Learning Hierarchical Cluster Structure of Graphs in Sublinear Time

Michael Kapralov
EPFL

Akash Kumar
EPFL

Silvio Lattanzi
Google Research

Aida Mousavifar
Google

Abstract

Learning graph cluster structure using few queries is a classical question in property testing, with the fundamental special case, namely expansion testing, considered in the seminal work of Goldreich and Ron[STOC'96]. The most recent results in this line of work design *clustering oracles* for (k, ϵ)-clusterable graphs, which are graphs that can be partitioned into k induced expanders with outer conductance bounded by $\epsilon \ll 1$. These oracles, given a graph whose vertex set can be partitioned into a disjoint union of k clusters (i.e., good expanders) with outer conductances bounded by $\epsilon \ll 1$, provide query access to an $O(\epsilon \log k)$-approximation to this ground truth clustering in time $\approx 2^{\text{poly}(k/\epsilon)} n^{1/2+O(\epsilon)}$ per query.

Motivated by the rising interest in learning hierarchical structures in large networks, in this paper we introduce (k, γ)-*hierarchically clusterable* graphs, a natural hierarchical analog of classical (k, ϵ)-clusterable graphs; intuitively, these are graphs that exhibit pronounced hierarchical structure. We give a *hierarchical clustering oracle* for this model, i.e. a small space data structure that provides query access to a good hierarchical clustering at cost $\approx \text{poly}(k) \cdot n^{1/2+O(\gamma)}$ per query; notably, the dependence on k is polynomial, in contrast to best known flat clustering oracles. The result relies on several structural properties of hierarchically clusterable graphs that we hope will be of independent interest in sublinear time spectral graph algorithms.

Contents

1 Introduction

Clustering graph data is an important algorithmic problem. It has applications in a wide variety of scientific disciplines from graph analysis to social science, statistics and more. The overall objective in these problems is to partition the vertex set of the graph into vertex disjoint subgraphs where each of the subsets induce a “well connected” graph and such that the subgraphs are sparsely connected to each other. A classically motivated measure for evaluating cluster quality uses the notion of *conductance*. One natural graph clustering objective motivated by conductance considers problem of partitioning the vertices of the graph into subsets (called clusters) which have large inner conductance and a sparse edge boundary. Many efficient algorithms [KVV04, NJW02, SM00, VL07] have been discovered for graph clustering which use this objective, many of them relying on spectral techniques. Motivated by applications in big data analysis, a lot of recent research has focused on developing sublinear time algorithms [CS04, MOP01, GKL$^+$21] to cluster graph data. Such algorithms can typically answer queries about the clustering without computing it explicitly at any point in time.

In this paper, we focus on the popular version of the problem where one assumes the existence of a planted solution, namely that the input graph $G = (V, E)$ admits a partitioning into a disjoint union of k induced expanders $C_1, \ldots, C_k$ with outer conductance bounded by $\epsilon \ll 1$. We refer to such instances as *k-clusterable* graphs and we (informally) define the *flat-clustering* problem as the task of recovering an approximation to $C_1, \ldots, C_k$ that is correct up to a small missclassification error on every cluster. This problem has been extensively studied in the property testing framework as well as local computation models. Its testing version, where one essentially wants to determine k, the number of clusters in G, in sublinear time, generalizes the well-studied problem of testing graph expansion, where one wants to distinguish between an expander (i.e. a good single cluster) and a graph with a sparse cut (i.e. at least two clusters). Goldreich and Ron [GR11] showed that expansion testing requires $\Omega(n^{1/2})$ queries, then [CS07, KS08, NS10] developed algorithms to distinguish an expander from a graph that is far from a graph with conductance ϵ in time $\approx n^{1/2+O(\epsilon)}$, which the recent work of [CKK$^+$18] showed to be tight. The setting of $k > 2$ has seen a lot of attention recently [CPS15, CKK$^+$18, Pen20, GKL$^+$21]. The latest result is due to [GKL$^+$21], where the authors design a *clustering oracle*, i.e. small space data structure that allows fast query access to the clustering. The clustering oracle of [GKL$^+$21] recovers every cluster up to $O(\epsilon \cdot \log k)$ misclassification error and only needs $\approx n^{1/2+O(\epsilon)}$ preprocessing and query time, which is close to optimal due to the aforementioned lower bound of [CKK$^+$18]. The focus of this paper is to understand the power of sublinear algorithms in discovering the *hierarchical* cluster structure of k-clusterable graphs.

The *hierarchical clustering* problem is the task of partitioning vertices of a graph into nested clusters, naturally represented by a rooted tree whose leaves correspond to the vertices and whose internal nodes represent clusters of their descendant leaves (such a tree is referred to as a hierarchical tree). Dasgupta [Das16] introduced a hierarchical clustering objective function and thereby initiated a line of work on developing algorithms that optimize the cost of the hierarchical tree. The main question we address in this work is:

> Is it possible to design a clustering oracle that recovers the underlying hierarchical structure of a k-clusterable graph in sublinear time (i.e., a *hierarchical clustering oracle*)?

Hierarchically-clusterable graphs: We say that graph G is *hierarchically-clusterable* if there exists a nested sequence of partitions into clusters of increasing outer conductance, each partition refining the previous one.

In particular, for every cluster S that belongs to the partition at level h, we assume that $\phi_{\text{in}}(S) \geq \varphi_h$ and $\phi_{\text{out}}(S) \leq O(\varphi_{h-1})$. Therefore, φ_h increases as we move from coarse partitions towards refined partitions. We say that graph G is (k, γ)-*hierarchically-clusterable* if the last partition (i.e the most refined one) consists of k clusters with constant inner conductance and the conductance of clusters increases by about a factor $1/\gamma$ as we go down the tree level by level, i.e., the conductance of clusters at level h is at least $\varphi_h = \varphi_0/\gamma^h$. Moreover, each cluster gets split into few subclusters of comparable size (see Definition 6 for the formal version).

We show that it is possible to recover the hierarchical structure of well-separated hierarchically clusterable graphs sufficiently precisely to obtain a constant factor approximation to Dasgupta cost:

Theorem 1. *[Informal version of Theorem 2] For sufficiently small constant $\gamma \in (0, 1)$ there exists a* **hierarchical clustering** *oracle with $\approx k^{O(1)} n^{1/2+O(\gamma)}$ preprocessing time and $\approx k^{O(1)} n^{1/2+O(\gamma)}$ query time that achieves a constant factor approximation to Dasgupta cost on (k, γ)-hierarchically clusterable graphs.*

In what follows we formally define our model and formally state the main result, then give an overview of the technical contributions, and finally give the detailed proofs.

1.1 Problem Statement and Main Definitions

We start by defining

Definition 1 (Clustering)**.** For a graph $G = (V, E)$ a *clustering* of G is a collection of disjoint subsets of the vertex set V of G.

