Local Correlation Clustering with Asymmetric Classification Errors

Jafar Jafarov [*1] **Sanchit Kalhan** [*2] **Konstantin Makarychev** [*2] **Yury Makarychev** [*3]

Abstract

In the Correlation Clustering problem, we are given a complete weighted graph G with its edges labeled as "similar" and "dissimilar" by a noisy binary classifier. For a clustering $\mathcal{C}$ of graph G, a similar edge is in disagreement with $\mathcal{C}$, if its endpoints belong to distinct clusters; and a dissimilar edge is in disagreement with $\mathcal{C}$ if its endpoints belong to the same cluster. The disagreements vector, dis, is a vector indexed by the vertices of G such that the v-th coordinate dis_v equals the weight of all disagreeing edges incident on v. The goal is to produce a clustering that minimizes the ℓ_p norm of the disagreements vector for $p \geq 1$. We study the ℓ_p objective in Correlation Clustering under the following assumption: Every similar edge has weight in the range of $[\alpha\mathbf{w}, \mathbf{w}]$ and every dissimilar edge has weight at least $\alpha\mathbf{w}$ (where $\alpha \leq 1$ and $\mathbf{w} > 0$ is a scaling parameter). We give an $O\left(\left(1/\alpha\right)^{1/2-1/2p} \cdot \log 1/\alpha\right)$ approximation algorithm for this problem. Furthermore, we show an almost matching convex programming integrality gap.

1. Introduction

Grouping objects based on the similarity between them is a ubiquitous and important task in machine learning. This similarity information between objects can be represented in many ways, some of them being pairwise distances between objects (objects which are closer are more similar) or the degree of similarity between pairs of objects (objects which are more similar have a higher degree of similarity). Bansal, Blum, and Chawla (2004) introduced the Correlation Clustering problem, a versatile model that elegantly captures this task of grouping objects based on similarity information. Since its introduction, the correlation clustering problem has found use in a variety of applications, such as co-reference resolution (see e.g., Cohen and Richman (2001; 2002)), spam detection (see e.g., Ramachandran et al. (2007), Bonchi et al. (2014)), image segmentation (see e.g., Wirth (2010)) and multi-person tracking (see e.g., Tang et al. (2016; 2017)). In the Correlation Clustering problem, we are given a set of objects with pairwise similarity information. Our goal is to partition the objects into clusters that agree with this information *as much as possible*. The pairwise similarity information is given as a weighted graph G with edges labeled as either "positive/similar" or as "negative/dissimilar" by a noisy binary classifier. For a clustering $\mathcal{C}$, a positive edge is in disagreement with $\mathcal{C}$, if its endpoints belong to distinct clusters; and a negative edge is in disagreement with $\mathcal{C}$ if its endpoints belong to the same cluster.

To ascertain the quality of the clustering produced, Bansal et al. (2004) studied the Correlation Clustering problem under two complimentary objectives. Over the years, the objective that has received the most attention is to find a clustering that minimizes the total weight of edges in disagreement. For the case of complete unweighted graphs, Bansal et al. (2004) gave a constant factor approximation algorithm for this objective. Ailon, Charikar, and Newman (2008) improved the approximation ratio to 3 by presenting a simple-yet-elegant combinatorial algorithm. They also presented a 2.5-approximation algorithm based on Linear Programming (LP) rounding which was later derandomized without any loss in approximation ratio by van Zuylen, Hegde, Jain, and Williamson (2007). Finally, Chawla, Makarychev, Schramm, and Yaroslavtsev (2015) gave an LP rounding algorithm which improved the approximation ratio to 2.06. The standard LP was shown to have an integrality gap of 2 by Charikar, Guruswami, and Wirth (2003) for the case of complete unweighted graphs. For the case of general weighted graphs, Charikar et al. (2003) and Demaine, Emanuel, Fiat, and Immorlica (2006) gave an $O(\log n)$-approximation algorithm.

Define the disagreements vector to be a vector indexed by the vertices of G. Given a clustering $\mathcal{P}$, $\mathrm{dis}(\mathcal{P}, E^+, E^-) \in \mathbb{R}^V$ is a $|V|$-dimensional vector where the u-th coordinate is equal to the weight of disagreements at u with respect to $\mathcal{P}$. That is, $\mathrm{dis}_u(\mathcal{P}, E^+, E^-) = \sum_{(u,v)\in E} w_{uv} \cdot$

[*]Equal contribution [1]University of Chicago, Chicago, IL, USA [2]Northwestern University, Evanston, IL, USA [3]TTIC, Chicago, IL, USA. Correspondence to: Jafar Jafarov <jafarov@uchicago.edu>, Sanchit Kalhan <skalhan@u.northwestern.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

$$
\begin{array}{lll}
\text{minimize} & \|y\|_p & \text{(P)} \\
\text{subject to} & y_u = \sum_{v:(u,v)\in E^+} w_{uv}x_{uv} + \sum_{v:(u,v)\in E^-} w_{uv}(1-x_{uv}) \quad \text{for all } u \in V & \text{(P1)} \\
& x_{v_1v_2} + x_{v_2v_3} \geq x_{v_1v_3} \quad \text{for all } v_1, v_2, v_3 \in V & \text{(P2)} \\
& x_{uv} = x_{vu} \quad \text{for all } u, v \in V & \text{(P3)} \\
& x_{uv} \in [0,1] \quad \text{for all } u, v \in V & \text{(P4)}
\end{array}
$$

Figure 1. Convex relaxation for Correlation Clustering with min ℓ_p objective for $p \geq 1$ or $p = \infty$.

$\mathbf{1}\{(u,v)$ is in disagreement with $\mathcal{P}\}$. Thus, minimizing the total weight of disagreements is equivalent to finding a clustering minimizing the ℓ_1 norm of the disagreements vector. Another objective for Correlation Clustering that has received attention recently is to minimize the weight of disagreements at the vertex that is worst off (also known as Min Max Correlation Clustering). This is equivalent to finding a clustering that minimizes the ℓ_∞ norm of the disagreements vector. Observe that minimizing the ℓ_1 norm is a global objective since the focus is on minimizing the total weight of disagreements. In contrast, for higher values of p (particularly $p = \infty$), minimizing the ℓ_p norm becomes a more local objective since the focus shifts towards minimizing the weight of disagreements at a single vertex. Minimizing the ℓ_2 norm of the disagreements vector can thus provide a balance between these global and local perspectives – it considers the weight of disagreements at all vertices but penalizes vertices that are worse off more heavily. The following scenario is a showcase that minimizing the ℓ_2 norm might be a more suitable objective than minimizing the ℓ_1 norm. Consider a recommender system such that input is a bipartite graph with left and right sides representing customers and services, respectively. A positive edge implies that a customer is satisfied with the service; whereas a negative edge implies that they are dissatisfied with or have not used the service. We may be interested in grouping customers and services so that the total and the individual dissatisfaction of customers are minimized.

