Robust Correlation Clustering

Devvrit
BITS Pilani, Goa Campus, Goa, India
devvrit.03@gmail.com

Ravishankar Krishnaswamy
Microsoft Research, Bengaluru, India
rakri@microsoft.com

Nived Rajaraman
IIT Madras, Chennai, India
nived.rajaraman@gmail.com

Abstract

In this paper, we introduce and study the Robust-Correlation-Clustering problem: given a graph $G = (V, E)$ where every edge is either labeled $+$ or $-$ (denoting similar or dissimilar pairs of vertices), and a parameter m, the goal is to delete a set D of m vertices, and partition the remaining vertices $V \setminus D$ into clusters to minimize the cost of the clustering, which is the sum of the number of $+$ edges with end-points in different clusters and the number of $-$ edges with end-points in the same cluster. This generalizes the classical Correlation-Clustering problem which is the special case when $m = 0$. Correlation clustering is useful when we have (only) qualitative information about the similarity or dissimilarity of pairs of points, and Robust-Correlation-Clustering equips this model with the capability to handle noise in datasets.

In this work, we present a *constant-factor* bi-criteria algorithm for Robust-Correlation-Clustering on complete graphs (where our solution is $O(1)$-approximate w.r.t the cost while however discarding $O(1)m$ points as outliers), and also complement this by showing that no finite approximation is possible if we do not violate the outlier budget. Our algorithm is very simple in that it first does a simple LP-based *pre-processing* to delete $O(m)$ vertices, and subsequently runs a particular Correlation-Clustering algorithm ACNAlg [2] on the residual instance. We then consider general graphs, and show $(O(\log n), O(\log^2 n))$ bi-criteria algorithms while also showing a hardness of α_{MC} on both the cost and the outlier violation, where α_{MC} is the lower bound for the Minimum-Multicut problem.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms; Theory of computation → Facility location and clustering

Keywords and phrases Correlation Clustering, Outlier Detection, Clustering, Approximation Algorithms

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2019.33

Category APPROX

1 Introduction

Clustering is one of the most widely used tools in various scientific disciplines (such as biology, computer science, machine learning and operations research to name a few) due to its wide applicability in these domains. Broadly speaking, the goal of clustering is to partition a given dataset into a number of clusters such that data items in the same cluster are more alike each other than data items in different clusters. In many application domains, the data items are naturally represented as points in a metric space, and the distance between the corresponding vectors is used as a measure of (dis)similarity. In such cases, clustering formulations such as k-median or k-means are the de-facto standards to utilize. However, there are also quite a few application domains where the information available to us is simply

whether different pairs of data items are similar or dissimilar to each other. Examples of such settings where there is only qualitative information include data items being web-pages on the internet, a collection of people on a social network or even a group of proteins. Motivated by such settings, Bansal et al. [3] formulated a problem known as *correlation clustering* (in fact, a similar problem was implicitly studied by Ben-Dor et al.[4] as 'Cluster Editing').

▶ **Problem 1** (Correlation-Clustering). *We are given a complete graph $G = (V, \binom{V}{2})$, and a labelling of each edge as either positive or negative, denoting whether the end vertices of the edge are similar to each other or dissimilar. In other words, the edge set $\binom{V}{2}$ is partitioned into $E_+ \dot\cup E_-$ where E_+ denotes the similar pairs and E_- denotes dissimiliar pairs. The goal is to compute a partition $\mathcal{C} = \{C_1, C_2, \ldots, C_r\}$ of V (so $V = \dot\cup_{1 \le i \le r} C_i$ is a disjoint union of the C_i's) to minimize the* cost *of the clustering, which is the total number of E_+ edges with end-points in different clusters and E_- edges with end-points in the same cluster.*

A nice modeling aspect of this problem is that the number of clusters is not specified as part of the input, and rather, left to the algorithm. This makes it a compelling problem when we do not have a priori knowledge of the number of clusters we seek in the final partitioning.

Since being introduced formally as an optimization problem, there have been numerous works trying to understand the computational complexity of the problem. Bansal et al. [3] show that the problem is APX-hard (ruling out the design of PTASes unless P=NP) and obtain a constant-factor *approximation algorithm* for this problem. Subsequently, there have been a series of works (see, e.g., the survey by Wirth [23]) getting better factors, with the current best bound being a factor of 2.06 due to Chawla et al. [8].

Despite the simplicity and elegance of the various clustering formulations described thus far, a significant shortcoming of most of them is that they are not robust to noisy points. For example, the presence of a few outliers in the data set can completely change the *cost* and *structure* of solutions obtained by running clustering algorithms for k-median, k-means, etc. Indeed, this has prompted much recent study in the CS, ML and statistics communities of *robust* versions of these problems [6, 10, 17]. Motivated by this observation, and the fact that real-world data sets are often noisy, we investigate the *robustness* of correlation clustering.

▶ **Problem 2** (Robust-Correlation-Clustering). *The input to this problem is identical to the correlation clustering instance as in Problem 1. Additionally, we are also given a parameter m, which denotes the number of points we can discard while clustering. The goal is to identify a set $D \subseteq V$ of* outliers *of size m, and cluster the remaining points $V \setminus D$ to minimize the cost of the resulting clustering, i.e., the total number of E_+ edges (resp. E_- edges) in $V \setminus D$ with end-points in different clusters (resp. same cluster).*

We note that Correlation-Clustering problem also makes sense when the edge set $E_+ \cup E_-$ is not the complete graph, since we often do not have complete information about the (dis)similarity of each pair of points (it could be expensive or even impossible to obtain such information like in the case of protein-protein interactions). Now the problem becomes much harder, and the current best known algorithms have approximation guarantees of a factor of $O(\log n)$. Moreover, there is an approximation-preserving reduction from the Minimum-Multicut problem, for which the best known approximation is an $O(\log n)$ factor [5]. In this paper, we also consider the Robust-Correlation-Clustering problem on general graphs, analogous to the study of Correlation-Clustering in general graphs [5].

▶ **Problem 3** (Robust-Correlation-Clustering on General Graphs). *The problem is identical to Problem 2, with the exception that the union of E_+ and E_- need not be $\binom{V}{2}$.*

1.1 Our Results

Having introduced the problem, the first question we address is whether the Correlation-Clustering objective is indeed susceptible to outliers in the dataset. That is, we seek to understand whether the solution cost and/or structure can change a lot by the removal of a few points in the dataset. Classical objectives such as k-median and k-means suffer from this drawback *even in the simplest of settings* when we are promised that after removing some m data-points, *the optimal clustering of the remaining points would have* 0 *cost.* In such cases, solving k-means objective on the original instance could yield very different solutions than the intended solution, which is the 0 cost (or perfect clustering).

Somewhat surprisingly, our first simple observation is that the correlation clustering objective is inherently robust to an extent, at least in the case when the cost of the clustering after removing m outliers becomes 0. We show that in this case, the optimal correlation clustering solution and the optimal robust correlation clustering solution are structurally identical upto $O(m)$ points.

▶ **Theorem 4.** *Consider an instance $\mathcal{I}$ of* Robust-Correlation-Clustering *on complete graphs such that $\mathsf{Opt}(\mathcal{I}) = 0$, i.e., there exists a set $D^* \subseteq V$ of m vertices deleting which, the subgraph induced by $V \setminus D^*$ admits a perfect clustering $\mathcal{C}^*$. Then, consider any optimal solution $\widetilde{\mathcal{C}}$ to* Correlation-Clustering *(Problem 1). There exists a set $\widetilde{D}$ of $O(m)$ vertices s.t. the cost of $\widetilde{\mathcal{C}} \setminus \widetilde{D}$[1] has objective function value 0.*

This theorem in fact sets apart the correlation clustering objective from other clustering objectives such as k-means and k-median where an analogous statement to Theorem 4 does not hold. Moreover, we believe that a similar result is true even when $\mathsf{Opt}(\mathcal{I}) \neq 0$ when comparing the optimal solutions of the robust and non-robust problems.

Now, while this exhibits the robustness of correlation clustering w.r.t. *optimal solutions*, the problem is APX-hard and hence we typically do not deal with optimal solutions. Hence, we next consider the same question, but for approximation algorithms.

▶ **Theorem 5.** *There exists an instance $\mathcal{I}$ of* Robust-Correlation-Clustering *on complete graphs which satisfies the following properites: (a) $\mathsf{Opt}(\mathcal{I}) = 0$, i.e., there exists a set $D \subseteq V$ of $m = O(\sqrt{n})$ vertices deleting which, the subgraph induced by $V \setminus D$ admits a perfect clustering, and (b) there exists a constant-factor approximately optimal solution $\mathcal{C}$ to the* Correlation-Clustering *objective function (1), such that, for any set S of $< n - 1$ vertices, the cost of the clustering $\mathcal{C} \setminus S$ is still non-zero.*

This then provides sufficient motivation for undertaking this study, with the main focus of whether we can design efficient approximation algorithms for Robust-Correlation-Clustering. Our first result in this direction is a negative result, which says that it is in fact NP-hard to obtain any finite approximation algorithm for Robust-Correlation-Clustering, even on complete graphs. This is in stark contrast to Problem 1, where we know very good constant-factor approximations.

