
Hierarchical Clustering via Sketches and Hierarchical Correlation Clustering

Danny Vainstein
Tel Aviv University
Google Research, NY

Vaggos Chatziafratis
Google Research, NY

Gui Citovsky
Google Research, NY

Anand Rajagopalan
Google Research, NY

Mohammad Mahdian
Google Research, NY

Yossi Azar
Tel Aviv University

Abstract

Recently, Hierarchical Clustering (HC) has been considered through the lens of optimization. In particular, two maximization objectives have been defined. Moseley and Wang defined the *Revenue* objective to handle similarity information given by a weighted graph on the data points (w.l.o.g., $[0, 1]$ weights), while Cohen-Addad et al. defined the *Dissimilarity* objective to handle dissimilarity information. In this paper, we prove structural lemmas for both objectives allowing us to convert any HC tree to a tree with constant number of internal nodes while incurring an arbitrarily small loss in each objective. Although the best-known approximations are 0.585 and 0.667 respectively, using our lemmas we obtain approximations arbitrarily close to 1, if not all weights are small (i.e., there exist constants ϵ, δ such that the fraction of weights smaller than δ , is at most $1 - \epsilon$); such instances encompass many metric-based similarity instances, thereby improving upon prior work. Finally, we introduce Hierarchical Correlation Clustering (HCC) to handle instances that contain similarity and dissimilarity information simultaneously. For HCC, we provide an approximation of 0.4767 and for complementary similarity/dissimilarity weights (analogous to $+/-$ correlation clustering), we again present nearly-optimal approximations.

1 INTRODUCTION

Clustering is a fundamental problem in unsupervised learning and has been widely and intensively explored. Classically, one considers a set of data points (with some notion of either similarity or dissimilarity between every pair) and then partitions these data points into sets. In order to differentiate between different partitions, many classical *flat* clustering objectives have been introduced, such as k -means, k -median and k -center. However, what if one would like a more granular view of the clusters (specifically, to understand the relations between data points within a given cluster)?

To explore these questions, the notion of *Hierarchical Clustering* (HC) has been introduced. One way of studying this notion is through the lens of optimization. Dasgupta (2016) initiated this line of work, inspiring others to consider several different objectives. Two notable objectives that we will consider in our paper are the Revenue and Dissimilarity objectives.

The problem is defined as follows. We are given a set of data points with some notion of similarity (or dissimilarity) between every pair of points which is defined by a weighted graph, $G = (V, E, w)$ such that V is our set of data points, $|V| = n$ and $w : E \rightarrow \mathbb{R}_{\geq 0}$. We then define an HC tree as a rooted tree with leaves in bijective correspondence with the original data points. Intuitively, we would expect a "good" HC tree T to split more similar data points towards the leaves of the tree. When we are given similarity weights, this corresponds to larger weights. Thus, Moseley and Wang (2017) proposed to maximize the *Revenue* objective:

$$rev_G(T) = \sum_{i < j} w_{ij}(n - |T_{ij}|), \quad (\text{Rev-HC})$$

where T_{ij} is the subtree rooted at the lowest common ancestor (LCA) of i and j , and $|T_{ij}|$ denotes the number of leaves of T_{ij} for any binary tree T . The

second objective we consider was defined within the dissimilarity realm by Cohen-Addad et al. (2018). In this case, larger weights corresponds to dissimilar data points. Therefore, a (binary) tree T should be rewarded for splitting larger weights towards its root and thus their *Dissimilarity* objective is to maximize:

$$dis_G(T) = \sum_{i < j} w_{ij} |T_{ij}|. \quad (\text{Dis-HC})$$

Note that when considering both objectives, we may (and will) assume w.l.o.g. that $w_{ij} \in [0, 1]$.

Since the objectives have been introduced, there has been a line of work designing approximation algorithms. For the **Rev-HC** objective, the best approximation ratio is 0.585 (Alon et al., 2020), while for the **Dis-HC** the best ratio is 0.667 (Charikar et al., 2019a). In terms of hardness, both problems have been proven to be APX-hard (Ahmadian et al., 2019; Chatziafratis et al., 2020) and thus do not admit optimal or even arbitrarily close to optimal approximations. Given these results, it seems natural to ask whether this hardness is inherent in the objectives, or rather can be somehow circumvented. Towards that end, we consider the following question:

Is there a large class of interesting instances that can be shown to have significantly better approximations?

Surprisingly, we show that if we consider instances with weights that are not all small (see Definition 3) then the above holds true. First, we obtain approximations arbitrarily close to optimal (specifically, Efficient Polynomial Time Randomized Approximation Schemes (Efficient-PRAS)) for both **Rev-HC** and **Dis-HC** objectives. Interestingly, in order to do so we first consider a tree’s *sketch* (defined as the tree resulting from removing all its leaves (and corresponding edges)). Even though it is well known that the optimal trees for these settings are binary (and therefore contain $n - 1 = \Omega(n)$ nodes), we show that there exist trees with constant sized (i.e., a constant number of nodes and edges) sketch, for both objectives, that approximate the optimal values arbitrarily good. **We stress that this holds true for any HC instance, and not only if not all input weights are small.** We then leverage the seminal work of Goldreich et al. (1998) in order to obtain approximations arbitrarily close to optimal, if not all weights are small.

Second, we show that many interesting, and formerly researched problems, are encapsulated by these types of instances. Specifically, we show that a large family of metric-based similarity instances (as defined by Charikar et al. (2019b) - see Subsection 3.3) are such instances, and thus admit approximations arbitrarily close to optimal. We note that this partially answers an open question raised in their work of whether there

exist good approximation algorithms for low dimensions. We also note that our results immediately provide an Efficient-PRAS for similarity instances defined by a Gaussian Kernel in high dimensions when the minimal similarity is $\delta = \Omega(1)$ which was specifically considered by Charikar et al. (2019b); improving the approximation from $\frac{1+\delta}{3}$ to an approximation that is arbitrarily close to optimal. Finally, we show that these results also provide an approximation that is arbitrarily close to optimal, for the +/- Hierarchical Correlation Clustering problem (defined next).

