Robust Correlation Clustering

A Project Report

submitted by

NIVED RAJARAMAN

in partial fulfilment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY AND MASTER OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.

MAY 2019

THESIS CERTIFICATE

This is to certify that the thesis titled **Robust Correlation Clustering**, submitted by **Nived Rajaraman**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelor of Technology and Master of Technology**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Ravishankar Krishnaswamy
Research Guide
Researcher
Foundations Group
Microsoft Research, 560 001

Prof. Andrew Thangaraj
Research Co-Guide
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: IIT Madras, Chennai

Date: 7^{th} May 2019

ACKNOWLEDGEMENTS

I would like to express my gratitude to my guide Dr. Ravishankar Krishnaswamy for giving me an opportunity to work on this project. Furthermore I would like to thank Devvrit for his enthusiasm as a co-worker on this project for the past couple of months. I am also grateful to Profs. Andrew Thangaraj and Rahul Vaze from whom I have learnt much over the past years, and have truly helped broaden my academic horizons. I am also full of gratitude for my parents for being a constant and unwavering support in my life, and I have nothing but admiration for my brother, Amit for being an example of someone who is truly passionate about what they do.

ABSTRACT

KEYWORDS: qualitative clustering, NP-hardness, multicut

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 an important problem when we have (only) qualitative information about the similarity or dissimilarity of pairs of points, and ROBUST-CORRELATION-CLUSTERING equips this model the capability to handle noise in the datasets.

In this work, our main result is 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 approximation is possible if we do not violate the outlier budget. A nice feature of our algorithm is that it first runs a particular CORRELATION-CLUSTERING algorithm ACNAlg Ailon *et al.* (2005), and then does a simple *post-processing* by deleting $O(m)$ vertices from the clustering output by ACNAlg. This in fact suggests that the ACNAlg algorithm is inherently robust to outliers! 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 NP-hardness factor for the MINIMUM-MULTICUT problem.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 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 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. Bansal *et al.* (2004) formulated a problem known as *correlation clustering* (in fact, a similar problem was implicitly studied by Ben-Dor et al.Ben-Dor and Yakhini (1999) as 'Cluster Editing').

Problem 1.0.1 (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 otfher 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.*

In addition to the problem requiring only qualitative (dis)similarity information between pairs of data points, another 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 optimization to infer. 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. Bansal *et al.* (2004) 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 Wirth (2010)) getting better factors, with the current best bound being a factor of 2.06 due to Chawla et al. Chawla *et al.* (2015).

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 Charikar *et al.* (2001); Chen (2008); Krishnaswamy *et al.* (2018). Motivated by this observation, and the fact that real-world data sets are often noisy, we investigate the *robustness* of correlation clustering.

Problem 1.0.2 (ROBUST-CORRELATION-CLUSTERING)**.** *The input to this problem is identical to the correlation clustering instance as in Problem 1.0.1. 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 Charikar

et al. (2005). In this paper, we also consider the ROBUST-CORRELATION-CLUSTERING problem on general graphs, analogous to the study of CORRELATION-CLUSTERING in general graphs Charikar *et al.* (2005).

Problem 1.0.3 (ROBUST-CORRELATION-CLUSTERING on General Graphs)**.** *The input and problem objective are identical to that in Problem 1.0.2, with the sole 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 1.1.1. *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 1.0.1). 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

[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 $\mathcal{C}$

objectives such as k-means and k-median where an analogous statement to Theorem 1.1.1 does not hold. Moreover, we believe that it is conceivable 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 1.1.2. *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 (1.0.1), 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 1.0.1, where we know very good constant-factor approximations.

Theorem 1.1.3. *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 to delete.*

We then turn our attention to obtaining *bi-criteria approximation algorithms*: an (a, b) bi-criteria approximation for ROBUST-CORRELATION-CLUSTERING is one where the cost of our solution is at most a times the optimal cost, and the number of points our solution discards is at most $b \cdot m$.

Theorem 1.1.4. *There is an efficient combinatorial bi-criteria $(6, 6)$-approximation algorithm for* ROBUST-CORRELATION-CLUSTERING *on complete graphs.*

A nice property of our algorithm in fact just executes a specific constant-factor approximation due to Ailon et al. Ailon *et al.* (2005) (henceforth called ACNAlg) for CORRELATION-

CLUSTERING on our given instance, and then performs a simple *post-processing* on the resulting clustering $\mathcal{C}$ which removes $O(m)$ points. We use a randomized dual fitting argument building to bound the approximation ratio of the overall algorithm. We remark that the post-processing is innately tied to the specific CORRELATION-CLUSTERING algorithm we use, and we simply cannot replace it with any constant-factor approximation algorithm for CORRELATION-CLUSTERING.

Our result in a sense says that ACNAlg is *inherently robust* to outliers. Indeed, note that if $\mathsf{Opt}(\mathcal{I})$ is small, that means that there exists a clustering on the entire set of points for which most of the mis-classified edges are incident on a few (m) vertices. And while this may not be true of other CORRELATION-CLUSTERING algorithms, our proof says that this structure is preserved in the solution output by ACNAlg, which is what we exploit to derive our final ROBUST-CORRELATION-CLUSTERING algorithm.

Finally, we turn our attention to ROBUST-CORRELATION-CLUSTERING on general graphs.

Theorem 1.1.5. *There is an efficient bi-criteria $(O(\log n), O(\log^2 n))$-approximation algorithm for* ROBUST-CORRELATION-CLUSTERING *on general graphs.*

While the CORRELATION-CLUSTERING problem is equivalent to MINIMUM-MULTICUT Demaine *et al.* (2006) and we can use any MINIMUM-MULTICUT algorithm to solve CORRELATION-CLUSTERING, we show that one specific technique based on *padded decompositions* of metric spaces is naturally the correct approach to solving the robust problem. Finally, we also show hardness results on general graphs.

Theorem 1.1.6. *It is NP-hard to obtain any bi-criteria (a, b)-approximation algorithm for* ROBUST-CORRELATION-CLUSTERING *on general graphs for $b < \alpha_{\mathrm{MC}}$ or $a < \alpha_{\mathrm{MC}}$ where α_{MC} is the inapproximability factor for the* MINIMUM-MULTICUT *problem.*

We leave it as an interesting open question to resolve the factor of violation in the number of outliers: while the hardness shows that $O(\log n)$ is necessary unless we improve the approximability of the classical MINIMUM-MULTICUT problem, our algorithm gets a bound of $O(\log^2 n)$.

1.2 Related Work

Since its introduction, CORRELATION-CLUSTERING has received much attention with focus on designing better algorithms (see the survey of Wirth (2010)), faster algorithms in the parallel and distributed Chierichetti *et al.* (2014) and streaming settings Ahn *et al.* (2015), stochastic/average-case settings Makarychev *et al.* (2015), and applications Cohen and Richman (2001, 2002); McCallum and Wellner (2003). There is also work on a related objective function of *maximizing* the number of classified edges Bansal *et al.* (2004). 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 Charikar *et al.* (2005); Swamy (2004). 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 Chen (2008); Krishnaswamy *et al.* (2018), and designing faster algorithms Chawla and Gionis (2013); Rujeerapaiboon *et al.* (2019); Gupta *et al.* (2017), and parallel and distributed algorithms in this model Chen *et al.* (2018); Li and Guo (2018). 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 Chapter 2. We then consider ROBUST-CORRELATION-CLUSTERING for complete graphs, and show our hardness of approximation in Chapter 3, followed by the combinatorial algorithm in Chapter 5. Finally, in Section 5.4 and chapter 4, we turn our attention to the case of general graphs and present our algorithm and hardness.

