Boston College

Department of Computer Science

Scholar of the College Thesis

submitted for the degree of

Bachelors of Science in Computer Science

A 3-approximation Algorithm for the Min Max Correlation Clustering Problem

by

Steven Roche

Submission Date: April 16, 2025

Supervisor: Professor Hsin-Hao Su

A 3-approximation Algorithm for the Min Max Correlation Clustering Problem

Steven Roche

Abstract

In this paper, we give a 3-approximation algorithm to the min max correlation clustering problem. Given a complete graph, vertices are related by positive and negative edges. A positive edge denotes *similar* vertices and a negative edge denotes *dissimilar* vertices, and the goal is to minimize the l_∞-norm of disagreements over all vertices. The 3-approximation is possible by observing a structural property of vertices with degree greater than or equal to 3ϕ, where ϕ is our guess of the optimal solution. A combinatorial argument demonstrates the correctness of the algorithm, which identifies the optimal ϕ and has a runtime of $O(n^2 D \log D \log n)$. This runtime includes determining the ϕ that corresponds to the optimal objective value over the graph.

1 Introduction

The correlation clustering problem is performed on a complete graph G where edges are labeled as either "$+$" or "$-$". These labels denote if vertices are "similar" or "dissimilar", or whether they should be clustered together or not. Given any clustering of the graph - any partition of the graph - an edge is considered to be in disagreement if it is a negative edge and its endpoints belong to the same cluster, or if it is a positive edge and its endpoints belong to different clusters. The total disagreement for a vertex v is denoted by $\rho(v)$, which counts the total number of edges incident to v that are in disagreement. The goal of the correlation clustering problem is to identify a clustering that minimizes an objective function measuring the disagreement of edges.

Puleo and Milenkovic [13] introduced an objective function aimed at minimizing the l_p-norm of disagreements over all vertices, defined as:

$$\left(\sum_{v \in V} \rho(v)^p\right)^{1/p}$$

Notable advancements have been made on the l_1-norm [5, 1, 6, 9, 8], which was the original correlation clustering proposed by Bansal, Blum, and Chawla [2] and seeks to minimize the total number of disagreements. A recent paper by Cao et al. presented a 1.437-algorithm for the l_1-norm [3].

For $p = \infty$, the objective focuses on minimizing the maximum disagreement per vertex, known as the min max correlation clustering problem, which remains less explored and is NP-hard in general [2]. This problem is of great interest, as the min max correlation clustering problem enables control over the similarity of clusters. This property distinguishes the l_∞-norm from other norms (such as the l_1-norm); the l_1-norm minimizes the total number of clustering errors at the cost of producing individual clusters that have little similarity. For example, consider a company that wants to ensure a minimum level of quality in the group of movie or product recommendations given to a user [13]. The l_1-norm will minimize the total number of product mismatches over all users, but some groups of users can receive many recommendations that are irrelevant to them. The l_∞-norm ensures that the number of irrelevant recommendations given to a group of users is upper bounded. An interesting example of min max correlation clustering is in gene expression data analysis; using the similarity of genes and conditions, Cheng and Church noted that gene-condition biclusters can produce clusters with high coherence [7]. Here, a bicluster refers to a simultaneous clustering of rows and columns in a matrix. In this particular case, each row of the matrix can correspond to a condition and each column can correspond to a gene. If genes and conditions form a bipartite graph as described by Cheng and Church, we can apply a min max correlation clustering algorithm to find groups of related genes and conditions. Moreover, the error in clusters will be bounded, ensuring all gene-condition clusters have a guaranteed level of coherence. This may be desirable over using the l_1 norm, as it can produce clusters where the conditions have little correlation to the genes. Other problems that form a bipartite graph fit this framework. Such an example is described by Puleo and Milenkovic [13]. Suppose viewers are on the left side of a bipartite graph and movies are on the right, and positive edges represent if a viewer likes the movie and negative edges represent if a viewer has not seen or dislikes a movie. Then running a min max correlation clustering problem will identify communities of viewers with similar movie taste. This information has clear applications in industry, where a company may desire to give movie recommendations to each identified group of viewers [13].

Turning back to the theory of min max correlation clustering, we review advancements on the problem. Puleo and Milenkovic [13] proposed an algorithm achieve a 48-approximation ratio, using metric linear programming followed by rounding. Using the same framework, this was improved to a 7-approximation by Charikar et al. [4], and Kalhan, Makarychev, and Zhou [12] further reduced it to a 5-approximation.

Davies, Moseley, and Newman [10] designed a combinatorial algorithm that achieves a 40-approximation ratio with a runtime of $O(n^2 \log n)$, where $n = |V|$ is the number of vertices in the graph. While it runs in polynomial time, it has a large approximation factor. The best known approximation ratio to date is 4, achieved by Heidrich, Irmai, and Andres [11], who also used a combinatorial approach with a runtime of $O(n^2 + nD^2)$, where D is the maximum degree of the graph.

Inspired by the Heidrich et al. paper, we make additional insights that enable a 3-approximation to be achieved with a runtime of $O(n^2 D \log D \log n)$. This is an advancement in this area of theoretical computer science, as this is the lowest bound achieved thus far on the min max correlation clustering problem. Note that throughout the paper, $|V| = n$ and m denotes the number of positive edges in the graph.