Definition 1.1. ***(Local Correlation Clustering)*** *Given an instance of Correlation Clustering $G = (V, E = E^+ \cup E^-)$ and $p \geq 1$, the local objective is to find a partitioning $\mathcal{P}$ that minimizes the ℓ_p norm.*

We use the standard definition of the ℓ_p norm of a vector x: $\|x\|_p = (\sum_u |x_u|^p)^{\frac{1}{p}}$. Since its introduction by Puleo and Milenkovic (2018), local objectives for Correlation Clustering have been mainly studied under two models (see Charikar, Gupta, and Schwartz (2017), Ahmadi, Khuller, and Saha (2019), Kalhan, Makarychev, and Zhou (2019)). We will refer to these models as (1) Correlation Clustering on Complete Graphs, and (2) Correlation Clustering with Noisy Partial Information. In the first model, the input graph G is complete and unweighted. For this model, the first approximation algorithm was by Puleo & Milenkovic (2018) with an approximation factor of 48 for minimizing the ℓ_p norm. This was later improved to 7 by Charikar et al. (2017). Lastly, Kalhan et al. (2019) provided a 5 approximation algorithm. In the second model, G is an arbitrary weighted graph with possibly missing edges. For minimizing the ℓ_∞ norm of the disagreements vector in this model, Charikar et al. (2017) provided a $O(\sqrt{n})$ approximation. Kalhan et al. (2019) gave an $O(n^{\frac{1}{2}-\frac{1}{2p}} \cdot \log^{\frac{1}{2}+\frac{1}{2p}} n)$-approximation algorithm for minimizing the ℓ_p norm of the disagreements vector.

We study local objectives in a different model – Correlation Clustering with Asymmetric Classification Errors – recently introduced by Jafarov, Kalhan, Makarychev, and Makarychev (2020). In this model, the input graph G is complete and weighted. Furthermore, the ratio of the smallest edge weight to the largest positive edge weight is at least $\alpha \leq 1$. Thus, for some $\mathbf{w} > 0$, each positive edge weight lies in the interval $[\alpha\mathbf{w}, \mathbf{w}]$ and each negative edge weight is at least $\alpha\mathbf{w}$. This model better captures the subtleties in real world instances than the standard models. Since real world instances rarely have equal edge weights, assumptions in the Correlation Clustering on Complete Graphs model are too strong. In contrast, in the Correlation Clustering with Noisy Partial Information model, we can have edge weights that are arbitrarily small or large, an assumption which is too weak. In many real world instances, the edge weights lie in some range $[a, b]$ with $a, b > 0$. For this model, Jafarov et al. (2020) gave a $(3 + 2\ln\frac{1}{\alpha})$ approximation for minimizing the ℓ_1 norm of the disagreements vector.

Definition 1.2. *Correlation Clustering with Asymmetric Classification Errors is a variant of Correlation Clustering on Complete Graphs. We assume that the weight of each positive edge lies in $[\alpha\mathbf{w}, \mathbf{w}]$ and the weight of each negative edge lies in $[\alpha\mathbf{w}, \infty)$, where $\alpha \in (0, 1]$ and $\mathbf{w} > 0$.*

Our Contributions. In this paper we study the task of minimizing local objectives (Definition 1.1) under the Correlation Clustering with Asymmetric Classification Errors model (Definition 1.2). Our main result is an

$O\left(\left(\frac{1}{\alpha}\right)^{\frac{1}{2}-\frac{1}{2p}} \cdot \log \frac{1}{\alpha}\right)$ approximation algorithm for minimizing the ℓ_p norm of the disagreements vector, which we now state.

Theorem 1.3. *There exists a polynomial-time $O\left(\left(\frac{1}{\alpha}\right)^{\frac{1}{2}-\frac{1}{2p}} \cdot \log \frac{1}{\alpha}\right)$-approximation algorithm for the ℓ_p objective in the Correlation Clustering with Asymmetric Classification Errors model.*

For $p = 1$, our algorithm provides an $O(\log \frac{1}{\alpha})$ approximation, which matches the approximation guarantee given by Jafarov et al. (2020) up to constant factors. Consider $p = 2$, that is, the ℓ_2 norm. If we ignored the edge weights and applied the state of the art algorithm in the Correlation Clustering on Complete Graphs model, we would get an $O(\frac{1}{\alpha})$ approximation. If we were to use the state of the art algorithm in the Correlation Clustering with Noisy Partial Information model, we would get an $\tilde{O}\left(n^{1/4}\right)$ approximation. However, by using our algorithm (Theorem 1.3), we obtain an $\tilde{O}\left(\left(1/\alpha\right)^{1/4}\right)$ approximation, which is a huge improvement when $1/\alpha \ll n$.

Corollary 1.4. *There exists a polynomial-time $O\left(\left(1/\alpha\right)^{1/4} \cdot \log \frac{1}{\alpha}\right)$-approximation algorithm for the ℓ_2 objective in the Correlation Clustering with Asymmetric Classification Errors model.*

Finally, we present the implication of our main result for the ℓ_∞ norm. For the ℓ_∞ norm, Kalhan et al. (2019) presented an $\tilde{O}(\sqrt{n})$ approximation under the Correlation Clustering with Noisy Partial Information model. Using our algorithm for Correlation Clustering under Asymmetric Classification Errors we obtain an $\tilde{O}(\sqrt{1/\alpha})$-approximation factor, which is a significant improvement to the approximation guarantee in this setting.

Corollary 1.5. *There exists a polynomial-time $O(\sqrt{1/\alpha} \cdot \log 1/\alpha)$-approximation algorithm for the ℓ_∞ objective in the Correlation Clustering with Asymmetric Classification Errors model.*

We emphasize that our approximation ratio for the ℓ_p norm is independent of the graph size and only depends on α.

Our algorithm relies on the natural convex programming relaxation for this problem (Section 2). We compliment our positive result (Theorem 1.3) by showing that it is likely to be the best possible based on the natural convex program, by providing an almost matching integrality gap.

Theorem 1.6. *The natural convex programming relaxation for the ℓ_p objective in the Correlation Clustering with Asymmetric Classification Errors model has an integrality gap of $\Omega\left(\left(1/\alpha\right)^{\frac{1}{2}-\frac{1}{2p}}\right)$.*

Organization of the paper. In Section 2, we describe the convex relaxation that we will use in our algorithm for Correlation Clustering. In Section 3, we introduce a novel technique for partitioning metric spaces. This forms the main technical basis for our algorithm for Correlation Clustering. In Section 4, we prove our main result, Theorem 1.3. In Section 5, we describe our metric space partitioning scheme and give a proof overview of its correctness. In Appendices A, B, C and D, we formally prove the correctness of our partitioning scheme, Theorem 3.1. In Appendix E, we prove our integrality gap result, Theorem 1.6.