CHAPTER 2

Is the CORRELATION-CLUSTERING objective inherently robust?

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 combined with that fact that correlation clustering is APX-hard Bansal *et al.* (2004) serves as a strong motivation for studying the ROBUST-CORRELATION-CLUSTERING problem.

2.1 Inherent Robustness of Optimal Solutions for CORRELATION-CLUSTERING

In this section, we exhibit the inherent robustness of the correlation clustering objective (1.0.1) 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 1.0.1) to instance $\mathcal{I}$. Our goal now is to investigate how these solutions compare with the optimal solution to Problem 1.0.2. Indeed, we show that there exist $O(m)$ points, deleting which, the objective indeed becomes 0 for the optimal CORRELATION-CLUSTERING clustering. This tells us that the optimal solutions to 1.0.2 and 1.0.1 are nearly identical to each other (upto $O(m)$ points), and hence, that the correlation clustering objective is inherently robust, unlike traditional clustering objectives such as k-median and k-means.

Proof of Theorem 1.1.1. 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 (1.0.1). 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 $\mathcal{C}^*$ also contain at most $O(m)$ points. Therefore, we once again add these points to $\widetilde{D}$. Finally, we add whatever remains of D^* (at most m points) to $\widetilde{D}$. It is easy to see that the resulting clustering $\widetilde{\mathcal{C}} \setminus \widetilde{D} = \mathcal{C}^* \setminus \widetilde{D}$. The full proof is in Section 6.1. □

2.2 Non-Robustness of Approximate Solutions

We next focus on *approximation algorithms* to CORRELATION-CLUSTERING, and show that they need not be robust to outliers (Theorem 1.1.2). 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.

CHAPTER 3

Hardness of ROBUST-CORRELATION-CLUSTERING on complete graphs

In this section, we prove Theorem 1.1.3. The proof follows by an approximation preserving reduction from vertex cover. Consider any unlabelled graph, $G = (V, E)$ on n vertices. Let $\mathsf{vc}(G)$ denote the set of vertices corresponding to the minimum vertex cover on G. We construct the ROBUST-CORRELATION-CLUSTERING instance, $\mathcal{I}_G$ from G as follows: for each vertex $v \in V$, we create two points v_1 and v_2. 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.

Note that the only mis-classified edges in the natural clustering $\mathcal{C} = \{\{v_1, v_2\} :, v \in V\}$ obtained by grouping the vertices v_1 and v_2 for each $v \in V$ are the (u_1, v_1) edges corresponding to $(u, v) \in E$. Hence, if there is a vertex cover S for G of at most m vertices, we may simply delete $\{u_1 : u \in S\}$ and obtain a clustering of 0 cost. Likewise, we can show that we can efficiently recover a vertex cover for G of size m' from any clustering which deletes m' vertices and has 0 cost. The formal proof is in Section 6.4.

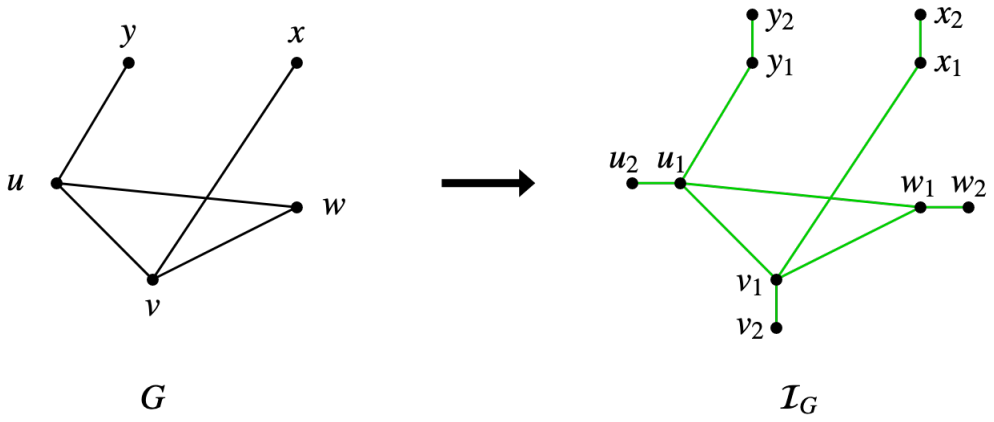

Figure 3.1: Reducing vertex cover instance G to $\mathcal{I}_G$

CHAPTER 4

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 $\Omega(\alpha_{\mathrm{MC}})$ Charikar *et al.* (2005). We show that it is NP-hard to get any (a, b)-approximation for ROBUST-CORRELATION-CLUSTERING with finite b when $a < \alpha_{\mathrm{MC}}$, for any $m > 0$.

Theorem 4.0.1. *It is NP-hard to have an* (a, b) *bi-criteria approximation to* ROBUST-CORRELATION-CLUSTERING *for any finite* b *and* $a < \alpha_{\mathrm{MC}}$.

Proof. The proof is via a reduction from MINIMUM-MULTICUT, similar to the proof for CORRELATION-CLUSTERING in Charikar *et al.* (2005). 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 4.0.2. *It is NP-hard to find an (a, b) bi-criteria approximation to* ROBUST-CORRELATION-CLUSTERING *for any finite a, and $b < \alpha_{\mathrm{MC}}$.*

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 4.0.3 (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 4.0.4. *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 $\mathrm{clique}(v_i) = \{v_{i1}, v_{i2}, \ldots, v_{in}\}$, 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 4.0.5. *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 4.0.1.

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 4.0.4, 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 4.0.1 and lemma 4.0.4, 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. □

CHAPTER 5

Algorithms for ROBUST-CORRELATION-CLUSTERING on Complete Graphs

In this section, we design efficient and *combinatorial* bi-criteria approximation algorithms for ROBUST-CORRELATION-CLUSTERING (Problem 1.0.2) and prove Theorem 1.1.4. 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 5.0.1 (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 the instance $\mathcal{I}$.

5.1 Recap of ACNAlg for CORRELATION-CLUSTERING Ailon *et al.* (2005)

Since the first step of our algorithm is ACNAlg for correlation clustering, we first present these details. In words, this very elegant algorithm picks a random unclustered vertex as a new cluster center, includes all other unclustered vertices it is similar to in its cluster, and iterates till all points are clustered.

Theorem 5.1.1. ACNAlg(V, E_+, E_-) *is a* 3 *approximation for Problem 1.0.1.*

Algorithm 1 $\mathsf{ACNAlg}(V, E_+, E_-)$

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

Before we prove Theorem 5.1.1, we begin with some crucial observations about this algorithm which will be useful in understanding our generalization to ROBUST-CORRELATION-CLUSTERING.

Definition 5.1.2. *A triangle $(u, v, w) \in \mathcal{B}$ is* touched *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 5.1.3. *At the end of Algorithm 1, every mis-clasified 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.*

We now prove Theorem 5.1.1. We remark that while this is directly not useful for us, we will prove some lemmas which we will use in our final analysis.

Proof of Theorem 5.1.1. The first step is the following LP-based lower bound on $\mathsf{Opt}(\mathcal{I})$. Indeed, we know that each bad triangle must have at least one mis-clasified edge, and so the LP is simply a linear relaxation for finding a maximal set of disjoint bad triangles.