2 Technical Overview

Our primary technical contribution is achieving a new approximation factor of 3. Given an estimate ϕ for the optimal objective value, Heidrich et al. [11] observed that when the optimal objective OPT satisfies $OPT \le \phi$, certain neighborhood properties emerge. Specifically, if the neighborhoods of vertices u and v share at least 2ϕ elements, u and v must belong to the same cluster in the optimal solution. Conversely, if u and v differ by more than 2ϕ elements, they belong to different clusters.

Building on these properties, they showed that clusters can be determined for vertices with degrees of at least 4ϕ. For any vertex x with $\deg(x) \ge 4\phi$, it holds that for every other vertex y, either $|N[x] \cap N[y]| \ge 2\phi$ or $|N[x] \Delta N[y]| \ge 2\phi$. Here, $N[x] = N(x) \cup \{x\}$ represents the closed neighborhood of x, and Δ is the symmetric difference - or XOR operator. The remaining vertices, which do not meet this high-degree condition, can be placed in singleton clusters, since their disagreements per vertex are upper-bounded by their degrees, 4ϕ. This leads to a clustering with disagreements bounded by 4ϕ, yielding a 4-approximation algorithm.

To improve this result and achieve a 3-approximation, we first observe that for any two vertices x and y with degrees greater than 3ϕ, it must also hold that either $|N[x] \cap N[y]| \ge 2\phi$ or $|N[x] \Delta N[y]| \ge 2\phi$. Thus, whether x and y belong to the same cluster is uniquely determined in the optimal solution. As a result, the clustering induced on these high-degree vertices (those with $\deg(x) > 3\phi$) is fully determined.

The remaining challenge is to handle the placement of low-degree vertices, i.e., those with degrees upper bounded by 3ϕ. Unlike the high-degree case, it is not immediately clear whether these low-degree vertices should form singleton clusters or join existing clusters, as including them incorrectly could increase the disagreement associated with high-degree vertices.

We address this by showing that low-degree vertices can be assigned to high-degree clusters while keeping the maximum disagreement below 3ϕ, provided that $OPT \leq \phi$. A crucial structural insight is that if a low-degree vertex w belongs to a cluster C in an optimal solution, no vertex v outside C can have a neighborhood similar to that of w, i.e., $|N[v]\Delta N[w]| > 2\phi$.

Using this structural result, we present the following algorithm, which constructs a clustering with a maximum disagreement of 3ϕ. This algorithm is described here with slight modifications to enhance intuition:

1. Form clusters for high-degree vertices by grouping any pair u, v where $|N[u]\Delta N[v]| \leq 2\phi$.

2. Select an arbitrary vertex u from each high-degree cluster and have it propose to low-degree neighbors whose neighborhoods closely resemble its own.

3. For each low-degree vertex that receives at least one proposal, arbitrarily accept one proposal and join the corresponding cluster.

4. Place any remaining low-degree vertices that do not receive a proposal into singleton clusters.

3 Preliminary Definitions

In this paper, a clustering is defined as an object which contains clusters. Clusters are defined as the groups in which vertices are assigned. Clusterings will be denoted as $\mathcal{C}$ throughout this paper, and clusters will be denoted as C. Let E^+ denote the positive edges in the graph, G.

DEFINITION 1. *Given $u \in G^+ = (V, E^+)$, let $N(u)$ denote the neighbors of u in G^+. Define*

$$N[u] = N(u) \cup \{u\}$$

DEFINITION 2. *Given any clustering $\mathcal{C}$ and any vertex u, C_u is defined to be the cluster of $\mathcal{C}$ containing u.*

DEFINITION 3. *Given clustering $\mathcal{C}$, define $\rho_{\mathcal{C}}(x)$ as the number of disagreements incident to vertex x. More precisely:*

$$\rho_{\mathcal{C}}(x) = |C_x \setminus N[x]| + |N[x] \setminus C_x| = |N[x]\Delta C_x| = |N(x) \ \Delta \ C_x| - 1$$

DEFINITION 4. *Given a clustering $\mathcal{C}$, the objective function of $\mathcal{C}$, obj$(\mathcal{C}) = \max_u \rho_{\mathcal{C}}(u)$, is defined as the maximum incident disagreements over every vertex $u \in V$, where V is the set of all vertices.*

Lemma 3.1. *Let A, B, C be clusters of nodes. Then:*

$$|A\Delta C| \leq |A\Delta B| + |B\Delta C|$$

Proof. Consider $(A\Delta B)\Delta(B\Delta C)$. Due to the associativity of the symmetric difference, we attain:

$$\begin{aligned}(A\Delta B)\Delta(B\Delta C) &= A\Delta(B\Delta B)\Delta C \\ &= A\Delta C\end{aligned}$$

Observe that:

$$\begin{aligned}&(A\Delta B)\Delta(B\Delta C) \subseteq (A\Delta B) \cup (B\Delta C) \\ &\implies |(A\Delta B)\Delta(B\Delta C)| \leq |(A\Delta B) \cup (B\Delta C)| \qquad (\text{as } X\Delta Y \subseteq X \cup Y)\end{aligned}$$

Thus, the proof is finished, for:

$$(A\Delta C) \leq |(A\Delta B) \cup (B\Delta C)| = |(A\Delta B)| + |(B\Delta C)|$$