Note that our notion of clustering allows for outliers, i.e. a clustering of G is not necessarily a partition of V. The main object of study in this paper is

Definition 2 (Hierarchical clustering)**.** A hierarchical clustering of a graph $G = (V, E)$ is a sequence $\mathcal{P} = (\mathcal{P}^0, \ldots, \mathcal{P}^H)$ of nested clusterings of V, where $\mathcal{P}^0 = \{V\}$. We say that the sequence $(\mathcal{P}^0, \ldots, \mathcal{P}^H)$ is nested if for every $h \in [H]$ and every $S \in \mathcal{P}^h$ there exists a unique $S^* \in \mathcal{P}^{h-1}$ such that $S \subseteq S^*$. We call such an S^* the *parent* of S in $\mathcal{P}$ and write $S^* = \textsc{Parent}_{\mathcal{P}}(S)$, often omitting the subscript when it is clear from context.

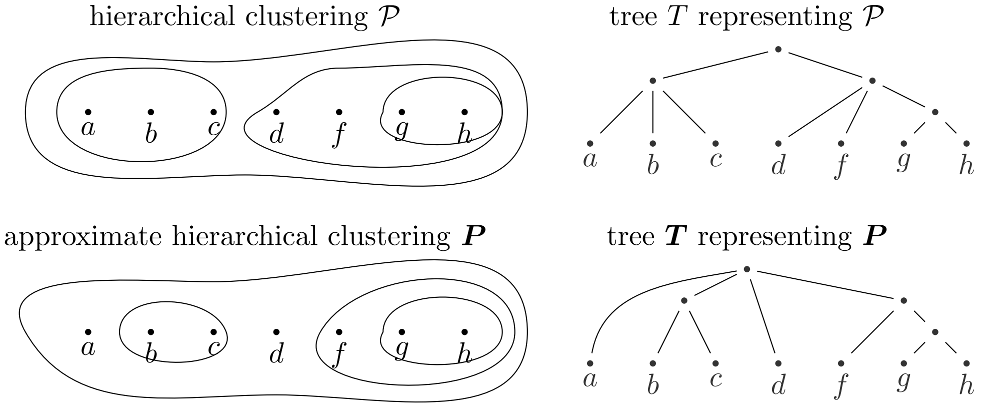
We write $S \in \mathcal{P}$ if $S \in \mathcal{P}^h$ for some $h \in [H]$. We say that a hierarchical clustering $\mathcal{P}$ is isomorphic to a hierarchical clustering $\boldsymbol{P}$ if there exists a one-to-one mapping σ of sets in $\mathcal{P}$ to sets in $\boldsymbol{P}$ such that if for some $S \in \mathcal{P}$ and $\mathbf{S} \in \boldsymbol{P}$ one has $\sigma(S) = \mathbf{S}$, then $\sigma(\textsc{Parent}_{\mathcal{P}}(S)) = \textsc{Parent}_{\boldsymbol{P}}(\mathbf{S})$.

Definition 3 (Tree representation of a hierarchical clustering)**.** Note that hierarchical clusterings $\mathcal{P} = (\mathcal{P}^0, \ldots, \mathcal{P}^H)$ are in one-to-one correspondence with rooted trees T whose leaves are vertices in V and internal nodes are clusters in $\mathcal{P}^h$ for some $h \in [H]$. See Fig. 1 for an illustration.

Dasgupta introduced a natural optimization framework for formulating hierarchical clustering tasks as an optimization problem [Das16]. We recall this framework now. Let T be any rooted tree whose leaves are vertices of the graph. For any node u of T, let $T[u]$ be the subtree rooted at u, and let $\textsc{leaves}(T[u]) \subseteq V$ denote the leaves of this subtree. For leaves $x, y \in V$, let $\text{LCA}(x, y)$ denote the lowest common ancestor of x and y in T. In other words, $T[\text{LCA}(x, y)]$ is the smallest subtree whose leaves contain both x and y.

Definition 4. (Dasgupta's cost [Das16]) Dasgupta's cost of the tree T for the graph $G = (V, E)$ is defined to be $\text{COST}(T) = \sum_{\{x,y\}\in E} |\textsc{leaves}(T[\text{LCA}(x, y)])|$.

We will sometimes write $\text{COST}(\mathcal{P})$ for a hierarchical clustering $\mathcal{P}$ to denote $\text{COST}(T)$ for the tree T corresponding to $\mathcal{P}$.


hierarchical clustering $\mathcal{P}$

tree T representing $\mathcal{P}$

approximate hierarchical clustering $\boldsymbol{P}$

tree $\boldsymbol{T}$ representing $\boldsymbol{P}$

Figure 1: A hierarchical clustering $\mathcal{P}$ of $\{a, b, c, d, e, f, g, h\}$ and the tree T representing $\mathcal{P}$ together with an approximate clustering $\boldsymbol{P}$ and its tree $\boldsymbol{T}$.

After its introduction, Dasgupta's cost received a lot of attention from the research community [CC17, MW17, CKMM19, CCN19] and a $O(\sqrt{\log n})$ algorithm is known for the problem [CC17]. Furthermore, [CC17] showed that it's impossible to approximate Dasgupta's cost within a constant factor in general graphs under the Small-Set Expansion hypothesis [CNC18]. A very interesting recent line of research started in [CKM17] and continued in [MS21] studies the approximability of Dasgupta's cost on the class of random graphs (akin to the stochastic block model) that exhibit planted hierarchical structure. For example, in [CKM17] the authors consider the hierarchical stochastic block model and prove that it is possible to obtain better approximation for it.

In this paper we initiate the study of sublinear time hierarchical clustering algorithms. We define a notion of hierarchically clusterable graphs, which is essentially a class of graphs with pronounced hierarchical clustering structure, and give a sublinear time hierarchical clustering oracle for this model. We consider d-regular graphs throughout the paper[1], and parameterize the cluster structure of the graph using the notions of inner and outer conductance, defined below and our main object of study is defined after:

Definition 5. (Inner and outer conductance) Let $G = (V, E)$ be a d-regular graph. For a set $C \subseteq V$ and a set $S \subseteq C$, let $E(S, C \setminus S)$ be the set of edges with one endpoint in S and the other in $C \setminus S$. The *conductance of S within C* is $\phi^G_C(S) = \frac{|E(S,C\setminus S)|}{d\cdot|S|}$. The *outer conductance* of C is defined to be $\phi^G_{\text{out}}(C) = \phi^G_V(C) = \frac{|E(C,V\setminus C)|}{d\cdot|C|}$. The *inner conductance* of $C \subseteq V$ is defined to be $\phi^G_{\text{in}}(C) = \min_{S\subseteq C, 0<|S|\leq\frac{|C|}{2}} \phi^G_C(S)$ if $|C| > 1$ and one otherwise.