▶ **Theorem 6.** *It is NP-hard to obtain any finite approximation factor for* Robust-Correlation-Clustering *on complete graphs, unless we violate the budget on the number of outliers.*

[1] We somewhat abuse notation to let $\mathcal{C} \setminus D$ to denote the clustering obtained by removing the points in D from the clustering $\mathcal{C}$.

We therefore seek to obtain *bi-criteria approximation algorithms*: an (a, b) bi-criteria approximation for Robust-Correlation-Clustering is one where the solution's cost is at most a times the optimal cost, and the number of outliers in our solution is at most $b \cdot m$.

▶ **Theorem 7.** *There is an efficient bi-criteria* $(6, 6)$*-approximation algorithm for* Robust-Correlation-Clustering *on complete graphs.*

Our algorithm is extremely simple: it essentially does a simple LP-based *pre-processing* step to prune out a set of $O(m)$ outliers, and then executes a classical algorithm for Correlation-Clustering [2] (henceforth called ACNAlg) on the remaining vertices. This approach works because the LP relaxation which [2] uses for solving Correlation-Clustering is a purely covering LP (as opposed to the more natural metric LP relaxation for Correlation-Clustering), and can easily be adapted to incorporating outliers. We remark that, owing to the pre-processing step, our overall algorithm requires solving an LP: it would be very interesting to develop a purely combinatorial algorithm for Robust-Correlation-Clustering on complete graphs. It might even be possible for a simple adaptation of the ACNAlg algorithm to be a constant-factor bi-criteria approximation. We leave this as an important avenue of future research.

Finally, we turn our attention to Robust-Correlation-Clustering on general graphs, where we show poly-logarithmic bi-criteria algorithms and logarithmic hardness results on both the cost as well as the outlier budget. While the Correlation-Clustering problem is equivalent to Minimum-Multicut [14] and we can use any Minimum-Multicut algorithm to solve the problem, we show that one specific technique based on *padded decompositions* of metric spaces naturally lends itself to solving the robust problem.

▶ **Theorem 8.** *There is an efficient bi-criteria* $(O(\log n), O(\log^2 n))$*-approximation algorithm for* Robust-Correlation-Clustering *on general graphs.*

▶ **Theorem 9.** *It is NP-hard to obtain any bi-criteria* (a, b)*-approximation algorithm for* Robust-Correlation-Clustering *on general graphs for* $b < \alpha_{\text{MC}}$ *or* $a < \alpha_{\text{MC}}$ *where* α_{MC} *is the inapproximability factor for the* Minimum-Multicut *problem.*

It would be interesting to resolve the gap between the $O(\log^2 n)$ upper bound and the $\Omega(\log n)$ lower bound for the outlier budget violation.

1.2 Related Work

Since its introduction, Correlation-Clustering has received much attention with focus on designing better algorithms (see the survey of [23]), faster algorithms in the parallel and distributed [11] and streaming settings [1], stochastic/average-case settings [19], and applications [12, 13, 20]. There is also work on a related objective function of *maximizing* the number of classified edges [3]. Being a maximization objective, it is easier to design simple constant-factor approximation algorithms like random partitions, etc. There are however, better SDP-based approximation algorithms [5, 22].

Recently there has also been a large body of work on the crucial problem of noise-resilient or *robust* clustering for distance-based clustering objectives such as k-means [10, 17], and designing faster algorithms [7, 21, 16], and parallel and distributed algorithms in this model [9, 18]. To the best of our knowledge, this is the first work to study the Correlation-Clustering problem from robustness point of view.

1.3 Paper Outline

We first describe the inherent robustness to outliers of *optimal solutions* for Correlation-Clustering in Section 2. We then consider Robust-Correlation-Clustering for complete graphs, and show our hardness of approximation in Section 3, followed by the bi-criteria algorithm in Section 4. Finally, in Section 5 and Appendix A, we turn our attention to the case of general graphs and present our algorithm and hardness.

2 Robustness of the Correlation-Clustering Objective

In this section, we show two simple but illuminating results. The first result explains how, in contrast to problems like k-median and k-means, the vanilla correlation clustering objective is in fact *inherently robust* to an extent, *when solved optimally*. The second result then shows this not to be true when considering solutions which are only approximately optimal. We remark that the second result and that fact that correlation clustering is APX-hard [3] serves as a strong motivation for studying the Robust-Correlation-Clustering problem.

2.1 Optimal Correlation-Clustering Solutions are Robust

In this section, we exhibit the inherent robustness of the correlation clustering objective (1) in a specialized scenario. Indeed, consider an instance $\mathcal{I}$ of Robust-Correlation-Clustering such that $\mathsf{Opt}(\mathcal{I}) = 0$, i.e., there exists a set of m points deleting which the remaining points are perfectly clusterable, i.e., have 0 cost. Now, imagine we obtain an optimal Correlation-Clustering solution (Problem 1) to instance $\mathcal{I}$. We show that there exist $O(m)$ points, deleting which, the cost indeed becomes 0 for this solution. This tells us that the optimal solutions to 2 and 1 are nearly identical to each other (upto $O(m)$ points), and hence, that the correlation clustering objective is inherently robust!

Proof of Theorem 4. We begin by recalling the theorem statement and setting up notation. Let $\mathcal{I}$ be an instance of Robust-Correlation-Clustering such that $\mathsf{Opt}(\mathcal{I}) = 0$, i.e., there exists a set $D^* \subseteq V$ of m vertices deleting which, the subgraph induced by $V \setminus D^*$ admits a perfect clustering $\mathcal{C}^*$. And consider any optimal solution $\widetilde{\mathcal{C}}$ to instance $\mathcal{I}$ w.r.t the Correlation-Clustering objective function (1). We would like to claim that there exists a set $\widetilde{D}$ of $O(m)$ vertices such that $\widetilde{\mathcal{C}} \setminus \widetilde{D}$ is identical to $\mathcal{C}^* \setminus \widetilde{D}$. We show this by showing that the cost of the clustering $\widetilde{\mathcal{C}} \setminus \widetilde{D}$ is 0, and hence it must be the same as $\mathcal{C}^* \setminus \widetilde{D}$.

To this end, let $\mathcal{C}^* = \{C_1^*, C_2^*, \ldots, C_r^*\}$ denote the optimal Robust-Correlation-Clustering clustering over vertices $V \setminus D^*$, and let $\widetilde{\mathcal{C}} = \{\widetilde{C}_1, \widetilde{C}_2, \ldots, \widetilde{C}_s\}$ denote the optimal Correlation-Clustering clustering over all vertices V. We divide the clusters in $\widetilde{\mathcal{C}}$ into two types:

(a) A cluster $\widetilde{C} \in \widetilde{\mathcal{C}}$ is a *mixed cluster* if it contains points from more than one cluster in $\mathcal{C}^*$, i.e., there exists i_1, i_2 s.t $|\widetilde{C} \cap C_{i_1}^*| > 0$ and $|\widetilde{C} \cap C_{i_2}^*| > 0$, and

(b) A cluster $\widetilde{C} \in \widetilde{\mathcal{C}}$ is an *isolated cluster* if it contains points from only one cluster in $\mathcal{C}^*$.

We then show that the total number of points in mixed clusters is $O(m)$, and can simply add all such points to $\widetilde{D}$. At this point, we would only be left with isolated clusters. Subsequently, we show that two isolated clusters composed of points from the same cluster in $\mathcal{C}^*$ can contain at most $O(m)$ points. Therefore, we once again add these points to $\widetilde{D}$. Finally, we add all the remaining set of at most m outliers to $\widetilde{D}$. It is easy to see that the resulting clustering $\widetilde{\mathcal{C}} \setminus \widetilde{D} = \mathcal{C}^* \setminus D^*$. These results are established in Lemmas 10 and 11. ◀

▶ **Lemma 10.** *Let $\widetilde{C}$ be a mixed cluster, and let $X = \widetilde{C} \cap D^*$ denote its overlap with the outlier set R^* in the optimal* Robust-Correlation-Clustering *clustering. Then we have $|\widetilde{C}| \leq O(1)|X|$.*