Up until now we have only considered instances handling either similarity or dissimilarity information, *but not both*. In many scenarios, however, both types of information are accessible simultaneously. These scenarios have been tackled within the realm of correlation clustering both in theory (e.g., Bansal et al. (2002); Swamy (2004); Charikar et al. (2005); Ailon et al. (2008); Chawla et al. (2015)) and in practice (e.g., Bonchi et al. (2014); Cohen and Richman (2001)). However, this line of work has been centered around flat clustering. With that in mind, it is natural to ask:

In presence of mixed information, how can we extend the notion of Correlation Clustering to hierarchies?

In order to answer the question, we introduce the Hierarchical Correlation Clustering objective. The objective interpolates naturally between the **Rev-HC** and **Dis-HC** objectives. Again, we are given a set of data points; however, in this case every pair of data points i and j are given a similarity weight w_{ij}^s and a dissimilarity weight w_{ij}^d . The objective is then defined as,

$$hcc_G(T) = \sum_{i < j} w_{ij}^s (n - |T_{ij}|) + \sum_{i < j} w_{ij}^d |T_{ij}|. \quad (\text{HCC})$$

Observe that this objective is a direct generalization of the **Rev-HC** and **Dis-HC** objectives simply by letting either $w_{ij}^d = 0$ or $w_{ij}^s = 0$ respectively. Moreover, it captures the fact that similar points (i.e., large w_{ij}^s) should be separated towards the tree’s leaves (yielding a large $n - |T_{ij}|$ coefficient), whereas dissimilar points (i.e., large w_{ij}^d) should be split towards the tree’s root (yielding a large $|T_{ij}|$ coefficient).

Finally, we consider the +/- variant of correlation clustering (Bansal et al., 2002) extended to hierarchies as well. We define this objective as the **HCC** objective reduced to instances that guarantee $w_{ij}^s = 1 - w_{ij}^d$ for all data points i and j . We will refer to this objective as the **HCC $^\pm$** objective. This may be motivated by the following folklore example: assume one is given a document classifier f that returns a confidence level in $[0, 1]$ corresponding to how certain it is that two documents are similar. Thus, 1 minus the confidence level

may be seen as how confident the classifier is that the two documents are dissimilar. For further comments regarding our formulation and how it is related to the correlation clustering objectives of [Bansal et al. \(2002\)](#) and of [Swamy \(2004\)](#), see Section 5.

Contributions of this paper. With respect to the [Rev-HC](#) and [Dis-HC](#) objectives:

- We present structural lemmas for the revenue and dissimilarity settings that provide a way of converting optimal trees in both settings such that the resulting trees (1) are of constant sketch size and (2) approximate the respective objectives arbitrarily close (see [Figure 1](#) for an example). Note that this result holds for **any** similarity/dissimilarity input graphs.
- We use the resulting trees in order to obtain Efficient-PRAS’s for revenue or dissimilarity instances with not all small weights (see [Definition 3](#)). We note that this includes an Efficient-PRAS for any similarity Gaussian Kernel based instances with minimal weight $\delta = \Omega(1)$ (specifically considered by [Charikar et al. \(2019b\)](#)).
- We show that many metric-based similarity instances in fact do not have all small weights, thus admitting Efficient-PRAS’s. We note that this partially solves the case where the metric’s dimension is constant (raised in [Charikar et al. \(2019b\)](#)).

With respect to the [HCC](#) objective:

- We present a 0.4767 approximation for the [HCC](#) objective by extending the proof of [Alon et al. \(2020\)](#) to include dissimilarity weights.
- We combine our Revenue and Dissimilarity algorithms to produce an Efficient-PRAS for the [HCC \$\pm\$](#) objective.

Finally, we provide hardness results in all (three) settings in [Section 6](#).

Techniques. In order to reduce HC trees to trees with constant sketch that approximate the [Rev-HC](#) and [Dis-HC](#) objectives arbitrarily closely, we use the following techniques. For both objectives the first step is to consider an optimal solution, T , and contract it (i.e., contract some subgraphs of T into single nodes) into an intermediate tree denoted as $K(T)$. Briefly, $K(T)$ is generated by recursively finding a constant-sized set of edges whose removal creates a set of trees, each containing a small and roughly equal number of data points. Thereafter, each such tree is contracted (within T) to a single node. This results in $K(T)$ that guarantees that (1) it contains a constant number of nodes and (2) its structure resembles that of T

which allows us to easily convert it to the final revenue/dissimilarity tree. Note that during this process of contraction, some data points may have been contracted as well (see [Figure 2](#)). Next we describe, at a high level, how to convert $K(T)$ to a proper revenue/dissimilarity tree.

Revenue setting. In the revenue setting we convert $K(T)$ to a tree denoted by T^R , such that T^R has a constant-sized sketch and approximates the revenue gained by T up to an arbitrarily small constant factor. In order to do so we replace each contracted node in $K(T)$ with a “star” structure (which is an auxiliary node with the contracted data points connected as its children) - see [Figure 3](#). Note that there is a trade-off between T^R ’s internal tree size and the revenue approximation factor guaranteed (see [Section 3](#) for formal details).

Dissimilarity setting. In the dissimilarity setting we convert $K(T)$ to a tree denoted by T^D such that T^D has a constant-sized sketch and approximates the dissimilarity gained by T up to an arbitrarily small constant factor. Instead of replacing the contracted node with a “star” structure as in the revenue case, we replace it with a random “comb” structure (formally defined in [Section 4](#) and depicted in [Figure 3](#)). Also here, there exists a trade-off between T^D ’s size and the approximation factor.

Related Work. HC has been extensively studied and therefore many variations have been considered (for a survey on the subject, see [Berkhin \(2006\)](#)). The work on HC trees began within the realm of phylogenetics ([Sneath and Sokal, 1962](#); [Jardine and Sibson, 1968](#)) but has since then expanded to many other domains (e.g., genetics, data analysis and text analysis - [Alon et al. \(1999\)](#); [Brown et al. \(1992\)](#); [Seo and Shneiderman \(2002\)](#)).