Definition 6 ((k,γ)-hierarchically clusterable graphs)**.** A graph $G = (V, E)$ is (k, γ)-hierarchically clusterable if there exists a hierarchical clustering $\mathcal{P} = (\mathcal{P}^0, \ldots, \mathcal{P}^H)$ such that: (i) every $\mathcal{P}^h$ is a partition of V; (ii) the bottom level partition $\mathcal{P}^H$ contains exactly k sets; (iii) for every level $h \in [H]$ and for every $S \in \mathcal{P}^h$ we have $\phi^G_{\text{in}}(S) \geq \varphi_h$ and $\phi^G_{\text{out}}(S) \leq O(\varphi_{h-1})$, where $\varphi_{h-1} = \gamma\cdot\varphi_h$ for $2 \leq h \leq H$ and $\varphi_0 \leq \gamma \cdot \varphi_1$. We let $\varphi = \varphi_H$ denote a lower bound on the inner conductance of the k base clusters.

We note that setting $H = 1$, we immediately recover the classical notion of a (k, ϵ)-clusterable graph, i.e. a graph that admits a partition into k induced expanders each with outer conductance upper bounded by $\epsilon \in (0, 1)$ (in our case we will have $\epsilon = \gamma$). This is a natural worst case (i.e., robust) analog of the stochastic block model, and has been studied extensively in the literature [CKM17]. Similarly, our model is a natural robust analog of the hierarchical stochastic block model. We show in Section A that a natural class of random graphs with hierarchical

[1] Note that if the input graph is bounded degree, it can be turned into a d-regular graph by adding an appropriate number of self-loops; this of course changes the notion of conductances somewhat.

structure akin to the stochastic block model is (k,γ)-clusterable as per our definition. To simplify notation throughout the paper we assume that the conductance of the base clusters satisfies $\varphi \geq \Omega(\gamma^{1/20})$, and that clusters in the ground truth hierarchical clustering $\mathcal{P}$ get partitioned into constant number of subclusters of comparable size, i.e., for every $S^* \in \mathcal{P}$ and every child S of S^* we have $|S| \geq \beta|S^*|$ for some $\beta \in (0,1)$. We assume $\beta \geq \Omega(\gamma^{1/30})$ [2].

We now introduce a natural notion of approximation for hierarchical clusterings.

Definition 7. (D-approximation of a hierarchical clustering) A hierarchical clustering $\boldsymbol{P}$ is a D-approximation of a hierarchical clustering $\mathcal{P}$ if (i) $\mathcal{P}$ is isomorphic to $\boldsymbol{P}$ (denote the isomorphism by $\sigma : \mathcal{P} \to \boldsymbol{P}$) and if (ii) For every every cluster $S \in \mathcal{P}$ at level $h \in [H]$ we have $|S \triangle \boldsymbol{S}| \leq D \cdot \varphi_{h-1} \cdot |S|$. where here and below for a cluster $S \in \mathcal{P}$ we use boldface $\boldsymbol{S}$ to denote its image under the isomorphism σ, i.e., let $\boldsymbol{S} = \sigma(S)$.

Before describing our main result it is interesting to describe few interesting properties of our definition of D-approximation of a hierarchical clustering.

Why the notion of (k,γ)-hierarchically clusterable is natural (i.e., why are instances with planted solutions interesting). The study of graph clustering with planted solutions has received a lot of attention from the research community on stochastic block model (see, e.g. [Abb18] and references therein) as well as property-testing literature and sublinear algorithms. The latter line of work was initiated by Goldreich and Ron [GR11] and has been extended to various graph clustering algorithms that aim to recover the underlying ground-truth solution [KPS08, NS10, CPS15, CKK$^+$18, Pen20, GKL$^+$21]. Another prominent example is the Nibble algorithm of [ST13] (with many follow-up works, e.g. [AGPT16], providing improvements) that guarantees finding sparse cuts in graphs when the sparsity is quite pronounced, yet works very well and is commonly used in practice, where the sought-after cuts are not nearly as sparse as the analysis requires. Our (k,γ)-hierarchically clusterable model is a natural extension of (k,φ)-clusterable graphs, and we hope that it will lead to interesting algorithmic insights into the hierarchical clustering problem.

Regularity assumption. Throughout the paper we consider d-regular graphs, which is a clean and standard setting that has been considered in the property testing literature [GR11, CPS15] as well as graph clustering community [GKL$^+$21]. Some of the results in testing cluster structure generalize to irregular graphs (e.g., [CKK$^+$18]), and ours probably does too. However, for simplicity in this paper we work with d-regular graphs.

Why the notion of D-approximate hierarchical clustering is natural. Note that the definition above is a natural extension of what is achievable in flat clustering that has been studied extensively (e.g. see [KVV04, NJW02]), where if every cluster has outer conductance φ_{out} and inner conductance φ_{in}, one expects to be able to classify all but $O(\varphi_{\text{out}}/\varphi_{\text{in}}^2))$ fraction of vertices per cluster (and the result of [GKL$^+$21] gives an algorithm that misclassifies at most $O(\varphi_{\text{out}} \log k)$ fraction of vertices). The guarantees of Definition 7 are quite a bit stronger, however, as we do not divide the outer conductance φ_{h-1} by the inner conductance of the cluster at the same level. Instead, the loss parameter D will be set to a polynomial in the inner conductance of the *base clusters* at level H.

Why significantly better approximation cannot be achieved. Note that one cannot achieve a significantly better reconstruction quality. For example, suppose that $H = 2$ and our input graph is a union of two induced expanders C_1, C_2, as well as a disjoint set Q of $\approx \varphi_{\text{out}} n$ vertices each of which has $d/2$ random neighbors in C_1 and $d/2$ random neighbors in C_2. Then with high probability the corresponding graph is (k,γ)-hierarchically clusterable, but

[2] Note that in our definition of (k,γ)-clusterable graphs we assume a lower bound on the conductance of the base clusters $\varphi = \varphi_H$ as a function of γ, as well as a lower bound on the worst case ratio β of the size of a child cluster to the size of the parent cluster, again as a function of γ. This is to alleviate notation in the rest of the paper. In particular, for a constant $c > 0$, we use the notation $O_{\beta,\varphi}(\gamma^c)$ to suppress factors of $\left(\frac{1}{\beta \cdot \varphi}\right)^{O(1)}$.

the hierarchical clustering is not unique: the set Q of 'outliers' can be assigned either to the first or the second cluster. Thus, one cannot recover the clustering up to a better than $O(\varphi_{\text{out}})$ precision.