2. Convex Relaxation

Our algorithm for minimizing local objectives is based on rounding the optimal solution to a suitable convex program (Figure 1). This convex program is similar to the relaxations used in Charikar et al. (2017) and Kalhan et al. (2019). In this convex program, we have a variable x_{uv} for every pair of vertices $u, v \in V$. The variable x_{uv} captures the distance between u and v in the "multicut metric". In the integral solution, $x_{uv} = 0$ if u and v are in the same partition and $x_{uv} = 1$ if u and v are in different partitions. In order to enforce that the partitioning is consistent, we add triangle inequality constraints between all triplets of vertices (P2). We also require that distance x_{uv} is symmetric (P3).

For every vertex $u \in V$, we use the variable y_u to denote the total weight of violated edges incident on u (P1). The objective of the convex program is thus to minimize $\|y\|_p$ – the ℓ_p norm of the vector y. Notice that each constraint in the convex program is linear, and the objective function $\|\cdot\|_p : \mathbb{R}^n \to \mathbb{R}$ is convex (by the Minkowski inequality).

What remains to be shown is that the relaxation presented in Figure 1 is valid. To this end, consider any partition $\mathcal{P} = (P_1, P_2, \ldots, P_k)$ of the set of vertices V. For every pair of vertices u, v, if u and v lie in the same partition, we assign the corresponding variable x_{uv} a value of 0, else we assign it a value of 1. Note that such an assignment satisfies the triangle inequality (P2). Variable y_u thus captures the total weight of violated edges incident on u; every similar edge (u, v) incident on u that crosses a partition contributes $w_{uv} \cdot x_{uv} = w_{uv}$ to y_u, and every dissimilar edge present within a cluster contributes $w_{uv} \cdot (1 - x_{uv}) = w_{uv}$ to y_u. Thus, y_u is equal to $\mathrm{dis}_u(\mathcal{P}, E^+, E^-)$. Hence, an integral convex program solution defined in such a manner is feasible and has the same cost as the partitioning. It is possible, however, that the cost of the optimal fractional solution is less than the cost of the optimal integral solution, and hence the convex program in Figure 1 is a relaxation to our problem. We note that our relaxation is simpler than the relaxation used in Kalhan et al. (2019). The additional variables in their convex program are not needed in our case because all edge weights belong to the interval $[\alpha\mathbf{w}, \mathbf{w}]$.

3. A New Technique for Partitioning Metric Spaces

We will use the following notation: Given expressions X and Y, we write $X \lesssim Y$ if $X \leq C \cdot Y$ for some constant $C > 0$ (that is, $X = O(Y)$). We define $\gtrsim$ similarly. Furthermore, let $X^+ = 0$ if $X < 0$ and $X^+ = X$ if $X \geq 0$. We use $\mathrm{Ball}(v, l) = \{u : d(u, v) \leq l\}$ to denote the set of vertices at a distance of at most l from v.

In this section, we describe our main technical tool – a novel probabilistic scheme for partitioning metric spaces which may be of independent interest. This partitioning scheme forms the basis of our algorithm (Algorithm 1) for Correlation Clustering. We begin by stating this technical result.

Theorem 3.1. *For every $q \geq 1$ there exists a $\beta_q^* = \Theta\big(\frac{1}{q \ln(q+1)}\big) < 1$ such that the following holds. Consider a finite metric space (X, d). Fix two positive numbers r and R such that $\beta = r/R \leq \beta_q^*$. Let $D_\beta = 2(q+1)\ln 1/\beta$. Then, there exists a probabilistic partitioning $\mathcal{P}$ satisfying properties (1), (2), and (3):*

(1) $\mathrm{diam}(P) \leq 2R$ *for every $P \in \mathcal{P}$ (always);*

(2) *For every point u in X, the following bound holds:*

$$\sum_{v \in \mathrm{Ball}(u,R)} \left(\Pr\big\{\mathcal{P}(u) \neq \mathcal{P}(v)\big\} - D_\beta \frac{d(u,v)}{R} \right)^+ \lesssim \lesssim \beta^q \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R},$$

where $\mathcal{P}(u)$ denotes the partition of $\mathcal{P}$ that contains u.

(3) *Moreover, for every u in X, we always have,*

$$\sum_{v \in \mathrm{Ball}(u,r)} \mathbf{1}\big\{\mathcal{P}(u) \neq \mathcal{P}(v)\big\} \lesssim \lesssim \beta \cdot D_\beta^2 \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R}.$$

The partitioning we construct in Theorem 3.1 resembles a $2D$–separating $2R$-bounded stochastic decomposition of a metric space (Bartal, 1996; Călinescu et al., 2000; Fakcharoenphol et al., 2004). Recall that a $2D$–separating $2R$-bounded stochastic decomposition satisfies property (1) of Theorem 3.1 and the $2D$-separating condition: for every $u, v \in X$,

$$\Pr\big\{\mathcal{P}(u) \neq \mathcal{P}(v)\big\} - D\frac{d(u,v)}{R} \leq 0. \qquad (3.1)$$

At a very high level, the goals of our partitioning and the $2D$–separating $2R$-bounded stochastic decomposition are similar: decompose a metric space in clusters of diameter at most $2R$ so that nearby points lie in the same cluster with high enough probability. However, the specific conditions are quite different. Loosely speaking, property (2) of Theorem 3.1 says that the decomposition satisfies (3.1) with $D = D_\beta$ on average up to an additive error term of $O(\beta^q) \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R}$. Crucially, property (3) provides an analogous guarantee not only in expectation, but also in the worst case (which a $2D$-separating decomposition does not satisfy).

Property (3) plays a key role in proving our main result, Theorem 1.3. For the standard objective function for Correlation Clustering (minimizing the ℓ_1 norm of the disagreements vector), properties (1) and (2) are sufficient since an upper bound on the expected weight of disagreements on a single vertex implies an upper bound on the expected weight of the total disagreements. The situation gets trickier when we consider minimizing arbitrary ℓ_p ($p > 1$) norms of the disagreements vector. For instance, having an upper bound on the expected weight of disagreements on a single vertex does not necessarily translate to an upper bound on the expected weight of disagreements on a worst vertex (ℓ_∞ norm). We overcome this nonlinear nature of the problem for higher values of p by using the deterministic (worst-case) guarantee given by property (3) of Theorem 3.1.