As stated earlier, Dasgupta elegantly linked the fields of approximation algorithms and HC trees, thereby initiating this line of work. Formally, given an HC tree, T , [Dasgupta \(2016\)](#) considered the problem of minimizing its cost, $cost_G(T) = \sum w_{ij}|T_{ij}|$. In his work, Dasgupta showed that recursively finding a sparsest cut results in a $O(\log^{1.5} n)$ approximation. This analysis was later improved to $O(\sqrt{\log n})$ ([Charikar and Chatziafratis, 2017](#); [Cohen-Addad et al., 2018](#)). [Charikar and Chatziafratis \(2017\)](#) also showed that no constant approximation exists (assuming the Small Set Expansion hypothesis).

Later, [Moseley and Wang \(2017\)](#) considered the [Rev-HC](#) objective. [Charikar et al. \(2019a\)](#) showed a 0.3364 approximation through the use of semi-definite programming. Later, [Ahmadian et al. \(2019\)](#) made use of the MAX-UNCUT BISECTION problem in order to prove a 0.4246 approximation. Finally, [Alon et al.](#)

(2020) improved upon this by showing a 0.585 approximation, by proving the existence of a bisection which yields large revenue.

Cohen-Addad et al. (2018) considered the **Dis-HC** objective (defined earlier). In their work they showed that the Average-Linkage algorithm is a $\frac{1}{2}$ approximation and then improved upon this by presenting a simple algorithm achieving a $\frac{2}{3}$ approximation. Charikar et al. (2019a) then showed a further improvement by presenting a more intricate algorithm that achieves a 0.6671 approximation.

Since the work of Bansal et al. (2002), correlation clustering has been extensively studied. Considering more theoretical settings, the work most relevant to ours is that of Swamy (2004), showing a 0.766-approximation for a maximization version of the problem, interpolating between roundings from multiple hyperplanes, instead of just one as in Goemans and Williamson (1995). The problem is also highly significant in practice as well - see e.g., spam filtering (Rama-chandran et al., 2007), image segmentation (Kim et al., 2011) and co-reference resolution (Cohen and Richman, 2002; Elmagarmid et al., 2006).

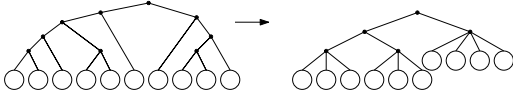


Figure 1: Converting an HC tree to a tree of constant Sketch while approximating the goal function.

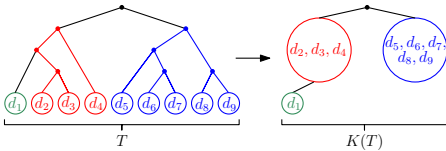


Figure 2: Converting an HC tree T to $K(T)$.

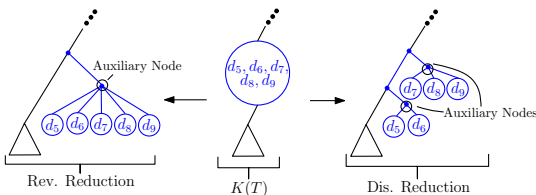


Figure 3: Converting $K(T)$ to an HC tree for each goal function.

2 PRELIMINARIES

We first consider several graph-specific definitions.

Definition 1. Given a tree T and a set of edges $F \subset E(T)$, let $T - F$ denote the set of trees that results from

removing F from $E(T)$. Furthermore, given a set of nodes $U \subset V(T)$, let $T - U$ denote the set of trees that results from removing U (and any edge that has a node in U) from T .

Definition 2. Given a graph G and a subset of edges $U \subset V(G)$ we define the contraction of U as the replacement of U within G with a single node attached to all edges which were formerly attached to U .

As pointed out by Charikar et al. (2019a), the average-linkage algorithm generates $\frac{(n-2)}{3} \sum w_{ij}$ revenue and $\frac{2(n-2)}{3} \sum w_{ij}$ dissimilarity, yielding the following facts:

Fact 1. $rev(T^O) \geq \frac{(n-2)}{3} \sum_{i < j} w_{ij}$, where T^O denotes the optimal revenue tree.

Fact 2. $dis(T^O) \geq \frac{2n}{3} \sum_{i < j} w_{ij}$, where T^O denotes the optimal dissimilarity tree.

Furthermore, as pointed out by Dasgupta (2016) all binary trees generate the same dissimilarity on instances defined by cliques (i.e., $w_{ij} = 1$ for all i and j).

Fact 3. $\sum_{i,j} |T_{ij}| = \frac{2n}{3} \binom{n}{2}$.

A note on non-binary HC trees. Even though the **Rev-HC** and **Dis-HC** objectives are defined for binary trees, we make use of star structures. A star structure is simply a node that contains more than two data points as children (and therefore leaves). We use these star structures as a proxy for any binary tree containing the same set of data points. More formally, by replacing the star structure (within some larger tree) with any binary tree containing the same set of data points and then rooting it in the same place within the original tree, the goal function would only increase.

In the revenue case this follows immediately. In the dissimilarity case, however, by following the definition of T_{ij} plainly, clearly attaching all data points to a single root results in an optimal tree. Therefore, we instead extend the dissimilarity definition to non-binary trees as follows. Given an HC tree T and internal node v , let $|T_v|$ denote the set of data points contained within the subtree rooted at v (in particular, for any 2 data points i and j , $|T_{ij}| = |T_{lca(i,j)}|$). We then define the dissimilarity as

$$dis_G(T) = \sum w_{ij} (|T_{v_i}| + |T_{v_j}|),$$

where v_i and v_j denote $lca(i, j)$'s children containing i and j in their subtree. We emphasize the fact that for binary HC trees, this definition coincides with the classic dissimilarity (since $|T_{v_i}| + |T_{v_j}| = |T_{ij}|$). Clearly any non-binary node may be replaced with a binary subgraph within the HC tree thereby only increasing the dissimilarity generated. Therefore, any of our algorithmic results apply to the binary setting (by performing

these replacements). Further, all of our approximation results are with respect to optimal binary trees and thus directly apply to the binary setting.

Finally, we will use the following definitions throughout the paper. (Recall that w.l.o.g. we may assume that all weights are in $[0, 1]$).

Definition 3. An HC instance is said to have not all small weights if there exist constants (with respect to $|V|$) ρ, τ such that the fraction of weights smaller than τ , is at most $1 - \rho$.

Definition 4. An algorithm is considered an Efficient-PRAS if for any $\epsilon > 0$ the algorithm runs in time $f(1/\epsilon)n^{O(1)}$ and approximates the optimal solution's value up to a factor of $1 - \epsilon$ with high probability.