Approximate hierarchical clustering vs Dasgupta cost. Our notion of approximation for hierarchical clusterings is strong enough to imply approximation of Dasgupta cost of corresponding trees:

Lemma 1. *Let $G = (V, E)$ be a (k, γ)-hierarchical clusterable graph and let $\mathcal{P}$ be the hierarchical-clustering. If $\boldsymbol{P}$ is a D-approximation of $\mathcal{P}$, then $COST(\boldsymbol{P}) \leq O\left(\frac{D}{\beta}\right) COST(\mathcal{P})$.*

In particular, a D-approximate hierarchical clustering is also a good approximation to the *optimal tree* in terms of Dasgupta cost:

Lemma 2. *Let $G = (V, E)$ be a (k, γ)-hierarchical clusterable graph and let $\mathcal{P}$ be the hierarchical-clustering. Suppose that $\phi_{in}(G) \geq \varphi_0$. Let $\mathcal{P}^*$ be a hierarchical clustering of the graph G that minimizes Dasgupta cost, then $COST(\mathcal{P}) \leq O\left(\frac{1}{\beta^2}\right) \cdot COST(\mathcal{P}^*)$.*

The above lemmas are close to best possible: in order to approximate Dasgupta cost to within a constant factor, a clustering essentially needs to be $O(1)$-approximate as per Definition 7. Indeed, suppose that the hierarchical clustering $\mathcal{P}$ of G is balanced and b-ary in the sense that at all levels clusters are partitioned into $b \geq 2$ equal size subclusters, and the conductance of the base clusters φ is constant. That way we have $H = \log_b k$, and the base clusters are of size n/k. Suppose further that $\varphi_h = \gamma^{H-h} \cdot \varphi$, where $\gamma = 1/b + \epsilon$ for every $h = 1, \ldots, H$, for a constant $\epsilon > 0$. One can verify that the Dasgupta cost of this instance is dominated by the cost of the root cut, and that any hierarchical clustering $\boldsymbol{P}$ that yields a constant factor approximation to Dasgupta cost on this instance must misclassify at most $O(\varphi_0)$ fraction of vertices across the root cut, i.e. be $O(1)$-approximate in the sense of Definition 7. Similar instances can be constructed in which Dasgupta cost is dominated by level h cuts for every $h \in \{1, \ldots, H\}$.[3]

1.2 Main Result

We are now ready to state our main result.

Definition 8. (Hierarchical clustering oracle) A randomized algorithm $\mathcal{O}$ is a D-approximate (k, γ)-hierarchical clustering oracle if, when given query access to a d-regular graph $G = (V, E)$ that admits a (k, γ)-hierarchical clustering $\mathcal{P}$, the algorithm $\mathcal{O}$ with high probability provides consistent query access to a D-approximate hierarchical clustering $\boldsymbol{P}$.

In the definition above the D-approximate clustering $\boldsymbol{P}$ is only a function of G and the random seed of the oracle. Our main result is such an oracle:

Theorem 2. *For every integer $k \geq 2$, every $H \in O(\log k)$, every $\beta, \varphi \in (0, 1)$, every $\gamma \leq O(\min(\varphi^{20}, \beta^{30}))$, every graph $G = (V, E)$ that admits (k, γ)-hierarchical clustering, there exists a D-approximate hierarchical clustering oracle (Definition 8) with $D = O\left(\frac{1}{\beta^4 \cdot \varphi^2}\right)$ such that*

- *has $\widetilde{O}\left((dk)^{O(1)} \cdot n^{1/2 + O_{\beta,\varphi}(\gamma)}\right)$ preprocessing time, query time, and space.*
- *uses $\widetilde{O}\left(k^{O(1)} \cdot n^{O_{\beta,\varphi}(\gamma)}\right)$ random bits.*

The dependency of running time on d is unnecessary and is an artifact of the running time of Algorithm 13 (from [CKK$^+$18]) that counts the number of clusters at every level. This could easily be adapted to run in time $\approx n^{1/2 + O(\gamma)}$ (without any dependency on d) for d-regular graphs. However, for the simplicity of presentation of Section F we didn't optimize factors d.

[3] One must note here that this particular instance does not satisfy our assumption that the ratio of child cluster size to parent cluster size is lower bounded by $\beta = \Omega(\gamma^{1/30})$, and a slight relaxation of our notion of D-approximate hierarchical clusterings would suffice under this assumption. However, such a relaxation would not be as clean as our notion of approximation (Definition 7).

1.3 Related Work

We briefly review developments in the area of algorithms for hierarchical clustering since the introduction of Dasgupta's objective function. Dasgupta designed an algorithm based on recursive sparsest-cut that provides $O(\log^{3/2} n)$ approximation for his objective function. This was improved by Charikar and Chatizafratis who showed that the recursive sparsest-cut algorithm already returns a tree with approximation guarantee $O(\sqrt{\log n})$ [CC17]. Furthermore, they showed that it's impossible to approximate Dasgupta's cost withing constant factor in general graphs under the Small-Set Expansion hypothesis. [CNC18] presents algorithms with improved approximations to Dasgupta Cost when the underlying hierarchy satisfies some nice constraints. [MW17, CKMM19, CCN19] give algorithms for hierarchical clustering that consider maximization variants of Dasgupta Objective.

Relation to work on hierarchical clustering of clusterable graphs. A related hierarchical version of the stochastic block model was studied in [CKM17]. In [CKM17] the input graph is generated by including every edge independently with probability that depends on the least common ancestor of the endpoints in an underlying hierarchical clustering of the vertices (edges whose least common ancestor is closer to the root have lower probability, corresponding to our notion of the outer conductance φ_h of level h cuts decreasing with h).

Furthermore for the algorithm of [CKM17] to work, the input graphs has to be very dense (at least $\approx \sqrt{n}$ vertex degrees). Their algorithm also does not recover the underlying tree (in fact, they do not have a separation assumption on the level conductances), and does not operate in sublinear time. Thus, the work is related in spirit, but not technically comparable to ours.

In a very recent and concurrent work [MS21] propose a new algorithm that obtains a $O(k^{22}/\varphi^{10})$ approximation of the Dasgupta cost assuming that the input graph is *k-clusterable.* Our algorithm differs on their work on several aspect. Our algorithm works in sublinear time and recovers the underlying tree (in fact, also in their case they do not have a separation assumption on the level conductances), our algorithm also obtains a better approximation factor although under different assumptions (they assume a larger gap between inner and outer conductance at the base level but do not have a separation assumption on the level conductances).

Two recent works, [AKLP22, ACL$^+$22] consider the problem of hierarchical clustering under Dasgupta objective in the streaming model. Both papers give a one pass $\widetilde{O}(n)$ memory streaming algorithm which finds a tree with Dasgupta cost within an $O(\sqrt{\log n})$ factor of the optimum in polynomial time. Additionally, [AKLP22] also considers this problem in the query model and presents a $O(\sqrt{\log n})$ approximate hierachical clustering using $\widetilde{O}(n)$ queries without making any clusterability assumptions of the input graph. On the other hand, our algorithms assume the graph is hierachically clusterable and run in sublinear time.

Relation to work on k-clusterable graphs. The class of *k-clusterable* graphs, i.e. graphs that can be partitioned into k expanders with small outer conductance, has been studied extensively since the seminal work of Goldreich and Ron on expansion testing [GR11]:

Definition 9. ((k, φ, ϵ)-clustering) A (k, φ, ϵ)-clustering of G is a partition of vertices V into disjoint subsets $C_1, \ldots, C_k$ such that for all $i \in [k]$, $\phi^G_{\text{in}}(C_i) \geq \varphi$, $\phi^G_{\text{out}}(C_i) \leq \epsilon$ (Definition 5). Graph G is called (k, φ, ϵ)-clusterable if there exists a (k, φ, ϵ)-clustering for G.