Also note that coefficients D_β and β do not depend on the size $|X|$ of the metric space (in our algorithm, they will only depend on α, which is defined as the ratio of the smallest edge weight to the largest positive edge weight). However, the optimal value of D in the $2D$-separating condition is $\Theta(\log |X|)$.

4. Correlation Clustering via Metric Partitioning

In this section, we will prove our main theorem, Theorem 1.3. Our algorithm (Algorithm 1) for minimizing local objectives for Correlation Clustering with Asymmetric Classification Errors begins by solving the convex relaxation in Figure 1 to obtain a solution $\{x_{uv}\}_{u,v \in V}$. It then defines a metric $d(\cdot, \cdot)$ on V by setting distances $d(u, v) = x_{uv}$.

We let $q = 2$. Let α^* be the solution of equation $3\sqrt{\alpha^*}/\ln 1/\alpha^* = \beta_2^*$ (note that α^* is an absolute constant). We assume that $\alpha \leq \alpha^*$. If $\alpha > \alpha^*$, we just redefine α as α^* (this will increase the approximation ratio only by a constant factor). We set $r = \sqrt{\alpha}/\ln 1/\alpha$ and $R = 1/3$. Note that $r/R \leq \beta_2^* < 1$.

At this point, the algorithm makes use of our key technical contribution – a new probabilistic scheme for partitioning metric spaces (Algorithm 2) – and outputs the partitioning thus obtained. Please refer to Algorithm 1 for a summary.

Algorithm 1 Correlation Clustering Algorithm

Input: $G = (V, E^+, E^-, \mathbf{w}, \alpha)$, $\{x_{uv}\}_{u,v \in V}$
Define a metric d on V such that $d(u, v) = x_{uv}$ for all $u, v \in V$.
Define $r = (\sqrt{\alpha}/\ln 1/\alpha)$, $R = 1/3$, $q = 2$.
$\mathcal{P}$ = Metric Space Partitioning Scheme(V, d, r, R, q).
Output $\mathcal{P}$.

To show that $\mathcal{P}$ has the desired approximation ratio in Theorem 1.3, we bound the weight of disagreements at every vertex $u \in V$ with respect to $\mathcal{P}$. To this end, we show that two useful quantities, the total weight of disagreements at u and the expected weight of disagreements at u can be bounded in terms of y_u, the cost paid by the convex program for vertex u. In Theorem 4.1, we make use of the properties of $\mathcal{P}$ given by Theorem 3.1 to get a bound on these two quantities for each vertex $u \in V$. Then, in Section 4.1, we use the bounds from Theorem 4.1 to complete the proof of Theorem 1.3: we show that if the total cost of disagreements and the expected cost of disagreements with respect to $\mathcal{P}$ are bounded for every $u \in V$, then the partitioning $\mathcal{P}$ achieves the desired approximation ratio in Theorem 1.3. We remind the reader that given a partitioning $\mathcal{P}$ of the vertex set and a vertex $u \in V$, $\mathrm{dis}_u(\mathcal{P}, E^+, E^-)$ denotes the weight of edges incident on u that are in disagreement with respect to $\mathcal{P}$. Moreover, y_u denotes the convex programming (CP) cost of the vertex u.

Define $A_1 = \ln 1/\alpha$ and $\mathrm{A}_\infty = \ln(\frac{1}{\alpha})/\sqrt{\alpha} = 1/r$. Our analysis focuses on bounding two key quantities related to a vertex $u \in V$. The first quantity, $\mathrm{dis}_u(\mathcal{P}, E^+, E^-)$, is the total weight of edges incident on u that are in disagreement with $\mathcal{P}$. We show that this quantity can be charged to the CP cost of u and is at most $\mathrm{A}_\infty \cdot y_u$. We then get a stronger bound for our second quantity of interest, $\mathbf{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)]$, the expected cost of a vertex u. In particular, we show that $\mathbf{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)] \le A_1 \cdot y_u$.

Theorem 4.1. *Given an instance of Correlation Clustering with Asymmetric Classification Errors (Definition 1.2), Algorithm 1 outputs a partitioning $\mathcal{P}$ of the vertex set such that the following holds for every vertex $u \in V$:*

(a) $\mathrm{dis}_u(\mathcal{P}, E^+, E^-) \lesssim A_\infty \cdot y_u$;

(b) $\mathbf{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)] \lesssim A_1 \cdot y_u$,

where $A_1 = \ln(1/\alpha)$ *and* $\mathrm{A}_\infty = \ln(\frac{1}{\alpha})/\sqrt{\alpha}$.

Proof. Without loss of generality we assume that the scaling parameter $\mathbf{w}$ is 1. Thus, for every positive edge $e^+ \in E^+$, $w_{e^+} \in [\alpha, 1]$, while for every negative edge $e^- \in E^-$, $w_{e^-} \ge \alpha$. Write the formula for $\mathrm{dis}_u(\mathcal{P}, E^+, E^-)$ for a given vertex $u \in V$,

$$\mathrm{dis}_u(\mathcal{P}, E^+, E^-) = \sum_{(u,v)\in E^+} w_{uv} \cdot \mathbf{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\} + \sum_{(u,v)\in E^-} w_{uv} \cdot \mathbf{1}\{\mathcal{P}(u) = \mathcal{P}(v)\}.$$

Let $E^{\ge r}$ be the set of positive edges (v, w) in E^+ with $x_{vw} \ge r$. Observe that

$$\begin{aligned}\mathrm{dis}_u(\mathcal{P}, E^+, E^-) = \ & \mathrm{dis}_u(\mathcal{P}, \varnothing, E^-) \\ & + \mathrm{dis}_u(\mathcal{P}, E^{\ge r}, \varnothing) \\ & + \mathrm{dis}_u(\mathcal{P}, E^+ \setminus E^{\ge r}, \varnothing).\end{aligned} \tag{4.1}$$

Recall that $\beta = r/R = 3\sqrt{\alpha}/\ln 1/\alpha$, $q = 2$, and $D_\beta = \Theta(\ln 1/\beta) = \Theta(\ln 1/\alpha)$. From Theorem 3.1, part (a), we know that the diameter of each partition P in $\mathcal{P}$ is at most $2R$. For any negative edge to be in disagreement, both its endpoints must lie in the same partition. Thus, the length x_{uv} for any such edge $(u, v) \in E^-$ is at most $2R$, and hence its CP contribution is at most $(1 - 2R) = 1/3$. Hence,

$$\mathrm{dis}_u(\mathcal{P}, \varnothing, E^-) = \sum_{(u,v)\in E^-} w_{uv} \mathbf{1}\{\mathcal{P}(u) = \mathcal{P}(v)\} \le 3y_u.$$

Then,

$$\mathrm{dis}_u(\mathcal{P}, E^{\ge r}, \varnothing) \le |\{v : (u, v) \in E^{\ge r}\}| \le \frac{y_u}{r} = \mathrm{A}_\infty\, y_u.$$