□

4 Algorithm

Clustering Algorithm - Algorithm 1

Let C_u denote the cluster containing vertex u. Let $\mathcal{L}_u$ denote the cluster in $\mathcal{L}$ that contains vertex u. Let V_{high} denote all $v \in V$ such that $deg(v) \geq 3\phi$ and let $V_{low} = V \setminus V_{high}$.

```
function FindClustering(G⁺ = (V, E⁺), φ)
  ▷ Initialization
    𝓛 = ∅
    V_1 = V_low
    for v ∈ V_high do
        marked[v] ← 0
    end for
    for u ∈ V_high do
        if marked[u] == 0 then
            C_u = ∅
            for v ∈ V_high do
                if |N[u] ∩ N[v]| > 2φ then
                    C_u ← C_u ∪ {v}.
                    marked[v] ← 1
                end if
                𝓛 ← 𝓛 ∪ {C_u}
            end for
        end if
    end for
    for i from 1 to |𝓛| do
        Choose a node u_i ∈ 𝓛_i
         Compute R(u_i) = {w ∈ V_i ∩ N[u_i] | |N[w]ΔN[u_i]| ≤ 2φ}
         C_i ← 𝓛_i ∪ R(u_i)
        V_{i+1} ← V_i \ R(u_i)
    end for
    if for some C ∈ 𝓒 there exists u_i such that ρ_𝓒(u_i) > 3φ then
        return "OPT > φ"
    else
        return 𝓒 = {C_i}_{i=1}^{|𝓛|} ∪ ⋃_{v∈V_{|𝓛|+1}} {{v}}
    end if
end function
```

THEOREM 4.1. *Suppose that* OPT $\leq \phi$, *Algorithm 1 outputs a clustering* $\mathcal{C}$ *with obj*$(\mathcal{C}) \leq 3\phi$. *Recall that* V_1 *is defined to be all vertices with degree less than* 3ϕ.

Proof. Let $\mathcal{C}^*$ be an optimal solution so obj$(\mathcal{C}^*) \leq \phi$ and let $\mathcal{L}_u$ denote the grouping in $\mathcal{L}$ that contains vertex u. We show the following four statements in the proceeding section:

1. (High-Degree Nodes Clustering). For any u such that deg$(u) > 3\phi$, $\mathcal{L}_u = C_u^* \cap V_{high}$.

2. (No Stealing on Low-Degree Nodes). For any $u \in V_{high}$. Let C_u^* be the cluster in the clustering $\mathcal{C}^*$ such that $\mathcal{L}_u \cap V_{high} = C_u^* \cap V_{high}$. We then have $C_u^* \cap V_i = C_u^* \cap V_1$.

3. (Low-Degree Nodes Inclusion). For any grouping $\mathcal{L}'$ in $\mathcal{L}$, let u_i be the designated vertex in Algorithm 1 that is in $\mathcal{L}'$. Then, $C_{u_i}^* \cap N[u_i] \subseteq C_{u_i}$.

4. (Closeness). For any grouping $\mathcal{L}'$ in $\mathcal{L}$, let u_i be the designated vertex in Algorithm 1 that is in $\mathcal{L}'$. Then $|N[u_i]\Delta C_{u_i}| \leq \phi$ and $|C_{u_i}^* \Delta C_{u_i}| \leq \phi$.

With the above four theorems, we show $\text{obj}(C) \leq 3\phi$. Let the clustering produced by Algorithm 1 be denoted $\mathcal{C}$. $\forall v \in V$ such that $\deg(v) \leq 3\phi$, a singleton cluster will suffice. For all other vertices, we consider the associated cluster C_{u_i} for a given vertex u_i. Let v be any vertex in C_{u_i}; our goal is to show $\rho_{\mathcal{C}}(v) \leq 3\phi$. By High-Degree Nodes Clustering, $\exists C_{u_i}^*$ such that $\mathcal{L}_{u_i} = C_{u_i}^* \cap V_{high}$.
Now suppose that $v \in (C_{u_i}^* \cap C_{u_i})$. Then:

$$
\begin{aligned}
\rho_{\mathcal{C}}(v) &= |N[v]\Delta C_{u_i}| \\
&\leq |N[v]\Delta C_{u_i}^*| + |C_{u_i}^* \Delta C_{u_i}| && \text{Lemma 3.1} \\
&= \rho_{\mathcal{C}^*}(v) + |C_{u_i}^* \Delta C_{u_i}| \\
&\leq \phi + \phi = 2\phi
\end{aligned}
$$

Otherwise, suppose that $v \notin (C_{u_i}^* \cap C_{u_i})$, implying that $v \in (C_{u_i} \backslash C_{u_i}^*)$. In this case, v must be a vertex in $R(u_i)$ from Line 21 in Algorithm 1. Therefore, $|N[v]\Delta N[u_i]| \leq 2\phi$, and we get:

$$
\begin{aligned}
\rho_{\mathcal{C}}(v) &= |N[v]\Delta C_{u_i}| \\
&\leq |N[v]\Delta N[u_i]| + |N[u_i]\Delta C_{u_i}| && \text{Lemma 3.1} \\
&\leq 2\phi + \phi = 3\phi
\end{aligned}
$$

□

5 High-Degree Nodes Clustering

Let ϕ be our guess of OPT, the optimal solution. If ϕ is an upper bound of OPT, then vertices with degree greater than 3ϕ are uniquely clustered in the optimal solution. Our algorithm reproduces this optimal clustering.

For the following lemmas and theorem, let u, v have degree greater than or equal to 3ϕ.

LEMMA 5.1. *If* $|N[u] \cap N[v]| > 2\phi$, *and* $\mathcal{C}$ *is a clustering where* u, v *are in different clusters, then* $obj(\mathcal{C}) > \phi$.