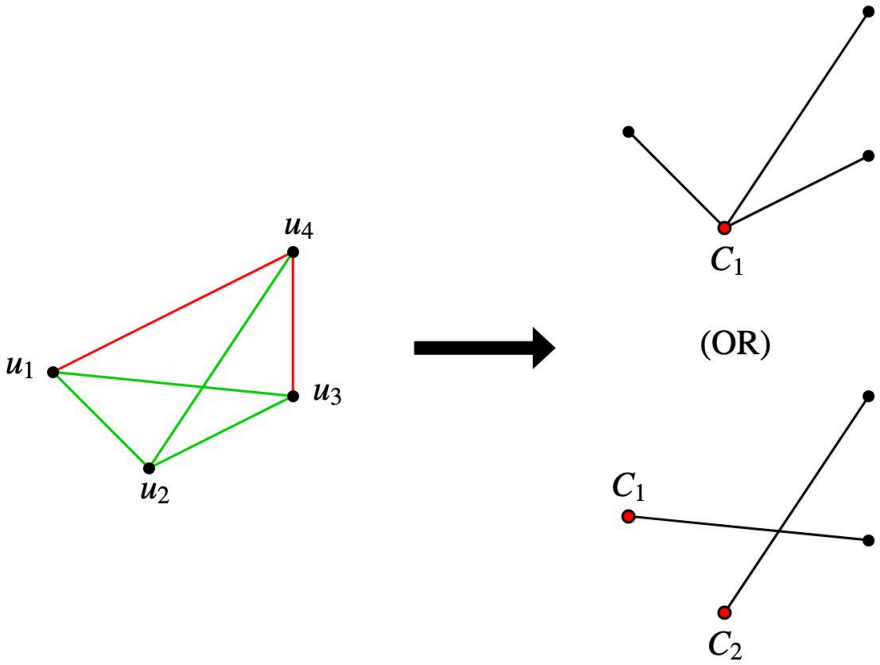

Figure 5.1: Clustering output by ACNAlg (V, E_+, E_-)

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

Now, for any triangle $t \equiv (u, v, w) \in \mathcal{B}$, let $\mathsf{touched}(t)$ denote the indicator random variable for whether triangle t is touched in the algorithm or not, and let $p_t = \mathbb{E}[\mathsf{touched}(t)]$. Note that by Lemma 5.1.3, we have that $\mathbb{E}[\mathsf{cost}(\mathcal{C})] = \sum_{t \in \mathcal{B}} p_t$, where $\mathsf{cost}(\mathcal{C})$ denotes the objective value of the clustering $\mathcal{C}$ output by Algorithm 1.

The crux of the proof is the following lemma.

Lemma 5.1.4. *The values $\{p_t/3 : t \in \mathcal{B}\}$ form a feasible solution to the LP relaxation 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 5.1.3 tells us that (u, v) will be mis-clasified 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: $1 \geq \mathbb{P}((u, v) \text{ is mis-clasified}) = \sum_{t \in \mathcal{B}_{u,v}} p_t/3$, thereby showing the LP feasibility of $\{p_t/3\}$. □

Lemma 5.1.4 coupled with the inequality bounding the cost $\mathbb{E}[\mathsf{cost}(\mathcal{C})] \leq \sum_t p_t$ then completes the proof. □

5.2 Our Algorithm for ROBUST-CORRELATION-CLUSTERING

The first step of our final algorithm runs ACNAlg to compute a clustering $\mathcal{C}$, which say has a set X of mis-classified edges. In the second step, *we fix the structure of the clustering* $\mathcal{C}$ and simply try to delete some $O(m)$ vertices such that the number of edges in X which are not deleted is minimized. This sub-problem is reminiscent of prize-collecting (or) partial vertex cover-type problems for which there are simple combinatorial primal-dual algorithms. The crucial part of the proof is to show that this strategy indeed works, i.e., the number of edges in X which are not deleted is at most $O(1)\mathsf{Opt}$. For this, we use a randomized duality-based argument building on the proof of Lemma 5.1.3.

Throughout this section, we assume that we know the value Opt of the optimal solution for the ROBUST-CORRELATION-CLUSTERING instance $\mathcal{I}$; we can easily handle this assumption using a standard guess-and-double strategy. Moreover, we also assume that $\mathsf{Opt} > m$.[1]

5.3 Outline of Proof of Theorem 1.1.4

The starting point is the following primal-dual pair, where we have lifted the budget constraint to the objective.

[1] Otherwise, we know that there exists $2m$ vertices deleting which, the remaining points are perfectly clusterable, i.e., there are no bad triangles. So we can simply solve a hitting set for all the bad triangles and obtain a set of $6m$ vertices to hit all bad triangles.

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

Initialization: $V_r \leftarrow \Phi$ ▷ V_r is the set of deleted/outlier vertices
Run ACNAlg(V, E_+, E_-) to get output clustering $\mathcal{C}$
Let $\mathcal{B}_{\text{touch}}$ denote the set of bad triangles "touched" by ACNAlg;
Mark all bad triangles in $\mathcal{B}_{\text{touch}}$ as unfrozen
for all $u \in V$ **do**
    Let $\mathcal{B}_u$ denote the set of triangles $t \in \mathcal{B}_{\text{touch}}$ such that $u \in t$ and u is not the cluster center
end for
while $\exists$ an unfrozen triangle in $\mathcal{B}_{\text{touch}}$ **do**
    Choose any unfrozen $t \in \mathcal{B}_{\text{touch}}$
    Raise Δ_t until 1 or until for some vertex u, $\sum_{t \in \mathcal{B}_u} \Delta_t = \frac{2\mathsf{Opt}}{m}$
    for all u such that $\sum_{t \in \mathcal{B}_u} \Delta_t = \frac{2\mathsf{Opt}}{m}$ **do**
        Mark all $\{t \in \mathcal{B}_{\text{touch}} : u \in t\}$ as frozen
    end for
end while
Define $V_r = \left\{u : \sum_{t \in \mathcal{B}_u} \Delta_t = \frac{2\mathsf{Opt}}{m}\right\}$
Return: Clustering $\mathcal{C}$, V_r

$$\text{Minimize} \quad \sum_{(u,v) \in \binom{V}{2}} z_{u,v} + \frac{\mathsf{Opt}}{m} \sum_u y_u, \qquad s.t., \tag{LP2}$$

$$y_u + y_v + y_w + z_{u,v} + z_{v,w} + z_{u,w} \geq 1, \qquad \forall t = (u,v,w) \in \mathcal{B},$$

$$z_{u,v} \in [0,1], \qquad \forall (u,v) \in \binom{V}{2},$$

$$y_u \in [0,1], \qquad \forall u \in V.$$

$$\text{Maximize} \quad \sum_{t \in \mathcal{B}} w_t, \qquad s.t., \tag{LP3}$$

$$\sum_{t \in \mathcal{B}: u,v \in t} w_t \leq 1, \qquad \forall (u,v) \in \binom{V}{2}, \tag{5.1}$$

$$\sum_{t: u \in t} w_t \leq \frac{\mathsf{Opt}}{m}, \qquad \forall v \in V, \tag{5.2}$$

$$w_t \geq 0, \qquad \forall v \in V.$$

Lemma 5.3.1. *The value of an optimal solution to Equation* (LP2) *is at most* $2 \cdot \mathsf{Opt}$ *where* Opt

is the objective value of an optimal solution to the ROBUST-CORRELATION-CLUSTERING *instance* $\mathcal{I}$.