To complete the proof of Theorem 4.1, part (a) we write:

$$\begin{aligned}\mathrm{dis}_u(\mathcal{P}, E^+ \setminus E^{\ge r}, \varnothing) &= \sum_{v\in \mathrm{Ball}(u,r)} w_{uv} \cdot \mathbf{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\} \\ &\le \sum_{v\in \mathrm{Ball}(u,r)} \mathbf{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\}.\end{aligned}$$

The inequality above holds because the weight of each positive edge is at most 1. Next, using the bound for $\sum_{v\in \mathrm{Ball}(u,r)} \mathbf{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\}$ from Theorem 3.1 part (c), we get,

$$\begin{aligned}\sum_{v\in \mathrm{Ball}(u,r)} \mathbf{1}\{\mathcal{P}(u) \neq \mathcal{P}(v)\} &\lesssim \beta \cdot D_\beta^2 \sum_{v\in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \\ &\lesssim \frac{\sqrt{\alpha}}{\ln(1/\alpha)} \cdot (\ln^2(1/\alpha)) \sum_{v\in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \\ &\lesssim \frac{\sqrt{\alpha}}{\ln(1/\alpha)} \cdot \ln^2(1/\alpha) \cdot \frac{y_u}{\alpha} = A_\infty \cdot y_u,\end{aligned}$$

where the last inequality follows from the fact that each positive edge weight is at least α. Thus, from (4.1) it follows:

$$\mathrm{dis}_u(\mathcal{P}, E^+, E^-) \lesssim \mathrm{A}_\infty \cdot y_u.$$

We now prove Theorem 4.1, part (b). We separately consider short and long positive edges. Let $E^{\le R}$ be the set of positive edges $(v,w) \in E^+$ with $x_{vw} \le R$. Note that

$$y_u \ge \sum_{v \in \mathrm{Ball}(u,R)} w_{uv} \min(d(u,v), 1-d(u,v)) \qquad (4.2)$$
$$= \sum_{v \in \mathrm{Ball}(u,R)} w_{uv} d(u,v) = \frac{1}{3} \sum_{v \in \mathrm{Ball}(u,R)} w_{uv} \frac{d(u,v)}{R}.$$

Therefore, we have

$$\begin{aligned}
&\mathbf{E}[\mathrm{dis}_u(\mathcal{P}, E^{\le R}, \varnothing) - 3D_\beta \cdot y_u] \le \\
&\le \mathbf{E}\Big[\sum_{v \in \mathrm{Ball}(u,R)} w_{uv} \cdot \mathbf{1}\{\mathcal{P}(u) \ne \mathcal{P}(v)\} \\
&\qquad - D_\beta \sum_{v \in \mathrm{Ball}(u,R)} w_{uv} \frac{d(u,v)}{R} \Big] \\
&= \sum_{v \in \mathrm{Ball}(u,R)} w_{uv} \Big(\Pr\{\mathcal{P}(u) \ne \mathcal{P}(v)\} - D_\beta \frac{d(u,v)}{R} \Big) \\
&\le \sum_{v \in \mathrm{Ball}(u,R)} w_{uv} \Big(\Pr\{\mathcal{P}(u) \ne \mathcal{P}(v)\} - D_\beta \frac{d(u,v)}{R} \Big)^+.
\end{aligned}$$

Since all edges (u,v) in $E^{\le R}$ are positive, we have $w_{uv} \le 1$. Consequently,

$$\begin{aligned}
&\mathbf{E}[\,\mathrm{dis}_u(\mathcal{P}, E^{\le R}, \varnothing) - 3D_\beta \cdot y_u] \\
&\le \sum_{\substack{v \in \mathrm{Ball}(u,R) \\ \text{s.t. } (u,v) \in E^+}} \Big(\Pr\{\mathcal{P}(u) \ne \mathcal{P}(v)\} - D_\beta \frac{d(u,v)}{R} \Big)^+.
\end{aligned}$$

We bound the right hand side using property (2) of Theorem 3.1:

$$\begin{aligned}
&\sum_{v \in \mathrm{Ball}(u,R)} \Big(\Pr\{\mathcal{P}(u) \ne \mathcal{P}(v)\} - D_\beta \frac{d(u,v)}{R} \Big)^+ \\
&\lesssim \beta^2 \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \lesssim \frac{\alpha}{\ln^2(1/\alpha)} \sum_{v \in \mathrm{Ball}(u,2R)} d(u,v) \\
&\le \frac{1}{\ln^2(1/\alpha)} \sum_{v \in \mathrm{Ball}(u,2R)} w_{uv} \cdot 2\min(d(u,v), 1-d(u,v)) \\
&\le \frac{2}{\ln^2(1/\alpha)} \cdot y_u.
\end{aligned}$$

Here, we used that $w_{uv} \ge \alpha$ and $d(u,v) \le 2(1-d(u,v))$ for $v \in \mathrm{Ball}(u,2R)$. Thus,

$$\mathbf{E}[\mathrm{dis}_u(\mathcal{P}, E^{\le R}, \varnothing)] \lesssim \big(\ln(1/\alpha) + \frac{1}{\ln^2(1/\alpha)} \big) y_u \lesssim A_1 \cdot y_u.$$

Furthermore, $\mathrm{dis}_u(\mathcal{P}, E^+ \setminus E^{\le R}, \varnothing) \le \frac{1}{R} \cdot y_u \le A_1 \cdot y_u$. Therefore, from (4.1) it follows that

$$\mathbf{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)] \lesssim A_1 y_u.$$

□

We now use Theorem 4.1 to prove Theorem 1.3.

4.1. Proof of Theorem 1.3

In this section, we show that the partitioning $\mathcal{P}$ output by Algorithm 1 achieves the desired approximation ratio – thereby proving our main theorem, Theorem 1.3. To show this, we will use the fact that $\mathcal{P}$ satisfies the properties in Theorem 4.1.