Proof. Assume the conditions hold and let U, V be the respective clusters for u, v. For each $x \in (N[u] \cap N[v])$, either:

- $x \in U$
- $x \in V$
- $x \in (U \cup V)^C$

In the first case, one disagreement is added to $\rho_{\mathcal{C}}(v)$. In the second case, one disagreement is added to $\rho_{\mathcal{C}}(v)$. In the third case, one disagreement is added to both $\rho_{\mathcal{C}}(v), \rho_{\mathcal{C}}(v)$. Therefore, at least one of the following holds:

- $\rho_{\mathcal{C}}(v) > \phi$
- $\rho_{\mathcal{C}}(u) > \phi$

Hence, $\text{obj}(\mathcal{C}) > \phi$. ◻

LEMMA 5.2. *If* $|N[u] \Delta N[v]| > 2\phi$, *and* $\mathcal{C}$ *is a clustering such that* u, v *are in the same cluster, then* $obj(\mathcal{C}) > \phi$.

Proof. Assume the conditions hold and let C' be the cluster containing u, v. For each $x \in (N[u] \Delta N[v])$, either:

- $x \in C'$
- $x \notin C'$

Consider the first case, or that $x \in C'$. Moreover, assume $x \in N[u]$. Then, one disagreement is added to $\rho_{\mathcal{C}}(v)$ because $x \notin N[v]$. On the other hand, if we assume that $x \in N[v]$, one disagreement is still added to $\rho_{\mathcal{C}}(u)$ because $x \notin N[u]$.

So, now consider the second case, or that $x \notin C'$. First, assume $x \in N[u]$. Then, it adds one disagreement to $\rho_{\mathcal{C}}(u)$ because $x \in N[u]$. Alternatively, if $x \in N[v]$, one disagreement is added to $\rho_{\mathcal{C}}(v)$ because $x \in N[v]$. In either case, at least one of the following holds, since there are more than 2ϕ disagreements added:

- $\rho_{\mathcal{C}}(v) > \phi$
- $\rho_{\mathcal{C}}(u) > \phi$

Hence, $\text{obj}(\mathcal{C}) > \phi$. ◻

THEOREM 5.1. *Assume $\phi = obj(\mathcal{C}^*)$. For any u such that $deg(u) > 3\phi$, $\mathcal{L}_u = C_u^* \cap V_{high}$.*

Proof. First, consider any $x \in \mathcal{L}_u$. Then, by Algorithm 1, $|N[x] \cap N[u]| > 2\phi$. Also, $x \in V_{high}$ by the algorithm, or it would not be in $\mathcal{L}_u$. Considering C_u^*, we know that $\text{obj}(\mathcal{C}^*) \le \phi$. So by inspection of Lemma 5.1, it must be that $x \in C_u^*$. So, $x \in C_u^* \cap V_{high}$, implying that $\mathcal{L}_u \subseteq C_u^* \cap V_{high}$.

Now consider that $x \in C_u^* \cap V_{high}$. Then the goal is to show $|N[x] \cap N[u]| > 2\phi$. Because x, u are in C_u^* and $\text{obj}(\mathcal{C}^*) \le \phi$, Lemma 5.1 implies that x, u are in the same cluster. Proceeding now to contradiction, let us assume $|N[x] \cap N[u]| \le 2\phi$. Then, since $\deg(u), \deg(x) > 3\phi$:

$$\begin{aligned} |N[x]\Delta N[v]| &= |N[x]\backslash N[v]| + |N[v]\backslash N[x]| \\ &> (3\phi - 2\phi) + (3\phi - 2\phi) \\ &> 2\phi \end{aligned}$$

Then, Lemma 5.2 says that $\text{obj}(\mathcal{C}^*) > \phi$. This is a contradiction, and so it must be that $|N[x] \cap N[u]| > 2\phi$. Hence, by Algorithm 1, we will have that $x \in \mathcal{L}_u$. So, $C_u^* \cap V_{high} \subseteq \mathcal{L}_u$.

Therefore, we have $\mathcal{L}_u = C_u^* \cap V_{high}$. ◻

6 No Stealing on Low-Degree Nodes

THEOREM 6.1. *Let u_i be a designated vertex as denoted in Algorithm 1. For any i, let $C_{u_i}^*$ be the cluster in the clustering $\mathcal{C}^*$ such that $\mathcal{L}_{u_i} \cap V_{high} = C_{u_i}^* \cap V_{high}$. We then have $C_{u_i}^* \cap V_i = C_{u_i}^* \cap V_1$.*

Before proving this theorem, we prove two lemmas.

LEMMA 6.1. *Suppose that for any nodes u, v with degree at least 3ϕ and $w \in C_u^*$ with $deg(w) < 3\phi$, such that $|N[w] \cap N[v] \cap N[u]| = \phi + x, x \in \mathbb{N}_0$. Also, suppose that $C_v^* \cap C_u^* = \varnothing$.*
Then, $|N[w]\Delta N[u]| \le 2(\phi - x)$.

