Approximating (Weighted) Chromatic Correlation Clustering via Cluster LP

Fateme Abbasi[1]
Institute of Computer Science, University of Wrocław, Poland

Hyung-Chan An[1]
Department of Computer Science and Engineering, Yonsei University, Seoul, South Korea

Jarosław Byrka[1]
Institute of Computer Science, University of Wrocław, Poland

Changyeol Lee[1]
Department of Computer Science and Engineering, Yonsei University, Seoul, South Korea

Yongho Shin[1]
Institute of Computer Science, University of Wrocław, Poland

Abstract

Correlation Clustering is a fundamental clustering problem that is generalized to Chromatic Correlation Clustering to incorporate categorical data. Both problems have been intensively studied, and recently, substantial improvements were obtained in the approximation algorithms for Correlation Clustering. At the heart of this success lies a new linear program (LP) formulation called the *cluster LP*; a natural question was whether this LP can be extended to Chromatic Correlation Clustering to enable similar success.

We answer this question in the affirmative by presenting a $(2 + \varepsilon)$-approximation algorithm for the problem using a *chromatic* cluster LP. We then consider Weighted Chromatic Correlation Clustering, in which edges have fractional weights satisfying the probability constraints, to show that our algorithm extends to this weighted version to yield the same approximation guarantee.

2012 ACM Subject Classification Theory of computation → Approximation algorithms analysis

Keywords and phrases Chromatic correlation clustering, Chromatic cluster LP, Preclustering

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2026.13

Category APPROX

Related Version *Full Version*: `https://arxiv.org/abs/2510.13446` [1]

Funding Supported by NCN grant number 2020/39/B/ST6/01641. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2021-II212068, Artificial Intelligence Innovation Hub). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00563707). This work was partly supported by an IITP grant funded by the Korean Government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University)).

Acknowledgements The authors thank the anonymous reviewers for their helpful comments. Part of this work was done while H.-C. An and C. Lee were visiting the University of Wrocław, Poland.

[1] Equal contribution

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2026).
Editors: Mohit Singh and Tom Gur; Article No. 13; pp. 13:1–13:20
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Given a complete graph where each edge is labeled by either "+" or "−", Correlation Clustering (CC) asks to find a vertex clustering that minimizes the total number of *disagreed edges*, where we say an edge is disagreed if it is a "−" edge within a cluster or a "+" edge across clusters. This problem is a fundamental combinatorial optimization problem that has been extensively studied in various forms across fields [28, 23, 15, 7, 12, 24, 6]. The readers are also referred to Wahid and Hassini [26] for a dedicated survey on this problem.

This problem is well-known to be NP/APX-hard [24, 6, 12, 10], and recently, it is further shown that the problem cannot be approximated within a constant smaller than 24/23 unless P = BPP [9]. There has also been a long line of work [6, 10, 2, 11, 14, 13, 9] on improving the approximation ratio for this problem that recently culminated in 1.485 due to Cao, Cohen-Addad, Lee, Li, Newman, and Vogl [9]. A key component of their positive result is in approximately solving an exponential-sized LP, called the *cluster LP*, in polynomial time and rounding this nearly optimal solution to an integral solution.

Chromatic Correlation Clustering (CCC) is a generalization of CC motivated by clustering data with categorical interactions [8, 5, 3]. In this generalization, "+" edges are labeled with colors. The objective is to find a clustering of vertices and a coloring of the clusters that minimize the total number of disagreements, where "+" edges within a cluster are deemed disagreed if their colors differ from that of the cluster. This chromatic version has been intensively studied in the literature [8, 4, 18, 27, 20], but previous approaches to this generalization were nevertheless limited to color-blind algorithms [18], reduction to the monochromatic version [4, 18, 27], and rounding algorithms using the standard LP relaxation [4]. The recent work by Lee, Fan, and Lee [20] achieving an approximation ratio of 2.15 for CCC is also based on reducing to monochromatic CC using an optimal solution to the standard LP relaxation for CCC.

In light of success of the cluster LP in CC, it is interesting to see if a chromatic version of the cluster LP helps improve the approximation algorithm for CCC. In this paper, we affirmatively answer this question: we show that a natural adaptation of the cluster LP, called the *chromatic cluster LP*, can be solved nearly optimally in polynomial time, and any feasible solution to the chromatic cluster LP can be rounded into an integral solution of cost within a factor of 2. These together yield a $(2+\varepsilon)$-approximation algorithm for CCC, improving upon 2.15 due to Lee et al. [20]. Furthermore, we identify how we can generalize this approach to the weighted case of CCC.

In addition to CCC, Weighted (Monochromatic) Correlation Clustering [6, 17, 2, 25, 11, 21] is another well-studied generalization of CC. In this problem, rather than "integrally" declaring an edge as "+" or "−", each edge is now assigned a fractional "+" weight and "−" weight. These two weights must be nonnegative and sum to 1. This condition is thereby called the *probability constraint*. The objective of the problem is to find a clustering of vertices that minimizes the total *weight* of disagreements. It is worth noting that there exists an approximation-preserving reduction to unweighted CC due to Chawla, Makarychev, Schramm, and Yaroslavtsev [11]. This reduction is based on replacing each vertex in the original weighted instance by a sufficiently large clique of "+" edges while the numbers of "+" and "−" edges between two cliques are proportional to the weights between the corresponding vertices.

In this paper, we study the weighted generalization of CCC, called Weighted Chromatic Correlation Clustering (WCCC), which is naturally defined as follows: each edge is assigned "+" weights, one for each color, and a (colorless) "−" weight. These weights must respect the probability constraint. The goal is to find a clustering that minimizes the total weighted disagreements. We demonstrate that our approach for CCC smoothly extends to this weighted generalization, yielding an approximation algorithm with the same factor of $(2+\varepsilon)$. The aforementioned approximation-preserving reduction for the monochromatic version [11] unfortunately does not easily extend to this chromatic version because it is unclear how one could assign colors to the edges within each clique originated from a vertex in the original instance.

Recently, Fan, Lee, and Lee [16] presented a 1.64-approximation algorithm for unweighted CCC. They also approximately solve the same chromatic cluster LP in polynomial time using a slightly different approach from ours, and then round this solution with both a cluster-based algorithm and a pivot-based algorithm, yielding the 1.64-approximation algorithm. One notable difference in solving the chromatic cluster LP lies in the construction of *admissible edges* (see Section 5.2): they adapted Cao et al.'s construction [9] into a color-sensitive version with parallel admissible edges introduced, while we show that the original construction is sufficient. Furthermore, we extend our results to the weighted setting, demonstrating the applicability of this approach to other settings. We note that our results for CCC presented in this paper were obtained independently, whereas our results for WCCC were obtained after the arXiv announcement [19] of the unweighted results of Fan et al [16].

2 Problem definitions

We begin with defining Chromatic Correlation Clustering (CCC). An instance of CCC consists of a set L of colors and a complete graph (V, E) where each edge is specified as either a $+$ edge or a $-$ edge, and each $+$ edge e is associated with a color $c_e \in L$. Let E^ℓ be the set of edges colored ℓ; let $E^+ := \bigcup_{\ell \in L} E^\ell$ and $E^- := E \setminus E^+$ denote the sets of $+$ and $-$ edges, respectively. Given an instance $(V, E, L, \{c_e\}_{e \in E^+})$ of the problem, a (feasible) solution consists of a partition $\mathcal{C}$ of V and a coloring function $\chi : \mathcal{C} \to L$ that assigns a color to each part. Given a solution $(\mathcal{C}, \chi)$, an edge $e = uv$ is called a *disagreement* if and only if one of the following conditions holds:

- e is a $-$ edge and both u and v belong to the same cluster $C \in \mathcal{C}$;
- e is a $+$ edge and u and v belong to different clusters $C \neq C' \in \mathcal{C}$; and
- e is a $+$ edge and both u and v belong to the same cluster $C \in \mathcal{C}$, but $\chi(C) \neq c_e$.

We also refer to a disagreement simply as a *disagreed* edge. If an edge is not a disagreement, we call it an *agreement* or an *agreed* edge. The goal of the problem is to find a solution that minimizes the total number of disagreements, which is also referred to as the cost of the solution.

We now turn our attention to Weighted Chromatic Correlation Clustering (WCCC). An instance of WCCC consists of a complete graph where each edge e is associated with $|L \cup \{-\}|$ *weights*, denoted by $\{w^\ell(e)\}_{\ell \in L \cup \{-\}}$. The weights of each edge satisfy the *probability constraint*, i.e., $w^\ell(e) \geq 0$ for every $\ell \in L \cup \{-\}$ and $\sum_{\ell \in L \cup \{-\}} w^\ell(e) = 1$. A (feasible) solution to a WCCC instance (V, E, L, w) also consists of a partition $\mathcal{C}$ of V and a coloring function $\chi : \mathcal{C} \to L$. To define the total cost of a solution $(\mathcal{C}, \chi)$, we define the cost incurred by an edge $e = uv$ in the solution, denoted by $\mathsf{cost}_{(\mathcal{C},\chi)}(e)$, as follows:

- $\mathsf{cost}_{(\mathcal{C},\chi)}(e) := \sum_{\ell \in L} w^\ell(e) = 1 - w^-(e)$ if u and v belong to different clusters in $\mathcal{C}$;
- $\mathsf{cost}_{(\mathcal{C},\chi)}(e) := \sum_{\ell \in L \cup \{-\} \setminus \{\chi(C)\}} w^\ell(e) = 1 - w^{\chi(C)}(e)$ if both u and v belong to a cluster $C \in \mathcal{C}$.

The cost of a solution $(\mathcal{C}, \chi)$ is then defined to be the sum of costs incurred by all edges, i.e., $\sum_{e \in E} \mathsf{cost}_{(\mathcal{C},\chi)}(e)$. We may omit the subscript $(\mathcal{C}, \chi)$ when it is clear from context. The objective of the problem is to find a solution of the minimum cost.

We further define the notation that will be used. Let $n := |V|$ denote the number of vertices. For the unweighted CCC problem, let $E^{-\ell} := E \setminus E^\ell$ for every $\ell \in L$. Given $F \subseteq E$, $u \in V$, and $S \subseteq V$, let $F(u, S) := \{uv \in F \mid v \in S\}$; for example, given $u \in V$ and $S \subseteq V$, $E^+(u, S)$ denotes the set of $+$ edges whose one endpoint is u and the other endpoint is in S.

On the other hand, for WCCC, given $U \subseteq L \cup \{-\}$ and $S, T \subseteq V$, let $w^U(S, T) := \sum_{\ell \in U} \sum_{uv: u \in S, v \in T} w^\ell(uv)$. For notational simplicity, let $w^U(u, T)$ and $w^U(u, v)$, respectively, denote $w^U(\{u\}, T)$ and $w^U(\{u\}, \{v\})$. For any edge $e = uv$, $w^U(u, v)$ can be written as $w^U(e)$. Note that $w^U(u, u) = 0$ by definition. For $\ell \in L$, let $w^{-\ell}(S, T) := w^{L \cup \{-\} \setminus \{\ell\}}(S, T)$. Similarly, let $w^+(S, T) := w^L(S, T)$.

3 A constant-factor approximation algorithm for WCCC

This section is devoted to a constant-factor approximation algorithm for WCCC. The algorithm is a reduction to CCC inspired by Bansal, Blum, and Chawla [6]. Recall that constant-factor approximation algorithms [4, 18, 27, 20] are known for CCC.

