} \alpha'_i = \min_{i\in[\ell]} \alpha_i/(1 - \alpha_i)$.

Proof. Let $i \in [\ell]$ be the color that violates the fairness constraint, i.e., $|V_i \cap T_u| > (1 + \epsilon)\alpha_i|T_u|$. Note, this i exists, since u is a degenerate cluster (the if condition in Algorithm 1 is false for u) but T_u is dense. Using these facts in addition to the LP fairness constraints gives Claim 1. The last step is to plug in Claim 2.

Claim 1. $\epsilon\alpha'_i|T_u| \leq \mathsf{LP}(V_i \cap T_u) + \alpha'_i \sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv})$ where $\alpha'_i = \frac{\alpha_i}{1-\alpha_i}$.

Proof. Since x is a feasible solution to FCC-LP:

$$\sum_{v\in V_i}(1 - x_{uv}) \leq \alpha_i \sum_{v\in V}(1 - x_{uv}) \Rightarrow$$
$$\sum_{v\in V_i}(1 - x_{uv}) \leq \frac{\alpha_i}{1-\alpha_i} \sum_{v\in V\setminus V_i}(1 - x_{uv}).$$

For ease of notation, let us refer to $\frac{\alpha_i}{1-\alpha_i}$ as α'_i so we have

$$\sum_{v\in V_i}(1 - x_{uv}) \leq \alpha'_i \sum_{v\in V\setminus V_i}(1 - x_{uv}) \Rightarrow$$
$$\sum_{v\in V_i\cap T_u}(1 - x_{uv}) \leq \alpha'_i \sum_{v\in V\setminus V_i}(1 - x_{uv}) \Rightarrow$$
$$|V_i \cap T_u| - \sum_{v\in V_i\cap T_u} x_{uv} \leq \alpha'_i \sum_{v\in V\setminus V_i}(1 - x_{uv})$$

$$= \alpha'_i \sum_{v\in T_u\setminus V_i}(1 - x_{uv}) + \alpha'_i \sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv})$$
$$\leq \alpha'_i|T_u\backslash V_i| + \alpha'_i \sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv}) \Rightarrow$$
$$|V_i \cap T_u| - \alpha'_i|T_u\backslash V_i|$$
$$\leq \sum_{v\in V_i\cap T_u} x_{uv} + \alpha'_i \sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv})$$
$$\leq_{\text{By Fact 2}} \mathsf{LP}(V_i \cap T_u) + \alpha'_i \sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv}).$$

Using the fact that fairness constraint is violated for V_i, meaning, $|V_i \cap T_u| > (1 + \epsilon)\alpha_i|T_u|$ and consequently $|T_u\backslash V_i| < (1 - (1 + \epsilon)\alpha_i)|T_u|$ we have

$$(1 + \epsilon)\alpha_i|T_u| - \alpha'_i(1 - (1 + \epsilon)\alpha_i)|T_u| \leq$$
$$\mathsf{LP}(V_i \cap T_u) + \alpha'_i \sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv}).$$

Recall, $\alpha'_i = \alpha_i/(1 - \alpha_i)$ so the coefficient of $|T_u|$ on the LHS can be simplified to $\epsilon\alpha'_i$. So we get

$$\epsilon\alpha'_i|T_u| \leq \mathsf{LP}(V_i \cap T_u) + \alpha'_i \sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv}).$$

□

The last step is to plug in Claim 2 into Claim 1 to get the lemma statement.

Claim 2. $\sum_{v\in(V\setminus T_u)\setminus V_i}(1 - x_{uv})$ is at most $\frac{(1-\rho)}{\rho}\mathsf{LP}((\mathcal{C}^1\backslash T_u)\backslash V_i) + \big(\frac{1}{\rho-\sigma} + 1\big)\mathsf{LP}(V\backslash\mathcal{C}^1)$.

Proof. We bound the sum by breaking it into two cases based on whether v is in $\mathcal{C}^1$ or not.

Case $v \in (\mathcal{C}^1\backslash T_u)\backslash V_i$**.** since $v \notin T_u$, $x_{uv} > \rho$ and so $1 - x_{uv} < (1 - \rho) \leq \frac{(1-\rho)}{\rho} x_{uv}$, therefore, $\sum_{v\in(\mathcal{C}^1\setminus T_u)\setminus V_i}(1 -$

x_{uv}) equals

$$\sum_{\substack{v\in(\mathcal{C}^1\setminus T_u)\setminus V_i:\\ v\in E^+}}(1-x_{uv})+\sum_{\substack{v\in(\mathcal{C}^1\setminus T_u)\setminus V_i:\\ v\in E^-}}(1-x_{uv})$$
$$\le\frac{(1-\rho)}{\rho}\sum_{\substack{v\in(\mathcal{C}^1\setminus T_u)\setminus V_i:\\ v\in E^+}}x_{uv}+\sum_{\substack{v\in(\mathcal{C}^1\setminus T_u)\setminus V_i:\\ v\in E^-}}(1-x_{uv})$$
$$\le\frac{(1-\rho)}{\rho}\mathsf{LP}((\mathcal{C}^1\setminus T_u)\setminus V_i). \tag{3}$$

Case $v\in(V\setminus\mathcal{C}^1)\setminus V_i$. Note that, $V\setminus\mathcal{C}^1$ is already partitioned into non-degenerate clusters in $\mathcal{C}$. Just for the sake of this proof, for any $v\in(V\setminus\mathcal{C}^1)$, let $C(v)$ denote the center of v's cluster. That is, $v\in T_{C(v)}\in\mathcal{C}$. Now using this new notation, let us divide the set $(V\setminus\mathcal{C}^1)\setminus V_i$ into three parts. The first part, is simply members with negative edges to u, that is, $N=\{v\in(V\setminus\mathcal{C}^1)\setminus V_i:(u,v)\in E^-\}$. For $v\in(V\setminus\mathcal{C}^1)$ with $(u,v)\in E^+$, it falls into either the second or third part, depending on $x_{uC(v)}$ where $C(v)$ is the center of v's cluster. That is, the second part is defined as $P_l=\{v\in(V\setminus\mathcal{C}^1)\setminus V_i:(u,v)\in E^+$ and $x_{uC(v)}\ge(1-\sigma)\}$ and the third part is $P_s=\{v\in(V\setminus\mathcal{C}^1)\setminus V_i:(u,v)\in E^+$ and $x_{uC(v)}<(1-\sigma)\}$.