Proof of Theorem 1.3. If $p = \infty$, then we get an $O(A_\infty) = O((1/\alpha)^{1/2} \ln 1/\alpha)$ approximation by Theorem 4.1, item (a), as desired. So we assume that $p < \infty$ below. Given the guarantees from Theorem 4.1, we observe,

$$\begin{aligned}
&\mathbf{E}\left[\sum_{u \in V} \mathrm{dis}_u(\mathcal{P}, E^+, E^-)^p \right] \\
&= \sum_{u \in V} \mathbf{E}[\mathrm{dis}_u(\mathcal{P}, E^+, E^-)^{p-1} \cdot \mathrm{dis}_u(\mathcal{P}, E^+, E^-)] \\
&\lesssim \sum_{u \in V} \mathbf{E}\left[(\mathrm{A}_\infty \cdot y_u)^{p-1} \cdot \mathrm{dis}_u(\mathcal{P}, E^+, E^-) \right] \\
&= \sum_{u \in V} (\mathrm{A}_\infty \cdot y_u)^{p-1} \mathbf{E}\left[\mathrm{dis}_u(\mathcal{P}, E^+, E^-) \right] \\
&\lesssim \sum_{u \in V} (\mathrm{A}_\infty \cdot y_u)^{p-1} \cdot A_1 \cdot y_u = \sum_{u \in V} A^p \cdot y_u^p,
\end{aligned}$$

where $A = \left(\mathrm{A}_\infty^{p-1} \cdot A_1 \right)^{\frac{1}{p}}$. Note that the desired approximation factor is $O(A)$. From Jensen's inequality, it follows that

$$\begin{aligned}
&\mathbf{E}\left[\Big(\sum_{u \in V} \mathrm{dis}_u(\mathcal{P}, E^+, E^-)^p \Big)^{\frac{1}{p}} \right] \le \\
&\qquad \le \left(\mathbf{E}\left[\sum_{u \in V} \mathrm{dis}_u(\mathcal{P}, E^+, E^-)^p \right] \right)^{\frac{1}{p}} \\
&\qquad \lesssim \left(\sum_{u \in V} A^p \cdot y_u^p \right)^{\frac{1}{p}} = A \cdot \|y\|_p.
\end{aligned}$$

This finishes the proof. □

5. Overview of Metric Partitioning Scheme

In this section we describe our partitioning scheme and give a proof overview of Theorem 3.1. More specifically, in Section 5.1 we reduce the problem to choosing a random set of particular interest as stated in Theorem 5.1. In Section 5.2 we describe an algorithm for choosing such a random set and give a proof overview of its correctness.

5.1. Iterative Clustering

Given a metric space (X,d), our partitioning scheme uses an iterative algorithm – Algorithm 2 to obtain $\mathcal{P}$. Let X_t denote the set of not-yet clustered vertices at the start of

Algorithm 2 Metric Space Partitioning Scheme

Input: Metric Space (X, d) and $r, R > 0$, $q \geq 1$.
Define $t = 0$, $X_1 = X$.
repeat
  $t = t + 1$.
  $P_t = \text{Cluster Select}(X_t, d, r, R, q)$.
  $X_{t+1} = X_t \setminus P_t$.
until $X_t = \varnothing$
Output $(P_1, P_2, \ldots, P_t)$.

iteration t of Algorithm 2. At step t, the algorithm finds and outputs random set $P_t \subseteq X_t$. It then updates the set of not-yet clustered vertices ($X_{t+1} = X_t \setminus P_t$), and repeats this step until all vertices are clustered. Algorithm 2 makes use of the following theorem in each iteration to find the random set P_t.

We need the following notation to state the theorem. Let $\delta_P(u, v)$ be the cut metric induced by the set P: $\delta_P(u, v) = 1$ if $u \in P$ and $v \notin P$ or $u \notin P$ and $v \in P$; $\delta_P(u, v) = 0$ if $u \in P$ and $v \in P$ or $u \notin P$ and $v \notin P$. Also, let $\vee_P(u, v)$ be the indicator of the event $u \in P$ or $v \in P$ or both u and v are in P. We denote $[k] = \{1, 2, \ldots k\}$.

Theorem 5.1. *For every $q \geq 1$ there exists a $\beta_q^* = \Theta\big(\frac{1}{q \ln(q+1)}\big) < 1$ such that the following holds. Consider a finite metric space (X, d). Fix two positive numbers r and R such that $\beta = r/R \leq \beta_q^*$. Let $D_\beta = 2(q+1)\ln 1/\beta$. Then, there exists an algorithm for finding a random set P satisfying properties (a), (b), and (c):*

(a) $\operatorname{diam}(P) \leq 2R$ *(always);*

(b) For every point u in X, the following bound holds:

$$\sum_{v \in \text{Ball}(u,R)} \Big(\Pr\{\delta_P(u,v) = 1\} - D_\beta \frac{d(u,v)}{R} \Pr\{\vee_P(u,v) = 1\} \Big)^+ \lesssim \beta^q \cdot \mathbf{E}\left[\sum_{v \in \text{Ball}(u,2R)} \frac{d(u,v)}{R} \cdot \vee_P(u,v) \right].$$

(c) Moreover, for every u in X, we always have

$$\sum_{v \in \text{Ball}(u,r)} \delta_P(u,v) \lesssim \beta \cdot D_\beta^2 \cdot \sum_{v \in \text{Ball}(u,2R)} \frac{d(u,v)}{R} \cdot \vee_P(u,v).$$

Informally, Theorem 5.1 is a "single-cluster" version of Theorem 3.1, and there is a one-to-one correspondence between their properties. In Appendix A, we show that Theorem 3.1 holds for $\mathcal{P}$ if we assume that each partition $P \in \mathcal{P}$ satisfies Theorem 5.1. Thus, to obtain Theorem 3.1, it remains to prove Theorem 5.1.

Algorithm 3 Cluster Select

Input: Metric space (X, d) and $r, R > 0$, $q \geq 1$
Define: $\beta = r/R$, $D_\beta = 2(q+1)\ln 1/\beta$.
Define: $R_0 = R/D_\beta$, $R_1 = R - R_0$, $\rho_q(\beta) = (1/\beta)^{q+1}$.
Select $z = \arg\max_{u \in X} |\operatorname{Ball}(u, R_0)|$.
if $|\operatorname{Ball}(z, R_1)| \geq \rho_q(\beta) \cdot |\operatorname{Ball}(z, R_0)|$ **then**
  Set $P = \operatorname{Ball}(z, R_1)$.
else
  Consider S as stated in Definition 5.2.
  Consider π_S^{inv} as stated in Definition 5.3.
  Let F be the cumulative distribution function stated in Definition 5.4.
  Choose a random $x \in [0, R/2]$ according to F.
  Set $P = \operatorname{Ball}(z, \pi_S^{inv}(x))$.
end if
Output P.

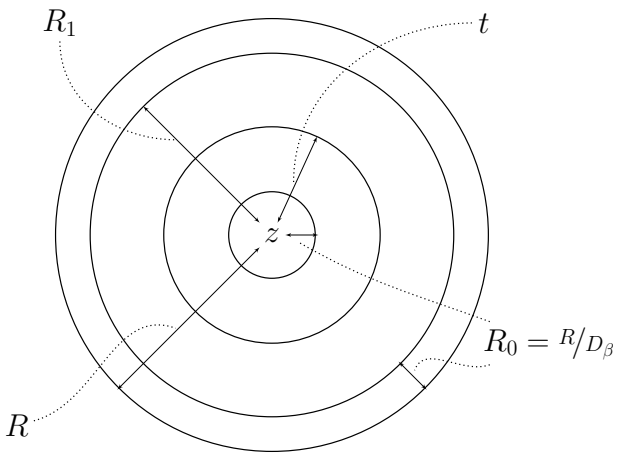

Figure 2. Balls with Different Radii
$R > r > 0$, $q \geq 1$, $\beta = r/R$, $D_\beta = 2(q+1)\ln 1/\beta$,
$R_0 = R/D_\beta$, $R_1 = R - R_0$.