Given an instance $I_{\mathsf{ori}} = (V, E, L, w_{\mathsf{ori}})$ of WCCC, we construct a new instance $I_{\mathsf{new}} = (V, E^+ \cup E^-, L, c)$ for CCC as follows:

$$E^- := \{e \in E \mid \arg\max_{\ell \in L \cup \{-\}} \{w^\ell_{\mathsf{ori}}(e)\} = -\} \text{ and}$$
$$c_e := \arg\max_{\ell \in L} \{w^\ell_{\mathsf{ori}}(e)\} \text{ for } e \in E^+ := E \setminus E^-,$$

where ties are broken arbitrarily. We then run a ρ-approximation algorithm for CCC on I_{new}, and return the output of this algorithm.

Let $\mathsf{cost}_{\mathsf{ori}}(\cdot)$ and $\mathsf{cost}_{\mathsf{new}}(\cdot)$ denote the total cost of a solution with respect to I_{ori} and I_{new}, respectively. Let $\mathrm{OPT}_{\mathsf{ori}}$ and $\mathrm{OPT}_{\mathsf{new}}$ denote optimal solutions to I_{ori} and I_{new}, respectively, and let SOL denote the final output of the algorithm. Note that $\mathsf{cost}_{\mathsf{new}}(\mathrm{SOL}) \le \rho\, \mathsf{cost}_{\mathsf{new}}(\mathrm{OPT}_{\mathsf{new}})$ by the execution.

▶ **Lemma 1.** $\mathsf{cost}_{\mathsf{ori}}(\mathrm{SOL}) \le \mathsf{cost}_{\mathsf{new}}(\mathrm{SOL}) + \mathsf{cost}_{\mathsf{ori}}(\mathrm{OPT}_{\mathsf{ori}})$.

Proof. Fix any edge $e \in E$. If e is disagreed by SOL in I_{new}, it incurs a unit cost in I_{new}. Since the cost incurred by any edge in I_{ori} is trivially bounded by 1, the total cost incurred by the edges disagreed by SOL is at most $\mathsf{cost}_{\mathsf{new}}(\mathrm{SOL})$.

On the other hand, if e is agreed by SOL in I_{new}, we cannot use the above argument because it incurs no cost in I_{new}. Observe however that this edge incurs $1 - w^\ell_{\mathsf{ori}}(e)$ in SOL in I_{ori}, where $\ell = \arg\max_{\ell' \in L \cup \{-\}} \{w^{\ell'}_{\mathsf{ori}}(e)\}$ by the construction of I_{new}. Note that, in I_{ori}, any solution must incur $1 - w^\ell_{\mathsf{ori}}(e)$ from e due to the choice of ℓ. Hence, the total cost incurred by the edges agreed by SOL is bounded from above by $\mathsf{cost}_{\mathsf{ori}}(\mathrm{OPT}_{\mathsf{ori}})$. ◀

▶ **Lemma 2.** *For any solution $\mathcal{S}$, $\mathsf{cost}_{\mathsf{new}}(\mathcal{S}) \le 2\, \mathsf{cost}_{\mathsf{ori}}(\mathcal{S})$.*

Proof. Fix any edge $e \in E$ disagreed by $\mathcal{S}$ in I_{new}. Let $\ell := -$ if $e \in E^-$ and $\ell := c_e$ if $e \in E^+$. Observe that $w^{\ell'}_{\mathsf{ori}}(e) \le \frac{1}{2}$ for any $\ell' \in L \cup \{-\} \setminus \{\ell\}$ due to the construction of I_{new}. Hence, the cost incurred by e in $\mathcal{S}$ in I_{ori} is then at least $\min_{\ell' \in L \cup \{-\} \setminus \{\ell\}} \{1 - w^{\ell'}_{\mathsf{ori}}(e)\} \ge \frac{1}{2}$. ◀

▶ **Theorem 3.** *This algorithm is a $(2\rho + 1)$-approximation algorithm for WCCC.*

Proof. We can derive that

$$\begin{aligned}
\mathsf{cost}_{\mathsf{ori}}(\mathrm{SOL}) &\le \mathsf{cost}_{\mathsf{new}}(\mathrm{SOL}) + \mathsf{cost}_{\mathsf{ori}}(\mathrm{OPT}_{\mathsf{ori}}) \\
&\le \rho\, \mathsf{cost}_{\mathsf{new}}(\mathrm{OPT}_{\mathsf{new}}) + \mathsf{cost}_{\mathsf{ori}}(\mathrm{OPT}_{\mathsf{ori}}) \\
&\le \rho\, \mathsf{cost}_{\mathsf{new}}(\mathrm{OPT}_{\mathsf{ori}}) + \mathsf{cost}_{\mathsf{ori}}(\mathrm{OPT}_{\mathsf{ori}}) \\
&\le (2\rho + 1)\, \mathsf{cost}_{\mathsf{ori}}(\mathrm{OPT}_{\mathsf{ori}}),
\end{aligned}$$

where the first inequality is due to Lemma 1, and the last is due to Lemma 2. ◀

4 Chromatic cluster LPs

In this section, we define the chromatic cluster LPs for CCC and WCCC, respectively, and present a rounding algorithm whose expected cost is at most twice the cost of the given LP solution.

Chromatic cluster LPs. Define the *chromatic cluster LP* for CCC as follows:

$$\begin{aligned} \min\ & \textstyle\sum_{S\subseteq V,\ell\in L}\left(\frac{|\delta^+(S)|}{2}+|E^{-\ell}[S]|\right)z_S^\ell & \\ \text{s.t.}\ & \textstyle\sum_{S\ni v,\ell\in L} z_S^\ell = 1, & \forall v\in V,\\ & z_S^\ell \ge 0, & \forall S\subseteq V\ \forall \ell\in L,\end{aligned}$$

where $z_S^\ell = 1$ indicates whether S is a cluster of color ℓ, $\delta^+(S) := \{(u,v)\in E^+ \mid u\in S, v\notin S\}$, and $E^{-\ell}[S] := \binom{S}{2}\cap E^{-\ell}$. Let $t_v^\ell := 1-\sum_{S\ni v} z_S^\ell$ indicate whether v is *not* colored ℓ and $x_{uv}^\ell := 1-\sum_{S\supseteq uv} z_S^\ell$ denote whether u and v are *not* together in a same cluster colored ℓ. Observe that

$$\textstyle\sum_{S\ni u,S\not\ni v} z_S^\ell = \sum_{S\ni u} z_S^\ell - \sum_{S\supseteq uv} z_S^\ell = x_{uv}^\ell - t_u^\ell, \qquad \forall \ell\in L\ \forall uv\in E \tag{1}$$

$$\textstyle\sum_{\ell\in L} t_v^\ell = |L| - \sum_{\ell\in L,S\ni v} z_S^\ell = |L|-1, \qquad \forall v\in V. \tag{2}$$

Using these equalities, we can rephrase the objective function as $\mathrm{obj}(x) := \sum_{uv\in E^+} x_{uv}^{c_{uv}} + \sum_{uv\in E^-}\sum_{\ell\in L}(1-x_{uv}^\ell)$. We defer its derivation to Appendix A.

The chromatic cluster LP for WCCC has the same sets of variables and constraints, while the objective function is replaced with

$$\textstyle\sum_{S\subseteq V,\ell\in L}\left(\frac{w^+(S,V\setminus S)}{2}+w^{-\ell}(S,S)\right)z_S^\ell.$$

Through a similar derivation to the unweighted version, the objective function can be rephrased as $\mathrm{obj}(x) := \sum_{uv\in E,\ell\in L}\left(w^\ell(uv)x_{uv}^\ell + w^-(uv)(1-x_{uv}^\ell)\right)$. The derivation can also be found in Appendix A.

Cluster-based algorithm. The algorithm is the same for CCC and WCCC. Given a solution z feasible to the chromatic cluster LP, consider the following algorithm. Let V' be the currently unclustered vertices (initially, $V' := V$). Until V' becomes empty, we randomly sample a cluster $S\subseteq V$ colored ℓ with probability $\frac{z_S^\ell}{\sum_{\ell'\in L,S'\subseteq V} z_{S'}^{\ell'}}$. We then cluster $V'\cap S$ colored ℓ and update $V' := V'\setminus S$ accordingly. This ends the description of the cluster-based algorithm.

We begin by analyzing the algorithm for CCC. Recall that $x_{uv}^\ell = 1-\sum_{S\supseteq uv} z_S^\ell$ for every $\ell\in L$ and $uv\in E$. For notational simplicity, let $x_{uv} := 1-\sum_{\ell\in L,S\supseteq uv} z_S^\ell$ indicate whether u and v are *not* clustered together (regardless of its color). Note that we have $1-x_{uv} = \sum_{\ell\in L}(1-x_{uv}^\ell)$, implying that, for any $\ell\in L$, $x_{uv}\le x_{uv}^\ell$. The following lemma is useful in showing the approximation ratios, whose proof is deferred to Appendix A.

▶ **Lemma 4.** *For every $uv\in E$ and color $\ell\in L$, the probability that uv is in a same cluster colored ℓ is $\frac{1-x_{uv}^\ell}{1+x_{uv}}$, and the probability that uv is separated is $\frac{2x_{uv}}{1+x_{uv}}$.*

▶ **Theorem 5.** *Given a feasible solution z to the chromatic cluster LP for CCC, the cluster-based algorithm outputs an integral solution $(\mathcal{C},\chi)$ whose expected cost is at most twice the objective value of z. The algorithm terminates in time* $\mathsf{poly}(n,|\mathsf{supp}(z)|)$ *with high probability, where* $\mathsf{supp}(z)$ *denotes the support set of z.*

Proof. From Lemma 4, we can derive that the probability that $uv \in E$ is not in a cluster colored ℓ is at most $1 - \frac{1-x^\ell_{uv}}{1+x_{uv}} \le \frac{2x^\ell_{uv}}{1+x_{uv}}$. Hence, the expected cost of $(\mathcal{C}, \chi)$ for CCC is bounded by

$$\begin{aligned}
&\textstyle\sum_{uv\in E^+} \Pr[uv \text{ is not in a same cluster } C \text{ of color } \chi(C) = c_{uv}] \\
&\qquad + \textstyle\sum_{uv\in E^-} \Pr[uv \text{ is not separated}] \\
&\le \textstyle\sum_{uv\in E^+} \frac{2x^{c_{uv}}_{uv}}{1+x_{uv}} + \sum_{uv\in E^-} \frac{1-x_{uv}}{1+x_{uv}} \\
&\le 2\textstyle\sum_{uv\in E^+} x^{c_{uv}}_{uv} + \sum_{uv\in E^-} (1-x_{uv}) \\
&\le 2\textstyle\sum_{uv\in E^+} x^{c_{uv}}_{uv} + \sum_{uv\in E^-} \sum_{\ell\in L}(1-x^\ell_{uv}) \le 2\,\mathrm{obj}(x).
\end{aligned}$$

To show that the algorithm terminates in time $\mathsf{poly}(n, |\mathsf{supp}(z)|)$, observe that, for any vertex $v \in V$, the probability of v being clustered in each iteration is $\frac{\sum_{S\ni v,\ell\in L} z^\ell_S}{\sum_{\ell\in L, S\subseteq V} z^\ell_S} \ge \frac{1}{|\mathsf{supp}(z)|}$. Hence, using the union bound, we can show the algorithm terminates in $\Theta(|\mathsf{supp}(z)|\log n)$ iterations with high probability. ◀

For WCCC, note that Lemma 4 is still satisfied because the algorithm is equivalent. We thus have the weighted version for Theorem 5 as below.