$$\sum_{v\in(V\setminus\mathcal{C}^1)\setminus V_i}(1-x_{uv})$$
$$=\sum_{v\in N}(1-x_{uv})+\sum_{v\in P_l}(1-x_{uv})+\sum_{v\in P_s}(1-x_{uv})$$
$$\le\mathsf{LP}(N)+|P_l|+|P_s|.$$

Take any non-degenerate cluster centered at a w, that is, $T_w\in\mathcal{C}$. T_w could intersect at most one of P_l or P_s based on x_{uw}, and depending on which, we use Lemmas 1 and 2 from w's perspective to bound $|P_l|$ and $|P_s|$ respectively.

$$\sum_{v\in(V\setminus\mathcal{C}^1)\setminus V_i}(1-x_{uv})$$
$$\le_{\text{(Lemma 1)}}\mathsf{LP}(N)+\frac{1}{1-\rho-\sigma}\mathsf{LP}(P_l)+|P_s|\le_{\text{(Lemma 2)}}$$
$$\mathsf{LP}(N)+\frac{1}{1-\rho-\sigma}\mathsf{LP}(P_l)+\frac{1}{\rho-\sigma}\sum_{\substack{C\in\mathcal{C}:\\ C\cap P_s\neq\emptyset}}\mathsf{LP}(C)$$
$$\le\big(\frac{1}{\rho-\sigma}+1\big)\mathsf{LP}(V\setminus\mathcal{C}^1).$$

□

This concludes the proof of the lemma. □

We now have the required ingredients for proving Theorem 3.

Proof of Theorem 3. This proof is obtained by combining three sets of inequalities: (i) inequalities for long positive edges by Fact 1, (ii) inequalities in Lemma 5 for degenerate u with sparse T_u, and (iii) inequalities in Lemma 6 for for degenerate u with dense T_u. We use the size of T_u as an upper-bound on $E(u,T_u)$. Take any two degenerates $u,v\in\mathcal{C}^1$ with a short edge uv, i.e., $x_{uv}<\rho$. Note that in this case, T_u and T_v intersect so using either of the bounds in Lemma 5 or Lemma 6 counts this edge twice, once from each of its endpoints. Therefore, we use a coefficient of .5 for Equation (1)(of Lemma 5) and Equation (2)(of Lemma 6). Putting all these together, each edge between degenerate clusters is counted exactly once on the left hand side and we only have edges incident to degenerate clusters on the right hand side, i.e., edge uu' for $u,u'\in\mathcal{C}^1$ or uv for $u\in\mathcal{C}^1,v\in V\setminus\mathcal{C}^1$. So let us summarize how we bound the cost of an edge uu' for $u,u\in\mathcal{C}^1$, based on the length of the edge and condition of endpoints. All these terms can be bounded by the max term in the lemma:

$$\begin{cases} x_{uu'}>\rho:\frac{1}{\rho}+\frac{1}{\epsilon}\frac{1-\rho}{\rho} & \text{Fact 1 \& Eq. 2 for } u,u'^{2},\\ T_u\&T_{u'}\text{ sparse}:\frac{1}{\sigma} & \text{Eq. 1 for } u,u',\\ T_u\oplus T_{u'}{}^{3}\text{sparse}:\frac{1}{2\sigma}+\frac{1}{2\epsilon}\frac{1}{\alpha^*} & \text{Eq. 1 \& 2 for } u,u',\\ T_u\&T_{u'}\text{ dense}:\frac{1}{\epsilon}\frac{1}{\alpha^*} & \text{Eq. 2 for } u,u', \end{cases}$$

For an edge uv for $u\in\mathcal{C}^1,v\in V\setminus\mathcal{C}^1$, it can only get charged if T_u is dense and since we only consider .5 of Equation 2, it gets charged at most $\frac{1}{2\epsilon}\cdot\frac{1-\rho}{\rho}$. Combining results for the discussed two cases, we get the statement of the lemma. □

4.3 Proof of Theorem 1

proof of Theorem 1. For the choice of $\rho=.5$ and $\sigma=.25$, we get that an edge incident to an endpoint in a non-degenerate cluster is charged at most 4 by Theorem 2 and $\frac{1}{2\epsilon}$ by Theorem 3, so the total of $4+\frac{1}{2\epsilon}$. For an edge between degenerate clusters, it is charged by at most

$$\max\{2+\frac{1}{\epsilon},4,2+\frac{1}{2\epsilon\alpha^*},\frac{1}{\epsilon\alpha^*}\}\le\max\{4+\frac{1}{\epsilon},\frac{1}{\epsilon\alpha^*}\}.\ \square$$

So an edge is charged at most $\max\{\frac{1}{\epsilon\alpha^*},4+\frac{1}{\epsilon}\}$.

5 Experiments

In this section, we present the results of our experiments, designed to measure the quality of our algorithm. Our key findings are: **(1)** Our clustering cost is considerably better than our proven approximation ratio, namely, at most 15% more than the optimal cost even for $\epsilon=0.01$. **(2)** The maximum fairness violation of our algorithm on non-singleton clusters is often much less than ϵ. On some datasets, with varying ϵ, violation is fixed on a small constant as early as $\epsilon=0.3$. Our codes are publicly available on GitHub.[4]

[2]This is in the worst case that both T_u and $T_{u'}$ are dense.

[3]This symbol denotes exclusive or, meaning that exactly one of T_u and $T_{u'}$ is sparse.

[4]`github.com/moonin12/improved_fair_correlation_clustering`

We start the section by describing our datasets, followed by quality measures and benchmarks. Additional comparison to these work is listed in Appendix B.1.