3 THE REVENUE CASE

In this section we consider the **Rev-HC** objective. In Subsection 3.1 we show how to create a tree with constant sized sketch which approximates the optimal revenue tree up to an arbitrarily small factor (for an overview see Techniques). Note that this result holds for **any** revenue instance and thus may be of independent interest. We then leverage this and in Subsection 3.2 we present an Efficient-PRAS for instances with not all small weights. Finally, in Subsection 3.3 we show that a large family of metric-based similarity instances have weights that are not all small - thereby admitting Efficient-PRAS's. We note that this partially solves an open question raised by Charikar et al. (2019b) regarding constant dimension instances and immediately provides Efficient-PRAS's for similarity instances defined by a Gaussian Kernel in high dimensions when the minimal similarity is $\delta = \Omega(1)$ which was specifically in their work as well.

3.1 A Reduction to Constant Sketches

We begin by first proving the existence of a tree with constant-sized sketch that approximates the optimal tree arbitrarily well.

Theorem 1. Let T^O denote the optimal revenue tree and assume it contains n leaves (i.e., data points). Then, for any $\epsilon > 0$, there exists a tree T^R such that (i) T^R contains $\Theta(1/\epsilon)$ internal nodes each with at most $3\epsilon n$ children, and (ii) $rev(T^R) \geq (1 - 19\epsilon)rev(T^O)$.

In order to construct T^R we use a two step process: we first create an intermediate tree, denoted as $K(T)$ (to be defined) and then convert that to our final tree. In fact, this process may be applied to any binary tree T (in particular, we will apply it to T^O). Before we can define the process that generates $K(T^O)$, we must first present several definitions and lemmas, the first of which was shown by Dasgupta (2016) (this was not

explicitly proven, and therefore we add the proof in the Appendix for completeness).

Lemma 1. Given a rooted binary tree T with n data points as leaves, there exists an edge whose removal creates two binary trees each with at least $\frac{n}{3}$ data points (and therefore at most $\frac{2n}{3}$). Furthermore this edge can be found in polytime.

Lemma 2. Given a rooted binary tree T with n data points, there exists a set of edges F such that $\frac{1}{3\epsilon} \leq |F| + 1 \leq \frac{1}{\epsilon}$ and the number of data points in each tree of $T - F$ is at least ϵn and at most $3\epsilon n$. Furthermore F can be found in polytime.

Proof sketch (full proof in Appendix). Apply Lemma 1 recursively and stop once each tree in $T - F$ contains $\leq 3\epsilon n$ data points. \square

The following is a straightforward but useful lemma. (See Appendix for proof).

Lemma 3. For an arbitrary tree T , let V_3 denote the set of vertices with degree ≥ 3 and L denote its set of leaves. Then, $|V_3| \leq |L| - 1$.

Definition 5. Given F as defined by Lemma 2 we define two sets of nodes: blue and green, denoted by \mathbf{B} and \mathbf{G} . A **blue node** is any node connected to any edge of F or that is T 's root. A **green node** is any node that is not blue and that has two children, each of which contains a blue node as its descendant.

Next we define the process that given a binary tree, contracts it compactly. Given an input T , we denote the process' output as $K(T)$, formally defined by Algorithm 1. (See Figure 2 for a pictorial example). We note that each contracted node might have originally contained data points. We therefore associate every contracted node, c with its set of data points, D_c . Finally, we define the process that given any binary tree T , outputs T^R - formally defined by Algorithm 2.

Algorithm 1 Algorithm to convert T to $K(T)$.

```

Obtain  $F$  as described in Lemma 2.
Color the nodes green or blue as in Definition 5.
for every tree  $T_i$  in  $T - (B \cup G)$  do
    Contract  $T_i$ .
Return the resulting tree as  $K(T)$ .

```

Algorithm 2 Algorithm to convert T to T^R .

```

 $K(T) \leftarrow$  Algorithm 1 applied to  $T$ .
for each node  $c \in K(T)$  and its set of data points  $D_c$  do
    Attach a (new) auxiliary node as  $c$ 's child (in  $K(T)$ ).
    Attach  $D_c$  as the auxiliary node's children.
Return the resulting tree as  $T^R$ .

```

Remark 1. We note that T^R remains binary (except the auxiliary nodes). This is in fact true since otherwise this internal node would have contained at least 2 children which are colored green/blue (since it may only have a single auxiliary node). Thus, there would have been a green node contained within this contracted component in contradiction to the definition of $K(T)$.

In what follows we show that for any binary tree T , (1) T^R has a constant sketch and (2) $|T_{ij}^R|$ is (approximately) upper bounded for any data points i and j (which in turn guarantees that $rev(T^R)$ is close to T^O when $T = T^O$). Due to lack of space we defer the proof of Lemma 4 to the Appendix.

Lemma 4. T^R contains $\Theta(1/\epsilon)$ internal nodes each with at most $3\epsilon n$ children.

Lemma 5. For any two data points i and j , $|T_{ij}^R| \leq |T_{ij}| + 6\epsilon n$.

Proof. Consider any three data points in T , i, j and k , such that $k \notin T_{ij}$. We will show that $k \notin T_{ij}^R$ for all but $6\epsilon n$ such k 's. In order to prove our lemma we first introduce the following notations. First, for any node u we denote the set of data points contained in its induced subtree as $L(u)$. Secondly we note that any node colored green or blue in T will not be contracted and therefore will appear in $V(T^R)$. Finally, we observe the following given our contraction process.

Observation 1. Let $v \in V(T)$ denote a child of a green/blue node and let $v^* \in V(T^R)$ denote the node that contracted v in T^R . Therefore, $L(v) = L(v^*)$.

Observation 2. Data points i and j appear under the same auxiliary node in T^R if and only if i and j were contained in the same tree of $T - (B \cup G)$.

Recall that our goal is to show that if $k \notin T_{ij}$ then $k \notin T_{ij}^R$. Towards that end, denote by v_{ij} (resp. v_{ik} and v_{jk}) i and j 's LCA in T . Therefore, $v_{ik} = v_{jk}$ and v_{ij} is a descendant of v_{ik} . Furthermore, let $\{T_\ell^{B \cup G}\}$ denote the set of trees defined by $T - (B \cup G)$ and let $T_i^{B \cup G}$ (resp. $T_j^{B \cup G}$ and $T_k^{B \cup G}$) denote the tree in $T - (B \cup G)$ containing i (resp. j and k).