▶ **Theorem 6.** *Given a feasible solution z to the chromatic cluster LP for WCCC, the cluster-based algorithm outputs an integral solution $(\mathcal{C}, \chi)$ whose expected cost is at most twice of the objective value of z. The algorithm terminates in time $\mathsf{poly}(n, |\mathsf{supp}(z)|)$ with high probability.*

Proof. For WCCC, observe that the objective value can be written as

$$\begin{aligned}
\mathrm{obj}(x) &= \textstyle\sum_{e\in E,\ell\in L} \left(w^\ell(e)x^\ell_e + w^-(e)(1-x^\ell_e)\right) \\
&= \textstyle\sum_{e\in E} \left[\sum_{\ell\in L} w^\ell(e)x^\ell_e + w^-(e)(1-x_e)\right].
\end{aligned}$$

It thus suffices to show that the expected cost incurred by each $e \in E$ is bounded by twice $\sum_{\ell\in L} w^\ell(e)x^\ell_e + w^-(e)(1-x_e)$. Observe that

$$\begin{aligned}
\mathbb{E}[\mathsf{cost}(e)] &= \textstyle\sum_{\ell\in L} \left[w^\ell(e)\Pr[e \text{ separated}] + (1-w^\ell(e))\Pr[e \subseteq C \wedge \chi(C) = \ell]\right] \\
&\le \textstyle\sum_{\ell\in L} \left[w^\ell(e)\,2x_e + (1-w^\ell(e))\,(1-x^\ell_e)\right] \\
&= \textstyle\sum_{\ell\in L} \left[w^\ell(e)(x_e + x^\ell_e) + (1-x^\ell_e) - w^\ell(e)(1-x_e)\right] \\
&= \textstyle\sum_{\ell\in L} w^\ell(e)(x_e + x^\ell_e) + w^-(e)(1-x_e) \\
&\le 2\textstyle\sum_{\ell\in L} w^\ell(e)x^\ell_e + w^-(e)(1-x_e),
\end{aligned}$$

where the first inequality follows from Lemma 4, the last equality from the fact that $w^-(e) = 1 - \sum_{\ell\in L} w^\ell(e)$ and $1 - x_e = \sum_{\ell\in L}(1-x^\ell_e)$, and the last inequality from the fact that $x_e \le x^\ell_e$ for any $\ell \in L$.

The same argument used in the proof of Theorem 5 for termination completes the proof. ◀

5 Preclustering

Due to Theorems 5 and 6, to obtain a $(2+\varepsilon)$-approximation algorithm, it suffices to argue that a nearly optimal solution to the chromatic cluster LPs can be found in polynomial time. To this end, we adapt the notion of a *preclustered instance* [13, 9] to the chromatic versions as follows.

▶ **Definition 7** (Preclustered instance). *Given an instance, a* preclustered instance *is defined by a triple* $(\mathcal{K}, \chi_{\mathsf{pre}}, E_{\mathsf{adm}})$ *such that*

- $\mathcal{K}$ *is a partition of* V*;*
- $\chi_{\mathsf{pre}} : \mathcal{K}_{>1} \to L$ *is a color assignment for every non-singleton part in* $\mathcal{K}$*, where we denote by* $\mathcal{K}_{>1} := \{K \in \mathcal{K} \mid |K| > 1\}$ *the set of non-singleton parts of* $\mathcal{K}$*; and*
- $E_{\mathsf{adm}} \subseteq \{uv \in \binom{V}{2} \mid \nexists K \in \mathcal{K} \text{ such that } uv \subseteq K\}$ *is a set of edges across* $\mathcal{K}$*.*

We call $\mathcal{K}$ *and* E_{adm} *a* preclustering *and* admissible *edges, respectively. Each part in* $\mathcal{K}$ *is called a* precluster*. Any pair* uv *of two vertices* u *and* v *in a same precluster is called an* atomic *edge. Any pair that is neither atomic nor admissible is called a* non-admissible *edge.*

▶ **Definition 8** (Respecting solution). *Given a preclustered instance* $(\mathcal{K}, \chi_{\mathsf{pre}}, E_{\mathsf{adm}})$ *and a feasible solution* $(\mathcal{C}, \chi)$*, we say the solution* respects *the preclustered instance when*

- $\mathcal{K}$ *subdivides* $\mathcal{C}$*, i.e., for every precluster* $K \in \mathcal{K}$*, there exists a cluster* $C \in \mathcal{C}$ *such that* $K \subseteq C$*;*
- *for every non-singleton precluster* $K \in \mathcal{K}_{>1}$ *and the cluster* $C \in \mathcal{C}$ *containing* K*,* $\chi_{\mathsf{pre}}(K) = \chi(C)$*; and*
- *for every non-admissible edge* uv*, the two vertices* u *and* v *are not in a same cluster in the solution.*

In this section, we prove the following theorem.

▶ **Theorem 9.** *For any sufficiently small constant* $\varepsilon > 0$*, there exists an algorithm that, in time* $n^{\mathsf{poly}(\varepsilon^{-1})}$*, outputs a preclustered instance* $(\mathcal{K}, \chi_{\mathsf{pre}}, E_{\mathsf{adm}})$ *such that there exists a solution respecting the preclustered instance whose cost is at most* $(1+\varepsilon)\,\mathrm{opt}$*, and* $|E_{\mathsf{adm}}| \leq O(\varepsilon^{-2})\,\mathrm{opt}$*, where* opt *denotes the cost of an optimal solution in the instance input to the algorithm.*

5.1 Constructing $\mathcal{K}$ and χ_{pre}

In this subsection, we present an algorithm for constructing a preclustering $\mathcal{K}$ and a color assignment $\chi_{\mathsf{pre}} : \mathcal{K}_{>1} \to L$. The algorithm is an adaptation of Cao et al.'s construction [9] to the chromatic versions. In particular, one subtlety in our algorithm is that we do not assume a self-loop for each vertex which is no loss of generality in the monochromatic version.

Algorithm description. We first describe the algorithm for CCC, and then state the changes needed for WCCC. We use parameters $\alpha, \beta \in (0, 1)$ such that $\eta := \frac{\alpha+\beta}{1-\beta} \in (0, \frac{1}{13})$. Note that $\alpha = \beta = 0.02$ is one feasible choice.[2]

At the very start, we obtain a solution $(\mathcal{C}_{\mathsf{init}}, \chi_{\mathsf{init}})$ by running any constant-factor approximation algorithm for CCC on the given instance. We then iterate over every vertex $u \in V$ in an arbitrary order, and mark u if either

$$|E^{-\ell}(u, C)| \geq \alpha(|C| - 1) \text{ or } |E^{+}(u, V \setminus C)| \geq \alpha(|C| - 1) \text{ (or both)}, \tag{3}$$

where we denote by $C \in \mathcal{C}_{\mathsf{init}}$ the cluster containing u and by $\ell := \chi_{\mathsf{init}}(C)$ its color. After marking all such vertices, we now iterate over every cluster $C \in \mathcal{C}_{\mathsf{init}}$, and if the number of marked vertices in C is at least $\beta(|C| - 1)$, we mark all the vertices in C. We obtain $\mathcal{K}$ by making all marked vertices singleton clusters from $\mathcal{C}_{\mathsf{init}}$. The color assignment χ_{pre} is constructed by inheriting χ_{init}, i.e., for any non-singleton $K \in \mathcal{K}_{>1}$, we have $\chi_{\mathsf{pre}}(K) = \chi_{\mathsf{init}}(C)$, where $C \supseteq K$ is the cluster in $\mathcal{C}_{\mathsf{init}}$ containing K.

[2] The choice of α and β affects the constant hidden in the exponent of the running time $n^{\mathsf{poly}(\varepsilon^{-1})}$ for constructing a nearly optimal solution to the chromatic cluster LPs: this constant decreases as $\alpha\beta$ increases.

For WCCC, we only need to change the condition (3) of $u \in V$ being marked in the first iteration to

$w^{-\ell}(u, C) \geq \alpha(|C|-1)$ or $w^{+}(u, V \setminus C) \geq \alpha(|C|-1)$ (or both).

Recall that we present a constant-factor approximation algorithm for WCCC in Section 3.

Sketch of analysis. Let us now sketch the analysis of this algorithm. We focus on CCC while we mention the changes needed for WCCC at the end. The full analysis for CCC is given in Appendix B.1; while the proof for WCCC can be found in the full version.

We aim at proving the first two conditions of Definition 8 are satisfied by any optimal solution with the constructed $\mathcal{K}$ and χ_{pre}. Fix an arbitrary optimal solution $(\mathcal{C}^\star, \chi^\star)$. The next two lemmas are useful in the proof of the main lemmas of this subsection.

▶ **Lemma 10.** *For any non-singleton precluster $K \in \mathcal{K}_{>1}$ with $\ell := \chi_{\mathsf{pre}}(K)$ and $u \in K$, we have $|E^{-\ell}(u, K)| < \eta(|K|-1)$ and $|E^{+}(u, V \setminus K)| < \eta(|K|-1)$.*

▶ **Lemma 11.** *For any non-singleton precluster $K \in \mathcal{K}_{>1}$, if a cluster $C_1^\star \in \mathcal{C}^\star$ colored $\chi_{\mathsf{pre}}(K)$ intersects K (i.e., $\chi^\star(C_1^\star) = \chi_{\mathsf{pre}}(K)$ and $K \cap C_1^\star \neq \emptyset$), we have $|C_1^\star \setminus K| < 2\eta|K|$.*

Intuitively speaking, the former lemma follows from the execution of the algorithm in a fairly straightforward manner. The latter can be derived because, otherwise, splitting $C_1^\star$ into $C_1^\star \setminus K$ and $C_1^\star \cap K$ in $\mathcal{C}^\star$ incurs less cost due to the second condition of Lemma 10.

The following is the main technical lemma for the first condition of Definition 8. We provide a proof sketch of this lemma while the full proof can be found in Appendix B.1.

▶ **Lemma 12.** *$\mathcal{K}$ subdivides $\mathcal{C}^\star$, i.e., for any $K \in \mathcal{K}$, there is $C^\star \in \mathcal{C}^\star$ such that $K \subseteq C^\star$.*

Proof sketch. Suppose toward contradiction that $\mathcal{C}^\star$ is not subdivided by $\mathcal{K}$. Then, there exists a non-singleton precluster $K \in \mathcal{K}_{>1}$ such that $K \setminus C^\star$ is nonempty for every cluster $C^\star \in \mathcal{C}^\star$. We choose an arbitrary such precluster $K \in \mathcal{K}_{>1}$, and let $\ell := \chi_{\mathsf{pre}}(K)$ be the color of K in the preclustering. Let ξ be a constant between $\frac{1+5\eta}{2}$ and $1-4\eta$. Note that we can always choose such ξ because $\frac{1+5\eta}{2} < 1-4\eta$ for $\eta \in (0, \frac{1}{13})$. Observe that $\frac{1}{2} < \xi < 1$.