Proof. Since $\widetilde{C} \in \widetilde{\mathcal{C}}$ is a mixed cluster, there exists $i_1 \neq i_2$ s.t $|\widetilde{C} \cap C^*_{i_1}| > 0$ and $|\widetilde{C} \cap C^*_{i_2}| > 0$. Now, since $\widetilde{\mathcal{C}}$ is an optimal solution for Correlation-Clustering, we have that the cost of the clustering must increase when we consider the following clustering $\widetilde{\mathcal{C}}_1 = (\widetilde{\mathcal{C}} \setminus \widetilde{C}) \cup (\widetilde{C} \cap C^*_{i_1}) \cup (\widetilde{C} \setminus C^*_{i_1})$ formed by replacing $\widetilde{C}$ with $(\widetilde{C} \cap C^*_{i_1})$ and $(\widetilde{C} \setminus C^*_{i_1})$. since $\mathcal{C}^*$ is an optimal clustering with cost 0, we know that all the edges between $C^*_{i_1}$ and C^*_i for $i \neq i_1$ belong to E_-. This, combined with the fact that the cost of this new clustering is more than that of $\widetilde{\mathcal{C}}$ gives us the following inequality:

$$|\widetilde{C} \cap C^*_{i_1}| \left(\sum_{i \neq i_1} |\widetilde{C} \cap C^*_i| \right) \leq |X||\widetilde{C} \cap C^*_{i_1}|$$
$$\implies \sum_{i \neq i_1} |\widetilde{C} \cap C^*_i| \leq |X| \tag{1}$$

A similar argument by replacing $\widetilde{C}$ with $(\widetilde{C} \cap C^*_{i_2})$ and $(\widetilde{C} \setminus C^*_{i_2})$ would yield $\sum_{i \neq i_2} |\widetilde{C} \cap C^*_i| \leq |X|$. Summing the two inequalities, we get that $|\widetilde{C} \setminus X| \leq 2|X|$, and so $|\widetilde{C}| \leq 3|X|$, completing the proof. ◀

▶ **Lemma 11.** *Let $\widetilde{C}_1$, $\widetilde{C}_2$ be two isolated clusters containing points from the same cluster $C^* \in \mathcal{C}^*$, and let $X_1 = \widetilde{C}_1 \cap D^*$ and $X_2 = \widetilde{C}_2 \cap D^*$ denote their intersections with the outlier set R^* in the optimal* Robust-Correlation-Clustering *clustering. Then we have $|\widetilde{C}_1 \cup \widetilde{C}_2| \leq O(1)|X_1 \cup X_2|$.*

Proof. Since $\widetilde{\mathcal{C}}$ is an optimal solution w.r.t the Correlation-Clustering objective, we know that if we modify $\widetilde{\mathcal{C}}$ by moving the points $\widetilde{C}_1 \cap C^*$ to cluster $\widetilde{C}_2$, the cost does not decrease. This gives us the following inequality, which uses the fact that all edges within C^* belong to E_+ due to the fact that cost of $\mathcal{C}^*$ is 0:

$$|\widetilde{C}_1 \cap C^*||\widetilde{C}_2 \cap C^*| \leq (|X_1| + |X_2|)|\widetilde{C}_1 \cap C^*|$$
$$\implies |\widetilde{C}_2 \cap C^*| \leq |X_1| + |X_2|$$

A similar argument would also give us $|\widetilde{C}_1 \cap C^*| \leq |X_1| + |X_2|$. Adding these inequalities gives us $|\widetilde{C}_1 \cap C^*| + |\widetilde{C}_2 \cap C^*| \leq 2(|X_1| + |X_2|)$, and adding back X_1 and X_2 will incur an additional cost of $|X_1| + |X_2|$, hence completing the proof. ◀

2.2 Approximate Solutions may not be Robust

We next focus on *approximation algorithms* to Correlation-Clustering, and show that they need not be robust to outliers (Theorem 5). Indeed, consider the following instance $\mathcal{I} = (V, E)$ of Robust-Correlation-Clustering with $n + \sqrt{n}$ points. Consider a $\sqrt{n} \times \sqrt{n}$ grid, such that all points lying on the same row are pairwise similar, i.e., belong to E_+ while any two points lying on different rows are dissimilar and belong to E_-. To this arrangement, $\sqrt{n}$ *bad points* are added, which are pairwise dissimilar to one another, but share a $+$ edge with each of the n points in the original $\sqrt{n} \times \sqrt{n}$ grid.

We first note that the optimal Correlation-Clustering solution to $\mathcal{I}$ has cost $\Omega(n\sqrt{n})$. Indeed, consider any triangle u, v, w where u is a bad point, and v and w belong to different rows. Note that there must at least be one mis-classified edge in this triangle in the optimal

solution. So, if we let $\mathcal{B}$ denote the set of all such bad triangles, the following is a valid lower bound on OPT: $\min \sum_{e\in t, t\in\mathcal{B}} z_e$ s.t $\sum_{e\in t} z_e \geq 1, \forall t \in \mathcal{B}$. The dual of this is $\max \sum_{t\in\mathcal{B}} y_t$ s.t $\sum_{t:e\in t, t\in\mathcal{B}} y_t \leq 1, \forall e \in E$. It is easy to see that the optimal value of the dual LP is at least $\Omega(n\sqrt{n})$ by setting $y_t = 1/n$ for all bad triangles in $\mathcal{B}$. Now consider a clustering $\mathcal{C}$ which *clusters each column of the grid* into a cluster, and puts the bad points in another cluster. The overall cost of the clustering is $O(n\sqrt{n})$, which is a constant-factor approximation. Moreover, note that the only way to get a 0 cost clustering from $\mathcal{C}$ (without altering the structure of $\mathcal{C}$) is by deleting all the n grid points.

3 Robust-Correlation-Clustering on Complete Graphs: Hardness

In this section, we give the proof of Theorem 6. The proof follows by an approximation preserving reduction from vertex cover. Consider an instance $\mathcal{I}_{vc}$ of vertex cover, given by a graph, $G = (V, E)$ on n vertices. We construct the Robust-Correlation-Clustering instance $\mathcal{I}$ as follows: for each vertex $v \in V$, we create two points v_1 and v_2, giving us a total of $2n$ vertices in $\mathcal{I}$. For every vertex $v \in V$, we make the edge $(v_1, v_2) \in E_+$. Similarly, for any pair of vertices $u, v \in V$ the edges (u_2, v_2), (u_1, v_2) and (u_2, v_1) all belong to E_-. Finally, we place edge $(u_1, v_1) \in E_+$ if the edge $(u, v) \in E$, and in E_- otherwise. The outlier budget is some parameter m, unrelated to the number of edges in G.

▶ **Lemma 12.** *There exists a solution of cost* 0 *for* $\mathcal{I}$ *if* G *has a vertex cover of size* m*.*

Proof. Let $S \subseteq V$ denote a vertex cover of size m for G, and let $S_1 = \{v_1 : v \in S\}$. Then, consider the natural clustering $\mathcal{C} = \{\{v_1, v_2\} : v \in V\}$ comprising of the pairs of vertices. The only mis-classified edges in this clustering are of the form (u_1, v_1) corresponding to edges (u, v) of G. But now, suppose we declare the points in S_1 as outliers, then it follows that the resulting clustering $\mathcal{C} \setminus S_1$ has 0 cost, since S is a vertex cover for G. ◀

▶ **Lemma 13.** *If there is a set* S *of* m *outliers such that the remaining points has a* 0 *cost clustering* $\mathcal{C}$ *in instance* $\mathcal{I}$*, then* G *has a vertex cover of size at most* m *in instance* $\mathcal{I}_{vc}$*.*

Proof. We construct a candidate vertex cover S' for G from the outlier-set S as follows: for each $v \in V$, include $v \in S'$ if either v_1 or v_2 is in S. We claim then that S' is a valid vertex cover for G. To the contrary, suppose an edge (u, v) is not covered by S'. Then, none of the four points u_1, u_2, v_1, v_2 are included in the outlier-set S in the robust clustering solution. Now, since clustering $\mathcal{C}$ has 0 cost, it must be that the four points u_1, u_2, v_1 and v_2 must belong to the same cluster in $\mathcal{C}$, or else, one of the edges in (u_1, u_2), (u_2, v_2), and (v_2, v_1), all of which belong to E_+, would be mis-classified. But now the edges (u_1, v_2) and (v_1, u_2) belong to E_- and would be mis-classified in $\mathcal{C}$, which contradicts the fact that $\mathcal{C}$ has 0 cost. ◀

Theorem 6 then follows from Lemmas 12 and 13.

4 Robust-Correlation-Clustering on Complete Graphs: Algorithms

In this section, we design a simple LP-rounding based bi-criteria approximation algorithm for Robust-Correlation-Clustering (Problem 2) and prove Theorem 7. We begin by recalling the problem setup: we are given an instance $\mathcal{I}$ consisting of a graph (V, E_+, E_-) on n points with $E_+ \cup E_- = \binom{V}{2}$. The goal is to identify a set of vertices D such that $|D| = m$, and a clustering $\mathcal{C}$ over $V \setminus D$ such that the total cost is minimized. We start with the following definition crucial to the design and analysis of our algorithm.