Datasets. We use the datasets from Ahmadian et al. (2020). `amazon` dataset (Leskovec et al., 2007) is publicly available on SNAP[5] that corresponds to 2,441,053 items on Amazon with $+$ edges between co-reviewed items and $-$ edges for the rest. We also use datasets publicly available on the UCI repository[6], `reuters` (Liu) and `victorian` (Gungor, 2018)[7] text data corresponding to up to 16 authors, 50 to 100 texts for each. The text is embedded into a 10 dimensional space using Gensim's Doc2Vec (Řehůřek and Sojka, 2010) then set the top $\theta = \{0.25, 0.5, 0.75\}$ fraction of edges in terms of cosine similarity as positive. See Appendix B.2 for additional datasets with overlapping colors.

Quality Measures. We report *cost ratio* which is the ratio of clustering cost to $|E|$. For the LP, this is the LP objective divided by $|E|$. For fairness analysis, we measure *maximum fairness violation* defined as $\max_{C\in\mathcal{C},i\in[\ell]} |V_i \cap C|/(\alpha_i|C|) - 1$ of non-degenerate clusters (note, this is bounded by ϵ).

Benchmarks. We compare with the fair clustering algorithm of Ahmadian et al. (2020) as well as results they report on two fairness-oblivious Correlation Clustering algorithms Loc, their internal local search algorithm for Correlation Clustering, as well as Piv(Ailon et al., 2008). Since their reported cost ratios are relative to problem size, we use the numbers without implementing their algorithm. As for Ahmadi et al. (2020) published on arXiv, their code is not available and their plots are not re-usable in this fashion.

To compare with Ahmadian et al. (2020), we set α's uniformly to $1/\ell$. To experiment on datasets with overlapping colors, used in Ahmadi et al. (2020), Bera et al. (2019), and Chierichetti et al. (2017), we set α's uniformly to 0.8 in accordance with the disparate impact doctrine (EEOC., 1979). To tune the parameters ρ and σ, for each dataset, we try 5 different values for ρ from 0.1 to 0.5, and 10 different values for σ from $0.1 \times \rho/2$ to $\rho/2$. Also, we shuffle the points 20 times, re-run the algorithm, and report the result with the best clustering cost. We solve the LP using CPLEX (IBM, 2021).

5.1 Cost Analysis

We compare our clustering cost to the LP objective which is a *lower bound* on the cost of the optimal solution. Our results demonstrate that we are much closer to the optimal solution than our theoretical results suggest. We experiment once with varying ϵ and once with varying $\alpha^{\min} := \min_{i\in[\ell]} \alpha_i$, scaling all others αs accordingly (see Figure 2). Figure 2 suggests that allowing the algorithm to violate fairness, our cost can even beat the optimal cost (which is constrained by fairness). As stated in Theorem 1, the approximation ratio of our algorithm drops significantly by increasing ϵ. Figure 2 depicts how the LP costs drop by relaxing fairness through increasing $\alpha^{\min}$ and that our cost almost matches the optimal cost even for $\epsilon = 0.01$. See Appendix B.3 for the full set of experiments.

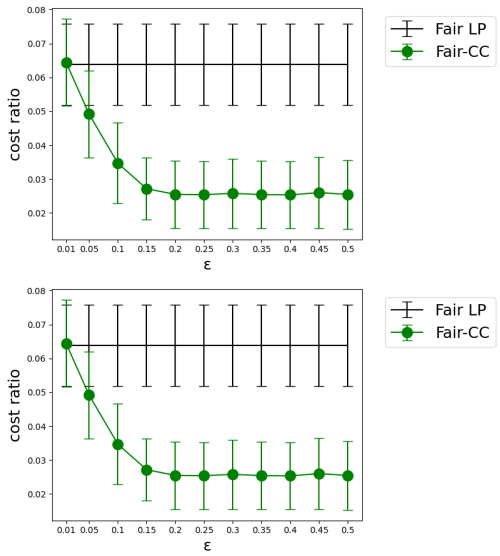


Figure 2: Cost ratio of our algorithm (Fair-CC) and the fair LP for `amazon` , 10 sub-samples of 200 each. The first plot uses varying ϵ from 0.01, to 0.5. The second plot is for $\epsilon = 0.01$, varying $\alpha^{\min}$ from its original value 0.5 to 1 (no fairness), scaling other α's accordingly.

Furthermore, we compare our cost ratio with that of Ahmadian et al. (2020). The LP solutions allows us to compare with optimum, but in contrast, Ahmadian et al. (2020) do not have any proxy for the optimal cost and just compare their cost ratio against popular fairness-oblivious correlation clusterings, i.e., Piv and Loc. For the sake of completeness, we also compare our results to that of Ahmadian et al. (2020) and algorithms therein. Table 1 demonstrates that even for ϵ as small as 0.01, our cost ratios are considerably better.

5.2 Fairness Analysis

In this section, we report result of study on how relaxing the allowed violation in fairness constraints affects the actual final violation of fairness in our output clusters. Figure 3 shows a case where maximum fairness violation of our algorithm is much less than ϵ (max allowed violation) and bounded by 0.11. On some datasets like `amazon` in Figure 3, the maximum violation curve demonstrates a few *elbows* for varying ϵ. See Appendix B.4 for the full set of

[5] `snap.stanford.edu/data/`

[6] `archive.ics.uci.edu/ml/datasets/`

[7] The datasets are available at `archive.ics.uci.edu/ml/datasets/Reuter_50_50` and `archive.ics.uci.edu/ml/datasets/Victorian+Era+Authorship+Attribution`

Table 1: Cost ratio comparison for datasets in Ahmadian et al. (2020) (AKEM) for the case of two colors, $\alpha_1 = \alpha_2 = .5$, $\epsilon = 0.01$. `amazon` is the average and standard deviation reported on 20 sub-samples of size 200. Loc and Piv are Correlation Clustering algorithms demonstrated to be unfair on these datasets (Ahmadian et al., 2020)

	Fair Algorithms		Unfair Algorithms	
Dataset	**Fair-CC**	AEKM	Loc	Piv
`amazon`	0.064 ± 0.013	0.064	0.010	0.011
`reuters` $\theta = 0.25$	**0.213**	0.230	0.096	0.161
`reuters` $\theta = 0.50$	**0.297**	0.350	0.181	0.231
`reuters` $\theta = 0.75$	**0.196**	0.199	0.188	0.241
`victorian` $\theta = 0.25$	0.217	**0.212**	0.109	0.158
`victorian` $\theta = 0.50$	**0.325**	0.348	0.183	0.268
`victorian` $\theta = 0.75$	**0.232**	0.237	0.203	0.280

experiments.