We break the proof into two cases depending on the existence of a cluster $C^\star \in \mathcal{C}^\star$ that "dominantly" intersects K, i.e., $|K \cap C^\star| \geq \xi|K|$. In each case, we respectively construct a new solution whose cost is strictly less than $(\mathcal{C}^\star, \chi^\star)$, leading to contradiction that $(\mathcal{C}^\star, \chi^\star)$ is optimal.

Case 1. No dominant intersections. Suppose that, for all $C^\star \in \mathcal{C}^\star$, $|K \cap C^\star| < \xi|K|$. In this case, we consider a new solution $(\mathcal{C}, \chi)$ by removing $K \cap C^\star$ from each $C^\star \in \mathcal{C}^\star$ and inserting K colored $\ell := \chi_{\mathsf{pre}}(K)$.

The "gain" of $(\mathcal{C}, \chi)$ over $(\mathcal{C}^\star, \chi^\star)$ is at least the total number of ℓ-colored edges within K but crossing $\mathcal{C}^\star$. For each $u \in K$, at least $(1-\xi)|K|$ edges are crossing $\mathcal{C}^\star$ inside K due to the condition of this case, while at most $\eta|K|$ edges can be colored other than ℓ due to Lemma 10, implying that the gain is at least $\frac{1-\xi-\eta}{2}|K|^2$.

On the other hand, the "loss" of $(\mathcal{C}, \chi)$ over $(\mathcal{C}^\star, \chi^\star)$ is due to either the appropriately colored $+$ edges crossing K but within a cluster in $\mathcal{C}^\star$ or the $-$ edges within K but crossing $\mathcal{C}^\star$. By Lemma 10, we can bound this loss from above by $\frac{3\eta}{2}|K|^2 < \frac{1-\xi-\eta}{2}|K|^2$, where the inequality is due to the choice of ξ.

Case 2. A dominant intersection exists. Suppose now there exists a cluster $C_1^\star \in \mathcal{C}^\star$ such that $|K \cap C_1^\star| \geq \xi|K|$. Note that we always have at most one such cluster since $\xi > \frac{1}{2}$. Let $K_1 := K \cap C_1^\star$.

Assume for now that $\chi^\star(C_1^\star) = \chi_{\mathsf{pre}}(K)$. We construct a solution $(\mathcal{C}, \chi)$ by transferring one arbitrary vertex $u \in K \setminus C_1^\star$ to $C_1^\star$. We can then see that the gain of $(\mathcal{C}, \chi)$ over $(\mathcal{C}^\star, \chi^\star)$ is due to the ℓ-colored edges between u and $C_1^\star$, whereas the loss is due to the appropriately colored + edges incident with u within the cluster containing u in $\mathcal{C}^\star$ and edges other than ℓ-colored ones between u and $C_1^\star$. Using the dominance condition of this case and Lemma 10, we can show that the gain is at least $(\xi - \eta)|K|$.

To bound the loss from above, we prove that the number of the first type of edges is at most $(1 - \xi + \eta)|K|$ using the fact that $|K \setminus C_1^\star| \leq (1 - \xi)|K|$ due to the case condition, and Lemma 10. On the other hand, the second type of edges can be partitioned into ones within K and ones incident with $C_1^\star \setminus K$; the former is bounded by $\eta|K|$ due to Lemma 10 while the latter is bounded by $2\eta|K|$ due to Lemma 11. Hence, the loss cannot exceed $(1 - \xi + 4\eta)|K| < (\xi - \eta)|K|$ by the choice of ξ.

Finally, for the case where $\chi^\star(C_1^\star) \neq \chi_{\mathsf{pre}}(K)$, we construct a solution $(\mathcal{C}, \chi)$ from $(\mathcal{C}^\star, \chi^\star)$ by splitting $C_1^\star$ into $K_1 := K \cap C_1^\star$ colored $\ell = \chi_{\mathsf{pre}}(K)$ and $C_1^\star \setminus K_1$ colored $\ell' := \chi^\star(C_1^\star)$. We observe that the gain of $(\mathcal{C}, \chi)$ against $(\mathcal{C}^\star, \chi^\star)$ is at least the number of the ℓ-colored edges within K_1 while the loss is at most the number of the ℓ'-colored edges within K_1 and between K_1 and $C_1^\star \setminus K_1$. Through a similar argument using Lemma 10, we can show that the gain is at least $\frac{1}{2}|K_1|(\xi|K| - 1 - \eta(|K| - 1))$ while the loss is at most $\frac{3}{2}\eta|K_1|(|K| - 1)$. The net gain is therefore at least $\frac{1}{2}|K_1| \cdot (\xi|K| - 1 - 4\eta(|K| - 1)) > 0$, where the inequality can be derived from the fact that K is non-singleton and the choice of ξ and η. ◀

The second condition of Definition 8 is also satisfied through a similar argument. The proof can be found in Appendix B.1.

▶ **Lemma 13.** *For every cluster $C^\star \in \mathcal{C}^\star$ and any non-singleton precluster $K \in \mathcal{K}$ contained in $C^\star$, we have $\chi_{\mathsf{pre}}(K) = \chi^\star(C^\star)$.*

For WCCC, we generalize the above analysis for CCC by substituting the number of edges satisfying a condition with the total sum of edge weights satisfying the same condition. In particular, instead of Lemma 10 for CCC, we have the following lemma for WCCC:

▶ **Lemma 14.** *For any non-singleton precluster $K \in \mathcal{K}_{>1}$ with $\ell := \chi_{\mathsf{pre}}(K)$ and $u \in K$, we have $w^{-\ell}(u, K) < \eta(|K| - 1)$ and $w^+(u, V \setminus K) < \eta(|K| - 1)$.*

Likewise, we have the counterparts of Lemmas 11, 12, and 13 for WCCC. The complete analysis can be found in the full version.

5.2 Constructing E_{adm}

We now present the construction of admissible edges E_{adm}. Interestingly, it turns out that a color-oblivious construction as in Cao et al. [9] suffices for the chromatic versions. Here we present its construction in the notation of WCCC.

For the unweighted version, it is noteworthy that the original construction [9] for the unweighted monochromatic version already incorporates weights. Hence, there exists a straightforward interpretation for CCC where we define $w^+(S, S') := |\{uv \in E^+ : u \in S, v \in S'\}|$ and $w^-(S, S') := |\{uv \in E^- : u \in S, v \in S'\}|$ for $S, S' \subseteq V$.

Given $\mathcal{K}$ from the previous section, the construction of E_{adm} is as follows. For $K \in \mathcal{K}$, let $d(K) := \frac{w^+(K, V \setminus K)}{|K|} + \frac{|K|}{2}$, $N_1(K) := \{K' \in \mathcal{K} \mid \varepsilon \cdot d(K') < d(K) < \frac{1}{\varepsilon} \cdot d(K')\}$, and $N_2(K) := \{K' \in \mathcal{K} \setminus \{K\} \mid K, K' \in N_1(K) \cap N_1(K')\}$. For distinct $K, K' \in \mathcal{K}$ and $\mathcal{K}' \subseteq \mathcal{K}$ such that $K, K' \in \mathcal{K}'$, we further define

$$W(K,K',\mathcal{K}') := \sum_{K''\in\mathcal{K}'\setminus\{K,K'\}} \left[|K''|\cdot\frac{w^+(K,K'')}{|K|\cdot|K''|}\cdot\frac{w^+(K',K'')}{|K'|\cdot|K''|}\right] + \frac{w^+(K,K')}{|K|} + \frac{w^+(K',K)}{|K'|}$$

Moreover, for two distinct preclusters $K, K' \in \mathcal{K}$, we denote by $K \sim K'$ if $K' \in N_2(K)$ and $W(K,K',N_1(K)\cap N_1(K')) > \varepsilon(d(K)+d(K'))$. We can then define $E_{\mathsf{adm}} := \{uv \in E \mid \exists K,K' \in \mathcal{K}$ such that $K \sim K', u \in K, v \in K'\}$.

This choice of E_{adm}, together with $\mathcal{K}$ and χ_{pre} from the previous section, satisfies the conditions of Theorem 9; the complete proof is given in the full version.

6 Solving the chromatic cluster LP

In this section, we present an algorithm for solving the chromatic cluster LP nearly optimally, using the preclustered instance $(\mathcal{K},\chi_{\mathsf{pre}},E_{\mathsf{adm}})$ satisfying Theorem 9. Our technique is obtained by introducing one more dimension of color to Cao et al.'s approach [9]. Recall that they first formulate the *bounded sub-cluster LP* making use of a constant-level Sherali-Adams hierarchy, present a procedure that samples one cluster using Raghavendra and Tan's rounding scheme [22], and lastly argue that the Monte Carlo method with sufficiently many samples from the procedure produces a nearly optimal solution with high probability.

In what follows, we describe the algorithm for solving the chromatic cluster LPs, adapted from Cao et al.'s approach to the chromatic versions. We remark that there are no differences in the algorithm and analysis between CCC and WCCC except the objective function of the bounded sub-cluster LP. The entire analysis is given in the full version.

Bounded sub-cluster LP. We first present the bounded sub-cluster LP for the chromatic versions. Given $\varepsilon > 0$ sufficiently small, we define $\varepsilon_1 := \varepsilon^3$, $\varepsilon_{\mathrm{rt}} := \varepsilon_1^2 = \varepsilon^6$, and $r := \Theta(\varepsilon_{\mathrm{rt}}^{-2}) = \Theta(\varepsilon^{-12})$. We assume that r is an integer where a large enough constant is hidden in its Θ-notation. For $u \in V$, let $K_u \in \mathcal{K}$ be the precluster containing u. For $u \in V$, let $N_{\mathsf{adm}}(u) := \{v \in V \mid uv \in E_{\mathsf{adm}}\}$; for $K \in \mathcal{K}$, let $N_{\mathsf{adm}}(K) := \{v \in V \mid \exists u \in K, uv \in E_{\mathsf{adm}}\}$. Then, the bounded sub-cluster LP is defined as follows:

$$\begin{aligned}
&\min\ \mathrm{obj}(x) \\
&\text{s.t. } y_S^\ell = \textstyle\sum_{s=1}^n y_S^{\ell,s}, && \forall \ell\in L\ \forall S\subseteq V : |S|\le r,\\
&\quad t_v^\ell = 1 - y_v^\ell, && \forall \ell\in L\ \forall v\in V,\\
&\quad x_{uv}^\ell = 1 - y_{uv}^\ell, && \forall \ell\in L\ \forall uv\in \tbinom{V}{2},\\
&\quad \textstyle\sum_{\ell\in L} y_v^\ell = 1, && \forall v\in V,\\
&\quad \textstyle\sum_{v\in V} y_{Sv}^{\ell,s} = s\, y_S^{\ell,s}, && \forall \ell\in L\ \forall s\in[n]\ \forall S\subseteq V : |S|\le r-1,\\
&\quad x_{uv}^{\chi_{\mathsf{pre}}(K)} = 0, && \forall uv\subseteq K\in\mathcal{K}_{>1},\\
&\quad x_{uv}^\ell = 1, && \forall uv\subseteq K\in\mathcal{K}_{>1}\ \forall \ell\ne\chi_{\mathsf{pre}}(K),\\
&\quad y_{uv}^{\ell,s} = y_u^{\ell,s}, && \forall \ell\in L\ \forall s\in[n]\ \forall u\in V\ \forall v\in K_u,\\
&\quad x_{uv}^\ell = 1, && \forall \text{non-admissible } uv\ \forall \ell\in L,\\
&\quad y_v^{\ell,s} = 0,\\
&\qquad\qquad \forall \ell\in L\ \forall v\in V\ \forall s\in[|K_v|-1]\cup[|K_v|+1, |K_v|+\varepsilon_1|N_{\mathsf{adm}}(v)|],
\end{aligned}$$