We note that our notion of a (k, γ)-hierarchically clusterable graphs is a natural extension of this classical definition. In fact, if the number of layers H is equal to 1, then a (k, γ)-hierarchically clusterable graph is exactly (k, φ, γ)-clusterable as per Definition 9. In particular, note that our hierarchical clustering oracle achieves preprocessing and query time $n^{1/2+O(\gamma)}$, which is the same as the $n^{1/2+O(\epsilon)}$ runtime that is achievable for the classical clustering version of the problem [GKL$^+$21].

Very recently [KKLM22] studied the problem of hierachical clustering of (k, φ, ϵ)-clusterable graphs in the query model *augmented with cluster id queries.* In this stronger model, [KKLM22] gives algorithms which on input a (k, φ, ϵ)-clusterable graph G, return a $O(\sqrt{\log k})$ approximation to the Dasgupta cost of G in sublinear time. This work is incomparable to ours, since the access model is stronger (it allows cluster id queries), but not hierarchical clusterability assumption is made and no *hierarchical clustering oracle* is obtained, i.e. only Dasgupta cost is estimated. The techniques used by [KKLM22] are very different from ours, modulo the fact that both use the dot product oracle of [GKL+21].

2 Technical Overview

In this section we give an overview of the main ideas that go into the proof of Theorem 2. We denote the input (k, γ)-hierarchically-clusterable graph by $G = (V, E)$ (Definition 6), the corresponding hierarchical clustering by $\mathcal{P} = (\mathcal{P}^h)_{h\in[H]}$ and the corresponding ground truth tree by T.[4] Our goal is to design an efficient local computation algorithm that provides fast query access to a hierarchical clustering $\boldsymbol{P}$ that D-approximates $\mathcal{P}$. Specifically, for every cluster $S \in \mathcal{P}$ at level h and the corresponding cluster $\boldsymbol{S} \in \boldsymbol{P}$ must satisfy

$$|\boldsymbol{S} \triangle S| \leq |S| \cdot O(\varphi_{h-1}). \tag{1}$$

In other words, the misclassification error at every level must be on the order of outer conductance of the corresponding level cuts. Let $\boldsymbol{T}$ be the tree corresponding to $\boldsymbol{P}$. Since $\boldsymbol{T}$ has n leaves, we never construct it explicitly, but rather obtain a local computation algorithm for it with small preprocessing and query time – we describe these phases of the algorithm below, and then discuss the main challenges in the analysis as well as the main ideas behind their resolution.

In the **preprocessing phase**, we construct a sketch $\widetilde{T}$ of the tree $\boldsymbol{T}$ by essentially sampling a few vertices in V (specifically, $\approx n^{O(\gamma)} \cdot k^{O(1)}$ vertices) and constructing a good hierarchical clustering on the sample – see ConstructTree (Algorithm 3 in Section 4.1). The construction proceeds top down, starting from the root of the tree, which corresponds to a single cluster that includes all vertices in V, and then iteratively refining the clusters – see RefinePartition (Algorithm 4 in Section 4.1).

In the **query phase** the approximate clustering $\boldsymbol{P}$ is then defined by essentially extending the clustering of the sample to the entire graph. This is done using the procedure Oracle (Algorithm 5 in Section 4.1). For every $h \in [H]$, let $\boldsymbol{P}^h = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{\kappa_h})$ denote the hierarchical clustering that we aim to construct, where $\kappa_h = |\mathcal{P}^h|$ is the number of clusters at level h in the ground truth clustering $\mathcal{P}$. The procedure Oracle proceeds top down. Having defined $\boldsymbol{P}^{h-1}$, it first calculates κ_h, the number of clusters at level h in the ground truth clustering [5] $\mathcal{P}^h$ and then for every $i \in [\kappa_h]$ defines the approximate cluster $\boldsymbol{S}_i \subseteq V$ by letting

$$\boldsymbol{S}_i = \left\{ z \in V : \textsc{Oracle}(G, z, \widetilde{T}, \mathcal{D}) = i \right\},$$

thereby defining $\boldsymbol{P}^h$. The runtime of Oracle is roughly $n^{1/2+O(\gamma)}$.

In what follows we outline the obstacles involved in designing ConstructTree and Oracle and how we overcome them. For sake of simplicity in the technical overview we explain how our techniques can be used to construct the tree $\boldsymbol{T}$ explicitly, even though this would take $\Omega(n)$ time. Then in the next section we will show how our approach leads to a sublinear preprocessing and query time.

Refining $\boldsymbol{P}^h$: how one could use known techniques and why they fail. Consider the h-th iteration of RefinePartition, where we would like to construct $\boldsymbol{P}^h$ that should

[4] Theorem 7 shows that the model of (k, γ)-hierarchically-clusterable graphs contains non-trivially interesting families of graphs. In particular, this theorem shows that an explicit family of graphs, which belong to the Hierarchical Stochastic Block Model of [CKM17], are (k, γ)-hierarchically clusterable.

[5] This is done using existing cluster structure testing results, namely [CKK+18] – see line 4 of Algorithm 3

approximate $\mathcal{P}^h$, the ground truth partition at level h. Let $\kappa_h = |\mathcal{P}^h|$ denote the number of clusters at level h in $\mathcal{P}^h$. As per (1) our goal is to recover every cluster $S \in \mathcal{P}^h$ up to $O(\varphi_{h-1})$ misclassification error, i.e. we need to solve the flat clustering problem for every level h up to error $O(\varphi_{h-1})$. The central challenge here is the fact that known techniques (e.g. [GKL+21]) only allow one to recover every cluster with precision $\approx \varphi_{h-1}/\varphi_h^2$, only ensures that for every cluster $S \in \mathcal{P}^h$ and the corresponding cluster $\boldsymbol{S} \in \boldsymbol{P}$ satisfy

$$|\boldsymbol{S} \triangle S| \leq |S| \cdot O(\varphi_{h-1}/\varphi_h^2).$$

This is because the partition $\mathcal{P}^h$ consists of clusters $S \in \mathcal{P}^h$ with inner conductance $\phi_{\text{in}}(S) \geq \varphi_h$ and outer conductance $\phi_{\text{out}}(S) \leq \varphi_{h-1}$. Not only is this clearly insufficient, but furthermore this guarantee becomes meaningless in our setting, since we are thinking of $\varphi_{h-1} \approx \gamma \cdot \varphi_h$ and φ_h small, so that $\varphi_{h-1}/\varphi_h^2$ is quite possibly larger than 1. This means that existing results on flat clustering are simply not applicable! The squaring of the inner conductance φ_h in the denominator comes from Cheeger's inequality and is clearly problematic. Our first step towards achieving our result is to show that such a loss is unnecessary for (k, γ)-hierarchically-clusterable graphs – the spectral gap turns out to be linear as opposed to quadratic in the inner conductance for such graphs:

Lemma 3. *[Linearity of the spectral gap] Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, and $S \in \mathcal{P}^h$ be a cluster at level h. Let $\chi_2(S)$ be the second smallest eigenvalue of L_S (Definition 13). Then we have*

$$\frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi^G_{in}(S) \leq \chi_2(S) \leq 2 \cdot \phi^G_{in}(S).$$

Note the φ^2 loss on the left hand side: since the bottom level clusters are arbitrary expanders with inner conductance lower bounded by φ, this loss is unavoidable. However, Lemma 3 shows that the eigenvalues of the Laplacian whose eigenvectors intuitively encode the partitions $\mathcal{P}^h$ are much closer to the corresponding inner conductance than to its square. Fortunately, these are the only eigenvalues and eigenvectors that we need to work with to obtain our main result.