5.2. Selecting a Single Cluster

We will use the following definitions. Let r and R be positive numbers with $r < R$. Define $\beta = r/R \leq \beta_q^*$ and $D_\beta = 2(q+1)\ln 1/\beta$ where $q \geq 1$. Let $R_0 = R/D_\beta$ and $R_1 = R - R_0$. We let $\rho_q(\beta) = (1/\beta)^{q+1}$ (see Figure 2). We choose β_q^* so that $r < R_0 < R$ (see Appendix B for details).

Given a metric space (X, d) and parameters r and R, our procedure for finding a random set $P \subseteq X$ begins by finding a pivot point z with a densely populated neighborhood – namely, z is chosen such that a ball of radius R_0 around z contains the maximum number of points. More formally,

$$z = \arg\max_{u \in X} |\operatorname{Ball}(u, R_0)|. \tag{5.1}$$

We refer to this ball of small radius around z as the "core" of the cluster. Our choice of the pivot z is inspired by the papers by Charikar et al. (2003); Puleo & Milenkovic (2018); Charikar et al. (2017). We then consider a ball of large radius R_1 around the pivot z and examine the following two cases – "Heavy Ball" and "Light Ball". If this ball of large radius around z is sufficiently populated, that is, if the number of points in $\mathrm{Ball}(z, R_1)$ is at least $(1/\beta)^{q+1}$ times the number of points in the core, we call this case "Heavy Ball". In the case of Heavy Ball, we will show that $P = \mathrm{Ball}(z, R_1)$ (a ball around z of radius slightly less than R) satisfies the properties of Theorem 5.1. In the case of "Light Ball", the ball of large radius around z is not sufficiently populated. In this case, the algorithm finds a radius t ($t \leq R$) such that $P = \mathrm{Ball}(z, t)$ satisfies the properties of Theorem 5.1. In the following subsections we provide an overview of the proof for these two cases. A formal proof of Theorem 5.1 can be found in Appendix B.

5.2.1. Heavy Ball

The Heavy Ball P is a ball of radius R_1 around z which contains many points. As the diameter of P is $2R_1 < 2R$, it is easy to see that a Heavy Ball satisfies property (a) of Theorem 5.1. We now focus on showing that properties (b) and (c) hold for Heavy Ball. Observe as z was chosen according to (5.1), for every point $u \in X \setminus \{z\}$, u has a less populated neighborhood of radius R_0 than that of z. This combined with the fact that $\mathrm{Ball}(z, R_1)$ is heavy, implies that for every u, there are sufficiently many points in P at a distance of at least R_0 from u. Thus, for any point $u \in X$, we can expect the sum of distances between u and the points in P to be large. In fact, we show that the left hand sides of properties (b) and (c) can be charged to $\sum_{v \in P} \frac{d(u,v)}{R}$, the sum of distances between u and the points in P. For points u such that $d(z, u) \leq R$, $P \subseteq \mathrm{Ball}(u, 2R)$ and hence, $\sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \vee_P (u, v) \geq \sum_{v \in P} \frac{d(u,v)}{R}$. Thus, for every $u \in X$, we can charge the left hand sides of properties (b) and (c) to the quantity $\sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u,v)}{R} \vee_P (u, v)$. This allows us to conclude that a Heavy Ball satisfies Theorem 5.1.

5.2.2. Light Ball

In this subsection, we consider the case of $|\mathrm{Ball}(z, R_1)| < \rho_q(\beta) \cdot |\mathrm{Ball}(z, R_0)|$, which we call Light Ball. In the case of Light Ball, we choose a random radius $t \in (0, R_1]$ and set $P = \mathrm{Ball}(z, t)$. Observe that property (a) of Theorem 5.1 holds trivially since the radius $t < R$.

Now consider property (c) of Theorem 5.1. Recall that for every point $u \in X$, property (c) gives a bound on the total number of points separated from u (by P) residing in a small ball $\mathrm{Ball}(u, r)$, i.e., $\sum_{v \in \mathrm{Ball}(u,r)} \delta_P(u, v)$. Note that property (c) gives a deterministic guarantee on P. Therefore, we choose a random radius $t \in (0, R_1]$ from the set of all radii for which property (c) of Theorem 5.1 holds. More specifically, we define the following set.

Definition 5.2. *Let S be the set of all radii s in $(3R_0, R_1]$ such that for every $u \in X$ set $P = \mathrm{Ball}(z, s)$ satisfies:*

$$\sum_{v \in \mathrm{Ball}(u,r)} \delta_P(u, v) \leq \leq 25\beta \cdot D_\beta^2 \cdot \sum_{v \in \mathrm{Ball}(u,2R)} \frac{d(u, v)}{R} \cdot \vee_P(u, v). \quad (5.2)$$

The set S can be computed in polynomial time since the number of distinct clusters $P = \mathrm{Ball}(z, t)$ is upper bounded by the size of the metric space, $|X|$. By the same token, S is a finite union of disjoint intervals.

Now we show why we can expect the set S to be large. Consider $P = \mathrm{Ball}(z, s)$ such that $s \in S$. As S is computed according to Definition 5.2, it implies that the boundary of P is somewhat sparsely populated – as for every $u \in X$, it bounds the number of points within a small neighborhood of $\mathrm{Ball}(u, r)$ that are separated from u (note that $\sum_{v \in \mathrm{Ball}(u,r)} \delta_P(u, v)$ is trivially 0 for points u that are not close to the boundary of P). Since $\mathrm{Ball}(z, R_1)$ does not contain many points, the number of points in $\mathrm{Ball}(z, s')$ cannot grow too quickly as we increase the radius s' from 0 to R_1. This suggests that for many of such radii s', the ball $P = \mathrm{Ball}(z, s')$ has a sparsely populated boundary, and hence the set S should be large. In fact, we use the above argument to show that the Lebesgue measure of the set S satisfies $\mu(S) \geq R/2$. This will allow us to define a continuous probability distribution on S.

What remains to be shown is that for a random radius $t \in S$, the set $P = \mathrm{Ball}(z, t)$ satisfies property (b) of Theorem 5.1. For this purpose we define a measure preserving transformation π_S that maps an arbitrary measurable set S to the interval $[0, \mu(S)]$.