▶ **Definition 14** (Bad Triangles). *A triplet (u, v, w) of points is said to be a* bad triangle *if exactly two of the three edges among $(u, v), (v, w), (u, w)$ belong to E_+ and one to E_-.*

Note a bad triangle captures the *smallest unit of inconsistency* in the similarity information among the points: either we delete one of the vertices as an outlier, or at least one of the edges must be mis-classified. In what follows, let $\mathcal{B}$ denote the set of all bad triangles in $\mathcal{I}$.

4.1 Recap of ACNAlg for Correlation-Clustering [2]

Since the crux of our algorithm is the ACNAlg for correlation clustering, we begin with a quick recap of ACNAlg. Essentially, the algorithm iteratively picks a *random* un-clustered vertex v as a new cluster center, and includes all other un-clustered vertices similar to v.

Algorithm 1 ACNAlg(V, E_+, E_-).

set $U = V$ and $C = \emptyset$ ▷ initialize set of un-clustered points and set of cluster centers
while $U \neq \Phi$ **do**
  sample $v \sim \text{Unif}(U)$
  update $C \leftarrow C \cup \{v\}$ ▷ random v is sampled as a cluster center
  let $C_v = \{u \in U : (u, v) \in E_+\} \cup \{v\}$ ▷ un-clustered vertices similar to v
  update $U \leftarrow U \setminus C_v$
end while
return: $\mathcal{C} = \{C_v \ : v \in C\}$

▶ **Theorem 15** ([2]). *ACNAlg(V, E_+, E_-) is a 3 approximation for* Correlation-Clustering.

In what follows, we outline the proof in [2] of ACNAlg, and describe a couple of definitions and lemmas which will be useful in understanding our overall analysis.

▶ **Definition 16.** *A bad triangle $(u, v, w) \in \mathcal{B}$ is said to be* touched*, denoted by* touched$(t) = 1$*, if there exists a point in the algorithm execution when all three vertices u, v, w belong to the un-clustered set U and one of u, v, w gets sampled as a cluster center.*

▶ **Lemma 17.** *At the end of Algorithm 1, every mis-classified edge (i.e., an E_- edge which is in a single cluster, or an E_+ edge which goes across clusters) is associated with a unique bad triangle which is touched. Moreover, the opposite vertex to the mis-classified edge must be sampled as the cluster center.*

Proof. Consider a stage of the algorithm when a vertex u gets chosen as a cluster center. Then any newly mis-classified edge (v, w) can be of two types: (i) $(v, w) \in E_-$ is mis-classified due to both (u, v) and (u, w) belonging to E_+; (ii) (v, w) belonging to E_+, with $(u, v) \in E_+$ and $(u, w) \in E_-$. In both cases we can associate the newly mis-classified edge (v, w) with the unique bad triangle (u, v, w) which gets touched. ◀

Proof of Theorem 15. The first step is the following LP-based lower bound on Opt$(\mathcal{I})$. Indeed, we know that each bad triangle must have at least one mis-classified edge, and so the LP is simply a linear relaxation for finding a maximal set of disjoint bad triangles.

$$\begin{aligned} \text{maximize} \quad & \sum_{t \in \mathcal{B}} w_t, && s.t., \\ & \sum_{t \in \mathcal{B}: u,v \in t} w_t \leq 1, && \forall e = (u, v) \in E, \\ & w_t \in [0, 1], && \forall t \in \mathcal{B}. \end{aligned} \tag{LP1}$$

Since it will be useful in the next section, we state the dual program, which is a relaxation for the hitting set for all bad triangles.

$$\begin{aligned} \text{minimize} \quad & \sum_{u,v} z_{u,v}, && s.t., \\ z_{u,v} + z_{v,w} + z_{u,w} &\geq 1, && \forall t \in \mathcal{B}, \\ z_{u,v} &\in [0,1], && \forall u, v \in \mathcal{B}. \end{aligned} \tag{LP2}$$

Now, let $p_t = \mathbb{E}[\mathsf{touched}(t)]$, where $\mathsf{touched}(t)$ is the indicator random variable for whether a bad triangle t is touched in the algorithm. The crux of the proof is the following lemma.

▶ **Lemma 18.** *The values* $\{\mathbb{E}[\mathsf{touched}(t)]/3 : t \in \mathcal{B}\}$ *form a feasible solution to LP1.*

Proof. To this end, consider any edge $e = (u, v)$ and the set of bad triangles $\mathcal{B}_{u,v} = \{(u, v, w) \in \mathcal{B}\}$ it is part of. Lemma 17 tells us that (u, v) will be mis-classified if and only if one of these bad triangles $t \equiv (u, v, w) \in \mathcal{B}_{u,v}$ is touched, and the third vertex w must be picked as a cluster center when the triangle is touched. Finally note that, for any triangle $t \equiv (u, v, w)$, the probability that w is picked as the cluster center conditioned on $\mathsf{touched}(t)$ is exactly $1/3$, since the algorithm selects the new cluster center uniformly at random from the un-clustered vertices. Thus we have that: $1 \geq \mathbb{P}((u, v) \text{ is mis-classified}) = \sum_{t \in \mathcal{B}_{u,v}} p_t/3$, thereby showing the LP feasibility of $\{p_t/3\}$. ◀

Also note that by Lemma 17, we have that $\mathbb{E}[\mathsf{cost}(\mathcal{C})] = \sum_{t \in \mathcal{B}} p_t$, where $\mathsf{cost}(\mathcal{C})$ is the objective value of the clustering $\mathcal{C}$. Lemma 18 coupled with this inequality bounding the cost completes the proof of Theorem 15. ◀

4.2 LP-rounding algorithm for Robust-Correlation-Clustering

We now present our constant-factor bi-criteria approximation for Robust-Correlation-Clustering which uses ACNAlg as a sub-routine. Since the ACNAlg algorithm analysis bounds the expected cost of the clustering in terms of the LP relaxation LP1, by duality, we can also infer that the expected cost of ACNAlg is bounded by the LP relaxation LP2. We use this intuition as our starting point: indeed, we can extend this covering LP to handle outliers in the following natural manner. Let $z_{u,v}$ denote whether an edge (u, v) is mis-classified, and y_u denote whether a vertex is deleted or not. Then the following LP3 is a valid LP relaxation for Robust-Correlation-Clustering on complete graphs.

$$\begin{aligned} \text{minimize} \quad & \sum_{(u,v) \in \binom{V}{2}} z_{u,v}, && s.t. \\ y_u + y_v + y_w + z_{u,v} + z_{v,w} + z_{u,w} &\geq 1, && \forall t = (u, v, w) \in \mathcal{B}, \qquad (2) \\ \sum_u y_u &\leq m, \\ z_{u,v} &\geq 0, && \forall (u, v) \in \binom{V}{2}, \\ y_u &\geq 0, && \forall u \in V. \end{aligned} \tag{LP3}$$

Equation (2) of LP3 states that at least a unit cost is incurred for any bad triangle in $\mathcal{B}$ if no vertices from this triangle are deleted. Let $\{y^*_u : u \in V\}, \left\{z^*_{u,v} : (u, v) \in \binom{V}{2}\right\}$ denote the optimal solution to LP3.

▶ **Lemma 19.** $\mathsf{Opt}(\mathcal{I}) \geq \sum_{(u,v) \in \binom{V}{2}} z^*_{u,v} = \mathsf{Opt}(LP3)$.

Proof. Indeed, consider any optimal solution to the Robust-Correlation-Clustering instance, and set $z_{u,v} = 1$ if (u, v) is mis-classified, and $y_u = 1$ if u is deleted. For any bad triangle $(u, v, w) \in \mathcal{B}$, note that either one of u, v or w must be deleted as an outlier in the optimal solution, or one of the three edges must be mis-classified. Hence the first LP constraint is satisfied. The second is true since the optimal solution deletes at most m outliers. Finally, the objective function captures the number of mis-classified edges. ◀

Algorithm 2 RCCAlg(V, E_+, E_-, m).

Initialization: $V_{\text{del}} \leftarrow \emptyset$ ▷ Set of deleted vertices
Let the optimal solution of LP3 be denoted as $\{y^*_u : u \in V\} \cup \{z^*_{uv} : (u, v) \in \binom{V}{2}\}$
$V_{\text{del}} \leftarrow \{v \in V : y^*_v \geq 1/6\}$ ▷ Delete vertices having $y^*_v \geq 1/6$
$V' \leftarrow V \setminus V_{\text{del}}$
return: ACNAlg(V', E'_+, E'_-) ▷ E'_+, E'_-: edges in $\binom{V}{2}$ not incident on V_{del}

4.3 Analysis

▶ **Theorem 20.** *RCCAlg(V, E_+, E_-, m) is a bi-criteria $(6, 6)$-approximation for* Robust-Correlation-Clustering.