We first assume $k \notin T_i^{B \cup G}$ and $k \notin T_j^{B \cup G}$. Therefore, a green or blue node must be either on the path $k \rightarrow v_{ik}$, or on the path $v_{ij} \rightarrow v_{ik}$. Otherwise there must be a green or blue node on the path $i \rightarrow v_{ij}$ and on the path $j \rightarrow v_{ij}$. We consider each case separately. (See Figure 4).

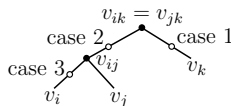


Figure 4: Explanation to proof of Lemma 5 (such that $v_a = a$ for $a \in \{i, j, k\}$).

Case 1. There exists a blue or green node on the path $k \rightarrow v_{ij}$: We further split this case into two cases. The first is that i and j are part of the same tree of $T - (B \cup G)$. In this case they will end up under the same auxiliary node and due to Observation 2 we are guaranteed that $k \notin T_{ij}^R$. The second case is that i and j are not part of the same tree and therefore there exists a blue/green node on the path $i \rightarrow j$. Thus, the node v_{ik} must be green or blue and due to Observation 1, i and j 's lca will remain lower than i and k 's in T^R . Therefore, $k \notin T_{ij}^R$.

Case 2. There exists a blue or green node on the path $v_{ij} \rightarrow v_{ik}$: In this case either v_{ik} is green/blue and due to Observation 1 we are done. Otherwise some other node along $v_{ij} \rightarrow v_{ik}$ is green/blue and then Observation 1 guarantees that k will not enter the subtree defined by i and j 's lca. Thus, in any case, $k \notin T_{ij}^R$.

Case 3. There exists a green or blue node on the paths $i \rightarrow v_{ij}$ and $j \rightarrow v_{ij}$: If v_{ij} is green/blue then Observation 1 guarantees that k will not enter the subtree defined by i and j 's lca. Otherwise, we are guaranteed to have two separate green/blue nodes, one on the path $i \rightarrow v_{ij}$ and one on the path $j \rightarrow v_{ij}$. Therefore, v_{ij} must be green/blue. Hence, in either case, $k \notin T_{ij}^R$.

Thus, we have shown that in all 3 cases if $k \notin T_i^{B \cup G}$ and $k \notin T_j^{B \cup G}$ then $k \notin T_{ij}^R$. Since the number of data points within both $T_i^{B \cup G}$ and $T_j^{B \cup G}$ is at most $3\epsilon n$ each, we get that at most $6\epsilon n$ such k 's may be contained in T_{ij}^R . Therefore, $|T_{ij}^R| \leq |T_{ij}| + 6\epsilon n$, concluding the proof. \square

Finally, combining Lemmas 4 and 5 for $T = T^O$ (i.e., the revenue optimal solution) with Fact 1, is enough to prove Theorem 1.

Proof of Theorem 1. Lemma 4 is enough to prove the first bullet. We consider the second bullet. It is a known fact that T^O may be taken to be binary. Therefore, due to Lemma 5 and Fact 1, we get,

$$\begin{aligned} rev(T^R) &= \sum_{i < j} w_{ij}(n - |T_{ij}^R|) \\ &\geq \sum_{i < j} w_{ij}(n - |T_{ij}^O| - 6\epsilon n) \\ &= rev(T^O) - 6\epsilon n \sum_{i < j} w_{ij} \\ &\geq (1 - 19\epsilon)rev(T^O), \end{aligned}$$

where the last inequality is due to Fact 1 and since n is assumed to be large enough. \square

3.2 An Efficient-PRAS for Revenue Instances with Not All Small Weights

In this section we consider the problem of finding an optimal revenue tree in instances with weights that are

not all small and present an Efficient-PRAS. We show that in a sense this is the best one could hope for, and complement our result by showing that the problem is NP-Complete and thus does not admit an optimal, polynomial solution unless $P = NP$ (see Theorem 8 in the Appendix).

Let $\epsilon > 0$, let $|V| = n$ and $k = \lceil \frac{1}{\epsilon} \rceil$. Finally, let T_ϵ^R denote the tree guaranteed by Theorem 1 for ϵ . We may define T_ϵ^R 's revenue as follows. For every one of T_ϵ^R 's internal nodes i , denote by D_i its set of children that are data points. Furthermore, let W_{ij} denote the total weight of the set of (similarity) edges crossing between D_i and D_j . Therefore, $rev(T_\epsilon^R) = \sum_{i < j} (|W_{ij}| \sum_{\ell} |D_\ell|)$, where the second summation is over all sets D_ℓ not contained in T_{ij}^R (as defined by T_ϵ^R 's sketch). We note that due to Theorem 1, the first summation is over at most $\Theta(k)$ entries (specifically, at most $20 \cdot k$).

Next, we consider the General Partitioning Property Tester of Goldreich et al. (1998). Given values α_i and β_{ij} (representing the sizes of the data point sets and the weight of edges between every pair of sets) the property tester allows us to test whether there exists a graph partition with set sizes α_i , and weight of edges crossing between the different sets β_{ij} . The property tester also takes as input ϵ_{err} and δ which define the error in α_i and β_{ij} and the probability of failing, respectively. Formally, we denote this as $PT(\{\alpha_i\}, \{\beta_{ij}\}, \epsilon_{err}, \delta)$. Thereafter, the property tester returns the following: if there exists a partition upholding the values α_i and β_{ij} then the tester returns this partition up to an additive error of $n\epsilon_{err}$ in the sizes of α_i and additive error of $n^2\epsilon_{err}$ in the sizes β_{ij} . If such a partition does not exist, the tester returns that such a partition does not exist.

Overall, this suggests an algorithm that guesses T_ϵ^R by guessing a tree of size $20 \cdot k$ (see Theorem 1) and guessing α_i and β_{ij} (simply through iteration). Unfortunately, guessing α_i and β_{ij} exactly would only yield a PRAS. To obtain an Efficient-PRAS, we guess α_i upto a factor of ϵ^2 and β_{ij} up to a factor of ϵ^3 . This yields Algorithm 3. Lemma 6 (proved in the Appendix) guarantees the approximation needed.