Now, recall from Definition 5.1.2 the definition of *touched* triangles in the execution of ACNAlg (V, E_+, E_-). We will use A_t to represent the event that triangle t is touched in some iteration. Also recall the value p_t, the probability that a triangle $t \in \mathcal{B}$ is touched during execution. Indeed, the proof of Theorem 5.1.1 proceeded by exhibiting showing that the solution $\{p_t : t \in \mathcal{B}\}$ satisfies equation eq. (5.1) and has expected cost equal to the dual objective. This was sufficient for the CORRELATION-CLUSTERING problem. However, there is no reason for this solution to satisfy equation eq. (5.2), which is needed in our case. Indeed, this is where our primal-dual step comes in. At a high level, we consider all the edges mis-classified by ACNAlg, and keep raising the dual variables Δ_t of the *unique* bad triangle associated with them that is touched (from Lemma 5.1.3) until either end-vertex becomes tight (i.e., total dual reaches $2\mathsf{Opt}/m$). In this case, we freeze all the bad edges incident on such tight vertices and proceed.

We then show that the collection $\{\mathbb{E}[\Delta_t/3]\}$ satisfies all the dual constraints. Indeed, since Δ_t is non-zero (and at most 1) only for triangles touched by ACNAlg, we have that $\{\mathbb{E}[\Delta_t/3]\}$ satisfies eq. (5.1) from Lemma 5.1.3. Moreover, $\{\Delta_t/2\}$ satisfies eq. (5.2) by definition even without expectations taken. As a result, we can infer that $\{\mathbb{E}[\Delta_t/3]\}$ is dual-feasible, and hence the sum $\sum_{t\in\mathcal{B}} \mathbb{E}[\Delta_t]$ is at most $O(1)\mathsf{Opt}$ using weak duality.

It remains to show that the number of mis-classified edges is at most $O(1)\mathsf{Opt}$, and also that the number of vertices deleted is at most $O(1)m$. To see the first property, note that for every edge mis-classified by ACNAlg, either $\Delta_t = 1$ for the unique bad triangle associated with it, or the dual constraint for one of the end-points becomes tight. In the latter, this edge will be deleted since we delete all tight vertices. Hence, the total number of mis-classified edges which remains can be upper bounded by $\sum_t \Delta_t$, and hence the expected value is at most $O(1)\mathsf{Opt}$. Finally, to see the deletion bound, note that for every vertex deleted, we have that $\sum_{t\in\mathcal{B}_u} \Delta_t = 2\mathsf{Opt}/m$, and also each Δ_t can contribute to at most 2 vertices (it is a part of exactly the $\mathcal{B}_u$ set for two vertices). Hence, we get that the expected number of vertices deleted can be at most $O(1)m$, again using the fact that $\sum_{t\in\mathcal{B}} \mathbb{E}[\Delta_t/3]$. The formal proof is

in Section 6.2.

5.4 Algorithms for ROBUST-CORRELATION-CLUSTERING on General Graphs

In this section, we prove Theorem 1.1.5. The starting point for our algorithm is the following LP relaxation for ROBUST-CORRELATION-CLUSTERING:

$$\text{Minimize} \quad \sum_{(u,v)\in E_+\cup E_-} z_{u,v}, \qquad s.t., \tag{LP4}$$

$$x_{u,v} + x_{v,w} \geq x_{u,w}, \qquad \forall u \neq v \neq w \tag{5.3}$$

$$y_u + y_v + z_{u,v} \geq 1 - x_{u,v}, \qquad \forall (u,v) \in E_- \tag{5.4}$$

$$y_u + y_v + z_{u,v} \geq x_{u,v}, \qquad \forall (u,v) \in E_+ \tag{5.5}$$

$$\sum_u y_u \leq m, \tag{5.6}$$

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

In simple terms, on imposing integer constraints, LP4 asks to find a clustering $\{x_{u,v} : (u,v) \in E_+ \cup E_-\}$, but only 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, the metric constraint in (5.3) is present to ensure that any integer solution to LP4 corresponds to a consistent clustering.

Lemma 5.4.1. *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.*

After solving the LP, we run the following *padded decomposition* scheme to partition the metric x^* to get clusters of diameter at most 0.25 using $\Delta = 0.25$.

Theorem 5.4.2 (Fakcharoenphol *et al.* (2004))**.** *For any finite metric space (X,d) and parameter $\Delta > 0$, there exists a randomized algorithm $\mathsf{PaddedClustering}(X,d,\Delta)$ 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 $\rho \in (0, \Delta/8)$,*

$$\text{Prob}(\text{Ball}_\rho(x) \not\subseteq \mathcal{C}(x)) \leq \alpha(x)\frac{\rho}{\Delta}, \tag{5.7}$$

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 $\mathcal{C}$.

We also delete all vertices v such that $y^*_v \geq 0.25$ for deletion. This will suffice to handle all the E_- edges which are mis-classified at this point. Indeed, by the diameter property of our clustering, we have that $x^*_{u,v}$ for all u, v in a single cluster is at most 0.25, and hence constraint eq. (5.4) implies that $y^*_u + y^*_v + z^*_{u,v} \geq 0.75$, and so, the total cost of E_- edges which are not deleted is at most $4\sum_{e \in E_-} z_e$.

Handling E_+ edges is trickier, and it is here where we use the full power of the padded decomposition scheme. Indeed, let E^b_+ denote the set of E_+ edges which are mis-classified by the padded decomposition clustering. Note that if an edge $e \in E_+$ has large z_e value, then this can be charged to the LP cost. So it remains to handle the edges with small z_e value which get separated. However, note that such an edges being separated happens with low probability due to the guarantees of the padded decomposition. To handle these edges, we *scale* each $\hat{y}_v = y^*_v/r_v$ variable, where r_v is the radius of the smallest ball around v which gets separated by the partitioning scheme, and then show that the $\hat{y}_v$ variables form a feasible solution to the *vertex cover problem* for the E^b_+ edges which have small z_e values. To bound the number of vertices delete, we again use the padded decomposition property to argue that the expected value of $\hat{y}_v$ is bounded by $O(\log^2 n)y^*_v$, and so overall we get the $O(\log^2 n)m$ bound on the number of vertices deleted. Please refer to Section 6.3 for complete details of this algorithm and analysis.

REFERENCES

1. **Ahn, K.**, **G. Cormode**, **S. Guha**, **A. McGregor**, and **A. Wirth**, Correlation clustering in data streams. *In* **F. Bach** and **D. Blei** (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*. PMLR, Lille, France, 2015. URL http://proceedings.mlr.press/v37/ahn15.html.

2. **Ailon, N.**, **M. Charikar**, and **A. Newman**, Aggregating inconsistent information: Ranking and clustering. *In Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05. ACM, New York, NY, USA, 2005. ISBN 1-58113-960-8. URL http://doi.acm.org/10.1145/1060590.1060692.

3. **Bansal, N.**, **A. Blum**, and **S. Chawla** (2004). Correlation clustering. *Mach. Learn.*, **56**(1-3), 89–113. ISSN 0885-6125. URL https://doi.org/10.1023/B:MACH.0000033116.57574.95.

4. **Ben-Dor, A.** and **Z. Yakhini**, Clustering gene expression patterns. *In Proceedings of the Third Annual International Conference on Computational Molecular Biology*, RECOMB '99. ACM, New York, NY, USA, 1999. ISBN 1-58113-069-4. URL http://doi.acm.org/10.1145/299432.299448.

5. **Charikar, M.**, **V. Guruswami**, and **A. Wirth** (2005). Clustering with qualitative information. *J. Comput. Syst. Sci.*, **71**(3), 360–383. ISSN 0022-0000. URL http://dx.doi.org/10.1016/j.jcss.2004.10.012.

6. **Charikar, M.**, **S. Khuller**, **D. M. Mount**, and **G. Narasimhan**, Algorithms for facility location problems with outliers. *In Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. ISBN 0-89871-490-7. URL http://dl.acm.org/citation.cfm?id=365411.365555.