Proof. The proof of this result follows from Lemmas 21 and 22. ◀

▶ **Lemma 21.** *At most $6m$ vertices are deleted by RCCAlg(V, E_+, E_-, m).*

Proof. Recall that RCCAlg(V, E_+, E_-, m) deletes those vertices having $y^*_u \geq 1/6$ in the optimal solution to LP3. Let the set of vertices deleted by RCCAlg(V, E_+, E_-, m) be denoted V_{del}. Then,

$$|V_{\text{del}}| = \sum_{u \in V} \mathbb{1}(y^*_u \geq 1/6) \leq \sum_{u \in V} 6y^*_u \leq 6m.$$

Therefore, the budget of vertices to remove is not exceeded by more than a factor of 6. ◀

We next bound the cost incurred by the clustering output by RCCAlg(V, E_+, E_-, m).

▶ **Lemma 22.** *The cost of the clustering output by RCCAlg(V, E_+, E_-, m) is at most 6 times the cost of the optimal clustering to $\mathcal{I}$.*

Proof. Since the first step deletes vertices in $V_{\text{del}} = \{v \in V : y^*_v \geq 1/6\}$, it suffices to consider the remaining vertices $V' = V \setminus V_{\text{del}}$ and show that ACNAlg has cost at most 6Opt on the residual instance. The proof is again very simple: indeed, each vertex $v' \in V'$ has $y^*_{v'} \leq 1/6$, we get that the optimal LP solution to LP3 satisfies $z^*_{u,v} + z^*_{v,w} + z^*_{u,w} \geq 1/2$ for all $(u, v, w) \in \mathcal{B}'$, where $\mathcal{B}'$ denotes the set of all bad triangles induced in the vertex set V'. Then by simply considering the scaled variables $2z^*_{u,v}$, we get that there exists a feasible solution to LP2 for the Correlation-Clustering instance induced in (V', E'_+, E'_-), of cost at most 2Opt. Hence, since the 3-approximation of ACNAlg guarantee holds against the dual LP LP1, we can use weak duality to complete the proof. ◀

5 Algorithms for Robust-Correlation-Clustering on General Graphs

In this section, we consider Robust-Correlation-Clustering on general graphs and prove Theorem 8. Given an instance $\mathcal{I}$, comprising of graph $G = (V, E_+ \cup E_-)$ and outlier budget m, we begin with the following LP relaxation:

$$\text{Minimize} \quad \sum_{(u,v)\in E_+\cup E_-} z_{u,v}, \qquad s.t., \tag{LP6}$$

$$x_{u,v} + x_{v,w} \geq x_{u,w}, \qquad \forall u \neq v \neq w \tag{3}$$

$$y_u + y_v + z_{u,v} \geq 1 - x_{u,v}, \qquad \forall (u,v) \in E_- \tag{4}$$

$$y_u + y_v + z_{u,v} \geq x_{u,v}, \qquad \forall (u,v) \in E_+ \tag{5}$$

$$\sum_u y_u \leq m, \tag{6}$$

$$x_{u,v}, z_{u,v}, y_u \in [0,1]$$

In simple terms, on imposing integer constraints, LP6 asks to find a clustering s.t. $x_{u,v} = 1$ if u and v belong to different clusters, and 0 otherwise. It is easy to check that such an assignment of $x_{u,v}$ satisfies the triangle inequality constraint Equation (3). The objective function charges a unit cost ($z_{u,v} = 1$) for dissimilar (resp. similar) pairs of points (u, v) placed in the same (resp. different) clusters, only if neither u nor v is deleted, i.e, if $y_u = y_v = 0$. In addition, Equation (6) ensures that at most m vertices are deleted in the intended solution. The following lemma is then an immediate consequence of the fact that the optimal integral solution to Robust-Correlation-Clustering instance $\mathcal{I}$ is feasible for Equation (LP6).

▶ **Lemma 23.** *The optimal solution $\{x^*, y^*, z^*\}$ to the LP above has objective value at most* $\mathsf{Opt}(\mathcal{I})$*, the cost of an optimal* Robust-Correlation-Clustering *solution. Moreover, we may slightly perturb this solution to ensure that (a) $\min_{(u,v):x^*_{u,v}\neq 0} x^*_{u,v} \geq 1/n^2$ and $\min_{u:y^*_u\neq 0} y^*_u \geq 1/n^2$, i.e., the smallest non-zero values among x^* and y^* variables is at least $1/n^2$, and (b) the perturbed solution has same objective value and satisfies all the LP inequalities except Equation* (6)*, which is satisfied up to $\sum_u y^*_u \leq (m + 1/n)$.*

We require the lower bound on the x^* and y^* variables for technical reasons which will become clear as the proof proceeds. However, for all practical purposes, the reader may assume that it is just the optimal solution to the LP. We begin by observing that the one of the techniques of solving the Correlation-Clustering problem is by reducing it to Minimum-Multicut problem (in fact, up to constant factors, the Correlation-Clustering problem on general graphs is *equivalent* to Minimum-Multicut on general graphs in [14]), and running the best known approximation to Minimum-Multicut to get $O(\log n)$ approximations to Correlation-Clustering. In our case, for Robust-Correlation-Clustering, just like how we used a specific approximation algorithm ACNAlg for Correlation-Clustering, it turns out that the right starting point for general graphs is the following beautiful partitioning scheme (Theorem 24) for metric spaces known as *padded decompositions*. At a high level, they randomly *partition* a metric space into regions of bounded diameter, such that the probability of a *ball of radius ρ around any vertex v* being separated by the partitioning is proportional to ρ. This generalizes the standard partitioning schemes which just guarantee that the probability that any pair u, v being separated is proportional to $d(u, v)$. While any scheme which satisfies the latter suffices to get good algorithms for Correlation-Clustering, we crucially use the stronger property in our algorithm for Robust-Correlation-Clustering.

▶ **Theorem 24** ([15]). *For any finite metric space (X, d) and parameter $\Delta > 0$, there exists a randomized algorithm* PaddedClustering(X, d, Δ) *which outputs a clustering $\mathcal{C}$ of points in X such that,*

- *Every cluster $C \in \mathcal{C}$ has diameter at most Δ,*
- *For every $x \in X$ and $\rho \in (0, \Delta/8)$,*

$$\text{Prob}(\text{Ball}_\rho(x) \not\subseteq \mathcal{C}(x)) \leq \alpha(x)\frac{\rho}{\Delta}, \tag{7}$$

where $\alpha(x) = \mathcal{O}(\log(\frac{|\text{Ball}_\Delta(x)|}{|\text{Ball}_{\Delta/8}(x)|})) = O(\log n)$ and $\mathcal{C}(x)$ denotes the points in the same cluster as x in $\mathcal{C}$.

5.1 Rounding Algorithm

Before we describe the algorithm in detail, we now provide an overview.

Step 1. We first compute a near-optimal solution $\{x^*, y^*, z^*\}$ for Equation (LP6) satisfying the conditions of Lemma 23.

Step 2. We run the padded decomposition scheme on x^* with $\Delta = 0.25$ to obtain a clustering $\mathcal{C}^*$ of the points. Indeed, we can interpret $\mathcal{C}^*$ as a *rounding* of the $x_{u,v}$ variables into an integral clustering: if $x^*_{u,v} \geq 0.25$, then u and v are definitely in different clusters of $\mathcal{C}^*$, and if $x^*_{u,v}$ is small, then the are in different clusters with probability $\propto O(\log n)x^*_{u,v}$.

Step 3. If a mis-classified edge in this clustering has $z^*_{u,v}$ at least some constant, say 0.25, then we can charge such edges to the LP objective.

Step 4a. It remains to consider mis-classified edges with small $z^*_{u,v}$. If $(u, v) \in E_-$, then again this is an easy case, since we know that $x^*_{u,v} \leq 0.25$ because (u, v) is mis-classified, hence it must belong to the same cluster, and all clusters have diameter at most 0.25 w.r.t the x^* metric. Hence, if $z^*_{u,v} \leq 0.25$ for such edges, we can infer that $y^*_u + y^*_v \geq 0.5$ from Equation (4), and we can handle all such edges by *deleting all vertices* with $y^*_u \geq 0.25$.

Step 4b. We are finally left with handling the case when $(u, v) \in E_+$, and $z^*_{u,v}$ is small. Here again, we are in good shape if $x^*_{u,v}$ is at least some constant, since from Equation (5) we know that at least one of y^*_u or y^*_v or $z^*_{u,v}$ must be large, so we can either delete an end-point of (u, v), or we can charge this mis-classified edge to the LP objective. On the other hand, if $x^*_{u,v}$ is small and (u, v) is mis-classified (and so u and v belong to different clusters since $(u, v) \in E_+$), we use the padded decomposition property that such an event occurred with very low probability, and we can actually afford to the scale variables by $x^*_{u,v}$ to get that $\frac{y^*_u}{x^*_{u,v}} + \frac{y^*_v}{x^*_{u,v}} + \frac{z^*_{u,v}}{x^*_{u,v}} \geq 1$. In expectation, the overall scaling factor would be bounded from Theorem 24, and moreover, for each mis-classified edge in E_+, we can either charge it to the scaled $z^*_{u,v}$ variable, or delete an end-point due to the scaled y^*_u or y^*_v being large. Of course, this is a simplified view since we cannot consider different scaling factors for different edges. In our actual algorithm, we scale each y^*_v by a quantity r_v, where r_v is the radius of the *smallest ball around v* w.r.t metric s^* which gets separated by the clustering $\mathcal{C}^*$. This is where our proof uses the stronger properties of the padded decomposition schemes.