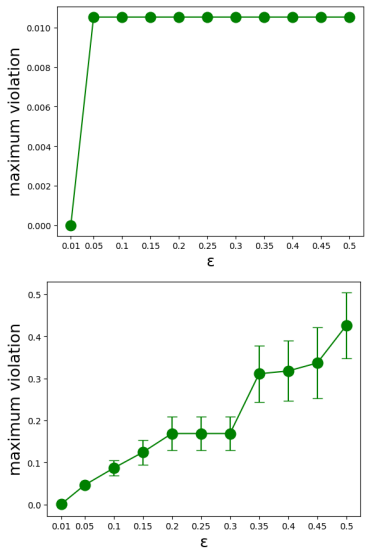


Figure 3: fairness violation of our algorithm (Fair-CC), varying ϵ from 0.01 to 0.5 on `reuters` $\theta = 0.75$ and `amazon`.

References

S. Ahmadi, S. Galhotra, B. Saha, and R. Schwartz. Fair correlation clustering. *arXiv:2002.03508*, 2020.

S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 267–275, 2019.

S. Ahmadian, A. Epasto, R. Kumar, and M. Mahdian. Fair correlation clustering. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 108:4195–4205, 2020.

N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55 (5):23:1–23:27, 2008.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica, May 23 2016.

N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine learning*, 56(1):89–113, 2004.

S. Bera, D. Chakrabarty, N. Flores, and M. Negahbani. Fair algorithms for clustering. In *Conference on Neural Information Processing Systems*, pages 4954–4965, 2019.

I. O. Bercea, M. Groß, S. Khuller, A. Kumar, C. Rösner, D. R. Schmidt, and M. Schmidt. On the cost of essentially fair clusterings. In *International Workshop on Approximation Algorithms for Combinatorial Optimization Problems*, pages 18:1–18:22, 2019.

S. Caton and C. Haas. Fairness in machine learning: A survey. *arXiv:2010.04053*, 2020.

L. E. Celis, L. Huang, and N. K. Vishnoi. Multiwinner voting with fairness constraints. In *IJCAI*, pages 144–151, 2018a.

L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *International Colloquium on Automata, Languages and Programming*, pages 28:1–28:15, 2018b.

M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, page 642–651, USA, 2001. Society for Industrial and Applied Mathematics. ISBN 0898714907.

M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. *Journal of Computer and System Sciences*, 71(3):360–383, 2005.

S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. *computational complexity*, 15(2):94–114, 2006.

X. Chen, B. Fain, C. Lyu, and K. Munagala. Proportionally fair clustering. In *Proc. 36th Proceedings, International Conference on Machine Learning (ICML)*, June 2019.

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Conference on Neural Information Processing Systems*, pages 5029–5037, 2017.

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvtiskii. Matroids, matchings, and fairness. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 2212–2220, 16–18 Apr 2019.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

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*, pages 475–480, 2002.

V. Cohen-Addad, E. Lee, and A. Newman. Correlation clustering with sherali-adams. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 651–661, 2022.

E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2-3):172–187, 2006.

N. Downing, P. J. Stuckey, and A. Wirth. Improved consensus clustering via linear programming. In *Proceedings of the Thirty-Third Australasian Conference on Computer Science-Volume 102*, pages 61–70. Citeseer, 2010.

J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4, 2018.

T. U. EEOC. Uniform guidelines on employee selection procedures, March 2 1979.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proc. 21st Annual SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

A. Gungor. Fifty victorian era novelists authorship attribution data. 2018.

IBM. Ibm ilog cplex, 2021.

S. Ji, D. Xu, M. Li, and Y. Wang. Approximation algorithms for two variants of correlation clustering problem. *Journal of Combinatorial Optimization*, 06 2020. doi: 10.1007/s10878-020-00612-1.

M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Conference on Neural Information Processing Systems*, pages 325–333, 2016.

T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *ICDM Workshops*, pages 643–650, 2011.

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

A. E. Khandani, A. J. Kim, and A. W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34:2767–2787, 2010.

M. Kleindessner, P. Awasthi, and J. Morgenstern. Fair k-center clustering for data summarization. In *ICML*, pages 3448–3457, 2019.

R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Annual SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1996.

J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *TWEB*, 1(1):5, 2007.

J. Li, K. Yi, and Q. Zhang. Clustering with diversity. In *ICALP*, pages 188–200, 2010.

Z. Liu. Donated by zhi liu from national engineering research center for e-learning technology, china.

R. Malhotra and D. K. Malhotra. Evaluating consumer loans using neural networks. *Omega*, 31:83–96, 2003.

A. McCallum and B. Wellner. Toward conditional models of identity uncertainty with application to proper noun coreference. pages 905–912, 2003.

R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

C. Rösner and M. Schmidt. Privacy preserving clustering with constraints. In *International Colloquium on Automata, Languages and Programming*, volume 107, pages 96:1–96:14, 2018.

B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

P. R. Sérgio Moro, Paulo Cortez. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, pages 22–31, 2014.

J. Van Gael and X. Zhu. Correlation clustering for crosslingual link detection. In *IJCAI*, pages 1744–1749, 2007.

A. Wirth. *Correlation Clustering*, pages 227–231. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_176.

K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proc. 29th International Conference on Scientific and Statistical Database Management*, page 22. ACM, 2017.

M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*, 44(6):1113–1122, 2011.

A Missing Proofs from Section 4

In this section, we restate Claims and Lemmas from Section 4 for which the proof is deferred to this section.

Lemma 1. For any vertex $v \in U \backslash T_u$ with $x_{uv} \geq (1-\sigma)$, $|E^+(v,T_u)| \leq \frac{1}{1-\rho-\sigma}\mathsf{LP}(E^+(v,T_u))$.