Definition 5.3. *Consider a measurable set $S \subset [0, R]$. Define function $\pi_S : [0, R] \to [0, \mu(S)]$ as follows $\pi_S(x) = \mu([0, x] \cap S)$. Also, for $y \in [0, \mu(S)]$, let*

$$\pi_S^{inv}(y) = \min\{x : \pi_S(x) = y\}.$$

Recall that the set S stated in Definition 5.2 is a finite union of disjoint intervals. In this case, what π_S does is simply pushing the intervals in S towards 0, and thus, allowing us to treat the set S as a single interval $[0, \mu(S)]$. For the rest of the proof overview, we assume that $S = [0, \mu(S)]$ and π_S is the identity. This simplifies the further analysis of Theorem 5.1 immensely.

Next, we define a cumulative distribution function F on $[0, R/2] \subseteq [0, \mu(S)]$:

Definition 5.4. *Let $F : [0, R/2] \to [0, 1]$ be a cumulative distribution function such that*

$$F(x) = \frac{1 - e^{-x/R_0}}{1 - e^{-R/2R_0}}. \tag{5.3}$$

We choose a random $x \in [0, R/2]$ according to F and set $P = \mathrm{Ball}(z, \pi_S^{inv}(x))$ (see Algorithm 3). Since we assume in this proof overview that π_S is the identity, $P = \mathrm{Ball}(z, x)$. Now, we show that the radius x chosen in such a manner guarantees that the cluster P satisfies property (b). Loosely speaking, the motivation behind our particular choice of cumulative distribution function F is the following: For two points $u, v \in X$, function F bounds the probability of u and v being separated by P, in terms of D_β times the probability that either u or v lies in P. Unfortunately, this bound does not hold for points u with $d(z, u)$ close to $R/2$. However, the choice of parameters for function F in Definition 5.4 gives us two desired properties. Without loss of generality assume that $d(z, u) \le d(z, v)$. Then, the probability that P separates the points u and v, $\Pr(\delta_P(u, v)) = \Pr(d(z, u) \le x \le d(z, v)) = F(d(z, v)) - F(d(z, u))$. Moreover, as $d(z, u) \le d(z, v)$, the probability that either u or v lies in P, $\Pr(\vee_P(u, v)) = 1 - F(d(z, u))$. Thus, choosing F according to Definition 5.4 ensures:

- (Property I) $F(d(z, v)) - F(d(z, u))$ is bounded in terms of D_β times $1 - F(d(z, u))$ (Please see Claim C.8 for a formal argument).
- (Property II) The probability that the cluster P includes points u such that $d(z, u) > R/2 - R_0$, is very small (please see Claim C.7).

In fact, (Property II) of function F is the reason why we are able to guarantee that property (b) satisfies (3.1) only on average, with the error term coming from our inability to guarantee (3.1) for points on the boundary. We refer the reader to Section C.2 for a formal proof. Thus, we conclude the case of Light Ball and show that it satisfies Theorem 5.1.

Acknowledgements

Jafar Jafarov and Yury Makarychev were supported by NSF CCF-1718820, CCF-1955173, and NSF TRIPODS CCF-1934843/CCF-1934813. Sanchit Kalhan and Konstantin Makarychev were supported by NSF CCF-1955351 and NSF TRIPODS CCF-1934931.

References

Ahmadi, S., Khuller, S., and Saha, B. Min-max correlation clustering via multicut. In *Proceedings of the Conference on Integer Programming and Combinatorial Optimization*, pp. 13–26, 2019.

Ailon, N., Charikar, M., and Newman, A. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23, 2008.

Bansal, N., Blum, A., and Chawla, S. Correlation clustering. *Machine learning*, 56(1-3):89–113, 2004.

Bartal, Y. Probabilistic approximation of metric spaces and its algorithmic applications. In *Proceedings of 37th Conference on Foundations of Computer Science*, pp. 184–193. IEEE, 1996.

Bonchi, F., García-Soriano, D., and Liberty, E. Correlation clustering: from theory to practice. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1972, 2014.

Călinescu, G., Karloff, H., and Rabani, Y. An improved approximation algorithm for multiway cut. *Journal of Computer and System Sciences*, 60(3):564–574, 2000.

Charikar, M., Guruswami, V., and Wirth, A. Clustering with qualitative information. In *Proceedings of the Symposium on Foundations of Computer Science*, 2003.

Charikar, M., Gupta, N., and Schwartz, R. Local guarantees in graph cuts and clustering. In *Proceedings of the Conference on Integer Programming and Combinatorial Optimization*, pp. 136–147, 2017.

Chawla, S., Makarychev, K., Schramm, T., and Yaroslavtsev, G. Near optimal LP rounding algorithm for correlation clustering on complete and complete k-partite graphs. In *Proceedings of the Symposium on Theory of Computing*, pp. 219–228, 2015.

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

Cohen, W. W. and Richman, J. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 475–480, 2002.

Demaine, E. D., Emanuel, D., Fiat, A., and Immorlica, N. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2-3):172–187, 2006.

Fakcharoenphol, J., Rao, S., and Talwar, K. A tight bound on approximating arbitrary metrics by tree metrics. *Journal of Computer and System Sciences*, 69(3):485–497, 2004.

Jafarov, J., Kalhan, S., Makarychev, K., and Makarychev, Y. Correlation clustering with asymmetric classification errors. In *Proceedings of the International Conference on Machine Learning*, pp. 4641–4650, 2020.

Kalhan, S., Makarychev, K., and Zhou, T. Correlation clustering with local objectives. In *Advances in Neural Information Processing System*, pp. 9341–9350, 2019.

Puleo, G. J. and Milenkovic, O. Correlation clustering and biclustering with locally bounded errors. *IEEE Transactions on Information Theory*, 64(6):4105–4119, 2018.

Ramachandran, A., Feamster, N., and Vempala, S. Filtering spam with behavioral blacklisting. In *Proceedings of the Conference on Computer and Communications Security*, pp. 342–351, 2007.

Tang, S., Andres, B., Andriluka, M., and Schiele, B. Multi-person tracking by multicut and deep matching. In *Proceedings of the European Conference on Computer Vision*, pp. 100–111, 2016.

Tang, S., Andriluka, M., Andres, B., and Schiele, B. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 3539–3548, 2017.

van Zuylen, A., Hegde, R., Jain, K., and Williamson, D. P. Deterministic pivoting algorithms for constrained ranking and clustering problems. In *Proceedings of the Symposium on Discrete Algorithms*, pp. 405–414, 2007.

Wirth, A. Correlation clustering. In *Encyclopedia of Machine Learning*, pp. 227–231. Springer, 2010.