7. **Chawla, S.** and **A. Gionis**, k-means-: A unified approach to clustering and outlier detection. *In SDM*. SIAM, 2013. ISBN 978-1-61197-283-2. URL http://dblp.uni-trier.de/db/conf/sdm/sdm2013.html#ChawlaG13.

8. **Chawla, S.**, **K. Makarychev**, **T. Schramm**, and **G. 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. ACM, New York, NY, USA, 2015. ISBN 978-1-4503-3536-2. URL http://doi.acm.org/10.1145/2746539.2746604.

9. **Chen, J.**, **E. S. Azer**, and **Q. 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.*. 2018. URL http://papers.nips.cc/paper/7493-a-practical-algorithm-for-distributed-clustering-and-outlier-detecti

10. **Chen, K.**, 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. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008. URL http://dl.acm.org/citation.cfm?id=1347082.1347173.

11. **Chierichetti, F.**, **N. Dalvi**, and **R. Kumar**, Correlation clustering in mapreduce. *In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14. ACM, New York, NY, USA, 2014. ISBN 978-1-4503-2956-9. URL http://doi.acm.org/10.1145/2623330.2623743.

12. **Cohen, W.** and **J. Richman**, Learning to match and cluster entity names. *In In ACM SIGIR-2001 Workshop on Mathematical/Formal Methods in Information Retrieval*. 2001.

13. **Cohen, W. W.** and **J. Richman**, Learning to match and cluster large high-dimensional data sets for data integration. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02. ACM, New York, NY, USA, 2002. ISBN 1-58113-567-X. URL http://doi.acm.org/10.1145/775047.775116.

14. **Demaine, E. D.**, **D. Emanuel**, **A. Fiat**, and **N. Immorlica** (2006). Correlation clustering in general weighted graphs. *Theoretical Computer Science*, **361**(2), 172 – 187. ISSN 0304-3975. URL http://www.sciencedirect.com/science/article/pii/S0304397506003227. Approximation and Online Algorithms.

15. **Fakcharoenphol, J.**, **S. Rao**, and **K. Talwar** (2004). Approximating metrics by tree metrics. *SIGACT News*, **35**(2), 60–70. ISSN 0163-5700. URL http://doi.acm.org/10.1145/992287.992300.

16. **Gupta, S.**, **R. Kumar**, **K. Lu**, **B. Moseley**, and **S. Vassilvitskii** (2017). Local search methods for k-means with outliers. *PVLDB*, **10**(7), 757–768. URL http://www.vldb.org/pvldb/vol10/p757-Lu.pdf.

17. **Krishnaswamy, R.**, **S. Li**, and **S. 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. ACM, New York, NY, USA, 2018. ISBN 978-1-4503-5559-9. URL http://doi.acm.org/10.1145/3188745.3188882.

18. **Li, S.** and **X. 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.*. 2018. URL http://papers.nips.cc/paper/8009-distributed-k-clustering-for-data-with-heavy-noise.

19. **Makarychev, K.**, **Y. Makarychev**, and **A. Vijayaraghavan**, Correlation clustering with noisy partial information. *In* **P. Grünwald**, **E. Hazan**, and **S. Kale** (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*. PMLR, Paris, France, 2015. URL http://proceedings.mlr.press/v40/Makarychev15.html.

20. **McCallum, A.** and **B. 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. AAAI Press, 2003. URL http://dl.acm.org/citation.cfm?id=3104278.3104294.

21. **Rujeerapaiboon, N.**, **K. Schindler**, **D. Kuhn**, and **W. Wiesemann** (2019). Size matters: Cardinality-constrained clustering and outlier detection via conic optimization. *SIAM Journal on Optimization*.

22. **Swamy, C.**, Correlation clustering: Maximizing agreements via semidefinite programming. volume 15. 2004.

23. **Wirth, A.**, Correlation clustering. *In Encyclopedia of Machine Learning*. 2010, 227–231. URL https://doi.org/10.1007/978-0-387-30164-8_176.

CHAPTER 6

APPENDIX

6.1 Proof of Theorem 1.1.1

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 (1.0.1). 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 6.1.1 and 6.1.2.

Lemma 6.1.1. *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:

$$
\begin{aligned}
|\widetilde{C} \cap C^*_{i_1}| \left(\sum_{i \neq i_1} |\widetilde{C} \cap C^*_i| \right) &\leq |X||\widetilde{C} \cap C^*_{i_1}| \\
\implies \sum_{i \neq i_1} |\widetilde{C} \cap C^*_i| &\leq |X| \qquad (6.1)
\end{aligned}
$$

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 6.1.2. *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:

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

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. □

6.2 ROBUST-CORRELATION-CLUSTERING on Complete Graphs

Proof of Lemma 5.1.3. Consider a stage of the algorithm when a vertex u gets chosen as a cluster center. Then the newly mis-classified edges fall into three categories: $(v, w) \in E_-$ with both (u, v) and (u, w) belonging to E_+, or $(v, w) \in E_+$ with $(u, v) \in E_+$ and $(u, w) \in E_-$, or $(v, w) \in E_+$ with $(u, w) \in E_+$ and $(u, v) \in E_-$. In all three cases we can associate the new mis-classified edge (v, w) with the unique bad triangle (u, v, w) which gets touched. □

6.2.1 Algorithms on Complete graphs

We now furnish the complete details of the proof of Theorem 1.1.4. Recall the algorithm definition RCCAlg from Chapter 5. It will also be useful to recall from Definition 5.1.2 the definition of *touched* triangles in the execution of ACNAlg (V, E_+, E_-). We will use A_t to represent the event that triangle t is touched in some iteration. Also recall the value p_t, the probability that a triangle $t \in \mathcal{B}$ is touched during execution. We stress that the criterion for a vertex u to be deleted in RCCAlg(V, E_+, E_-, m) is that,

$$\sum_{t \in \mathcal{B}_u} \Delta_t = \frac{\mathsf{Opt}}{m} \tag{6.2}$$

In other words, the vertex u is deleted if Δ_t summed over bad triangles $t = (u, v, w)$, conditioned on t being touched by v or w.

Lemma 6.2.1. *The cost of the clustering $\mathcal{C}$ returned by* RCCAlg(V, E_+, E_-, m) *accounted over $V \setminus V_r$ does not exceed $\sum_{t \in \mathcal{B}} \mathbb{1}(A_t)\mathbb{1}(\Delta_t = 1)$.*

The proof of this result follows from the following claim.

Claim 6.2.2. *Consider a bad triangle $t = (u_1, u_2, u_3) \in \mathcal{B}_{\text{touch}}$. Without loss of generality, let u_3 be the vertex chosen as cluster center in the iteration when t is touched during the execution of* ACNAlg (V, E_+, E_-). *If Δ_t is set as 0 in* RCCAlg(V, E_+, E_-, m), *then it necessarily means that at least one of u_1 and u_2 are deleted.*

Lemma 6.2.3. *The collection $\{w_t = \frac{p_t}{3}\mathbb{1}(\Delta_t = 1)\}$ satisfies equation* (5.2).