▶ **Theorem 25.** *RCC-general(V, E_+, E_-, m) is a randomized $(\mathcal{O}(\log n), \mathcal{O}(\log^2 n))$ bi-criteria approximation for* Robust-Correlation-Clustering *on general graphs.*

Proof. We begin by introducing some notation that will be useful for the analysis of the algorithm. Consider the clustering $\mathcal{C}^*$ output by PaddedClustering$(V, x^*, 0.25)$ in RCC-general(V, E_+, E_-, m). We slightly abuse notation and let $\mathcal{C}^*(v)$ denote the set of all vertices

Algorithm 3 RCC-general(V, E_+, E_-, m).

1: Let $\{x^*, y^*, z^*\}$ denote the (perturbed) optimal solution to LP6 obtained in Lemma 23
2: Compute $\mathcal{C}^* = \mathsf{PaddedClustering}(V, x^*, 0.25)$
3: Define $V_b^- = \{v \in V : \exists u \in \mathcal{C}^*(v) \text{ such that } (u,v) \in E_-\}$ ▷ candidate vertices for deletion: have a $-$ edge to at least one other vertex in the same cluster
4: Define $V_{\text{del}}^- = \{v \in V_b^- \ : \ y_v^* \geq 1/4\}$
5: Set $V' \leftarrow V \setminus V_{\text{del}}^-$
6: Define $V_b^+ = \{v \in V' : \exists u \in V' \setminus \mathcal{C}^*(v) \text{ such that } (u,v) \in E_+\}$ ▷ candidate vertices for deletion: have a $+$ edge to at least one vertex in a different cluster
7: For each $u \in V_b^+$, define

$$\hat{y}_u \stackrel{\text{def}}{=} 2^r \cdot y_u^*, \quad \text{where } \frac{1}{2^r} < \min_{v \in V' \setminus \mathcal{C}^*(u)} x_{u,v}^* \leq \frac{1}{2^{r-1}}$$

8: Define $V_{\text{del}}^+ = \{v \in V_b^+ : \hat{y}_v \geq 1/3\}$
9: **Return:** $D_{\mathsf{alg}} = V_{\text{del}}^- \cup V_{\text{del}}^+$ as outliers and the clustering $\mathcal{C}_{\mathsf{alg}} = \mathcal{C}^* \setminus D$

which are in the same cluster as v in the clustering $\mathcal{C}^*$. Define E_b^- as the set of $-$ edges between vertices in V in the same cluster in $\mathcal{C}^*$, $E_b^- \stackrel{\text{def}}{=} \{(u,v) \in E_- \ : \ u \in \mathcal{C}^*(v)\}$. In addition, define E_b^+ to be the set of $+$ edges between vertices in V' lying in different clusters in $\mathcal{C}^*$, i.e., $E_b^+ \stackrel{\text{def}}{=} \{(u,v) \in E_+ \ : \ u \in V' \setminus \mathcal{C}^*(v)\}$. Let $\mathsf{cost}(\mathsf{alg})$ denote the cost of the clustering output by RCC-general(V, E_+, E_-, m) and let $V_{\text{del}} = V_{\text{del}}^- \cup V_{\text{del}}^+$ denote the set of vertices deleted. Observe that any edge that contributes to $\mathsf{cost}(\mathsf{alg})$ belongs to either E_b^+ or E_b^- and is not incident on any vertex in V_{del}. Therefore, $\mathsf{cost}(\mathsf{alg})$ can be decomposed as

$$\mathsf{cost}(\mathsf{alg}) \leq \mathsf{cost}(\mathsf{alg})^- + \mathsf{cost}(\mathsf{alg})^+. \tag{8}$$

where $\mathsf{cost}(\mathsf{alg})^-$ denotes the cost associated with edges in E_b^- that are not incident on vertices in V_{del}^-, and $\mathsf{cost}(\mathsf{alg})^+$ denotes the cost associated with edges in E_b^+ that are not incident on vertices in $V_{\text{del}}^- \cup V_{\text{del}}^+$.

Let Opt^* denote the cost of the optimal solution to LP6. To bound the cost of our solution, we show in Lemmas 28 and 33 respectively that $\mathsf{cost}(\mathsf{alg})^-$ is upper-bounded by $4\mathsf{Opt}^*$, while $\mathbb{E}\left[\mathsf{cost}(\mathsf{alg})^+\right]$ is upper-bounded by $O(\log n)\mathsf{Opt}^*$.

On the other hand, to bound the number of vertices deleted by RCC-general(V, E_+, E_-, m), we follow a similar strategy. Since, $|V_{\text{del}}| = |V_{\text{del}}^+| + |V_{\text{del}}^-|$, we separately upper bound V_{del}^- and $\mathbb{E}[V_{\text{del}}^+]$ in Lemmas 27 and 32 by $4m$ and $\mathcal{O}(\log^2 n)m$ respectively. ◀

Recall that the optimal solution of LP6 is denoted as $\left(\{y_u^*\}, \{x_{u,v}^*\}, \{z_{u,v}^*\}\right)$. We begin by establishing some basic properties of the clustering $\mathcal{C}^*$.

▷ Claim 26. For any edge $(u,v) \in E_b^-$, $y_u^* + y_v^* + z_{u,v}^* \geq 0.75$.

Proof. Recall that E_b^- denotes the set of dissimilar points in V that are placed in the same cluster by $\mathcal{C}^*$. Since, $E_b^- \subseteq E_-$, the optimal solution to LP6 must satisfy the negative edge-constraint (4) for edge (u,v), and so $y_u^* + y_v^* + z_{u,v}^* \geq 1 - x_{u,v}^*$. Now, note that $x_{u,v}^* \leq 0.25$, since u and v belong to the same cluster in $\mathcal{C}^*$ and the diameter of any cluster in $\mathsf{PaddedClustering}(X, d, \Delta)$ is at most Δ from Theorem 24. ◁

▶ **Lemma 27.** *The set of vertices, V_{del}^- satisfies $|V_{\text{del}}^-| \leq 4\sum_{v \in V} y_v^* \leq 4(m + 1/n)$.*

Proof. Recall that V_{del}^- is the set of vertices, $v \in V_b^-$ such that $y_v^* \geq 1/4$. This, combined with the fact that $\{y_v^*\}$ satisfies $\sum_u y_u^* \leq m + 1/n$ from Lemma 23, completes the proof. ◀

▶ **Lemma 28.** *The cost of mis-classified E_- edges $\mathsf{cost}(\mathsf{alg})^-$ is at most $4\sum_{(u,v)\in E^-} z^*_{u,v}$.*

Proof. Observe that $\mathsf{cost}(\mathsf{alg})^-$ accrues unit cost only for edges in E_b^- which are not incident on a vertex in V_{del}^-. This implies that $y^*_u \le 0.25$ for all vertices incident on such edges. This, combined with Claim 26 completes the proof. ◀

We now move onto the analysis of $\mathsf{cost}(\mathsf{alg})^+$ and $|V^+_{\text{del}}|$, which are slightly more involved. In this respect, define

$$\hat{z}_{u,v} \stackrel{\text{def}}{=} \begin{cases} \frac{z^*_{u,v}}{x^*_{u,v}}, & v \notin \mathcal{C}^*(u), \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

We demonstrate some useful facts about $\hat{z}_{u,v}$ and $\hat{y}_u$, which recall is defined previously as,

$$\hat{y}_u = 2^r \cdot y^*_u, \quad \text{where, } r \ : \ \frac{1}{2^r} < \min_{v\in V'\setminus\mathcal{C}^*(u)} x^*_{u,v} \le \frac{1}{2^{r-1}}$$

▷ Claim 29. For any edge $(u,v) \in E_b^+$, $\mathbb{E}[\hat{z}_{u,v}] \le \mathcal{O}(\log n) z^*_{u,v}$.

Proof. Observe that if two points belong to different clusters, then we must necessarily have for $\rho = x^*_{u,v}$ that $\text{Ball}_\rho(u) \not\subseteq \mathcal{C}(u)$). Therefore, from Theorem 24,

$$\text{Prob}(u \notin \mathcal{C}^*(v)) \le \mathcal{O}(\log n)\frac{x^*_{u,v}}{0.25}.$$

Therefore, from the definition of $\hat{z}_{u,v}$ in (9), it follows that, $\mathbb{E}[\hat{z}_{u,v}] \le \mathcal{O}(\log n)\frac{x^*_{u,v}}{0.25}\frac{z^*_{u,v}}{x^*_{u,v}} + 0 = \mathcal{O}(\log n) z^*_{u,v}$. ◁

▷ Claim 30. For any vertex $v \in V_b^-$, $\mathbb{E}\left[\hat{y}_u\right] \le \mathcal{O}(\log^2 n)\cdot y^*_u$.