Proof. Fix an edge $vw \in E^+(v,T_u)$ (w may or may not be equal to u). Roughly speaking, since v is far from u and w is close to u (w belongs to T_u), the length of edge vw cannot be too short. Formally, by $\triangle$-ineq,

$$x_{vw} \geq x_{uv} - x_{uw} \geq 1-\sigma-\rho \qquad \text{since } x_{uw} \leq \rho,$$

and so the cost of this edge can be just charged to the LP cost the edge. □

Lemma 2. For any vertex $v \in U \backslash T_u$ with $x_{uv} < (1-\sigma)$, $|E^+(v,T_u)| \leq \frac{1}{\rho-\sigma}\mathsf{LP}(E(v,T_u))$.

Proof. Let $P := E^+(v,T_u)$ with $p = |P|$ and let $N := E^-(v,T_u)$ with $n = |N|$. Analyzing the LP cost of all crossing edges, we have

$$\begin{aligned}
\mathsf{LP}(E(v,T_u)) &= \sum_{vw \in P} x_{vw} + \sum_{vw \in N} 1 - x_{vw} \\
&\geq \sum_{vw \in P} x_{uv} - x_{uw} + \sum_{vw \in N} 1 - x_{uv} - x_{uw} \\
&= p x_{uv} + n(1-x_{uv}) - \sum_{w \in T_u} x_{uw} \\
&> \rho p + \sigma n - \sigma(p+n) = (\rho-\sigma)p,
\end{aligned}$$

where the first inequality follows from $\triangle$-ineq, and the final inequality uses $\rho \leq x_{uv} \leq 1-\sigma$ and density of T_u along with $|T_u| = |E(v,T_u)| = p+n$. As $p = |E^+(v,T_u)|$ this concludes the proof. □

Lemma 3. Let $N_s := \{vw \in E^-(T_u) : x_{uv}, x_{uw} \leq \frac{\rho}{2}\}$, then $|N_s| \leq \frac{1}{1-\rho}\mathsf{LP}(N_s)$.

Proof. Fix the edge $vw \in N_s$. Using ($\triangle$-ineq), $x_{vw} \leq x_{uv} + x_{uw} \leq \rho$ thus $1 - x_{vw} \geq 1-\rho$ and so the cost of the edge can be charged to the LP cost of the edge. □

Lemma 4. For any $v \in T_u$ with $\rho/2 < x_{uv}$, $|E^-(v, T_u^<(v))| \leq \frac{1}{1-\rho-\sigma}\mathsf{LP}(E(v,T_u^<(v)))$.

Proof. Let p and n be the number of vertices before v in the ordering with positive and negative edge to v, respectively, i.e., $p = |E^+(v,T_u^<(v))|$ and let $n = E^-(v,T_u^<(v))$.

$$\begin{aligned}
\mathsf{LP}(E(v,T_u^<(v))) &= \sum_{vw \in E^+(v,T_u^<(v))} x_{vw} + \sum_{vw \in E^-(v,T_u^<(v))} (1-x_{vw}) \\
&\geq \sum_{vw \in E^+(v,T_u^<(v))} (x_{uv} - x_{uw}) + \sum_{vw \in E^-(v,T_u^<(v))} (1 - x_{uv} - x_{uw}) \qquad \text{(using ($\triangle$-ineq))} \\
&= p x_{uv} + n(1-x_{uv}) - \sum_{w \in T_u^<(v)} x_{uw} \\
&> \frac{\rho}{2} p + (1-\rho)n - \sum_{w \in T_u^<(v)} x_{uw}, \qquad (4)
\end{aligned}$$

Where the last inequality is by the given assumption $\frac{\rho}{2} < x_{uv} \leq \rho$. To bound the last term, we use the fact that T_u is dense

(i.e. $\sum_{w\in T_u} x_{uw} \le \sigma|T_u|$), as follows.

$$\begin{aligned}
\sum_{w\in T_u} x_{uw} &\le \sigma|T_u| \Rightarrow \\
\sum_{w\in T_u^<(v)} x_{uw} &\le \sigma|T_u| - \sum_{\substack{w\in T_u:\\ w\ge v}} x_{uw} \\
&= \sigma(p+n+|\{w\in T_u : w\ge v\}|) - \sum_{\substack{w\in T_u:\\ w\ge v}} x_{uw} \\
&< \sigma(p+n+|\{w\in T_u : w\ge v\}|) - \frac{\rho}{2}|\{w\in T_u : w\ge v\}| && \text{(definition of ordering)} \\
& && \text{and } x_{uv} > \rho/2) \\
&\le \sigma(p+n). && \text{(since } \sigma \le \rho/2)
\end{aligned}$$

Next, we substitute this into Equation (4).

$$\begin{aligned}
\mathsf{LP}(E(v, T_u^<(v))) &> \frac{\rho}{2}p + (1-\rho)n - \sum_{w\in T_u^<(v)} x_{uw} \\
&> \frac{\rho}{2}p + (1-\rho)n - \sigma(p+n) \ge (1-\rho-\sigma)n,
\end{aligned}$$

since $\sigma \le \rho/2$. □

This concludes the analysis of edges incident with non-degenerate clusters and enables the following proof.

Proof of Theorem 2. First, for any internal edge $vw \in E(C)$ for some $C := T_u \in \mathcal{C}$ (u may be the same as v or w), it gets charged either through Lemma 3 or Lemma 4. Based on the length of x_{uv} and x_{vw} and the defined ordering on vertices in T_u, $\mathsf{LP}(vw)$ is charged at most by one of these lemmas and so it is charged at most by $\max\{\frac{1}{1-\rho}, \frac{1}{1-\rho-\sigma}\} = \frac{1}{1-\rho-\sigma}$ as $\sigma > 0$.

For any crossing edge $vw \in E(C, V\backslash C)$ with $v \notin C$ and $C := T_u \in \mathcal{C}$ (u may be the same as v), it can gets charged either through Lemma 1 or Lemma 2. Based on the length of x_{uv}, $\mathsf{LP}(vw)$ is charged at most by one of these lemmas and so it is charged at most $\max\{\frac{1}{1-\rho-\sigma}, \frac{1}{\rho-\sigma}\}$. □

B Complimentary Experiments

B.1 Additional Comparison with Previous Work

In this section, we bring our comparison results with previous work for 4 and 8 colors in Tables 2 and 3 respectively.