Proof. The proof of this statement follows by contradiction. Recall that a vertex v is deleted if the sum of Δ_t over all triangles $t \in \mathcal{B}_v$ equals $2\mathsf{Opt}/m$. Therefore, if both u_1 and u_2 are undeleted, for $i = 1, 2$,

$$\sum_{t' \in \mathcal{B}_{u_i}} w_{t'} < \frac{2\mathsf{Opt}}{m} \tag{6.3}$$

In addition, noting that $\mathcal{B}_{u_3}$ does not include the triangle t, we remark that the algorithm would increase Δ_t until the contraints in (6.3) become tight. This contradicts the initial assumption, $\Delta_t = 0$. □

Proof of Lemma 6.2.1. From Claim 6.2.2, it follows that every $t \in \mathcal{B}_{\text{touch}}$ such that $\Delta_t = 0$ does not add to the cost of $\mathsf{RCCAlg}(V, E_+, E_-, m)$. Therefore, $\sum_{t \in \mathcal{B}_{\text{touch}}} \mathbb{1}(\Delta_t = 1)$ is an upper bound to the cost accrued by $\mathsf{RCCAlg}(V, E_+, E_-, m)$. □

Lemma 6.2.4. *The collection* $\{w_t = \frac{p_t}{3}\mathbb{1}(\Delta_t = 1)\}$ *satisfies equation* (5.1).

Proof. This constraint is satisfied for free from Lemma 5.1.4. □

Proof of Lemma 6.2.3. Recall that a vertex u is deleted if the sum of Δ_t's over all triangles $t = (u, v, w)$ conditioned on t being touched by either of v or w, exceeds $2\mathsf{Opt}/m$. The probability that a triangle t is touched by either v or w is $2p_t/3$. Therefore, for every u,

$$\sum_{t : u \in t} \frac{2p_t}{3}\mathbb{1}(\Delta_t = 1) \leq \frac{2\mathsf{Opt}}{m}.$$

Dividing both sides by 2 concludes the proof. □

We next bound the cost of our solution.

Lemma 6.2.5. *The expected cost of the solution output by* $\mathsf{RCCAlg}(V, E_+, E_-, m))$ *is at most* $6\mathsf{Opt}$.

Proof. From Lemma 6.2.1, it follows that the expected cost of $\mathsf{RCCAlg}(V, E_+, E_-, m))$, denoted alg does not exceed $\sum_{t \in \mathcal{B}} p_t \mathbb{1}(\Delta_t = 1)$. Recall that the dual solution we consider in LP3

is $w_t = (p_t/3)\mathbb{1}(\Delta_t = 1)$. By the duality of LP2, we get

$$\begin{aligned}\mathbb{E}[\mathsf{alg}] &\leq 3\sum_{t\in\mathcal{B}} \frac{p_t}{3}\mathbb{1}(\Delta_t = 1)\\ &\leq 3\mathsf{Opt}(LP2)\\ &\leq 6\mathsf{Opt}.\end{aligned}$$

□

Lemma 6.2.6. *The expected number of deleted vertices by* $\mathsf{RCCAlg}(V, E_+, E_-, m)$ *is* $\leq 6m$.

Proof. Recall for every vertex $u \in V$ which is deleted,

$$\sum_{t=(u,v,w)\in\mathcal{B}} \mathbb{1}(t \text{ is touched by } v \text{ or } w)\mathbb{1}(\Delta_t = 1) = \frac{2\mathsf{Opt}}{m} \tag{6.4}$$

Summing (6.4) over the set of all deleted vertices, V_r, it follows that

$$\sum_{u\in V_r}\sum_{t=(u,v,w)\in\mathcal{B}} \mathbb{1}(t \text{ is touched by } v \text{ or } w)\mathbb{1}(\Delta_t = 1) = \frac{2\mathsf{Opt}}{m}|V_r|.$$

Observe that for any touched triangle t, $\mathbb{1}(\Delta_t = 1)$ appears at most twice upon expanding the LHS double summation. This is because, corresponding to the vertex $u \in t$ which is chosen as the cluster center $\mathbb{1}(t$ is touched by v or $w)$ would be 0. Therefore,

$$2\sum_{t\in\mathcal{B}} \mathbb{1}(A_t)\mathbb{1}(\Delta_t = 1) \geq \frac{2\mathsf{Opt}}{m}|V_r|. \tag{6.5}$$

Taking expectation on both sides of (6.5), it follows that,

$$\begin{aligned}\frac{\mathsf{Opt}}{m}\mathbb{E}[|V_r|] &\leq 3\sum_{t\in\mathcal{B}} \frac{p_t}{3}\mathbb{1}(\Delta_t = 1)\\ &\leq 3\mathsf{Opt}(LP2).\end{aligned} \tag{6.6}$$

where the last inequality follows by the duality of LP2, and noting that $\{(p_t/3)\mathbb{1}(\Delta_t = 1)\}$ is a feasible solution to LP3 as established in Lemmas 6.2.3 and 6.2.4. The proof concludes by noting from Lemma 5.3.1 that $\mathsf{Opt}(LP2) \leq 2\mathsf{Opt}$ and simplifying (6.6). □

6.3 ROBUST-CORRELATION-CLUSTERING on General Graphs

In this section, we prove theorem 1.1.5. The starting point for our algorithm is the following LP relaxation for ROBUST-CORRELATION-CLUSTERING:

$$\text{Minimize} \quad \sum_{(u,v)\in E_+\cup E_-} z_{u,v}, \qquad s.t., \tag{LP4}$$

$$x_{u,v} + x_{v,w} \geq x_{u,w}, \qquad \forall u \neq v \neq w \tag{6.7}$$

$$y_u + y_v + z_{u,v} \geq 1 - x_{u,v}, \qquad \forall (u,v) \in E_- \tag{6.8}$$

$$y_u + y_v + z_{u,v} \geq x_{u,v}, \qquad \forall (u,v) \in E_+ \tag{6.9}$$

$$\sum_u y_u \leq m, \tag{6.10}$$

$$x_{u,v}, \in [0,1], \qquad \forall u \neq v$$

$$z_{u,v} \in [0,1], \qquad \forall (u,v) \in E_+ \cup E_-$$

$$y_u \in [0,1], \qquad \forall u \in V$$

In simple terms, on imposing integer constraints, LP4 asks to find a clustering $\{x_{u,v} : (u,v) \in E_+ \cup E_-\}$, but only 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, the metric constraint in (6.7) is present to ensure that any integer solution to LP4 corresponds to a consistent clustering.

Lemma 6.3.1. *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 eq.* (6.10)*, 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 tech-

niques 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 Demaine *et al.* (2006)), 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 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 6.3.2 (Fakcharoenphol *et al.* (2004))**.** *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 $\rho \in (0, \Delta/8)$,*

$$\text{Prob}(\text{Ball}_\rho(x) \nsubseteq \mathcal{C}(x)) \leq \alpha(x)\frac{\rho}{\Delta}, \tag{6.11}$$

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 set of points in the same cluster as x in $\mathcal{C}$.

6.3.1 Algorithms on General Graphs

Given a clustering $\mathcal{C}$, recall that $\mathcal{C}(v)$ denotes the set of points in the same cluster as v.

Theorem 6.3.3. 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.*

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

1: Let $\{x^*, y^*, z^*\}$ denote the (perturbed) optimal solution to LP4 obtained in Lemma 5.3.1
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 $v \in V_b^+$, define

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

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 $\mathcal{C}_{\mathsf{alg}} = \mathcal{C}^* \setminus D$

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

$$E_b^+ \stackrel{\text{def}}{=} \{(u, v) \in E_+ \ : \ u \in V' \setminus \mathcal{C}^*(v)\}.$$

Let $\mathsf{alg}_{\mathsf{cost}}$ 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 $\mathsf{alg}_{\mathsf{cost}}$ belongs to either E_b^+ or E_b^- and is not incident on any vertex in V_{del}. Therefore, $\mathsf{alg}_{\mathsf{cost}}$ can be decomposed as

$$\mathsf{alg}_{\mathsf{cost}} \leq \mathsf{alg}_{\mathsf{cost}}^- + \mathsf{alg}_{\mathsf{cost}}^+. \tag{6.12}$$

where $\mathsf{alg}_{\mathsf{cost}}^-$ denotes the cost associated with edges in E_b^- that are not incident on vertices in

V_{del}^-, and $\mathsf{alg}_{\mathsf{cost}}^+$ 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 LP4. Subsequently, in Lemmas 6.3.6 and 6.3.11 respectively, we show that $\mathsf{alg}_{\mathsf{cost}}^-$ is upper-bounded by $4\mathsf{Opt}^*$, while $\mathbb{E}\left[\mathsf{alg}_{\mathsf{cost}}^+\right]$ is upper-bounded by $O(\log n)\mathsf{Opt}^*$.