$$y_S^{\chi_{\mathsf{pre}}(K),|K|} = y_v^{\chi_{\mathsf{pre}}(K),|K|}, \qquad \forall v \in K \in \mathcal{K}_{>1}\ \forall S \subseteq K : |S| \le r,$$
$$y_S^{\chi_{\mathsf{pre}}(K),|K|} = 0, \qquad \forall K \in \mathcal{K}_{>1}\ \forall S \not\subseteq K : |S| \le r, S \cap K \neq \emptyset,$$
$$\textstyle\sum_{T' \subseteq T} (-1)^{|T'|} y_{S \cup T'}^{\ell,s} \in [0, y_S^{\ell,s}],$$
$$\forall \ell \in L\ \forall s \in [n]\ \forall S, T \subseteq V : S \cap T = \emptyset, |S \cup T| \le r,$$
$$y_S^{\ell,s} \ge 0, \qquad \forall \ell \in L\ \forall s \in [n]\ \forall S \subseteq V : |S| \le r,$$

where $y_S^{\ell,s} = 1$ indicates whether S is a subset of a cluster of size s colored ℓ. The only distinction between CCC and WCCC is in the objective function, $\mathrm{obj}(x)$. Since the size of the bounded sub-cluster LP is indeed bounded by $n^{\mathsf{poly}(\varepsilon^{-1})}$, we can solve this LP in time $n^{\mathsf{poly}(\varepsilon^{-1})}$.

Sampling one colored cluster. We now present the procedure for sampling one cluster from the bounded sub-cluster LP. Let y be an optimal solution to the bounded sub-cluster LP. The procedure randomly samples a color $\ell \in L$ with probability $y_\emptyset^\ell / y_\emptyset$ where $y_\emptyset := \sum_{\ell' \in L} y_\emptyset^{\ell'}$, a cardinality $s \in [n]$ with probability $y_\emptyset^{\ell,s}/y_\emptyset^\ell$, and a pivot $u \in V$ with probability $y_u^{\ell,s}/(s \cdot y_\emptyset^{\ell,s})$. With these sampled ℓ, s, and u, define y' as $y'_S := y_{Su}^{\ell,s}/y_u^{\ell,s}$ for every $S \subseteq V$ of size $|S| \le r-1$. The procedure then runs Raghavendra and Tan's rounding scheme [22] upon y' to sample a cluster $C \subseteq V$, and return C colored $\chi(C) := \ell$.

Monte Carlo method. Finally, we describe the Monte Carlo method for constructing a nearly optimal solution. Let $\Delta := \Theta\left(\frac{n^4 \log n}{\varepsilon_1^2 |E_{\mathsf{adm}}|}\right)$ with a sufficiently large hidden constant. (Here, we assume $E_{\mathsf{adm}} \neq \emptyset$; otherwise, $(\mathcal{K}, \chi_{\mathsf{pre}})$ given by Theorem 9, with arbitrary colors assigned to singleton preclusters, is already a $(1+\varepsilon)$-approximate integral solution.) In particular, we assume that $\Delta y_\emptyset$ is an integer and $\Delta \ge \frac{1}{\varepsilon - \varepsilon_1} = \frac{1}{\varepsilon - \varepsilon^3}$. The method independently samples colored clusters from the above procedure for $\Delta y_\emptyset$ times. Let $C_1, \ldots, C_{\Delta y_\emptyset}$ be the sampled colored clusters. Let $\ell_1, \ldots, \ell_{\Delta y_\emptyset}$ denote their corresponding colors. For each vertex $v \in V$, let $R_v := \{t \in \{1, 2, \ldots, \Delta y_\emptyset\} \mid v \in C_t\}$, and let $\tilde{R}_v$ be the set of the $\lceil (1-\varepsilon)\Delta \rceil$ smallest indices in R_v. For every $t \in \{1, 2, \ldots, \Delta y_\emptyset\}$, we define $\tilde{C}_t := \{v \in V \mid t \in \tilde{R}_v\} \subseteq C_t$. The final output is $\tilde{z}_S^\ell := \frac{|\{t | \tilde{C}_t = S \wedge \ell_t = \ell\}|}{\lceil (1-\varepsilon)\Delta \rceil}$ for every $\ell \in L$ and $S \subseteq V$.

References

1 Fateme Abbasi, Hyung-Chan An, Jarosław Byrka, Changyeol Lee, and Yongho Shin. Approximating (weighted) chromatic correlation clustering via cluster LP. *arXiv preprint*, 2026. `arXiv:2510.13446`.

2 Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM*, 55(5):1–27, 2008. `doi:10.1145/1411509.1411513`.

3 Ilya Amburg, Nate Veldt, and Austin Benson. Clustering in graphs and hypergraphs with categorical edge labels. In *Proceedings of The Web Conference (WWW)*, pages 706–717, 2020. `doi:10.1145/3366423.3380152`.

4 Yael Anava, Noa Avigdor-Elgrabli, and Iftah Gamzu. Improved theoretical and practical guarantees for chromatic correlation clustering. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 55–65, 2015. `doi:10.1145/2736277.2741629`.

5 Eric Angel, Evripidis Bampis, A. Kononov, Dimitris Paparas, Emmanouil Pountourakis, and Vassilis Zissimopoulos. Clustering on k-edge-colored graphs. *Discrete Applied Mathematics*, 211:15–22, 2016. `doi:10.1016/J.DAM.2016.04.017`.

6 Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004. `doi:10.1023/B:MACH.0000033116.57574.95`.

7 Amir Ben-Dor and Zohar Yakhini. Clustering gene expression patterns. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB)*, pages 33–42, 1999. doi:10.1145/299432.299448.

8 Francesco Bonchi, Aristides Gionis, Francesco Gullo, Charalampos E. Tsourakakis, and Antti Ukkonen. Chromatic correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 9(4):1–24, 2015. doi:10.1145/2728170.

9 Nairen Cao, Vincent Cohen-Addad, Euiwoong Lee, Shi Li, Alantha Newman, and Lukas Vogl. Understanding the cluster linear program for correlation clustering. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1605–1616, 2024. doi:10.1145/3618260.3649749.

10 Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. Clustering with qualitative information. *Journal of Computer and System Sciences*, 71(3):360–383, 2005. doi:10.1016/J.JCSS.2004.10.012.

11 Shuchi Chawla, Konstantin Makarychev, Tselil Schramm, and Grigory Yaroslavtsev. Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing (STOC)*, pages 219–228, 2015. doi:10.1145/2746539.2746604.

12 Zhi-Zhong Chen, Tao Jiang, and Guohui Lin. Computing phylogenetic roots with bounded degrees and errors. *SIAM Journal on Computing*, 32(4):864–879, 2003. doi:10.1137/S0097539701389154.

13 Vincent Cohen-Addad, Euiwoong Lee, Shi Li, and Alantha Newman. Handling correlated rounding error via preclustering: A 1.73-approximation for correlation clustering. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1082–1104. IEEE, 2023. doi:10.1109/FOCS57990.2023.00065.

14 Vincent Cohen-Addad, Euiwoong Lee, and Alantha Newman. Correlation clustering with Sherali-Adams. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 651–661. IEEE, 2022. doi:10.1109/FOCS54457.2022.00068.

15 Patrick Doreian and Andrej Mrvar. A partitioning approach to structural balance. *Social Networks*, 18(2):149–168, 1996.

16 Chenglin Fan, Dahoon Lee, and Euiwoong Lee. 1.64-approximation for chromatic correlation clustering via chromatic cluster LP. In *Integer Programming and Combinatorial Optimization (IPCO)*, pages 81–96. Springer Nature Switzerland, 2026. doi:10.1007/978-3-032-28691-8_6.

17 Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1):4–es, 2007.

18 Nicolas Klodt, Lars Seifert, Arthur Zahn, Katrin Casel, Davis Issac, and Tobias Friedrich. A color-blind 3-approximation for chromatic correlation clustering and improved heuristics. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, pages 882–891, 2021. doi:10.1145/3447548.3467446.

19 Dahoon Lee, Chenglin Fan, and Euiwoong Lee. 1.64-approximation for chromatic correlation clustering via chromatic cluster LP. *arXiv preprint*, 2025. doi:10.48550/arXiv.2507.15417.

20 Dahoon Lee, Chenglin Fan, and Euiwoong Lee. Improved approximation algorithms for chromatic and pseudometric-weighted correlation clustering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

21 Mojtaba Ostovari and Alireza Zarei. Improved combinatorial approximations for weighted correlation clustering. *Discrete Mathematics & Theoretical Computer Science*, 27(Discrete Algorithms), 2025. doi:10.46298/DMTCS.12424.

22 Prasad Raghavendra and Ning Tan. Approximating CSPs with global cardinality constraints using SDP hierarchies. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 373–387. SIAM, 2012. doi:10.1137/1.9781611973099.33.

23 S. Regnie. On some mathematical aspects of automatic classification problems. *Math and Human Sciences*, 82:13–29, 1965.

24 Ron Shamir, Roded Sharan, and Dekel Tsur. Cluster graph modification problems. *Discrete Applied Mathematics*, 144(1-2):173–182, 2004. doi:10.1016/J.DAM.2004.01.007.

25 Anke van Zuylen and David P. Williamson. Deterministic pivoting algorithms for constrained ranking and clustering problems. *Mathematics of Operations Research*, 34(3):594–620, 2009. doi:10.1287/MOOR.1090.0385.

26 Dewan F. Wahid and Elkafi Hassini. A literature review on correlation clustering: Cross-disciplinary taxonomy with bibliometric analysis. *Operations Research Forum*, 3(3), 2022. doi:10.1007/S43069-022-00156-6.

27 Qing Xiu, Kai Han, Jing Tang, Shuang Cui, and He Huang. Chromatic correlation clustering, revisited. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:26147–26159, 2022.

28 Charles T. Zahn, Jr. Approximating symmetric relations by equivalence relations. *Journal of the Society for Industrial and Applied Mathematics*, 12(4):840–847, 1964.

A Deferred analysis from Section 4

We can rephrase the objective function for CCC as follows:

$$\begin{aligned}
&\textstyle\sum_{S\subseteq V,\ell\in L}\left(\frac{|\delta^+(S)|}{2}+|E^{-\ell}[S]|\right)z_S^\ell\\
&=\textstyle\sum_{uv\in E^+}\left[\frac{1}{2}\sum_{\ell\in L}\left(\sum_{S\ni u,S\not\ni v}z_S^\ell+\sum_{S\not\ni u,S\ni v}z_S^\ell\right)+\sum_{S\supseteq uv,\ell\neq c_{uv}}z_S^\ell\right]\\
&\qquad+\textstyle\sum_{uv\in E^-}\sum_{\ell\in L}\sum_{S\supseteq uv}z_S^\ell\\
&=\textstyle\sum_{uv\in E^+}\left[\sum_{\ell\in L}x_{uv}^\ell-\sum_{\ell\in L}\frac{t_u^\ell+t_v^\ell}{2}+\sum_{\ell\neq c_{uv}}\sum_{S\supseteq uv}z_S^\ell\right]\\
&\qquad+\textstyle\sum_{uv\in E^-}\sum_{\ell\in L}\sum_{S\supseteq uv}z_S^\ell\\
&=\textstyle\sum_{uv\in E^+}\left[\sum_{\ell\in L}x_{uv}^\ell-(|L|-1)+\sum_{\ell\neq c_{uv}}(1-x_{uv}^\ell)\right]\\
&\qquad+\textstyle\sum_{uv\in E^-}\sum_{\ell\in L}\sum_{S\supseteq uv}z_S^\ell\\
&=\textstyle\sum_{uv\in E^+}x_{uv}^{c_{uv}}+\sum_{uv\in E^-}\sum_{\ell\in L}(1-x_{uv}^\ell)=:\mathrm{obj}(x),
\end{aligned}$$

where the second equality comes from (1) and the third from (2).