Proof. Observe that $x^*_{u,v} \in [n^{-2}, 1]$. Therefore, r takes values from the set $\{0, 1, 2, \ldots, 2\log n\}$. By definition of $\hat{y}_u$,

$$\begin{aligned} \mathbb{E}[\hat{y}_u] &= \sum_{r=0}^{2\log n} 2^r \left(y^*_u\right) \text{Prob}\left(\frac{1}{2^r} < \min_{v\in V\setminus\mathcal{C}^*(u)} x^*_{u,v} \le \frac{1}{2^{r-1}}\right), \\ &\le \sum_{r=0}^{2\log n} 2^r \left(y^*_u\right) \text{Prob}\left(\min_{v\in V\setminus\mathcal{C}^*(u)} x^*_{u,v} \le \frac{1}{2^{r-1}}\right). \end{aligned} \tag{10}$$

Next, observe that the event $\min_{v\in V'\setminus\mathcal{C}^*(u)} x^*_{u,v} \le 2^{-(r-1)}$ can only occur if the ball of radius $2^{-(r-1)}$ centered at u does not lie entirely within $\mathcal{C}(u)$. Therefore, from Theorem 24,

$$\text{Prob}\left(\min_{v\in V\setminus\mathcal{C}^*(u)} x^*_{u,v} \le \frac{1}{2^{r-1}}\right) \le \mathcal{O}(\log n)\frac{1}{2^{r-1}}.$$

Plugging this into (10) gives, $\mathbb{E}[\hat{y}_u] \le \mathcal{O}(\log n)\sum_{r=0}^{2\log n} y^*_u = \mathcal{O}(\log^2 n)\cdot y^*_u$. ◁

▷ Claim 31. For any edge $(u,v) \in E_b^+$, we have that $\hat{y}_u + \hat{y}_v + \hat{z}_{u,v} \ge 1$.

Proof. Since $E_b^+ \subseteq E_+$, every $(u,v) \in E_b^+$ must satisfy the positive edge-constraint (5) $y^*_u + y^*_v + z^*_{u,v} \ge x^*_{u,v}$. The proof then concludes by dividing both sides by $x^*_{u,v}$, and using the definitions of $\hat{y}_u$ and $\hat{z}_{u,v}$. ◁

▶ **Lemma 32.** *The set of vertices V^+_{del} satisfies,* $\mathbb{E}\left[\left|V^+_{\text{del}}\right|\right] \leq \mathcal{O}(\log^2 n)\ m$.

Proof. Recall that V^+_{del} is defined as the set of vertices $v \in V^+_b$ such that $\hat{y}_v \geq 1/3$. Therefore $|V^+_{\text{del}}| = \sum_{v \in V^+_b} \mathbb{1}(\hat{y}_v \geq 1/3)$. Since $\mathbb{1}(\hat{y}_v \geq 1/3) \leq 3\hat{y}_v$, it follows that $|V^+_{\text{del}}| \leq 3\sum_{v \in V^+_b} \hat{y}_v$. Taking expectation on both sides, and using Claim 30, $\mathbb{E}[|V^+_{\text{del}}|] \leq \mathcal{O}(\log^2 n) \sum_{v \in V^+_b} y^*_v$. The proof concludes by relaxing the summation $v \in V^+_b$ to $v \in V$, and using Lemma 23 to claim that $\sum_{v \in V} y^*_v \leq m + \frac{1}{n} \leq 2m$. ◀

▶ **Lemma 33.** *The expected cost of the mis-classified E_+ edges $\mathbb{E}\left[\mathsf{cost}(\mathsf{alg})^+\right]$ is at most $\mathcal{O}(\log n) \sum_{(u,v) \in E^+} z^*_{u,v}$.*

Proof. $\mathsf{cost}(\mathsf{alg})^+$ is the cost corresponding to edges in E^+_b which are not incident on any vertex in V_{del}. Recall that a vertex $v \in V'$ belongs to V_{del} only if $\hat{y}_v \geq 1/3$. Following a similar proof as Lemma 28, we get that,

$$\mathsf{cost}(\mathsf{alg})^+ \leq \sum_{(u,v) \in E^+_b} \mathbb{1}(\hat{z}_{u,v} \geq 1/3) \leq 3 \sum_{(u,v) \in E^+_b} \hat{z}_{u,v},$$

Taking expectations on both sides, using Claim 29 to upper bound $\mathbb{E}[\hat{z}_{u,v}]$ by $\mathcal{O}(\log n) z^*_{u,v}$, and relaxing the summation to $(u,v) \in E_+$ completes the proof. ◀

References

1. KookJin Ahn, Graham Cormode, Sudipto Guha, Andrew McGregor, and Anthony Wirth. Correlation Clustering in Data Streams. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2237–2246, Lille, France, 2015. PMLR. URL: `http://proceedings.mlr.press/v37/ahn15.html`.
2. Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating Inconsistent Information: Ranking and Clustering. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pages 684–693, New York, NY, USA, 2005. ACM. `doi:10.1145/1060590.1060692`.
3. Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation Clustering. *Mach. Learn.*, 56(1-3):89–113, June 2004. `doi:10.1023/B:MACH.0000033116.57574.95`.
4. Amir Ben-Dor and Zohar Yakhini. Clustering Gene Expression Patterns. In *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, RECOMB '99, pages 33–42, New York, NY, USA, 1999. ACM. `doi:10.1145/299432.299448`.
5. Moses Charikar, Venkatesan Guruswami, and Anthony Wirth. Clustering with Qualitative Information. *J. Comput. Syst. Sci.*, 71(3):360–383, October 2005. `doi:10.1016/j.jcss.2004.10.012`.
6. Moses Charikar, Samir Khuller, David M. Mount, and Giri Narasimhan. Algorithms for Facility Location Problems with Outliers. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, pages 642–651, Philadelphia, PA, USA, 2001. Society for Industrial and Applied Mathematics. URL: `http://dl.acm.org/citation.cfm?id=365411.365555`.
7. Sanjay Chawla and Aristides Gionis. k-means-: A Unified Approach to Clustering and Outlier Detection. In *SDM*, pages 189–197. SIAM, 2013. URL: `http://dblp.uni-trier.de/db/conf/sdm/sdm2013.html#ChawlaG13`.
8. Shuchi Chawla, Konstantin Makarychev, Tselil Schramm, and Grigory Yaroslavtsev. Near Optimal LP Rounding Algorithm for CorrelationClustering on Complete and Complete K-partite Graphs. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pages 219–228, New York, NY, USA, 2015. ACM. `doi:10.1145/2746539.2746604`.

9 Jiecao Chen, Erfan Sadeqi Azer, and Qin Zhang. A Practical Algorithm for Distributed Clustering and Outlier Detection. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2253–2262, 2018. URL: http://papers.nips.cc/paper/7493-a-practical-algorithm-for-distributed-clustering-and-outlier-detection.

10 Ke Chen. A Constant Factor Approximation Algorithm for K-median Clustering with Outliers. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '08, pages 826–835, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics. URL: http://dl.acm.org/citation.cfm?id=1347082.1347173.

11 Flavio Chierichetti, Nilesh Dalvi, and Ravi Kumar. Correlation Clustering in MapReduce. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 641–650, New York, NY, USA, 2014. ACM. doi:10.1145/2623330.2623743.

12 William Cohen and Jacob Richman. Learning to Match and Cluster Entity Names. In *In ACM SIGIR-2001 Workshop on Mathematical/Formal Methods in Information Retrieval*, 2001.

13 William W. Cohen and Jacob Richman. Learning to Match and Cluster Large High-dimensional Data Sets for Data Integration. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 475–480, New York, NY, USA, 2002. ACM. doi:10.1145/775047.775116.

14 Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2):172–187, 2006. Approximation and Online Algorithms. doi:10.1016/j.tcs.2006.05.008.

15 Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. Approximating Metrics by Tree Metrics. *SIGACT News*, 35(2):60–70, June 2004. doi:10.1145/992287.992300.

16 Shalmoli Gupta, Ravi Kumar, Kefu Lu, Benjamin Moseley, and Sergei Vassilvitskii. Local Search Methods for k-Means with Outliers. *PVLDB*, 10(7):757–768, 2017. doi:10.14778/3067421.3067425.

17 Ravishankar Krishnaswamy, Shi Li, and Sai Sandeep. Constant Approximation for K-median and K-means with Outliers via Iterative Rounding. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 646–659, New York, NY, USA, 2018. ACM. doi:10.1145/3188745.3188882.