On the other hand, to bound the number of vertices deleted by $\mathsf{RCC\text{-}general}(V, E_+, E_-, m)$, we follow a similar strategy. Since,

$$|V_{\text{del}}| = |V_{\text{del}}^+| + |V_{\text{del}}^-|, \tag{6.13}$$

we separately upper bound V_{del}^- and $\mathbb{E}[V_{\text{del}}^+]$ in Lemmas 6.3.5 and 6.3.10 by $4m$ and $\mathcal{O}(\log^2 n)m$ respectively. In conjunction with (6.13), this proves that $\mathsf{RCC\text{-}general}(V, E_+, E_-, m)$ does not exceed the budget of the number of vertices to delete by more than a factor of $\mathcal{O}(\log^2 n)$. □

Recall that the optimal solution of LP4 is denoted $\left(\{y_u^* : u \in V\}, \left\{x_{u,v}^* : (u,v) \in \binom{V}{2}\right\}, \{z_{u,v}^* : (u,v) \in \binom{V}{2}\right.$ We begin by establishing some basic properties of the clustering $\mathcal{C}^*$.

Claim 6.3.4. *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 LP4 must satisfy the negative edge-constraint (6.8) for edge (u, v),

$$y_u^* + y_v^* + z_{u,v}^* \geq 1 - x_{u,v}^*. \tag{6.14}$$

From Theorem 6.3.2, the diameter of any cluster in $\mathsf{PaddedClustering}(X, d, \Delta)$ is at most Δ. Since u and v belong to the same cluster in $\mathcal{C}^* = \mathsf{PaddedClustering}(V, x^*, 0.25)$, it follows that $x_{u,v}^* \leq 0.25$. Substituting this into (6.14) completes the proof. □

Lemma 6.3.5. *The set of vertices, V^-_{del} satisfies,*

$$|V^-_{\text{del}}| \leq 4 \sum_{v \in V} y^*_v \leq 4m, \tag{6.15}$$

Proof. Recall that V^-_{del} is defined as the set of vertices, $v \in V^-_b$ such that $y^*_v \geq 1/4$. Therefore, $|V^-_{\text{del}}| = \sum_{v \in V^-_b} \mathbb{1}(y^*_u \geq 1/4)$. The proof concludes using the fact that $\mathbb{1}(y^*_u \geq 1/4) \leq 4y^*_u$ and relaxing the summation $v \in V^-_b$ to $v \in V$.

Lemma 6.3.6.

$$\mathsf{alg}^-_{\mathsf{cost}} \leq 4 \sum_{(u,v) \in E^-} z^*_{u,v}$$

Recall that a vertex v belongs to V^-_{del} only if $y^*_v \geq 1/4$. Since $\mathsf{alg}^-_{\mathsf{cost}}$ accrues unit cost for every edge in E^-_b which is not incident on a vertex in V^-_{del}, we have that,

$$\mathsf{alg}^-_{\mathsf{cost}} = \sum_{(u,v) \in E^-_b} \mathbb{1}(y^*_u \leq 1/4,\ y^*_v \leq 1/4).$$

From Claim 6.3.4, it follows that for any edge $(u, v) \in E^+_b$, if $y^*_u \leq 1/4$ and $y^*_v \leq 1/4$, then $z^*_{u,v}$ must be at least $1/4$. Therefore,

$$\mathsf{alg}^-_{\mathsf{cost}} \leq \sum_{(u,v) \in E^-_b} \mathbb{1}(z^*_{u,v} \geq 1/4). \tag{6.16}$$

Thereafter, by simplifying $\mathbb{1}(z^*_{u,v} \geq 1/4) \leq 4z^*_{u,v}$, it follows from (6.16) that,

$$\mathsf{alg}^-_{\mathsf{cost}} \leq 4 \sum_{(u,v) \in E^-_b} z^*_{u,v} \leq 4 \sum_{(u,v) \in E_-} z^*_{u,v}.$$

□

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

$$\hat{z}_{u,v} \stackrel{\text{def}}{=} \begin{cases} \frac{z^*_{u,v}}{x^*_{u,v}}, & v \notin \mathcal{C}^*(u), \\ 0 & \text{otherwise.} \end{cases} \tag{6.17}$$

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

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

Claim 6.3.7. *For any edge $(u,v) \in E^+_b$,*

$$\mathbb{E}[\hat{z}_{u,v}] \leq \mathcal{O}(\log n) z^*_{u,v}$$

Proof. Observe that if two points belong to different clusters, then we must necessarily have for $\rho = x^*_{u,v}$ that $\mathrm{Ball}_\rho(u) \nsubseteq \mathcal{C}(u))$. Therefore, from Theorem 6.3.2,

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

Therefore, from the definition of $\hat{z}_{u,v}$ in (6.17), it follows that,

$$\begin{aligned}\mathbb{E}[\hat{z}_{u,v}] &\leq \mathcal{O}(\log n) \frac{x^*_{u,v}}{0.25} \frac{z^*_{u,v}}{x^*_{u,v}} + 0 \\ &= \mathcal{O}(\log n) z^*_{u,v}.\end{aligned}$$

□

Claim 6.3.8. *For any vertex $v \in V^-_b$,*

$$\mathbb{E}\left[\hat{y}_u\right] \leq \mathcal{O}(\log^2 n) \cdot y^*_u \,.$$

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 $\hat{y}_u$,

$$\begin{aligned}\mathbb{E}[\hat{y}_u] &= \sum_{r=0}^{2\log n} 2^r \left(y^*_u\right) \mathrm{Prob}\left(\frac{1}{2^r} < \min_{v \in V \setminus \mathcal{C}^*(u)} x^*_{u,v} \leq \frac{1}{2^{r-1}}\right), \\ &\leq \sum_{r=0}^{2\log n} 2^r \left(y^*_u\right) \mathrm{Prob}\left(\min_{v \in V \setminus \mathcal{C}^*(u)} x^*_{u,v} \leq \frac{1}{2^{r-1}}\right). \end{aligned} \tag{6.18}$$

Observe that the event $\min_{v \in V \setminus \mathcal{C}^*(v)} x^*_{u,v} \leq 2^{-(r-1)}$ can only occur if the ball of radius $2^{-(r-1)}$

centered at u lies entirely within $\mathcal{C}(u)$. Therefore, from Theorem 6.3.2, we have that,

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

Plugging this into (6.18) gives,

$$\begin{aligned}\mathbb{E}[\hat{y}_u] &\leq \mathcal{O}(\log n) \sum_{r=0}^{2\log n} y^*_u, \\ &= \mathcal{O}(\log^2 n) \cdot y^*_u.\end{aligned}$$