Using a similar derivation, we can also rephrase the objective function for WCCC as follows:

$$\begin{aligned}
&\sum_{S\subseteq V,\ell\in L}\left(\frac{w^+(S,V\setminus S)}{2}+w^{-\ell}(S,S)\right)z_S^\ell\\
&=\sum_{uv\in E}\left[\frac{w^+(u,v)}{2}\sum_{\ell\in L}\left(\sum_{S\ni u,S\not\ni v}z_S^\ell+\sum_{S\not\ni u,S\ni v}z_S^\ell\right)+\sum_{S\supseteq uv}\sum_{\ell\in L}w^{-\ell}(u,v)\cdot z_S^\ell\right]\\
&=\sum_{uv\in E}\left[w^+(u,v)\cdot\left(\sum_{\ell\in L}x_{uv}^\ell-\sum_{\ell\in L}\frac{t_u^\ell+t_v^\ell}{2}\right)+\sum_{S\supseteq uv}\sum_{\ell\in L}w^{-\ell}(u,v)\cdot z_S^\ell\right]\\
&=\sum_{uv\in E}\left[w^+(u,v)\cdot\left(\sum_{\ell\in L}x_{uv}^\ell-(|L|-1)\right)+\sum_{\ell\in L}w^{-\ell}(u,v)\cdot(1-x_{uv}^\ell)\right]\\
&=\sum_{uv\in E}\left[\sum_{\ell\in L}(w^+(u,v)-w^{-\ell}(u,v))\cdot x_{uv}^\ell-w^+(u,v)(|L|-1)+\sum_{\ell\in L}w^{-\ell}(u,v)\right]\\
&=\sum_{uv\in E}\left[\sum_{\ell\in L}(w^\ell(u,v)-w^-(u,v))\cdot x_{uv}^\ell+w^-(u,v)|L|\right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{uv\in E}\sum_{\ell\in L}\left(w^{\ell}(u,v)x^{\ell}_{uv} + w^{-}(u,v)(1-x^{\ell}_{uv})\right)\\
&=: \mathrm{obj}(x),
\end{aligned}
$$

where the second equality comes from (1), the third from (2), and the fifth from the definition of w^+ and $w^{-\ell}$.

Proof of Lemma 4. Consider the set S such that $S \cap \{u, v\} \neq \emptyset$ for the first time. We have

$$
\begin{aligned}
\textstyle\sum_{\ell'\in L, S: S\cap uv\neq\emptyset} z^{\ell'}_S &= \textstyle\sum_{\ell'\in L, S\ni u} z^{\ell'}_S + \sum_{\ell'\in L, S\ni v} z^{\ell'}_S - \sum_{\ell'\in L, S\supseteq uv} z^{\ell'}_S\\
&= 1 + x_{uv}.
\end{aligned}
$$

For the first statement, recall that $\sum_{S\supseteq uv} z^{\ell}_S = 1 - x^{\ell}_{uv}$. The second statement immediately follows from the fact that $\sum_{\ell'\in L, S\supseteq uv} z^{\ell'}_S = 1 - x_{uv}$. ◀

B Deferred analysis from Section 5

B.1 Constructing $\mathcal{K}$ and χ_{pre} for CCC

Recall that we use parameters $\alpha, \beta \in (0, 1)$ such that $\eta := \frac{\alpha+\beta}{1-\beta} \in (0, \frac{1}{13})$.

▶ **Lemma 15** (Restatement of Lemma 10). *For any non-singleton precluster $K \in \mathcal{K}_{>1}$ with $\ell := \chi_{\mathsf{pre}}(K)$ and $u \in K$, we have*

$$|E^{-\ell}(u, K)| < \eta(|K| - 1) \text{ and } |E^{+}(u, V \setminus K)| < \eta(|K| - 1).$$

Proof. Let $C \in \mathcal{C}_{\mathsf{init}}$ be the cluster containing $K \subseteq C$. Since K is non-singleton, the vertices in K remain unmarked during the execution of the algorithm, implying that

$$|C \setminus K| < \beta(|C| - 1). \tag{4}$$

Since $|C \setminus K| = (|C| - 1) - (|K| - 1)$, rearranging the terms yields

$$|C| - 1 < \frac{1}{1-\beta}(|K| - 1). \tag{5}$$

Let $\ell := \chi_{\mathsf{pre}}(K) = \chi_{\mathsf{init}}(C)$. We can also observe that

$$|E^{-\ell}(u, C)| < \alpha(|C| - 1) \text{ and } |E^{+}(u, V \setminus C)| < \alpha(|C| - 1) \tag{6}$$

since $u \in K$ is unmarked. We therefore have

$$|E^{-\ell}(u, K)| \leq |E^{-\ell}(u, C)| < \alpha(|C| - 1) < \frac{\alpha}{1-\beta}(|K| - 1) < \eta(|K| - 1),$$

where the first inequality follows from the fact that $E^{-\ell}(u, K) \subseteq E^{-\ell}(u, C)$ and the second-to-last inequality is due to Inequality (5). Furthermore, we have

$$
\begin{aligned}
|E^{+}(u, V \setminus K)| &= |E^{+}(u, C \setminus K)| + |E^{+}(u, V \setminus C)| \leq |C \setminus K| + |E^{+}(u, V \setminus C)|\\
&< \beta(|C| - 1) + \alpha(|C| - 1) < \frac{\alpha+\beta}{1-\beta}(|K| - 1) = \eta\,(|K| - 1),
\end{aligned}
$$

where the second inequality is derived from Inequalities (4) and (6) and the last from Inequality (5). ◀

In what follows, we may instead use the following slightly weaker inequalities in most cases:

$|E^{-\ell}(u, K)| < \eta|K|$ and $|E^{+}(u, V \setminus K)| < \eta|K|$.

Now, fix any optimal solution $(\mathcal{C}^\star, \chi^\star)$ to the given instance I.

▶ **Lemma 16** (Restatement of Lemma 11 for CCC)**.** *For any non-singleton precluster $K \in \mathcal{K}_{>1}$, if a cluster $C_1^\star \in \mathcal{C}^\star$ colored $\chi_{\mathsf{pre}}(K)$ intersects K (i.e., $\chi^\star(C_1^\star) = \chi_{\mathsf{pre}}(K)$ and $K \cap C_1^\star \neq \emptyset$), we have $|C_1^\star \setminus K| < 2\eta|K|$.*

Proof. Let $\ell := \chi_{\mathsf{pre}}(K) = \chi^\star(C_1^\star)$ and $K_1 := K \cap C_1^\star$. We define a solution $(\mathcal{C}, \chi)$ constructed from $(\mathcal{C}^\star, \chi^\star)$ by splitting $C_1^\star$ into K_1 and $C_1^\star \setminus K_1$. That is,

$$\mathcal{C} := \mathcal{C}^\star \setminus \{C_1^\star\} \cup \{K_1, C_1^\star \setminus K_1\} \text{ and } \chi(C) := \begin{cases} \ell, & \text{if } C \in \{K_1, C_1^\star \setminus K_1\}, \\ \chi^\star(C), & \text{if } C \in \mathcal{C} \cap \mathcal{C}^\star. \end{cases}$$

Suppose toward contradiction that $|C_1^\star \setminus K| \geq 2\eta|K|$. We show that the number of disagreements in $(\mathcal{C}, \chi)$ is less than that of $(\mathcal{C}^\star, \chi^\star)$. To this end, we classify all edges as follows, and determine whether each edge is a disagreement or not in $(\mathcal{C}, \chi)$ and in $(\mathcal{C}^\star, \chi^\star)$:

1. Each $-$ edge across K_1 and $C_1^\star \setminus K_1$ is an agreement in $(\mathcal{C}, \chi)$, but a disagreement in $(\mathcal{C}^\star, \chi^\star)$.
2. Each $-$ edge across K_1 and $V \setminus C_1^\star$ is an agreement in both solutions.
3. Each $+$ edge colored ℓ across K_1 and $C_1^\star \setminus K_1$ is a disagreement in $(\mathcal{C}, \chi)$, but an agreement in $(\mathcal{C}^\star, \chi^\star)$.
4. Each $+$ edge colored differently from ℓ across K_1 and $C_1^\star \setminus K_1$ is a disagreement in both solutions.
5. Each $+$ edge across K_1 and $V \setminus C_1^\star$ is a disagreement in both solutions.
6. Each edge within K_1 has the same status in both solutions.
7. The remaining edges outside $C_1^\star$ also have the same statuses in both solutions, respectively.

We first lower bound the "gain" in cost of the constructed solution $(\mathcal{C}, \chi)$ – the number of edges that are agreed in $(\mathcal{C}, \chi)$, but not in $(\mathcal{C}^\star, \chi^\star)$. Observe that the type 1 edges contribute to this gain. For any $u \in K_1$, note that

$$\begin{aligned} |E^-(u, C_1^\star \setminus K_1)| &= |E(u, C_1^\star \setminus K_1)| - |E^+(u, C_1^\star \setminus K_1)| \\ &= |E(u, C_1^\star \setminus K_1)| - |E^+(u, C_1^\star \setminus K)| \\ &\geq |E(u, C_1^\star \setminus K_1)| - |E^+(u, V \setminus K)| \\ &= |C_1^\star \setminus K_1| - |E^+(u, V \setminus K)| \\ &> |C_1^\star \setminus K_1| - \eta|K|, \end{aligned}$$

where the last equality comes from $u \notin C_1^\star \setminus K_1$, and the last inequality from Lemma 15 (recall that K is non-singleton). We can thus derive that the gain of $(\mathcal{C}, \chi)$, the total number of type 1 edges, is strictly greater than $|K_1| \cdot (|C_1^\star \setminus K_1| - \eta|K|) = |K_1| \cdot (|C_1^\star \setminus K| - \eta|K|)$.

We now upper bound the "loss" in cost of $(\mathcal{C}, \chi)$ – the number of edges that are agreed in $(\mathcal{C}^\star, \chi^\star)$, but not in $(\mathcal{C}, \chi)$. Note that the edges of type 3 contribute to this loss. For any $u \in K_1$, we have

$$|E^\ell(u, C_1^\star \setminus K_1)| = |E^\ell(u, C_1^\star \setminus K)| \leq |E^+(u, V \setminus K)| < \eta|K|$$

where the last inequality is again due to Lemma 15, yielding that the loss of $(\mathcal{C}, \chi)$ is strictly less than $|K_1| \cdot \eta|K|$.