In the following subsection we first establish a key property of hierarchically-clusterable graphs i.e., hierarchically concentration of cluster *centers* around their ancestor cluster *centers*, then using this property we sketch the proof of the linearity of the spectral gap (Lemma 3).

2.1 Linearity of the Spectral Gap

We outline the proof of Lemma 3. Fix a level h of the hierarchical partition, and let $\kappa = |\mathcal{P}^h|$ denote the number of clusters in $\mathcal{P}^h$. For every $x \in V$ let $f_x^\kappa \in \mathbb{R}^\kappa$ denote the κ-dimensional spectral embedding of x, i.e. the vector whose coordinates are the values of the bottom κ eigenvectors of the (normalized) Laplacian of G at x[6]. We then define κ-dimensional cluster centers as follows:

Definition 10. (**κ-dimensional center of a cluster**) For any set $S \subseteq V$ and any $\kappa \in [n]$ we define the κ-dimensional center of S as $\mu_S = \frac{1}{|S|} \sum_{x \in S} f_x^\kappa$, where $f_x^\kappa \in \mathbb{R}^\kappa$ for vertex $x \in V$ is the κ-dimensional spectral embedding of x.

Using standard techniques in the literature (see Lemma 5 in Section 4.2) one can show that for every vector $\alpha \in \mathbb{R}^\kappa$ with $||\alpha||_2 = 1$ vertices are concentrated around their respective cluster centers along direction α:

$$\sum_{S \in \mathcal{P}^h} \sum_{x \in S} \langle f_x^\kappa - \mu_S, \alpha \rangle^2 \leq O\left(\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}\right). \tag{2}$$

[6] One notes that f_x is only well-defined up to orthogonal transformations, but this is enough – see Definition 11 for a more formal treatment.

In this subsection first by induction we assume that the spectral gap is linear for clusters at every level $h \geq h^*$, i.e., $\chi_2(S) \geq \Omega(\varphi_h)$ for every $S \in \mathcal{P}^h$. Then using this inductive assumption we prove that centers of refined clusters at level $h > h^*$ are concentrated around the centers of their ancestor at level h^*, i.e., for every $S^* \in \mathcal{P}^{h^*}, S \in \mathcal{P}^h$ where S is a descendent of S^* we have

$$||\mu_{S^*} - \mu_S||_2^2 \leq \frac{\gamma^{1/4}}{|S^*|}, \tag{3}$$

(see Lemma 10 in Section 4.3). Then using (3) we prove the linearity of the spectral gap for the next level (i.e., $h^* - 1$).

Concentration of cluster centers around their ancestors' centers (proof sketch): Using the directional variance bound (2) one can show that the sum of outer products of embeddings $f_x \in \mathbb{R}^\kappa$ (i.e., $\sum_{x \in V} f_x^\kappa f_x^{\kappa T} = I_{k \times k}$) is well approximated spectrally by $\sum_{x \in V} \mu_x \mu_x^T = I_{k \times k}$, where μ_x is the center of cluster that contains vertex x (see Lemma 6 and Lemma 7):

$$\left|\left| \sum_{S \in \mathcal{P}^h} |S| \mu_S \mu_S^T - I_{\kappa \times \kappa} \right|\right|_2 \leq 2 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}}, \tag{4}$$

We now show that for every $S^* \in \mathcal{P}^{h^*}, S \in \mathcal{P}^{h'}$, $h^* < h'$, such that S is a descendent of S^* we have $||\mu_{S^*} - \mu_S|| \leq \frac{\gamma^{1/4}}{|S^*|}$, where $\kappa = |\mathcal{P}^{h^*}|$ and μ_{S^*}, μ_S are κ-dimensional center of S^* and S respectively. Let $S^* = S^{h^*}, S^{h^*+1}, \ldots, S^{h'-1}, S^{h'} = S$ denote the path from the cluster $S^* \in \mathcal{P}^{h^*}$ to the cluster $S \in \mathcal{P}^h$ in tree T associated with $\mathcal{P}$. For every cluster S, let $\Delta_S = \mu_S - \mu_{\text{PARENT}(S)}$ be the difference of the κ-dimensional center of a cluster to the κ-dimensional center of its parent. In what follows we show that for every $h \geq h^* + 1$

$$||\Delta_{S^h}||_2^2 \leq \frac{2}{|S|} \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}}, \tag{5}$$

which implies (3) by the triangle inequality and by summing a geometric series:

$$||\mu_{S^*} - \mu_S||_2^2 \leq \sum_{h=h^*+1}^{h'} ||\Delta_{S^h}||_2^2 = \frac{2}{|S|} \sum_{h=h^*+1}^{h'} \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}} = O\left(\frac{\gamma^{1/4}}{|S|}\right).$$

It remains to establish (5) to complete the proof. (5) essentially follows because the subspace spanned by the centers of the refined clusters is a close to the subspace spanned by the centers of the coarse clusters. In particular, for every $h \geq h^* + 1$ we have

$$\left|\left| \sum_{S \in \mathcal{P}^h} |S| \mu_S {\mu_S}^T - \sum_{S \in \mathcal{P}^{h-1}} |S| \mu_S {\mu_S}^T \right|\right|_2 \leq \left|\left| \sum_{S \in \mathcal{P}^h} |S| \mu_S \mu_S^T - I_{\kappa \times \kappa} \right|\right|_2 + \left|\left| \sum_{S \in \mathcal{P}^{h-1}} |S| \mu_S \mu_S^T - I_{\kappa \times \kappa} \right|\right|_2$$
$$\leq 2 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}} + 2 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^{h-1}} \chi_2(S)}} \leq 2 \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}}, \tag{6}$$

where, the second inequality holds by (4) and the last inequality holds by the inductive assumption of the linearity of the spectral gap on every level $h \geq h^*$ (i.e., $\chi_2(S) \geq \Omega(\varphi_h)$ for $S \in \mathcal{P}^h$) and also by $\lambda_k \leq O(\varphi_{h^*-1})$ according to the multiway Cheeger inequalities. Now we show that $\left|\left|\sum_{S \in \mathcal{P}^h} |S| \mu_S {\mu_S}^T - \sum_{S \in \mathcal{P}^{h-1}} |S| \mu_S {\mu_S}^T \right|\right|_2 = \left|\left|\sum_{S \in \mathcal{P}^{h+1}} |S| \Delta_S {\Delta_S}^T \right|\right|_2$ as follows:

$$\begin{aligned}\sum_{S\in\mathcal{P}^h}|S|\mu_S\mu_S^T &= \sum_{S'\in\mathcal{P}^{h-1}}\sum_{S\in\text{CHILDREN}(S')}|S|\left(\mu_{S'}+\Delta_S\right)\left(\mu_{S'}+\Delta_S\right)^T\\ &= \sum_{S'\in\mathcal{P}^{h-1}}\sum_{S\in\text{CHILDREN}(S')}|S|\Delta_S\Delta_S^T+|S|\mu_{S'}{\mu_{S'}}^T+\mu_{S'}\left(\sum_{S\in\text{CHILDREN}(S')}|S|\Delta_S\right)+\left(\sum_{S\in\text{CHILDREN}(S')}|S|\Delta_S\right)\mu_{S'}^T\\ &= \sum_{S'\in\mathcal{P}^{h-1}}\sum_{S\in\text{CHILDREN}(S')}|S|\Delta_S\Delta_S^T+|S|\mu_{S'}{\mu_{S'}}^T \qquad (7)\end{aligned}$$

where the first equality holds as $\Delta_S=\mu_S-\mu_{\text{PARENT}(S)}=\mu_S-\mu_{S'}$ and the last equality holds as $\mu_S=\mathbb{E}_{S'\in\text{CHILDREN}(S)}[\mu_{S'}]$, hence, $\sum_{S\in\text{CHILDREN}(S')}|S|\Delta_S=0$. Therefore, by (6) and (7) we get

$$|S|\cdot||\Delta_{S^h}||_2^2\le\left|\left|\sum_{S\in\mathcal{P}^h}|S|\Delta_S{\Delta_S}^T\right|\right|_2\le 2\cdot\sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}},$$

as required in (5). The details are presented in Section 4.3 (Lemma 10).

Linearity of the spectral gap (proof sketch): We now outline the proof of Lemma 3. The proof inductively proceeds from the most refined partition (at level H) to the crudest partition (at level 0). For simplicity of the argument suppose that $S^*\in\mathcal{P}^{h^*}$ is a cluster with two children $S,S'\in\mathcal{P}^{h^*+1}$. By induction we assume that the spectral gap is linear for all levels $h\ge h^*+1$, hence, for S,S'. Let u_2 be the second eigenvector of L_S (i.e., the normalized Laplacian of the induced subgraph on S). Let $\mu_S,\mu_{S'}$ be the 2-dimentional center of cluster S and S' respectively. By linearity of the spectral gap for S,S' and by (4) applied to the subgraph $G[S^*]$, one can verify that[7]

$$\left|\left||S|\mu_S\mu_S^T+|S'|\mu_{S'}\mu_{S'}^T-I_{2\times 2}\right|\right|_2\le 2\cdot\sqrt{\frac{\chi_2(S^*)}{\min(\chi_2(S),\chi_2(S'))}}\le O\left(\sqrt{\frac{\varphi_{h^*}}{\varphi_{h^*+1}}}\right)\ll 1, \qquad (8)$$

i.e. μ_S and $\mu_{S'}$ are almost orthogonal. Now let $\nu(S)=\frac{\sum_{x\in S}u_2(x)}{|S|}$ be the mean of second eigenvector u_2 for vertices in cluster S (and similarly for S'). By near orthogonality of μ_S and $\mu_{S'}$ we have that $\nu(S)$ is far from $\nu(S')$ (i.e., $|\nu(S)-\nu(S')|\ge\frac{1}{2\sqrt{|S|}}$). Moreover, by (3) we have that the centers of descendant clusters are concentrated around the center of their ancestor, so for every cluster $C\in\mathcal{P}^H$ that is descendant of S and every cluster $C'\in\mathcal{P}^H$ that is a descendant of S' we have $|\nu(C)-\nu(C')|\ge\frac{1}{3\sqrt{|S|}}$.

We say that vertex x is *bad* if $u_2(x)$ is far from the center of base-cluster $C\in\mathcal{P}^H$ containing vertex x (i.e., $(u_2(x)-\nu(C))^2>\frac{1}{100\cdot|S|}$). By the directional variance bound (2) applied to the partition $\mathcal{P}^H$ we have

$$\sum_{\substack{C\in\mathcal{P}^H\\ C\subseteq S^*}}\sum_{x\in C}(u_2(x)-\nu_C)^2\le\frac{\chi_2(S^*)}{\min_{\substack{C\in\mathcal{P}^H\\ C\subseteq S^*}}\chi_2(C)}\le O(\chi_2(S^*)), \qquad (9)$$

where, the second inequality holds as $\chi_2(C)\ge\Omega(1)$ for $C\in\mathcal{P}^H$. Therefore, the number of bad vertices in S^* is bounded by $O(|S^*|\cdot\chi_2(S^*))$. Letting $E_{S^*}^{\text{bad}}$ denote the set of bad edges in S^*, edges with at least one bad endpoint, we get

$$|E_{S^*}^{\text{bad}}|\le O(d\cdot|S^*|\cdot\chi_2(S^*)). \qquad (10)$$

[7] Note that (8) requires a lower bound of $\Omega(\varphi_{h^*+1})$ on $\chi_2(S)$ and $\chi_2(S')$. This lower bound follows by linearity of spectral gap on level h^*, which we argue by induction in the actual proof – see Section 4.4 for more details.

On the other hand, as we show now, $|E_{S^*}^{\text{bad}}|$ can be lower bounded in terms of the inner conductance of S^*:

$$|E_{S^*}^{\text{bad}}| = \Omega(|S^*| \cdot d \cdot \phi_{in}^G(S^*)). \tag{11}$$

Combining (10) with (11) yields $\chi_2(S^*) = \Omega(\phi_{in}^G(S^*))$, as required. Let $E^{good}(S, S')$ be the set of edges with one endpoint in S and the other in S' such that both endpoints are good vertices. We have that any edge $(x, y) \in E^{good}(S, S')$ is long:

$$|u_2(x) - u_2(y)| \geq |\nu(C) - \nu(C')| - |\nu(C) - u_2(x)| - |\nu(C') - u_2(y)| \geq \frac{1}{3\sqrt{|S|}} - \frac{2}{10 \cdot \sqrt{|S|}},$$

where, $C, C' \in \mathcal{P}^H$ are base-clusters containing x and y respectively. Now, given the total length of edges in E_{S^*} is bounded $\sum_{(x,y)\in E_{S^*}} (u_2(x) - u_2(y))^2 = d \cdot \chi_2 (S^*)$, only small fraction of edges in $E(S, S')$ could be long. Therefore, most of edges in $E(S, S')$ are short and connected to at least one bad vertex. This implies that $|E_{S^*}^{\text{bad}}| \geq \frac{|E(S,S')|}{2} = \Omega(|S^*| \cdot d \cdot \phi_{in}^G(S^*))$, establishing (11) and completing the proof. The details are presented in Section 4.4 (Lemma 3).