Algorithm 3 EPRAS for Revenue case.

```

Enumerate over all trees,  $T$ , with  $k$  internal leaves.
for each such  $T$  do
  for  $\{\alpha_i\}_{i \leq k} \subset \{i\epsilon^2 n : i \in \mathbb{N} \wedge i \leq \frac{3}{\epsilon}\}$  do
    for  $\{\beta_{ij}\}_{i \leq k, j \leq k} \subset \{i\epsilon^3 n^2 : i \in \mathbb{N} \wedge i \leq \frac{9}{\epsilon}\}$  do
      Run  $PT(\{\alpha_i\}, \{\beta_{ij}\}, \epsilon_{err} = \epsilon^3, \delta)$ .
    Compute the revenue given  $T$  and  $PT$ 's output.
  Return the maximal revenue tree encountered.

```

Lemma 6. *For every $\epsilon > 0$, Algorithm 3 guarantees an approximation factor of $(1 - 18\epsilon - \frac{12\epsilon}{\rho\tau})$.*

We note that the error from the property tester is offset by the revenue from the optimal solution.

Theorem 2. *Algorithm 3 is an Efficient-PRAS.*

Proof. Lemma 6 guarantees that there exists $\hat{\epsilon} > 0$ (specifically, $\hat{\epsilon} = 18\epsilon + \frac{12\epsilon}{\rho\tau}$) such that our algorithm is a $1 - \hat{\epsilon}$ approximation. The property tester runs in time, $exp(\log(\frac{1}{\delta\epsilon_{err}})(\frac{O(1)}{\epsilon_{err}})^{k+1}) + O(\frac{\log(k/(\epsilon_{err}\delta))}{\epsilon_{err}^2})n$.

Further, we call the tester $k^k \cdot (3/\epsilon)^k \cdot (9/\epsilon)^{k^2}$ times. Now, since $\epsilon < \hat{\epsilon}$, if $\epsilon_{err} = \epsilon^3$ then the algorithm is an Efficient-PRAS. \square

3.3 Metric-Based Similarity Instances

We follow the definitions as seen in Charikar et al. (2019b). Suppose that our data points lie on a metric M with doubling dimension $D(M)$. Define a non-increasing function $g : \mathbb{R}_{\geq 0} \rightarrow [0, 1]$. Given two data points i and j let d_{ij} denote their distance as defined by our metric. Furthermore, we define the metric-based similarity weights $w_{ij} = g(d_{ij})$.

Define $A(\epsilon) = A$ to be the tree generated by the algorithm that adds a constant ϵ to all weights and then runs Algorithm 3 for ρ, τ -weighted instances. We note that A is well defined since the altered weights define a graph with not all small weights for $\tau = \epsilon$ and $\rho = 0$.

The following theorem shows that for a large class of functions g and metrics M , algorithm A is in fact an Efficient-PRAS. We defer its proof to the appendix.

Theorem 3. *Assume the metric's doubling dimension guarantees $D(M) = O(1)$ and g is scale invariant and ℓ -Lipschitz continuous for $\ell = O(1)$. Then, A is an Efficient-PRAS for the induced Revenue instance.*

4 THE DISSIMILARITY CASE

4.1 A Reduction to Constant Sketches

In this section we show how to create a tree that approximates the optimal dissimilarity value. This tree is produced by taking $K(T^O)$ for the optimal tree, T^O (as defined earlier) and altering it. As opposed to the revenue case, this theorem guarantees $O(1/\epsilon^2)$ internal nodes while maintaining a $(1 - \epsilon)$ approximation. Note that this result holds for any dissimilarity instance and thus may be of independent interest. For an overview we refer the reader to our Techniques section.

Theorem 4. *Let T^O denote the optimal dissimilarity tree and assume it contains n leaves (i.e., data points). Then, for any $\epsilon > 0$, there exists a tree T^D such that (i) T^D contains $\Theta(1/\epsilon^2)$ internal nodes, each with at most $3\epsilon^2 n$ children, and (ii) $dis(T^D) \geq (1 - \epsilon)dis(T^O)$.*

In order to obtain T^D given a binary tree, T , we use $K(T)$ (as defined in Section 3). We then convert $K(T)$ to T^D , by randomly partitioning each contracted node's data points into $1/\epsilon$ clusters and attaching them in a "comb"-like structure. The process is defined in Algorithm 4 (see Figure 3 for an example).

Algorithm 4 Algorithm to convert T to T^D .

$K(T) \leftarrow$ Algorithm 1 applied to T .
for each node $c \in K(T)$ and its data points D_c **do**
 Partition D_c into $1/\epsilon$ random sets of equal sizes,
 $P = \{P_1, \dots, P_{1/\epsilon}\}$.
 for $P_i \in P$ **do**
 Create a new auxiliary node, u_i .
 Attach P_i as u_i 's children.
 Create a new node ℓ_i , and attach it between c
 and its parent.
 Attach u_i as ℓ_i 's child.
 Return the resulting tree as T^D .

Note that $D_c = \emptyset$ if c is the root (since the root is blue) and therefore ℓ_i is indeed only defined for c 's that have a parent. Also note that as in Remark 1, T^D remains binary if we disregard the auxiliary nodes. Next we show that T^D is of constant size and that $|T_{ij}^D|$ is (approximately) lower bounded.

Lemma 7. T^D contains at most $20/\epsilon^2$ and at least $2/\epsilon^2$ internal nodes with at most $3\epsilon^2 n$ children.

Lemma 8. The resulting tree, T^D , guarantees in expectation that, $|T_{ij}^D| \geq (1 - \epsilon)|T_{ij}| - 6\epsilon n$.

We defer the proofs of Lemmas 7 and 8 to the Appendix. Finally, combining Lemmas 7 and 8 for $T = T^O$ with Fact 2, is enough to prove Theorem 4. (For the formal proof, see Appendix).