We can thus conclude that the "net gain" in cost of the constructed solution $(\mathcal{C}, \chi)$ against $(\mathcal{C}^\star, \chi^\star)$ is strictly greater than

$$|K_1| \cdot (|C_1^\star \setminus K| - \eta|K|) - |K_1| \cdot \eta|K| = |K_1| \cdot (|C_1^\star \setminus K| - 2\eta|K|) \geq 0,$$

where the inequality is due to the assumption that $|C_1^\star \setminus K| \geq 2\eta|K|$, contradicting the optimality of $(\mathcal{C}^\star, \chi^\star)$. ◀

▶ **Lemma 17** (Restatement of Lemma 12). *$\mathcal{K}$ subdivides $\mathcal{C}^\star$, i.e., for any $K \in \mathcal{K}$, there is $C^\star \in \mathcal{C}^\star$ such that $K \subseteq C^\star$.*

Proof. Suppose toward contradiction that $\mathcal{C}^\star$ is not subdivided by $\mathcal{K}$. Then, there exists a non-singleton precluster $K \in \mathcal{K}_{>1}$ such that $K \setminus C^\star$ is nonempty for every cluster $C^\star \in \mathcal{C}^\star$. We choose an arbitrary such precluster $K \in \mathcal{K}_{>1}$, and let $\ell := \chi_{\mathsf{pre}}(K)$ be the color of K in the preclustering. Let ξ be a constant between $\frac{1+5\eta}{2}$ and $1 - 4\eta$. Note that we can always choose such ξ because $\frac{1+5\eta}{2} < 1 - 4\eta$ for $\eta \in (0, \frac{1}{13})$. Observe that $\frac{1}{2} < \xi < 1$.

We break the analysis into two cases depending on the existence of a cluster $C^\star \in \mathcal{C}^\star$ that "dominantly" intersects K, i.e., $|K \cap C^\star| \geq \xi|K|$. The overall structure of the proof for each case is similar to the proof of Lemma 16: we respectively construct a solution whose cost is strictly less than $(\mathcal{C}^\star, \chi^\star)$, leading to contradiction that $(\mathcal{C}^\star, \chi^\star)$ is optimal.

Case 1. No dominant intersections. Suppose that, for all $C^\star \in \mathcal{C}^\star$, $|K \cap C^\star| < \xi|K|$. Consider a new solution $(\mathcal{C}', \chi')$ constructed by inserting K colored $\ell = \chi_{\mathsf{pre}}(K)$. More precisely, we have

$$\mathcal{C}' := \{K\} \cup \{C^\star \setminus K \mid C^\star \in \mathcal{C}^\star \text{ with } |C^\star \setminus K| > 0\} \text{ and}$$
$$\chi'(C) := \begin{cases} \ell, & \text{if } C = K; \\ \chi^\star(C^\star), & \text{if } C = C^\star \setminus K. \end{cases}$$

We classify each edge and determine whether it is a disagreement or not in both solutions.

1. Each $-$ edge across K and $V \setminus K$ is agreed in $(\mathcal{C}', \chi')$, and possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
2. Each $+$ edge across K and $V \setminus K$ is disagreed in $(\mathcal{C}', \chi')$, but possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
3. Each $-$ edge within K is disagreed in $(\mathcal{C}', \chi')$, but possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
4. Each $+$ edge e of color $c_e = \ell$ within $K \cap C^\star$ for some $C^\star \in \mathcal{C}^\star$ is agreed in $(\mathcal{C}', \chi')$, and possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
5. Each $+$ edge e of color $c_e \neq \ell$ within $K \cap C^\star$ for some $C^\star \in \mathcal{C}^\star$ is disagreed in $(\mathcal{C}', \chi')$, but possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
6. Each $+$ edge e of color $c_e = \ell$ that is within K but across clusters in $\mathcal{C}^\star$ is agreed in $(\mathcal{C}', \chi')$, but disagreed in $(\mathcal{C}^\star, \chi^\star)$.
7. Each $+$ edge e of color $c_e \neq \ell$ that is within K but across clusters in $\mathcal{C}^\star$ is disagreed in both solutions.
8. The remaining edges outside K have the same statuses in both solutions, respectively.

Observe that the gain in cost of $(\mathcal{C}', \chi')$ against $(\mathcal{C}^\star, \chi^\star)$ is lower-bounded by the number of type 6 edges while the loss in cost of $(\mathcal{C}', \chi')$ is upper-bounded by the number of edges of types 2, 3, and 5.

To give a lower bound on the number of type 6 edges, we consider any $u \in K$ and let $C_u^\star \in \mathcal{C}^\star$ be the cluster containing u in $\mathcal{C}^\star$. We then have

$$\begin{aligned} |E^\ell(u, K \setminus C_u^\star)| &= |E(u, K \setminus C_u^\star)| - |E^{-\ell}(u, K \setminus C_u^\star)| \\ &\geq |E(u, K \setminus C_u^\star)| - |E^{-\ell}(u, K)| \\ &= |K \setminus C_u^\star| - |E^{-\ell}(u, K)| > |K \setminus C_u^\star| - \eta|K| \\ &\geq (1 - \xi)|K| - \eta|K| = (1 - \xi - \eta)|K|, \end{aligned}$$

where the second equality comes from $u \notin K \setminus C_u^\star$, the second inequality from Lemma 15, and the last inequality from the condition of this case. Hence, the number of type 6 edges is bounded from below by

$$\frac{1}{2}\sum_{u\in K}|E^{\ell}(u, K\setminus C_u^\star)| > \frac{1-\xi-\eta}{2}|K|^2.$$

This is also a lower bound for the gain in cost of $(\mathcal{C}', \chi')$ against $(\mathcal{C}^\star, \chi^\star)$.

We now consider the edges of types 2, 3 and 5. Due to the second inequality of Lemma 15, we can upper bound the number of type 2 edges by $|K|\cdot\eta|K| = \eta|K|^2$. On the other hand, the total number of type 3 and type 5 edges is bounded from above by

$$\frac{1}{2}\sum_{u\in K}|E^{-\ell}(u, K)| < \frac{\eta}{2}|K|^2$$

due to the first inequality of Lemma 15. Hence, the loss in cost of $(\mathcal{C}', \chi')$ against $(\mathcal{C}^\star, \chi^\star)$ is strictly less than $\frac{3\eta}{2}|K|^2$. Since $\xi \le 1-4\eta$, we have

$$\frac{1-\xi-\eta}{2}|K|^2 - \frac{3\eta}{2}|K|^2 \ge 0,$$

implying that $(\mathcal{C}', \chi')$ has fewer disagreements than $(\mathcal{C}^\star, \chi^\star)$. This contradicts the optimality of $(\mathcal{C}^\star, \chi^\star)$.

Case 2. A dominant intersection exists. Suppose now there exists a cluster $C_1^\star \in \mathcal{C}^\star$ such that $|K \cap C_1^\star| \ge \xi|K|$. Note that we always have at most one such cluster since $\xi > \frac{1}{2}$. Let $K_1 := K \cap C_1^\star$.

We first claim that $\chi^\star(C_1^\star) \ne \chi_{\mathsf{pre}}(K)$. Otherwise, if $\chi^\star(C_1^\star) = \chi_{\mathsf{pre}}(K)$, we construct a solution $(\mathcal{C}'', \chi'')$ by transferring one arbitrary vertex $w \in K \setminus C_1^\star$ to $C_1^\star$. In other words,

$$\mathcal{C}'' := \{C_1^\star \cup \{w\}\} \cup \{C^\star \setminus \{w\} \mid C^\star \in \mathcal{C}^\star \setminus \{C_1^\star\}\} \text{ and}$$
$$\chi''(C) = \begin{cases} \chi^\star(C_1^\star), & \text{if } C = C_1^\star \cup \{w\}, \\ \chi^\star(C^\star), & \text{if } C = C^\star \setminus \{w\}. \end{cases}$$

Since $K \setminus C_1^\star$ is nonempty, we can always choose $w \in K \setminus C_1^\star$ in the above construction. Recall also that $\ell = \chi_{\mathsf{pre}}(K) = \chi^\star(C_1^\star)$. Consider the following classification of edges:

1. Each edge in $E^{\ell}(w, C_1^\star)$ is agreed in $(\mathcal{C}'', \chi'')$, but disagreed in $(\mathcal{C}^\star, \chi^\star)$.
2. Each edge in $E^{-\ell}(w, C_1^\star)$ is disagreed in $(\mathcal{C}'', \chi'')$, but possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
3. Each edge in $E^{+}(w, V \setminus C_1^\star)$ is disagreed in $(\mathcal{C}'', \chi'')$, but possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
4. Each edge in $E^{-}(w, V \setminus C_1^\star)$ is agreed in $(\mathcal{C}'', \chi'')$, and possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
5. All the edges that are not incident with w have the same statuses in both solutions, respectively.

Hence, the gain in cost of $(\mathcal{C}'', \chi'')$ against $(\mathcal{C}^\star, \chi^\star)$ is bounded from below by $|E^{\ell}(w, C_1^\star)|$, whilst the loss in cost of $(\mathcal{C}'', \chi'')$ is bounded from above by $|E^{-\ell}(w, C_1^\star)| + |E^{+}(w, V \setminus C_1^\star)|$. For the type 1 edges, we have

$$\begin{aligned} |E^{\ell}(w, C_1^\star)| &\ge |E^{\ell}(w, K_1)| = |E(w, K_1)| - |E^{-\ell}(w, K_1)| \\ &= |K_1| - |E^{-\ell}(w, K_1)| \ge |K_1| - |E^{-\ell}(w, K)| \\ &> |K_1| - \eta|K| \ge \xi|K| - \eta|K| = (\xi-\eta)|K|, \end{aligned}$$

where the second equality comes from $w \notin K_1$, and the second-to-last inequality from Lemma 15. For type 2 edges,

$$\begin{aligned}|E^{-\ell}(w, C_1^\star)| &= |E^{-\ell}(w, K_1)| + |E^{-\ell}(w, C_1^\star \setminus K_1)| \\ &\leq |E^{-\ell}(w, K)| + |C_1^\star \setminus K_1| \\ &= |E^{-\ell}(w, K)| + |C_1^\star \setminus K| \\ &< \eta|K| + 2\eta|K| = 3\eta|K|,\end{aligned}$$

where the last inequality comes from Lemma 15 and Lemma 16. Lastly, for type 3 edges,

$$\begin{aligned}|E^{+}(w, V \setminus C_1^\star)| &= |E^{+}(w, K \setminus C_1^\star)| + |E^{+}(w, (V \setminus K) \setminus C_1^\star)| \\ &< |K \setminus C_1^\star| + |E^{+}(w, V \setminus K)| \\ &\leq (1-\xi)|K| + |E^{+}(w, V \setminus K)| \\ &< (1-\xi)|K| + \eta|K| = (1-\xi+\eta)|K|,\end{aligned}$$

where the second inequality comes from that $|K \cap C_1^\star| \geq \xi|K|$, and the last inequality is due to Lemma 15. Hence, the net gain in cost of $(\mathcal{C}'', \chi'')$ against $(\mathcal{C}^\star, \chi^\star)$ is strictly greater than

$$(\xi - \eta)|K| - (1 - \xi + 4\eta)|K| \geq 0,$$

where the inequality follows from the choice of $\xi > \frac{1+5\eta}{2}$, contradicting the optimality of $(\mathcal{C}^\star, \chi^\star)$. This completes the proof of our claim that $\chi^\star(C_1^\star) \neq \chi_{\mathsf{pre}}(K)$.