Proof. Let $S = (N[w] \cap N[v] \cap N[u])$. If $|S| = \phi + x$, then we first show that $|S \cap C_u^*| \le \phi$. Proceeding to contradiction, assume that $|S \cap C_u^*| > \phi$. Because S is bounded by $\phi + x$ and by assumption we have

$(C_u^* \cap C_v^*) = \varnothing$, it must be that $|S \cap C_v^*| \leq x$ and $|S \backslash C_v^*| > \phi$. Now, recall that there are $\phi + x$ elements of $N[v]$ in S. So:

$$
\begin{aligned}
& |S \backslash C_v^*| > \phi \\
\implies \ & \rho_{\mathcal{C}^*}(v) = |N[v] \Delta C_v^*| \\
& \geq |N[v] \backslash C_v^*| \\
& \geq |S \backslash C_v^*| && S \subseteq N[v] \\
& > \phi
\end{aligned}
$$

This contradicts the fact that $\rho_{\mathcal{C}^*}(w) \leq \phi$ for any w. Therefore, our original assumption on $|S \cap C_u^*|$ is incorrect, and $|S \cap C_u^*| \leq \phi$.

Now with the facts that $|S \cap C_u^*| \leq \phi$ and $w \in C_u^*$, we have:

$$
\begin{aligned}
& |S \backslash C_u^*| \geq x \\
& |S \backslash C_w^*| \geq x && C_w^* = C_u^*
\end{aligned}
$$

Each of the vertices in $S \backslash C_u^*$ can contribute to $\rho_{\mathcal{C}^*}(u)$. But each such vertex is in both $N[w]$ and $N[u]$, and thus they are not in $(N[w] \Delta N[u])$.

Now, we claim that $|N[u] \Delta N[w]| \leq 2(\phi - x)$. Proceeding to contradiction, assume otherwise. Recall that $(S \backslash C_w^*) = (S \backslash C_u^*) \notin (N[w] \Delta N[u])$. Let $T = (S \backslash C_u^*)$.

Then:

$$
\rho_{\mathcal{C}^*}(u) = |T| + |(N[u] \Delta C_u^*) \backslash T|
$$

$$
\rho_{\mathcal{C}^*}(w) = |T| + |(N[w] \Delta C_u^*) \backslash T|
$$

Summing them:

$$\begin{aligned}
\rho_{\mathcal{C}^*}(u) + \rho_{\mathcal{C}^*}(w) &= 2|T| + |(N[u]\Delta C_u^*)\backslash T| + |(N[w]\Delta C_u^*)\backslash T| \\
&= 2|T| + |(N[u]\backslash T)\Delta C_u^*| + |(N[w]\backslash T)\Delta C_u^*| && \text{definition of. } T \\
&\geq 2|T| + |(N[u]\backslash T)\Delta(N[w]\backslash T)| && \text{Lemma 3.1} \\
&\geq 2|T| + |(N[u]\Delta N[w])\backslash T| \\
&\geq 2|T| + |N[u]\Delta N[w]| && T \notin (N[u]\Delta N[w]) \\
&> 2x + 2(\phi - x) \\
&> 2\phi
\end{aligned}$$

Therefore, either $\rho_{\mathcal{C}^*}(u) > \phi$ or $\rho_{\mathcal{C}^*}(w) > \phi$, which is a contradiction on the clustering $\mathcal{C}^*$.

So, $|N[u]\Delta N[w]| \leq 2(\phi - x)$. □

Lemma 6.2. *Let u, v have degree at least 3ϕ and $w \in C_u^*$ with $deg(w) < 3\phi$. Let $S = (N[w] \cap N[v] \cap N[u])$. Then:*

$$|N[w]\Delta N[v]| + |N[w]\Delta N[u]| \geq |N[v]| + |N[u]| - 2|S|$$

Proof. We first go about showing two set containments, which directly implies the desired statement. Observe, by definition that:

$$(N[w]\Delta N[v]) \cup (N[w]\Delta N[u]) = (N[v]\backslash N[w]) \cup (N[w]\backslash N[v]) \cup (N[w]\backslash N[u]) \cup (N[u]\backslash N[w])$$

We now focus on two of the terms. See that:

$$(N[v]\backslash N[w]) \cup (N[w]\backslash N[u]) = (N[v]\backslash(N[v] \cap N[w])) \cup (N[w]\backslash(N[w] \cap N[u]))$$

Consider some $x \in (N[v]\backslash S)$. We show that:

$$\begin{aligned}
&(N[v]\backslash S) \subseteq (N[v]\backslash N[w]) \cup (N[w]\backslash N[u]) \\
\implies\ & |N[v]| - |S| \subseteq |N[v]\backslash N[w]| + |N[w]\backslash N[u]|
\end{aligned}$$

Take such an x. If we suppose $x \in (N[v] \backslash N[w])$, we are done. If we instead suppose that $x \notin (N[v] \backslash N[w])$, then $x \in N[w]$. This is because $x \in N[v]$ and $x \notin (N[v] \backslash N[w])$. Thus, $x \in N[w]$ also. With this, we get:

$$
\begin{aligned}
& x \in N[w] \cap N[v] \\
& x \notin N[u] \qquad\qquad\qquad\qquad x \notin S \\
& \implies x \in N[w] \backslash N[u]
\end{aligned}
$$

Therefore, we obtain the set inclusion. Now considering the other case, that $x \in (N[u] \backslash S)$, a similar argument shows that:

$$
\begin{aligned}
& (N[u] \backslash S) \subseteq (N[w] \backslash N[v]) \cup (N[u] \backslash N[w]) \\
& \implies |N[u]| - |S| \leq |N[w] \backslash N[v]| + |N[u] \backslash N[w]|
\end{aligned}
$$

So we have shown the set containment, giving the sought after implication:

$$
\begin{aligned}
|N[v]| + N[u] - 2|S| &\leq |N[v] \backslash N[w]| + |N[w] \backslash N[u]| + |N[w] \backslash N[v]| + |N[u] \backslash N[w]| \\
&= |N[v] \Delta N[w]| + |N[u] \Delta N[w]|
\end{aligned}
$$

□

Now we are in a position to prove Theorem 6.1.