□

Claim 6.3.9. *For any edge $(u, v) \in E^+_b$, we have that*

$$\hat{y}_u + \hat{y}_v + \hat{z}_{u,v} \geq 1 \tag{6.19}$$

Proof. Observe that E^+_b is a subset of E_+. Therefore every $(u, v) \in E^+_b$ must satisfy the positive edge-constraint (6.9) $y^*_u + y^*_v + z^*_{u,v} \geq x^*_{u,v}$. Dividing both sides by $x^*_{u,v}$, the proof concludes by using the definitions of $\hat{y}_u$ and $\hat{z}_{u,v}$. □

Lemma 6.3.10. *The set of vertices, V^+_{del} satisfies,*

$$\mathbb{E}\left[|V^+_{\text{del}}|\right] \leq \mathcal{O}(\log^2 n)\, m.$$

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

$$|V^+_{\text{del}}| \leq 3 \sum_{v \in V^+_b} \hat{y}_v \tag{6.20}$$

Taking expectation on both sides of (6.20), and using Claim 6.3.8,

$$\mathbb{E}[|V^+_{\text{del}}|] \leq \mathcal{O}(\log^2 n) \sum_{v \in V^+_b} y^*_v.$$

The proof concludes by relaxing the summation $v \in V^+_b$ to $v \in V$, and using Lemma 6.3.1 to

claim that $\sum_{v \in V} y^*_v \leq m + \frac{1}{n} \leq 2m$. □

Lemma 6.3.11.

$$\mathbb{E}\left[\mathsf{alg}^+_{\mathsf{cost}}\right] \leq \mathcal{O}(\log n) \sum_{(u,v) \in E^+} z^*_{u,v}.$$

Proof. $\mathsf{alg}^+_{\mathsf{cost}}$ 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 $\hat{y}_v \geq 1/3$. Therefore,

$$\mathsf{alg}^+_{\mathsf{cost}} = \sum_{(u,v) \in E^+_b} \mathbb{1}(\hat{y}_u \leq 1/3, \hat{y}_v \leq 1/3). \tag{6.21}$$

From Claim 6.3.9, it follows that if both $\hat{y}_u$ and $\hat{y}_v$ are at most $1/3$, then $\hat{z}_{u,v} \geq 1/3$. Therefore, from (6.21),

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

where inequality (i) uses the fact that $\mathbb{1}(\hat{z}_{u,v} \geq 1/3) \leq 3\hat{z}_{u,v}$. Taking expectations on both sides and using Claim 6.3.7 to upper bound $\mathbb{E}[\hat{z}_{u,v}]$ by $\mathcal{O}(\log n) z^*_{u,v}$,

$$\mathbb{E}[\mathsf{alg}^+_{\mathsf{cost}}] \leq \mathcal{O}(\log n) \sum_{(u,v) \in E^+_b} z^*_{u,v}.$$

Relaxing the summation to $(u, v) \in E_+$ concludes the proof. □

Having established these results, it is straightforward to show that $\mathsf{RCC\text{-}general}(V, E_+, E_-, m)$ is an $(\mathcal{O}(\log n), \mathcal{O}(\log^2 n))$ bi-criteria approximation for m-Robust Correlation Clustering on general graphs.

Lemma 6.3.12.

$$\mathsf{alg}_{\mathsf{cost}} \leq \mathcal{O}(\log n) \sum_{(u,v) \in E_+ \cup E_-} z^*_{u,v}.$$

Proof. The proof of this result is a direct consequence of substituting Lemmas 6.3.6 and 6.3.11 into (6.12). □

Lemma 6.3.13. *The expected number of vertices deleted by* $\mathsf{RCC\text{-}general}(V, E_+, E_-, m)$ *is* $\leq \mathcal{O}(\log^2 n)\ m$.

Proof. The proof of this result follows by using Lemmas 6.3.5 and 6.3.10 in conjunction with (6.13). □

6.4 Hardness of ROBUST-CORRELATION-CLUSTERING on complete graphs

In this section, we furnish the complete details of the proof of Theorem 1.1.3. The proof follows by an approximation preserving reduction from vertex cover. Consider any unlabelled graph, $G = (V, E)$ on n vertices. Let $\mathsf{vc}(G)$ denote the set of vertices corresponding to the minimum vertex cover on G. We construct the ROBUST-CORRELATION-CLUSTERING instance $\mathcal{I}_G$ from G as follows: for each vertex $v \in V$, we create two points v_1 and v_2. 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.

In Lemma 6.4.1, we show that the optimal clustering on instance $\mathcal{I}_G$ has 0 cost if and only if the budget of vertices to be deleted, m is at least $|\mathsf{vc}(G)|$. Given a graph G and some m, since it is NP-hard to decide if $m \geq b|\mathsf{vc}(G)|$ or $m \leq |\mathsf{vc}(G)|$ for $b < \alpha_{\mathrm{VC}}$, it therefore follows that it is NP-hard to decide if the optimal solution to $\mathcal{I}_G$ has 0 cost unless $b \geq \alpha_{\mathrm{VC}}$. By contradiction, this in turn rules out the existence of any efficient finite approximation factor algorithm to ROBUST-CORRELATION-CLUSTERING in the cost of the clustering when $b < \alpha_{\mathrm{VC}}$.

Lemma 6.4.1. *For* ROBUST-CORRELATION-CLUSTERING *on $\mathcal{I}_G$, the cost of the optimal solution is 0 if and only if $m \geq |\mathsf{vc}(G)|$.*

Proof of Lemma 6.4.1. The proof of this statement follows from Claims 6.4.2 and 6.4.3. □

Claim 6.4.2. *If $m \geq |\mathsf{vc}(G)|$ the optimal solution to $\mathcal{I}_G$ has 0 cost.*

Proof. In order to show this result, it suffices to construct an explicit clustering, $\mathcal{C}'$ such that the minimum number of vertices that need to be removed from $\mathcal{C}'$ to bring its cost to 0 is $|\mathsf{vc}(G)|$.

To this end, consider the clustering $\mathcal{C}'$ such that for each vertex $v \in V$, v_1 and v_2 together form an independent cluster. Observe that for this clustering, each $+$ edge separating two

vertices v_1 and u_1 in $\mathcal{I}_G$, the contributes a unit cost. Removing the set of vertices $\{v_1 \ : \ v \in \mathsf{vc}(G)\}$ from the instance $\mathcal{I}_G$ is guaranteed to reduce the cost of the clustering $\mathcal{C}$ to 0 since every edge that contributes a unit cost has at least one of its endpoints deleted. □

Claim 6.4.3. *If $m < |\mathsf{vc}(G)|$, the optimal solution to $\mathcal{I}_G$ has strictly positive cost.*

In order to prove this result, we make the following claim.

Claim 6.4.4. *For every pair of bad vertices (v_1, u_1), such that u_1 and v_1 are similar to each other, at least one vertex must be removed from either $\{u_1, u_2\}$ or $\{v_1, v_2\}$ for the cost of any clustering to be 0.*

Proof. For the cost of any clustering to be 0, the 4 vertices in $\{u_1, u_2, v_1, v_2\}$ must belong in the same cluster since the pairs (u_2, u_1), (u_1, v_1) and (v_1, v_2) are similar. However, even in this case u_2 and v_2 are dissimilar and a cost of at least 1 is incurred. □

The proof of Claim 6.4.3 is straightforward using this result. This is because each pair of points (u_1, v_1) that are similar, corresponds to an edge $(u, v) \in E$. So from Claim 6.4.4, for every such edge $(u, v) \in E$ at least one vertex from $\mathcal{I}_G$ must be deleted. Therefore, it is not possible to delete fewer than $|\mathsf{vc}(G)|$ vertices for there to exist a clustering of $\mathcal{I}_G$ with 0 cost.