Table 2: Cost ratio comparison for datasets in Ahmadian et al. (2020) (AKEM) for the case of four colors and all α's equal to .25. We use $\epsilon = 0.01$. `amazon` and `victorian` are sub-sampled to 200 stratified on color combinations. We report the average and standard deviation over 20 sub-samples. Loc is a Correlation Clustering algorithm demonstrated to be unfair on these datasets Ahmadian et al. (2020).

	Fair Algorithms		Unfair Algorithm
Dataset	**Fair-CC**	AEKM	Loc
`reuters` $\theta = 0.25$	0.250	**0.244**	0.120
`reuters` $\theta = 0.50$	**0.306**	0.336	0.191
`reuters` $\theta = 0.75$	**0.218**	0.227	0.211
`victorian` $\theta = 0.25$	0.240 ± 0.012	**0.210**	0.141
`victorian` $\theta = 0.50$	0.322 ± 0.014	**0.311**	0.228
`victorian` $\theta = 0.75$	$\mathbf{0.240 \pm 0.007}$	0.245	0.225

Table 3: Cost ratio comparison for datasets in Ahmadian et al. (2020) (AKEM) for the case of eight colors and all α's equal to .125. We use $\epsilon = 0.01$. Datasets are sub-sampled to 200 stratified on color combinations. We report the average and standard deviation over 20 sub-samples for each. Loc is a Correlation Clustering algorithm demonstrated to be unfair on these datasets Ahmadian et al. (2020).

	FAIR ALGORITHMS		UNFAIR ALGORITHMS
DATASET	**FAIR-CC**	AEKM	LOC
reuters $\theta = 0.25$	$\mathbf{0.250 \pm 0.000}$	0.252	0.133
reuters $\theta = 0.50$	0.5 ± 0.000	**0.426**	0.239
reuters $\theta = 0.75$	$\mathbf{0.244 \pm 0.003}$	0.250	0.237
victorian $\theta = 0.25$	0.250 ± 0.000	**0.212**	0.161
victorian $\theta = 0.50$	0.370 ± 0.085	**0.319**	0.249
victorian $\theta = 0.75$	$\mathbf{0.242 \pm 0.006}$	0.246	0.218

B.2 Experiments with Overlapping Colors

We use datasets publicly available on the UCI repository[8] **(1)** bank Sérgio Moro (2014) with 4,521 points, corresponding to phone calls from a marketing campaign by a Portuguese banking institution. **(2)** census Kohavi (1996) with 32,561 points, representing information about individuals extracted from the 1994 US census. **(3)** diabetes Strack et al. (2014) with 101,766 points, extracted from diabetes patient records.

bank , census , and diabetes are used in Ahmadian et al. (2019); Bera et al. (2019); Chierichetti et al. (2017) and have 5, 7, and 8 colors respectively where each node has exactly two colors Table 4. These are sub-sampled to 200 stratified on color combinations. Since the previous work does not handle overlapping colors, we only compare with the LP cost in Table 5

Table 4: Detailed description of the datasets bank , census , and diabetes . For each dataset, the coordinates are the numeric attributes used to determined the position of each record in the Euclidean space. The sensitive attributes determines protected groups.

Dataset	**Coordinates**	**Sensitive attributes**	**Protected groups**
bank	age, balance, duration	marital	married, single, divorced
		default	yes, no
census	age, education-num,	sex	female, male
	final-weight, capital-gain,	race	Amer-ind, asian-pac-isl,
	hours-per-week		black, other, white
diabetes	gender, age, race,	gender	female, male
	time-in-hospital	race	6 groups

[8] archive.ics.uci.edu/ml/datasets/

Table 5: Cost ratio comparison with the LP cost for datasets with overlapping colors, used in Ahmadian et al. (2019); Bera et al. (2019); Chierichetti et al. (2017). all α's are set to 0.8 in accordance with the DI doctrine EEOC. (1979). We use $\epsilon = 0.01$. Datasets are sub-sampled to 200 stratified on color combinations.

DATASET	FAIR-CC COST	FAIR LP COST
bank $\theta = 0.25$	0.249	0.248
bank $\theta = 0.50$	0.498	0.497
bank $\theta = 0.75$	0.749	0.746
census $\theta = 0.25$	0.135	0.108
census $\theta = 0.50$	0.226	0.190
census $\theta = 0.75$	0.268	0.276
diabetes $\theta = 0.25$	0.100	0.077
diabetes $\theta = 0.50$	0.143	0.122
diabetes $\theta = 0.75$	0.130	0.119

B.3 Cost Analysis

In this section, we have the set of complete cost analysis experiments over all the datasets except `amazon` , which can be found in Section 5. For each dataset we have a pair of figures depicting the cost ratio of our algorithm (Fair-CC) and the fair LP for that dataset. On the left, varying ϵ from 0.01 (with completely fair clusters), to 0.5. On the right, for $\epsilon = 0.01$ and varying $\alpha^{\min}$ from its original value 0.5 to 1 (no fairness), scaling other α's accordingly.

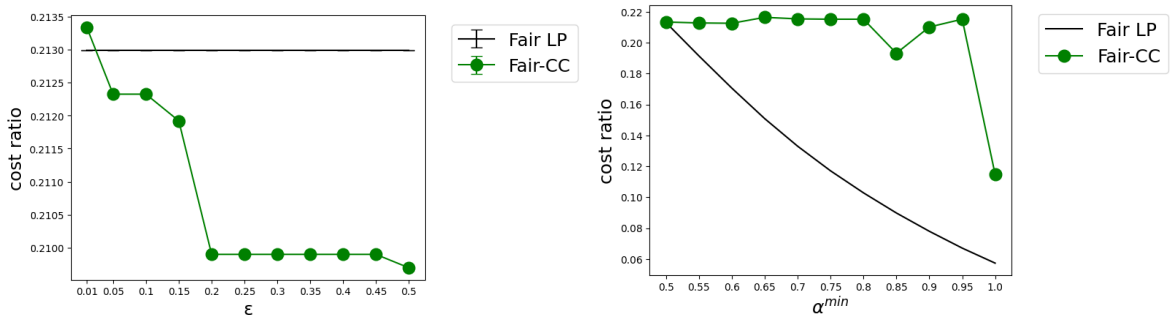


Figure 4: `reuters` $\theta = 0.25$, cost ratios of our algorithm (fair-CC) and fair LP for varying ϵ.