4.2 An Efficient-PRAS for Dissimilarity Instances with Not All Small Weights

In this section we consider the problem of finding an optimal dissimilarity tree in instances with weights that are not all small and present an Efficient-PRAS. As in the revenue case, again we show that this is the best one could hope for, and complement our result by showing that the problem is NP-Complete and thus does not admit an optimal, polynomial solution (see Theorem 9 in the Appendix)

Let $\epsilon > 0$ and let T_ϵ^D denote the tree guaranteed by Theorem 4 for ϵ . As in the revenue case, for an internal node of T^D , i , let D_i denote the set of data points that are i 's children and let W_{ij} denote the set of (dissimilarity) edges crossing between D_i and D_j . Therefore, $dis(T_\epsilon^D) = \sum_{i,j \in S} (W_{ij} \sum_{\ell \in S} |D_\ell|) + b$, where

the second sum is over all sets D_ℓ contained in T_{ij}^D (as defined by T_ϵ^D 's sketch). Furthermore, b is defined as the dissimilarity gained by nodes within the same "star" structure. Theorem 4 guarantees that $|D_i|$ is small - therefore, since our instance has weights that are not all small (and by Fact 2 the optimal solution is large) this dissimilarity is negligible and we may assume $b = 0$ since we already lose a factor of $1 - \epsilon$. Finally, recall that $|S| \leq 20k$.

Our Efficient-PRAS follows as in the revenue case and is therefore deferred to the Appendix (Algorithm 7). The following theorem is proven identically to the revenue case and is therefore omitted.

Theorem 5. Algorithm 7 is an EPRAS for dissimilarity instances with weights that are not all small.

5 HIERARCHICAL CORRELATION CLUSTERING

In this section we consider the case where the collected data may contain both similarity and dissimilarity information. We first show a worst case approximation and thereafter show an Efficient-PRAS for HCC $^\pm$.

5.1 Worst Case Guarantees for HCC

Here we consider two separate algorithms which, if combined properly, will yield our approximation. The first is a simple greedy algorithm whereas the second optimizes for the MAX-UNCUT BISECTION problem for its top most cut and then continues with the greedy algorithm. We first show baseline guarantees of the greedy algorithm and then use the work of Alon et al. (2020) in order to obtain guarantees on the second algorithm with respect to the HCC objective. We defer the following proof to the appendix.

Proposition 1. There exists a greedy algorithm, denoted by ALG_{GRE} , that returns an HC tree T_1 guaranteeing,

$$hcc(T_1) \geq \frac{1}{3}(n-2) \sum_{ij} w_{ij}^s + \frac{2}{3}n \sum_{ij} w_{ij}^d.$$

Denote by ALG_{MUB} the algorithm that generates an HC tree by first cutting according to MAX-UNCUT BISECTION based on the similarity weights of the instance and then running ALG_{GRE} on each of the two resulting sides. Let $OPT = OPT_s + OPT_d$ be the value of the optimum HCC tree where $OPT_s = \sum w_{ij}^s (n - |O_{ij}|)$ and $OPT_d = \sum w_{ij}^d |O_{ij}|$, defined such that O_{ij} denotes the number of leaves in the subtree rooted at the LCA of i and j in the tree of OPT .

Lemma 9. Let T_2 denote the HC tree returned by ALG_{MUB} . Therefore,

$$hcc_G(T_2) \geq 0.585 \cdot OPT_s + \frac{1}{3} \cdot OPT_d$$

Proof. For ease of exposition let $T_2 = T$. The top-split of T is a bisection which means that $|L| = |R| = \frac{n}{2}$. For ease of notation let:

$$W_L^s = \sum_{i,j \in L} w_{ij}^s \text{ and } W_L^d = \sum_{i,j \in L} w_{ij}^d$$

Similarly, we define W_R^s and W_R^d . Notice that for the L side, GREEDY will contribute at least $\frac{2}{3} \cdot \frac{n}{2} \cdot W_L^d$ to $\sum w_{ij}^d |T_{ij}|$, as per Proposition 1. Similarly, for the R side. This means that in the tree T , any edge contributes either $\frac{2}{3} \cdot \frac{n}{2}$ (if it was cut by GREEDY) or n (if it was cut at the top-split of MAX-UNCUT BISECTION). In any case, we have:

$$\sum w_{ij}^d |T_{ij}| \geq \frac{2}{3} \cdot \frac{n}{2} \sum w_{ij}^d \geq \frac{1}{3} \text{OPT}_d \quad (1)$$

by using the upper bound $\text{OPT}_d \leq n \sum w_{ij}^d$.

We now deal with OPT_s . Observe that:

$$\begin{aligned} \sum w_{ij}^d (n - |T_{ij}|) &\geq W_L^+ \left(\frac{n}{2} + \frac{1}{3} \frac{n}{2} \right) + W_R^s \left(\frac{n}{2} + \frac{1}{3} \frac{n}{2} \right) \\ &\geq \frac{2}{3} n (W_L^s + W_R^s) \end{aligned}$$

since every edge within L will contribute $\frac{n}{2}$ due to the bisection, plus an extra $\frac{1}{3} \frac{n}{2}$ due to the greedy step. The same is true for edges in R .

Finally, since we used a 0.8776 for MAX-UNCUT BISECTION, it holds directly from Alon et al. (2020) that:

$$\sum w_{ij}^d (n - |T_{ij}|) \geq \frac{2}{3} \cdot 0.8776 \cdot \text{OPT}_s \geq 0.585 \cdot \text{OPT}_d \quad (2)$$

The lemma follows by summing eq. (1) and (2). \square

Finally, we combine Proposition 1 and Lemma 9 in order to yield the following Theorem (whose proof is deferred to the appendix).

Theorem 6. *Running ALG_{GRE} with probability p and otherwise ALG_{MUB} guarantees an approximation of 0.4767 for the HCC objective, when $p = 0.43$.*

5.2 An Efficient-PRAS for HCC on complete graphs

Here we consider the HCC^\pm objective (as defined earlier in the introduction) and show an Efficient-PRAS. We also complement our results and show that in fact this problem is NP-Complete and thus we cannot hope for an optimal, polynomial solution (see Theorem 10 in the Appendix).

Let ALG^\pm denote the algorithm that runs Algorithm 3 and Algorithm 7 simultaneously and returns the tree maximizing the HCC^\pm objective. We prove that ALG^\pm is in fact an Efficient-PRAS for the HCC^\pm objective. We defer the theorem's proof to the appendix.