Proof. Let u_i be as in Algorithm 1. Choose $u_j \in (V \backslash V_1)$ so that $deg(v) \geq 3\phi$ and $C^*_{u_i} \cap C^*_{u_j} = \emptyset$. Let $deg(w) < 3\phi$ and $|S| \leq \phi$. Then, using Lemma 6.2.:

$$
\begin{aligned}
|N[u_j] \Delta N[w]| + |N[w] \Delta N[u_i]| &\geq |N[u_j]| + |N[u_i]| - 2|S| \\
&\geq 3\phi + 3\phi - 2\phi \\
&\geq 4\phi
\end{aligned}
$$

We know that $|N[w] \Delta N[u_i]| \leq 2\phi$ by Algorithm 1. This implies that $|N[w] \Delta N[u_j]| > 2\phi$. and w cannot be 'stolen' by u_j during the algorithm for $|S| \leq \phi$.

The other case is that $|S| \leq \phi + t$, for some $t \geq 0$. It suffices to show that $w \notin R(u_j)$. Using the result

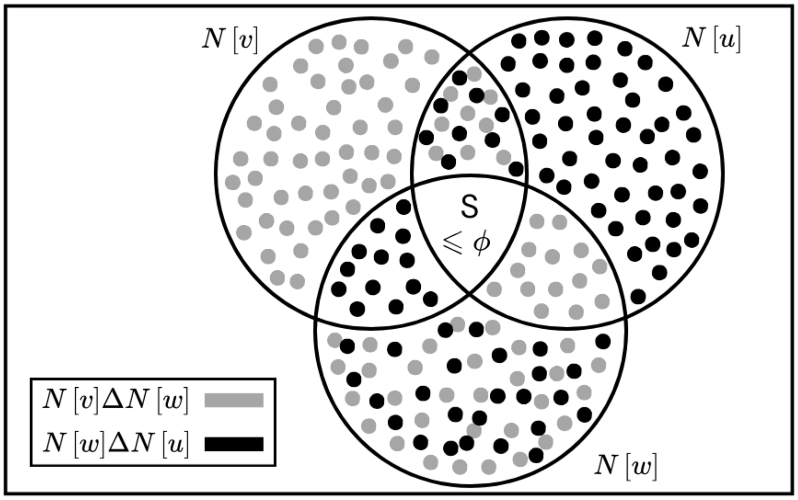

Figure 1: Venn Diagram of Lemma 6.2

of Lemma 6.2:

$$\begin{aligned}
&|N[w]\Delta N[u_j]| + |N[w]\Delta N[u_i]| \geq |N[u_j]| + |N[u_i]| - 2|S| \\
\implies &|N[w]\Delta N[u_j]| + |N[w]\Delta N[u_i]| \geq 3\phi + 3\phi - 2(\phi + t) \\
\implies &|N[w]\Delta N[u_j]| + |N[w]\Delta N[u_i]| \geq 2(2\phi - t)
\end{aligned}$$

Then, Lemma 6.1 informs us that $|N[u_i]\Delta N[w]| \leq 2(\phi - t)$. Hence, $|N[w]\Delta N[u_j]| > 2\phi$.

Because in all cases we observe that $|N[w]\Delta N[u_j]| > 2\phi$, w cannot be 'stolen' by u_j during the algorithm. Thus, all $w \in C^*_{u_i}$ with $\deg(w) < 3\phi$ will be clustered into the appropriate cluster C_{u_i}. Thus, we know for any $w \in (C^*_{u_i} \cap V_1)$, then $w \in (C^*_{u_i} \cap V_i)$ because w cannot be selected to join any $R(u_j)$ when $C^*_{u_j} \neq C^*_{u_i}$ and $i \neq j$. So we have:

$$C^*_{u_i} \cap V_1 \subseteq C^*_{u_i} \cap V_i$$

Then, because $V_i \subseteq V_1$, we observe that:

$$C^*_{u_i} \cap V_i \subseteq C^*_{u_i} \cap V_1$$

Hence, we derive $C^*_{u_i} \cap V_1 = C^*_{u_i} \cap V_i$. ◻

7 Low-Degree Nodes Inclusion

THEOREM 7.1. *For any grouping $\mathcal{L}'$ in $\mathcal{L}$, take u_i as denoted in Algorithm 1. Then, $C^*_{u_i} \cap N[u_i] \subseteq C_{u_i}$.*

Proof. Let $h \in (C^*_{u_i} \cap N[u_i])$. If $h = u_i$, then $u_i \in C_{u_i}$ by definition. If otherwise, assume $deg(h) \geq 3\phi$ and $h \neq u$. By Theorem 5.1, for $\phi = obj(\mathcal{C}^*)$, $\mathcal{L}_{u_i} = C^*_{u_i} \cap V_{high}$. Now, $h \in C^*_{u_i}$ and $h \in V_{high}$, implying that $h \in \mathcal{L}_{u_i}$. Thus, by line 22 in Algorithm 1, $h \in C_{u_i}$. So we attain:

$$C^*_{u_i} \cap N[u_i] \subseteq C_{u_i}$$

◻

8 Closeness

From the above arguments, for $v \in V$ such that $deg(v) \geq 3\phi$, we know $v \in C^*_v$ and $v \in C_v$ This follows from Theorem 5.1, since $\mathcal{L}_v = C^*_v \cap V_{high}$. Let u_i be as denoted in Algorithm 1. The following shows that for vertices of lower degree, the symmetric difference between $N[u_i], C_{u_i}$, and $C^*_{u_i}$ is 'small', i.e. less than or equal to ϕ.