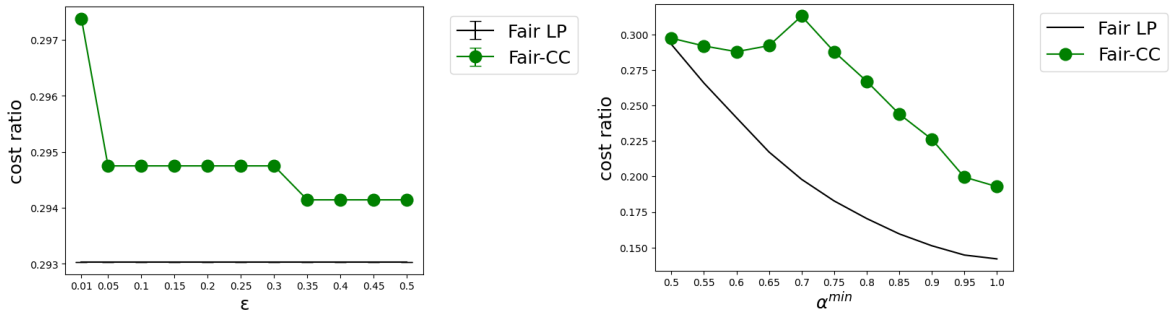


Figure 5: `reuters` $\theta = 0.5$, cost ratios of our algorithm (fair-CC) and fair LP for varying ϵ.

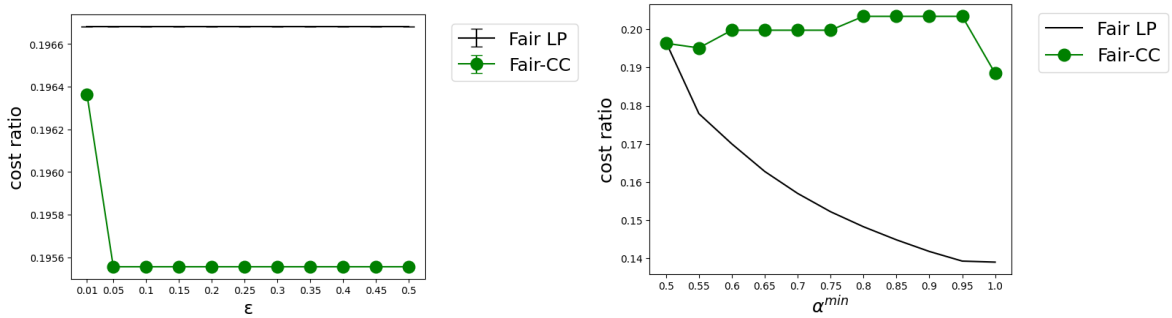


Figure 6: `reuters` $\theta = 0.75$, cost ratios of our algorithm (fair-CC) and fair LP for varying ϵ.

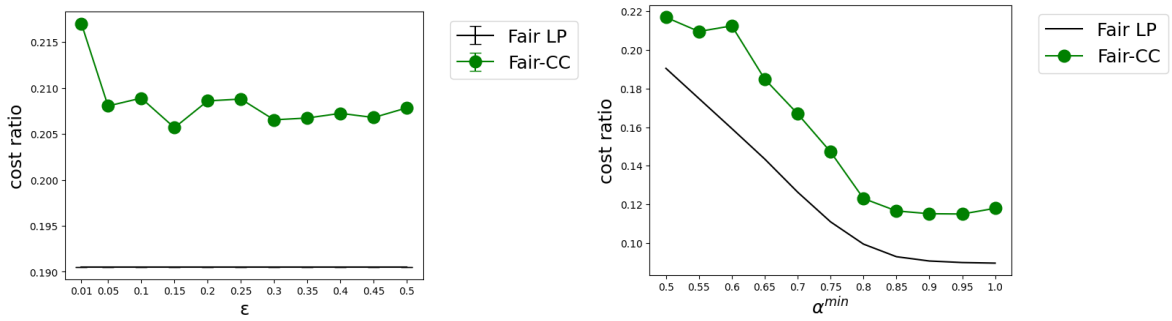


Figure 7: victorian $\theta = 0.25$, cost ratios of our algorithm (fair-CC) and fair LP for varying ϵ.

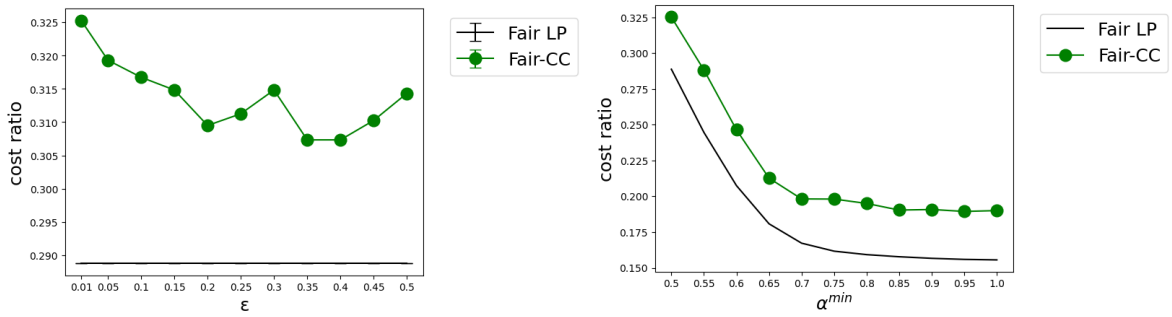


Figure 8: victorian $\theta = 0.5$, cost ratios of our algorithm (fair-CC) and fair LP for varying ϵ.