Theorem 7. *ALG^\pm is an Efficient-PRAS for the HCC^\pm objective.*

6 HARDNESS RESULTS FOR INSTANCES WITH NOT ALL SMALL WEIGHTS

When considering instances with weights that are not all small, we have only shown Efficient-PRAS's up until now. To complement our results, we show that we can not hope for optimal, polynomial algorithms, assuming the Small Set Expansion (SSE) hypothesis. (For a formal definition of SSE see Charikar and Chatziafratis (2017)). In fact, it is enough to show that these objectives are NP-complete assuming the instances are (1) unweighted and (2) guarantee that $\sum_{i < j} w_{ij} = \Omega(n^2)$. We call such instances *dense* instances. We begin by stating the results (and thereafter providing proofs for each).

Theorem 8. *The Revenue objective for dense instances is in NPC (assuming SSE).*

Theorem 9. *The Dissimilarity objective for dense instances is in NPC (assuming SSE).*

Theorem 10. *The HCC^\pm objective is in NPC (assuming SSE).*

7 CONCLUSION

In this paper we show that to optimize for the Rev-HC and Dis-HC objectives, it suffices to consider HC trees with constant-sized sketches, thereby greatly simplifying these problems. This result can be applied to both the heuristic setting (since it greatly reduces the range of optimal solutions that need to be considered) and the approximation setting. Specifically, an approximation algorithm may iterate over all constant sized trees. Thereafter, it will need to partition the data points into the leaves of the constant-sized tree - thus reducing our problem to the well-studied realm of graph partitioning problems.

We then consider the family of instances with weights that are not all small. We show Efficient-PRAS's for both Rev-HC and Dis-HC objectives. Furthermore, we show that this family of instances encompasses many metric-based similarity instances. Finally, we introduce the HCC objective which we hope will provide a better connection between the realms of correlation and hierarchical clustering. We then show a worst case approximation of 0.4767 and show an Efficient-PRAS for the HCC^\pm objective that leverages our algorithms presented for the Rev-HC and Dis-HC objectives for instances with weights that are not all small.

8 ACKNOWLEDGEMENTS

The authors would like to deeply thank Claudio Gentile and Fabio Vitale for their helpful discussions and

insights regarding the connection to metric-based similarity instances. We also thank Sara Ahmadian and Alessandro Epasto for interesting discussions during early stages of our work. This research was further supported in part by the Israel Science Foundation (grant No. 2304/20 and grant No. 1506/16).

References

- S. Ahmadian, V. Chatziafratis, A. Epasto, E. Lee, M. Mahdian, K. Makarychev, and G. Yaroslavtsev. Bisect and conquer: Hierarchical clustering via max-uncut bisection. *CoRR*, abs/1912.06983, 2019.
- N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):1–27, 2008.
- N. Alon, Y. Azar, and D. Vainstein. Hierarchical clustering: A 0.585 revenue approximation. In J. D. Abernethy and S. Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 153–162. PMLR, 2020. URL <http://proceedings.mlr.press/v125/alon20b.html>.
- U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999. ISSN 0027-8424. doi: 10.1073/pnas.96.12.6745. URL <https://www.pnas.org/content/96/12/6745>.
- N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *43rd Symposium on Foundations of Computer Science (FOCS 2002), 16-19 November 2002, Vancouver, BC, Canada, Proceedings*, page 238, 2002.
- P. Berkhin. A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pages 25–71, 2006.
- F. Bonchi, D. Garcia-Soriano, and E. Liberty. Correlation clustering: from theory to practice. In *KDD*, page 1972, 2014.
- P. F. Brown, V. J. D. Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.
- M. Charikar and V. Chatziafratis. Approximate hierarchical clustering via sparsest cut and spreading metrics. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 841–854, 2017.
- M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. *Journal of Computer and System Sciences*, 71(3):360–383, 2005.
- M. Charikar, V. Chatziafratis, and R. Niazadeh. Hierarchical clustering better than average-linkage. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 2291–2304, 2019a.
- M. Charikar, V. Chatziafratis, R. Niazadeh, and G. Yaroslavtsev. Hierarchical clustering for euclidean data. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2721–2730, 2019b. URL <http://proceedings.mlr.press/v89/charikar19a.html>.
- V. Chatziafratis, N. Gupta, and E. Lee. Inapproximability for local correlation clustering and dissimilarity hierarchical clustering. *arXiv preprint arXiv:2010.01459*, 2020. URL <https://arxiv.org/abs/2010.01459>.
- S. Chawla, K. Makarychev, T. Schramm, and G. Yaroslavtsev. Near optimal lp rounding algorithm for correlation clustering on complete and complete k-partite graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 219–228, 2015.
- W. Cohen and J. Richman. Learning to match and cluster entity names. In *ACM SIGIR-2001 Workshop on Mathematical/Formal Methods in Information Retrieval*, 2001.
- W. W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480, 2002.
- V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu. Hierarchical clustering: Objective functions and algorithms. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 378–397, 2018.
- S. Dasgupta. A cost function for similarity-based hierarchical clustering. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 118–127, 2016.
- A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, 19(1):1–16, 2006.
- M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satis-

- fiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.
- O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.
- N. Jardine and R. Sibson. A model for taxonomy. *Mathematical Biosciences*, 2(3-4):465–482, 1968.
- S. Kim, S. Nowozin, P. Kohli, and C. D. Yoo. Higher-order correlation clustering for image segmentation. In *Advances in neural information processing systems*, pages 1530–1538, 2011.
- B. Moseley and J. Wang. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3094–3103, 2017.
- A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proceedings of the 14th ACM conference on Computer and communications security*, pages 342–351, 2007.
- J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *IEEE Computer*, 35(7): 80–86, 2002. doi: 10.1109/MC.2002.1016905. URL <https://doi.org/10.1109/MC.2002.1016905>.
- P. H. Sneath and R. R. Sokal. Numerical taxonomy. *Nature*, 193(4818):855–860, 1962.
- C. Swamy. Correlation clustering: maximizing agreements via semidefinite programming. In J. I. Munro, editor, *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2004, New Orleans, Louisiana, USA, January 11-14, 2004*, pages 526–527. SIAM, 2004. URL <http://dl.acm.org/citation.cfm?id=982792.982866>.