18 Shi Li and Xiangyu Guo. Distributed k-Clustering for Data with Heavy Noise. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 7849–7857, 2018. URL: http://papers.nips.cc/paper/8009-distributed-k-clustering-for-data-with-heavy-noise.

19 Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Correlation Clustering with Noisy Partial Information. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1321–1342, Paris, France, 2015. PMLR. URL: http://proceedings.mlr.press/v40/Makarychev15.html.

20 Andrew McCallum and Ben Wellner. Toward Conditional Models of Identity Uncertainty with Application to Proper Noun Coreference. In *Proceedings of the 2003 International Conference on Information Integration on the Web*, IIWEB'03, pages 79–84. AAAI Press, 2003. URL: http://dl.acm.org/citation.cfm?id=3104278.3104294.

21 Napat Rujeerapaiboon, Kilian Schindler, Daniel Kuhn, and Wolfram Wiesemann. Size Matters: Cardinality-Constrained Clustering and Outlier Detection via Conic Optimization. *SIAM Journal on Optimization*, 2019.

22 Chaitanya Swamy. Correlation Clustering: Maximizing Agreements via Semidefinite Programming. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, volume 15, pages 526–527, January 2004.

23 Anthony Wirth. Correlation Clustering. In *Encyclopedia of Machine Learning*, pages 227–231. Springer, 2010. doi:10.1007/978-0-387-30164-8_176.

A Hardness of Robust-Correlation-Clustering on General Graphs

Firstly, when $m = 0$, Robust-Correlation-Clustering is simply Correlation-Clustering, for which is known NP-hardness of $\Omega(\alpha_{\text{MC}})$ [5]. We show that it is NP-hard to get any (a, b)-approximation for Robust-Correlation-Clustering with finite b when $a < \alpha_{\text{MC}}$, for any $m > 0$.

▶ **Theorem 34.** *It is NP-hard to have an (a, b) bi-criteria approximation to* Robust-Correlation-Clustering *for any finite b and $a < \alpha_{\text{MC}}$.*

Proof. The proof is via a reduction from Minimum-Multicut, similar to the proof for Correlation-Clustering in [5]. Consider the Minimum-Multicut instance problem $\mathcal{I} = \{G(V, E), \{(s_i, t_i), 1 \leq i \leq k\}\}$, where $(s_i, t_i), 1 \leq i \leq k$ represent k source-sink pairs. We construct the Robust-Correlation-Clustering problem instance $\mathcal{I}^*$ as follows. The edges in G become + edges in $\mathcal{I}^*$. For each $i, 1 \leq i \leq k$, we add a negative edge between (s_i, t_i) of weight $-W$, for some large positive integer W, say $W = n^3$. We can make the instance unweighted by replacing a negative edge of weight $-W$ by W parallel length two paths; each path has a fresh intermediate vertex, with one + edge and one − edge. Clearly, the minimum cost clustering must have (s_i, t_i) in different clusters $\forall 1 \leq i \leq k$. In addition, introduce m more vertices which act like outliers, represented by set $U = \{u_1, u_2, \ldots, u_m\}$ in $\mathcal{I}^*$. Connect each $u_i, 1 \leq i \leq m$ to every vertex $q, q \in V(\mathcal{I}^*) \setminus U$ with an edge of weight $-W$ and an edge of weight W. We can make the instance unweighted by replacing the negative edge as described before, and the positive edge of weight W by W parallel length two paths; each path has a fresh intermediate vertex, with both edges +.

Due to the above construction, the vertices $(q, u_i), q \in V(\mathcal{I}^*) \setminus U, 1 \leq i \leq k$ add a high cost irrespective of whether they lie in the same cluster or not.

Hence, the optimal solution to Robust-Correlation-Clustering on the problem instance $\mathcal{I}^*$ removes vertices $u_1, u_2, \cdots, u_m$, and the corresponding optimal cost is same as the Minimum-Multicut optimal cost on instance $\mathcal{I}$. ◀

We next establish that unless the budget of vertices to be removed is violated by a certain factor, it is NP-hard to find any approximation to the cost of the optimal solution to Robust-Correlation-Clustering.

▶ **Theorem 35.** *It is NP-hard to find an (a, b) bi-criteria approximation to* Robust-Correlation-Clustering *for any finite a, and $b < \alpha_{\text{MC}}$.*

Proof. The proof of this result once again follows via a reduction from Minimum-Multicut. Indeed, consider the Minimum-Multicut instance problem $\mathcal{I} = \{G(V, E), \{(s_i, t_i), 1 \leq i \leq k\}\}$, where $(s_i, t_i), 1 \leq i \leq k$ represent k source-sink pairs. We now define an intermediate problem which will simplify our overall reduction. ◀

▶ **Definition 36** (Vertex-Multicut)**.** *Given a problem instance $\mathcal{I} = \{H, \{(s_i, t_i), 1 \leq i \leq k\}\}$, where $(s_i, t_i), 1 \leq i \leq k$ represent k source-sink pairs, the* Vertex-Multicut *problem is to find the minimum set of vertices $S \subseteq V(H)$ such that no source-sink pair lie in the same connected component in the graph induced on $V(H) \setminus S$.*

▶ **Lemma 37.** *There exists an approximation preserving reduction from* Minimum-Multicut *to* Vertex-Multicut*.*

Proof. The idea is to reduce the MINIMUM-MULTICUT problem instance $\mathcal{I}$ to a VERTEX-MULTICUT problem instance $\mathcal{I}' = \{H(V', E'),\ \{(s'_i, t'_i),\ 1 \leq i \leq l\}\}$. Consider the graph $G = (V, E)$ as defined above. Reduce each vertex $v_i \in V$ into a clique of large size, say n, where $n = |V|$. Let clique$(v_i) = \{v_{i1}, v_{i2}, \ldots, v_{in}\}$, where $v_i \in V, 1 \leq i \leq n$ represent the clique in H. For every $(s_i, t_i),\ 1 \leq i \leq k$ source-sink pair in $\mathcal{I}$, let each of $(s_{ia}, t_{ib})\ \forall 1 \leq a, b \leq n$ be a source sink pair in instance $\mathcal{I}'$. Hence, instance $\mathcal{I}'$ will contain kn^2 source-sink pairs in comparison with the k pairs in $\mathcal{I}$. We now define the edges in $\mathcal{I}'$. E' is composed of two components, $\cup_{i \leq n} E_{\text{clique}(v_i)}$ and E_{across}, where $E_{\text{clique}(v_i)} = \{(v_{ia}, v_{ib}),\ 1 \leq i, a, b \leq n, a \neq b\}$, and $E_{\text{across}} = \{(v_{ij}, v_{ji})\ :\ (v_i, v_j) \in E\}$.
We now have a VERTEX-MULTICUT problem instance $\mathcal{I}'$. We claim that the reduction from $\mathcal{I}$ to $\mathcal{I}'$ is an approximation preserving reduction. Let S denote the optimal solution to problem instance $\mathcal{I}'$, that is, S denotes the optimal set of vertices to remove to disconnect the source-sink pairs. Let $v_{ij} \in S,\ 1 \leq i, j \leq n$. Removing the edge $(v_i, v_j) \in E$ in instance $\mathcal{I}$ is equivalent to removing the vertex v_{ij} (or v_{ji}) in $\mathcal{I}'$ where $(u_i, v_j) \in E'$. Hence solving the VERTEX-MULTICUT problem solves MINIMUM-MULTICUT problem as well. ◀

▶ **Lemma 38.** *There exists an approximation preserving reduction from* VERTEX-MULTICUT *to approximating the budget of number of vertices to remove in* ROBUST-CORRELATION-CLUSTERING *problem.*

Proof. Given a VERTEX-MULTICUT problem instance $\mathcal{I}' = \{H, \{(s_i, t_i) 1 \leq i \leq k, \}\}$, we construct a ROBUST-CORRELATION-CLUSTERING problem instance $\mathcal{I}''$. The edges in H becomes positive edges in $\mathcal{I}''$. In addition, add a negative edge between each (s_i, t_i) pair of weight $-W$, for some large positive integer W, say $W = n^3$. The graph can be made unweighted as discussed in the proof to Theorem 34.

Consider the instance $\mathcal{I}''$. The minimum set of vertices R such that the graph induced on remaining vertices has a 0 cost clustering is identical to the optimal solution to the instance $\mathcal{I}'$. From Lemma 37, it follows that if $\mathcal{I}'$ can be solved optimally, the underlying MINIMUM-MULTICUT problem instance $\mathcal{I}$ can be solved optimally. Therefore from Theorem 34 and Lemma 37, it follows that it is NP-hard to violate the budget of number of vertices to remove by a factor $< \alpha_{\text{MC}}$ such that the cost of the output clustering is a finite approximation to the optimal cost. ◀