We now construct another solution $(\mathcal{C}, \chi)$ from $(\mathcal{C}^\star, \chi^\star)$ by splitting $C_1^\star$ into K_1 colored $\ell = \chi_{\mathsf{pre}}(K)$ and $C_1^\star \setminus K_1$ colored $\chi^\star(C_1^\star)$, i.e.,

$$\mathcal{C} := \mathcal{C}^\star \setminus \{C_1^\star\} \cup \{K_1, C_1^\star \setminus K_1\} \quad \text{and} \quad \chi(C) := \begin{cases} \ell, & \text{if } C = K_1, \\ \chi^\star(C_1^\star), & \text{if } C = C_1^\star \setminus K_1, \\ \chi^\star(C), & \text{if } C \in \mathcal{C} \cap \mathcal{C}^\star. \end{cases}.$$

Consider the following breakdown of the edges:

1. Each $-$ edge across K_1 and $V \setminus K_1$ is agreed in $(\mathcal{C}, \chi)$, and possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
2. Each $+$ edge across K_1 and $C_1^\star \setminus K_1$ is disagreed in $(\mathcal{C}, \chi)$, but possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
3. Each $+$ edge across K_1 and $V \setminus C_1^\star$ is disagreed in both solutions.
4. Each $-$ edge within K_1 is disagreed in both solutions.
5. Each $+$ edge e of color $c_e = \ell$ within K_1 is agreed in $(\mathcal{C}, \chi)$, but disagreed in $(\mathcal{C}^\star, \chi^\star)$.
6. Each $+$ edge e of color $c_e \neq \ell$ within K_1 is disagreed in $(\mathcal{C}, \chi)$, but possibly agreed in $(\mathcal{C}^\star, \chi^\star)$.
7. All the remaining edges outside K_1 have the same statuses in both solutions, respectively.

We can thus see that the gain in cost of $(\mathcal{C}, \chi)$ against $(\mathcal{C}^\star, \chi^\star)$ is at least the number of type 5 edges, while the loss is at most the total number of type 2 and type 6 edges.

To lower-bound the number of type 5 edges, we first fix an $u \in K_1$ and consider the type 5 edges incident to u. We then have

$$\begin{aligned}|E^{\ell}(u, K_1)| &= |E(u, K_1)| - |E^{-\ell}(u, K_1)| = |K_1| - 1 - |E^{-\ell}(u, K_1)| \\ &\geq |K_1| - 1 - |E^{-\ell}(u, K)| > |K_1| - 1 - \eta(|K| - 1) \\ &\geq \xi|K| - 1 - \eta(|K| - 1),\end{aligned}$$

where the second inequality comes from Lemma 15, and the last inequality comes from the condition of K_1. Hence, the number of type 5 edges is at least

$$\frac{1}{2}\sum_{u\in K_1}|E^\ell(u,K_1)| > \frac{1}{2}|K_1|\cdot(\xi|K|-1-\eta(|K|-1)).$$

We now upper-bound the number of edges of types 2 and 6. For any $u \in K_1$, we have

$$|E^+(u,C_1^\star\setminus K_1)| = |E^+(u,C_1^\star\setminus K)| \le |E^+(u,V\setminus K)| < \eta(|K|-1),$$

where the last inequality is due to Lemma 15, implying that the number of type 2 edges is strictly less than $|K_1|\cdot\eta(|K|-1)$. On the other hand, we also have that, for every $u \in K_1$,

$$|E^+\cap E^{-\ell}(u,K_1)| \le |E^{-\ell}(u,K)| < \eta(|K|-1)$$

where the last inequality comes from Lemma 15, yielding that the number of type 6 edges is strictly upper-bounded by $\frac{1}{2}|K_1|\cdot\eta(|K|-1)$.

We can thus derive that the net gain in cost of $(\mathcal{C},\chi)$ against $(\mathcal{C}^\star,\chi^\star)$ is greater than

$$\frac{1}{2}|K_1|\cdot(\xi|K|-1-\eta(|K|-1)) - \frac{1}{2}|K_1|\cdot 3\eta(|K|-1)$$
$$= \frac{1}{2}|K_1|\cdot(\xi|K|-1-4\eta(|K|-1)).$$

Observe that $\xi > \frac{1+5\eta}{2} > \frac{1}{2} > \frac{4}{13} > 4\eta$ by the choice of ξ and η. We can thus see

$$\xi|K|-1-4\eta(|K|-1) = (\xi-4\eta)|K|-1+4\eta \ge 2\xi-4\eta-1 > \eta,$$

where the first inequality comes from the fact that K is non-singleton and $\xi > 4\eta$, and the last inequality from the fact that $\xi > \frac{1+5\eta}{2}$. This again contradicts the optimality of $(\mathcal{C}^\star,\chi^\star)$, completing the entire proof of this case. ◀

The above lemma shows that each precluster $K \in \mathcal{K}$ is fully contained in some cluster $C^\star \in \mathcal{C}^\star$. We can also show that every non-singleton precluster must be colored the same as the cluster in $\mathcal{C}^\star$ containing it.

▶ **Lemma 18** (Restatement of Lemma 13 for CCC)**.** *For every cluster $C^\star \in \mathcal{C}^\star$ and any non-singleton precluster $K \in \mathcal{K}$ contained in $C^\star$, we have $\chi_{\mathsf{pre}}(K) = \chi^\star(C^\star)$.*

Proof. Suppose toward contradiction that there exists a non-singleton K and $C_1^\star$ such that $K \subseteq C_1^\star$ but $\chi_{\mathsf{pre}}(K) \neq \chi^\star(C_1^\star)$. We show that a solution $(\mathcal{C},\chi)$ obtained by splitting $C_1^\star$ into K of color $\chi_{\mathsf{pre}}(K)$ and $C_1^\star\setminus K$ of color $\chi^\star(C_1^\star)$ has fewer disagreements than $(\mathcal{C}^\star,\chi^\star)$. The proof indeed follows from exactly the same argument as in the latter half of Case 2 in the proof of Lemma 17. ◀

▶ **Lemma 19.** *$(\mathcal{K},\chi_{\mathsf{pre}})$ is a constant-approximate solution.*

Proof. We show that the number of disagreements in $(\mathcal{K},\chi_{\mathsf{pre}})$ is within a constant factor of that in $(\mathcal{C}_{\mathsf{init}},\chi_{\mathsf{init}})$; since $(\mathcal{C}_{\mathsf{init}},\chi_{\mathsf{init}})$ is a constant-approximate solution, this completes the proof. To this end, we charge the edges newly disagreed in $(\mathcal{K},\chi_{\mathsf{pre}})$ due to the algorithm to the edges already disagreed in $(\mathcal{C}_{\mathsf{init}},\chi_{\mathsf{init}})$.

Consider any non-singleton cluster $C \in \mathcal{C}_{\mathsf{init}}$ of color $\ell := \chi_{\mathsf{init}}(C)$. Observe that, if a vertex $u \in C$ becomes a singleton precluster in $\mathcal{K}$, the set of newly disagreed edges is precisely $E^\ell(u,C)$. Let

$$B_1 := \{u \in C \mid |E^{-\ell}(u,C)| \ge \alpha(|C|-1)\} \text{ and}$$
$$B_2 := \{u \in C \mid |E^+(u,V\setminus C)| \ge \alpha(|C|-1)\}.$$

Note that $B_1 \cup B_2$ is the set of marked vertices in C right after iterating over V. We break the analysis into two cases depending on the size of $B_1 \cup B_2$.

Case 1. $|B_1 \cup B_2| < \beta(|C| - 1)$. In this case, the vertices in $C \setminus (B_1 \cup B_2)$ remain unmarked, and hence, it suffices to consider the vertices in $B_1 \cup B_2$. For any $u \in B_1$, observe that

$$\begin{aligned}|E^{\ell}(u,C)| &= |E(u,C)| - |E^{-\ell}(u,C)| = |C| - 1 - |E^{-\ell}(u,C)| \\ &\le (1-\alpha)(|C|-1) \le \frac{1-\alpha}{\alpha}|E^{-\ell}(u,C)| < \frac{1}{\alpha}|E^{-\ell}(u,C)|,\end{aligned}$$

where the first and second inequalities are due to the definition of B_1; on the other hand, for $u \in B_2$, we have

$$|E^{\ell}(u,C)| \le |C| - 1 \le \frac{1}{\alpha}|E^{+}(u, V \setminus C)|,$$

where the last inequality is due to the definition of B_2. Hence, for any $u \in B_1 \cup B_2$, it is sufficient to cover $|E^{\ell}(u,C)|$ by charging $\frac{1}{\alpha}$ to each edge in $E^{-\ell}(u,C) \cup E^{+}(u, V \setminus C)$. Hence the total charged amount for the cluster C is at most

$$\frac{1}{\alpha}\left(2\,|\{uv \in E^{-\ell} \mid uv \subseteq C\}| + |\{uv \in E^{+} \mid |uv \cap C| = 1\}|\right).$$

Case 2. $|B_1 \cup B_2| \ge \beta(|C| - 1)$. Recall that, in this case, the algorithm marks all vertices in C. The total number of newly disagreed edges is at most $\binom{|C|}{2}$. Note that, by the definition of B_1,

$$|\{uv \in E^{-\ell} \mid uv \subseteq C\}| \ge \frac{1}{2}\sum_{u \in B_1}|E^{-\ell}(u,C)| \ge \frac{1}{2}|B_1| \cdot \alpha(|C|-1),$$

and, by the definition of B_2,

$$|\{uv \in E^{+} \mid |uv \cap C| = 1\}| \ge \sum_{u \in B_2}|E^{+}(u, V \setminus C)| \ge |B_2| \cdot \alpha(|C|-1).$$

We can thus derive that

$$\begin{aligned}&2\,|\{uv \in E^{-\ell} \mid uv \subseteq C\}| + |\{uv \in E^{+} \mid |uv \cap C| = 1\}| \\ &\ge \alpha(|B_1| + |B_2|)(|C|-1) \\ &\ge \alpha\beta(|C|-1)^2,\end{aligned}$$

where the last inequality is due to the condition that $|B_1| + |B_2| \ge |B_1 \cup B_2| \ge \beta(|C|-1)$. This yields

$$\begin{aligned}\binom{|C|}{2} &= \frac{|C|(|C|-1)}{2} \le (|C|-1)^2 \\ &\le \frac{1}{\alpha\beta}\left(2\,|\{uv \in E^{-\ell} \mid uv \subseteq C\}| + |\{uv \in E^{+} \mid |uv \cap C| = 1\}|\right),\end{aligned}$$

where the first inequality is because C is non-singleton. It therefore suffices to cover the new disagreements due to C by charging $\frac{2}{\alpha\beta}$ to each edge in $\{uv \in E^{-\ell} \mid uv \subseteq C\}$ and $\frac{1}{\alpha\beta}$ to each edge in $\{uv \in E^{+} \mid |uv \cap C| = 1\}$, respectively.

Due to the above charging argument, we can see that every disagreement in $(\mathcal{C}_{\mathsf{init}}, \chi_{\mathsf{init}})$ is charged at most $\frac{2}{\alpha\beta}$ while sufficiently covering the cost incurred by the new disagreements in $(\mathcal{K}, \chi_{\mathsf{pre}})$. ◀