THEOREM 8.1. *For any grouping $\mathcal{L}'$ in $\mathcal{L}$, let u_i be the node in $\mathcal{L}'$ with maximum degree. Then $|N[u_i]\Delta C_{u_i}| \leq \phi$ and $|C^*_{u_i}\Delta C_{u_i}| \leq \phi$.*

Proof. We begin with $|N[u_i]\Delta C_{u_i}| \leq \phi$. First, we show $(N[u_i]\backslash C_{u_i}) \subseteq (N[u_i]\backslash C^*_{u_i})$. From Theorem 7.1, we have $C^*_{u_i} \cap N[u_i] \subseteq C_{u_i}$. So:

$$N[u_i]\backslash C_{u_i} \subseteq N[u_i]\backslash(C^*_{u_i} \cap N[u_i]) = N[u_i]\backslash C^*_{u_i} \tag{8.1}$$

Therefore:

$$\begin{aligned}
|V_{low} \cap (N[u_i]\Delta C_{u_i})| &= |V_{low} \cap (N[u_i] \setminus C_{u_i})| + |V_{low} \cap (C_{u_i} \setminus N[u_i])| \\
&= |V_{low} \cap (N[u_i] \setminus C_{u_i})| && (V_{low} \cap C_{u_i}) \subseteq N[u_i], \text{ Alg. 1, line 21-22} \\
&\leq |V_{low} \cap (N[u_i] \setminus C^*_{u_i})| && (8.1) \\
&\leq |V_{low} \cap (N[u_i]\Delta C^*_{u_i})|
\end{aligned}$$

Also, since $C^*_{u_i} \cap V_{high} = C_{u_i} \cap V_{high}$, it must be the case that

$$|(N[u_i]\Delta C_{u_i}) \cap V_{high}| = |(N[u_i]\Delta C^*_{u_i}) \cap V_{high}| \tag{8.2}$$

Now, putting the above together:

$$\begin{aligned}
|N[u_i]\Delta C_{u_i}| &= |(N[u_i]\Delta C_{u_i}) \cap V_{low}| + |(N[u_i]\Delta C_{u_i}) \cap V_{high}| \\
&\leq |(N[u_i]\Delta C^*_{u_i}) \cap V_{low}| + |(N[u_i]\Delta C^*_{u_i}) \cap V_{high}| && (8.2) \\
&= |N[u_i]\Delta C^*_{u_i}| \leq \phi && u_i \in C^*_{u_i} \text{ and } \rho_{\mathcal{C}^*}(u) \leq \phi
\end{aligned}$$

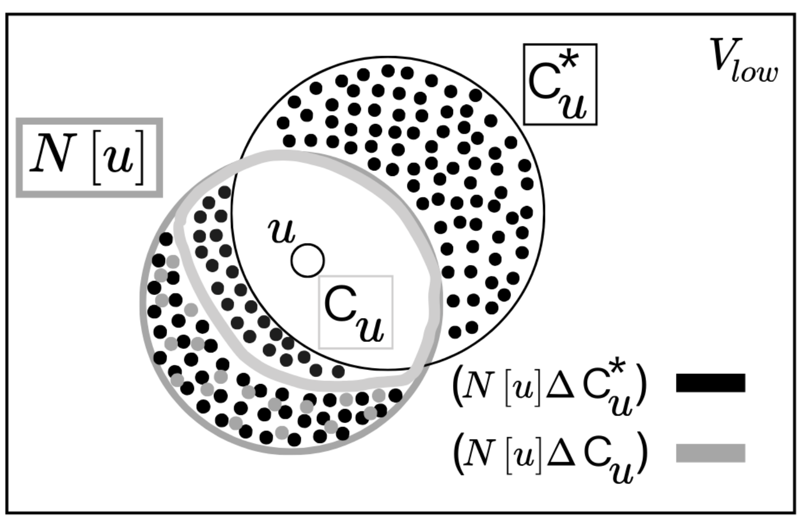

Figure 2: $|N[u_i]\Delta C_{u_i}| \leq |N[u_i]\Delta C^*_{u_i}|$

Moving on to the next part of the proof, we show $|C^*_{u_i} \Delta C_{u_i}| \leq \phi$.