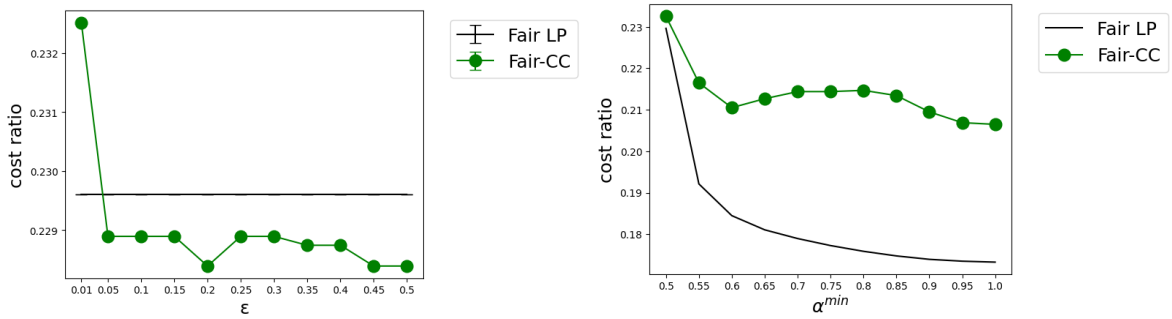


Figure 9: victorian $\theta = 0.75$, cost ratios of our algorithm (fair-CC) and fair LP for varying ϵ.

B.4 Fairness Analysis

In this section we compare the fairness violation of our algorithm (Fair-CC) with the allowed fairness violation ϵ, varying ϵ from 0.01 (with completely fair clusters) to 0.3. We have included all the datasets except `amazon` which is brought in Section 5. `reuters` and `victorian` , each dataset has three plots associated with $\theta = 0.25, 0.5$, and 0.75 on the top-left, top-right, and bottom respectively. `bank` , `census` , and `diabetes` have trivial max violations for this range of ϵ as $\alpha^{\min}$ is too low for an increase in ϵ to be able to make a change in cluster structure. There are some colors that are extremely rare in these datasets e.g. "Amer-ind" for race in `census` .

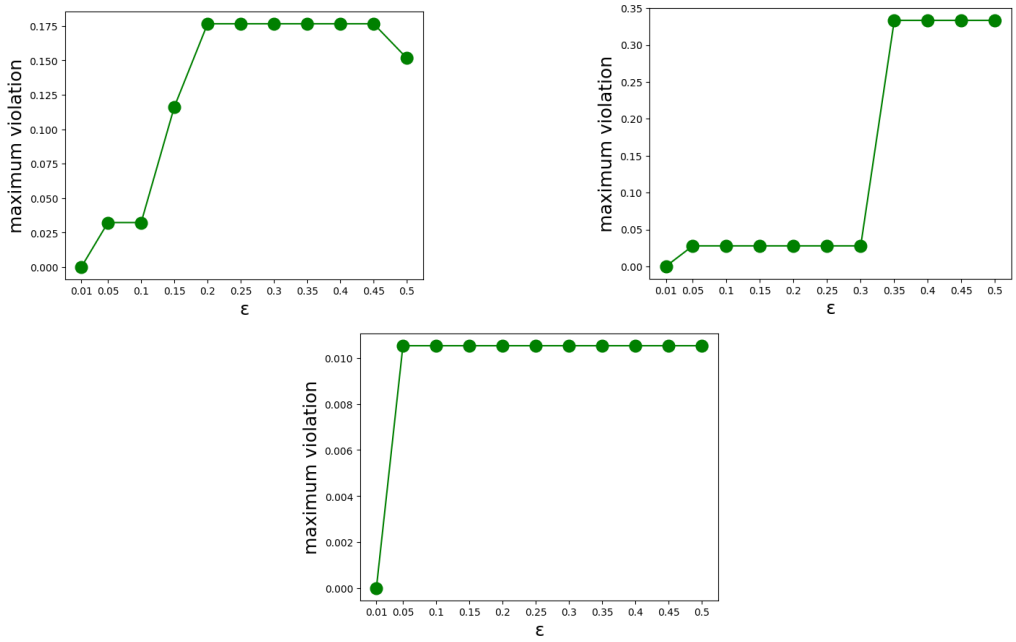


Figure 10: `reuters` , maximum violation of our algorithm for varying ϵ.

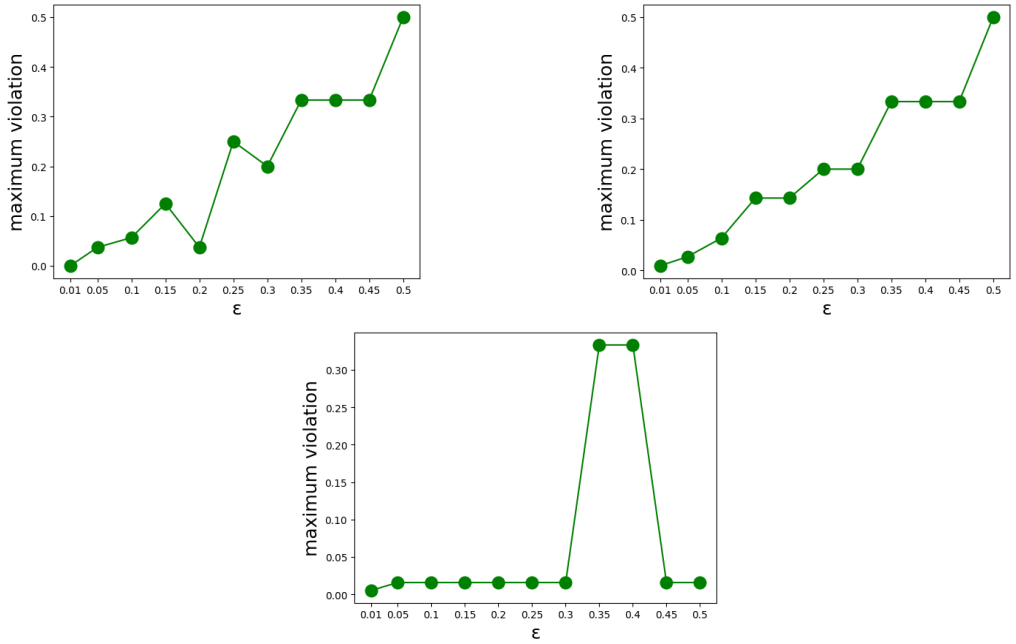


Figure 11: `victorian` , maximum violation of our algorithm for varying ϵ.