$$
\begin{aligned}
|C^*_{u_i} \Delta C_{u_i}| &= |V_{low} \cap (C^*_{u_i} \Delta C_{u_i})| && C^*_i \cap V_{high} \\
&&& = C_{u_i} \cap V_{high} \\
&= |V_{low} \cap (C^*_{u_i} \setminus C_{u_i})| + |V_{low} \cap (C_{u_i} \backslash C^*_{u_i})| && \text{by symm. diff} \\
&= |V_{low} \cap ([C^*_{u_i} \cap N[u]] \setminus C_{u_i})| + |V_{low} \cap ([C^*_{u_i} \setminus N[u_i]] \setminus C_{u_i})| + |V_{low} \cap (C_{u_i} \backslash C^*_{u_i})| \\
&&& \text{by Partition Theorem} \\
&= |V_{low} \cap ([C^*_{u_i} \cap N[u_i]] \setminus C_{u_i})| + |V_{low} \cap (C^*_{u_i} \setminus N[u_i])| + |V_{low} \cap (C_{u_i} \backslash C^*_{u_i})| && V_{low} \cap C_{u_i} \subseteq N[u_i] \\
&&& \text{(Alg 1., line 21-22)} \\
&= |V_{low} \cap (C^*_{u_i} \setminus N[u_i])| + |V_{low} \cap (C_{u_i} \backslash C^*_{u_i})| && C^*_{u_i} \cap N[u_i] \subseteq C_{u_i} \\
&&& \text{(Thm 7.1)} \\
&\leq |V_{low} \cap (C^*_{u_i} \setminus N[u_i])| + |V_{low} \cap (N[u_i] \backslash C^*_{u_i})| && V_{low} \cap C_{u_i} \subseteq N[u_i] \\
&= |V_{low} \cap (C^*_{u_i} \Delta N[u_i])| && \text{by symm. diff} \\
&\leq |C^*_{u_i} \Delta N[u_i]| \\
&\leq \phi && obj(\mathcal{C}^*) = \phi
\end{aligned}
$$

□

9 Runtime Analysis

Calculating V_{low} on line 3 of Algorithm 1 requires iterating over all $v \in V$ to check the degree of each vertex. This takes $O(m)$, where $m = |E^+|$. This also gives us V_{high}. To complete line 11, where we check $|N[u] \cap N[v]| > 2\phi$. For any $h \in V$, we represent $N[h]$ as a binary tree. This is achieved by giving every vertex a number that corresponds to its identity and constructing the binary tree using these identity values. Using a data structure such as a red-black tree, creating all the balanced trees takes $O(n \log D)$, as each vertex has at most D neighbors and there are n vertices.

Let the maximum degree in V be D. Given any $x \in N[v]$, checking if $x \in N[u]$ takes $O(\log D)$ because u has at most D neighbors. Hence, calculating $|N[u] \cap N[v]|$ takes $O(D \log D)$. We pre-calculate this $\forall u, v \in V$, taking $O(n^2 D \log D)$. This will gives us everything needed for line 11.

Now we consider the runtime of lines 19 through 24, which is dominated by line 21. Observe that for any $u, v \in V$, we have:

$$|N[u]\Delta N[v]| = deg(u) + deg(v) + 2 - |N[u] \cap N[v]|$$

Therefore, calculating $|N[u]\Delta N[v]|$ relies on calculating $|N[u] \cap N[v]|$, which takes $O(1)$ because we pre-calculated this earlier. We also maintain a bitmask for V_1, which iteratively becomes each V_i in the algorithm with deletion taking $O(1)$ via hashing into the bitmask. Then, finding whether $w \in V_i$ for each $w \in N[u_i]$ takes $O(D)$. Therefore, the running time of line 21 is:

$$O(D) + O(1) = O(D)$$

Thus, the total runtime of these lines takes:

$$O(|\mathcal{L}|D) = O(nD)$$

The remaining lines 15-29 take $O(n^2)$ to check the disagreement on each vertex. Therefore, the total running time of Algorithm 1 is $O(n^2 D \log D)$. Taking into account a binary search on ϕ, the total runtime required to find a 3-approximation is $O(n^2 D \log D \log n)$.

References

[1] Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5):23:1–23:27, 2008. 1

[2] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56(1-3):89–113, 2004. 2

[3] 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. 2

[4] Moses Charikar, Neha Gupta, and Roy Schwartz. Local guarantees in graph cuts and clustering. In *Integer Programming and Combinatorial Optimization (IPCO)*, pages 136–147, 2017. 2

[5] Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. Clustering with qualitative information. *Journal of Computer and System Sciences*, 71(3):360 – 383, 2005. 1

[6] 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 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 219–228, 2015. 1

[7] Yizong Cheng and George Church. Biclustering of expression data. In *Proceedings of the International Conference of Intelligent Systems for Molecular Biology*, volume 8, pages 93–103, 2000. 2

[8] Vincent Cohen-Addad, Euiwoong Lee, Shi Li, and Alantha Newman. Handling correlated rounding error via preclustering: A 1.73-approximation for correlation clustering. In *64rd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2023. to appear. 1

[9] Vincent Cohen-Addad, Euiwoong Lee, and Alantha Newman. Correlation clustering with sherali-adams. In *63rd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 651–661. IEEE, 2022. 1

[10] Sami Davies, Benjamin Moseley, and Heather Newman. Fast combinatorial algorithms for min max correlation clustering. In *International Conference on Machine Learning*, pages 7205–7230. PMLR, 2023. 2

[11] Holger SG Heidrich, Jannik Irmai, and Bjoern Andres. A 4-approximation algorithm for min max correlation clustering. In *AISTATS*, pages 1945–1953, 2024. 3

[12] Sanchit Kalhan, Konstantin Makarychev, and Timothy Zhou. Correlation clustering with local objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. 2

[13] Gregory J. Puleo and Olgica Milenkovic. Correlation clustering and biclustering with locally bounded errors. *IEEE Trans. Inf. Theory*, 64(6):4105–4119, 2018. 1, 2