2.2 Strong Concentration of Vertices Around Their Centers

Once linearity of the spectral gap is established, known flat clustering techniques (e.g., [GKL+21]) give a local computation algorithm with $\approx \varphi_{h-1}/\varphi_h$ misclassification rate per cluster. While this is nontrivial, it is very far from useful: since $\varphi_{h-1}/\varphi_h = \gamma$, this is not nearly enough to achieve (1), and in particular not enough to obtain meaningful guarantees for Dasgupta's cost. We now explain how the $\approx \varphi_{h-1}/\varphi_h$ misclassification rate can be achieved, and outline the main challenges in improving it to $\approx \varphi_{h-1}$, as well as our resolution of these challenges. Using standard techniques in the literature (see Lemma 5 in Section 4.2) one can show that for every $\alpha \in \mathbb{R}^\kappa$, $||\alpha||_2 = 1$

$$\sum_{S\in\mathcal{P}^h} \sum_{x\in S} \langle f_x^\kappa - \mu_S, \alpha\rangle^2 \leq O\left(\frac{\lambda_\kappa}{\min_{S\in\mathcal{P}^h} \chi_2(S)}\right) = O\left(\frac{\varphi_{h-1}}{\varphi_h}\right) = O(\gamma), \tag{12}$$

where we used Lemma 3 to lower bound $\min_{S\in\mathcal{P}^h} \chi_2(S)$. One can also convert (12) into an upper bound on the typical Euclidean distance squared from the cluster center, (by summing over α in an orthonormal basis of $\mathbb{R}^\kappa$), getting

$$\sum_{S\in\mathcal{P}^h} \sum_{x\in S} ||f_x^\kappa - \mu_S||_2^2 \leq O(\kappa \cdot \gamma). \tag{13}$$

It is then reasonable to define, for every cluster $S \in \mathcal{P}^k$ a candidate cluster $\widehat{S}$ to be the set of all vertices that are close to μ_S (we assume that we know cluster centers for simplicity of overview) as follows:

$$\widehat{S} = \left\{x \in V : ||f_x^\kappa - \mu_S||_2^2 \leq \frac{1}{100 \cdot |S|}\right\}. \tag{14}$$

While the above definition is reasonable, it suffers from a major deficiency: given (13), it is completely possible that $\Omega(\kappa)$ clusters defined by (14) are not even $O(\gamma)$-approximations to their respective clusters S, let alone a $O(\varphi_{h-1})$! This was resolved by using a recursive subspace partitioning approach in [GKL+21], where $O(\gamma)$ **per cluster** misclassification rate was achieved, but it's not clear how that approach can yield our target $O(\varphi_{h-1})$-misclassification rate. In this subsection, we outline the main ideas to achieve $O(\varphi_{h-1})$-approximation to clusters.

One can write $||f_x^\kappa - \mu_S||_2^2$ as $\sum_{i=1}^\kappa \langle f_x^\kappa - \mu_S, \alpha_i\rangle^2$ where $\alpha_1, \ldots, \alpha_\kappa$ are orthonormal basis. We start by showing that for every fixed direction $\alpha \in \mathbb{R}^\kappa$ with $||\alpha||_2 = 1$, at least $1 - O(\varphi_{h-1})$ fraction of vertices in S are concentrated around the center of cluster S along direction α (i.e., $\langle f_x^\kappa - \mu_S, \alpha\rangle^2 \leq \frac{1}{100\cdot|S|}$).

Let $x \in S$, and $C \in \mathcal{P}^H$ be the cluster in level H that contains vertex x. Note that cluster C is the the bottom-most descendant of the cluster S that contains vertex x. We define $\mu_x = \mu_C$ as the κ-dimensional center of the cluster C. Therefore, we have

$$\langle f_x^\kappa - \mu_S, \alpha\rangle = \langle f_x^\kappa - \mu_C, \alpha\rangle + \langle \mu_C - \mu_S, \alpha\rangle \tag{15}$$

We now show that both terms on rhs of (15) are appropriately small. By (3) for every cluster $C \in \mathcal{P}^H$ that is a descendent of cluster S we have

$$\langle \mu_C - \mu_S, \alpha\rangle \leq ||\mu_C - \mu||_2 \leq \sqrt{\frac{\gamma^{1/4}}{|S|}}. \tag{16}$$

By a **stronger** version of the variance bound (12), applied to the κ-dimensional embedding of vertices at level h and to the bottom level clustering (i.e., $\mathcal{P}^H$), we have that the embedding of vertices at level h are highly concentrated around the center of their clusters at level $H \gg h$:

$$\sum_{C \in \mathcal{P}^H} \sum_{x \in C} \langle f_x^\kappa - \mu_x, \alpha\rangle^2 \leq O\left(\frac{\lambda_\kappa}{\min_{C \in \mathcal{P}^H} \chi_2(C)}\right) = O(\varphi_{h-1}), \tag{17}$$

where, the last inequality holds as $\lambda_k \leq O(\varphi_{h-1})$ and $\chi_2(C) \geq \Omega(1)$. The strong improvement in (17) is because we consider the concetration of κ-dimensional embedding of vertices (where, $\kappa = \mathcal{P}^h$) around their respective cluster centers at level H (as opposed to level h). Thus, the gap between the conductcance of clusters at level h and level H allows to achieve $O(\varphi_{h-1})$-approximation. This is a novel tool comparing to the standard flat clustering techniques where the dimension of the embeddings is often the same as the number of clusters in the partition. Therfore, by (17) we have

$$\left|\left\{x \in S : \langle f_x^\kappa - \mu_x, \alpha\rangle^2 > \frac{1}{400 \cdot |S|}\right\}\right| \leq |S| \cdot O(\varphi_{h-1}). \tag{18}$$

Thus, by (16) and (18) for all but an $O(\varphi_{h-1})$ fraction of vertices in S we have

$$\langle f_x^\kappa - \mu_S, \alpha\rangle = \langle f_x^\kappa - \mu_C, \alpha\rangle + \langle \mu_C - \mu_S, \alpha\rangle \leq \sqrt{\frac{\gamma^{1/4}}{|S|}} + \sqrt{\frac{1}{400 \cdot |S|}}. \leq \frac{1}{10\sqrt{|S|}}. \tag{19}$$

2.3 Putting it Together

So far we showed at least $1 - O(\varphi_{h-1})$ fraction of vertices in S are concentrated around their center along direction α as in (19). Turning this fact into an actual hierarchical clustering oracle is not immediate, and requires another idea. Basically, we would like to turn strong concentration bounds in any fixed direction that we just proved into concentration of cluster vertices around corresponding centers in the Euclidean metric. This relation is lossy in high-dimensional spaces. However, it turns out that the problem of *refining* a cluster at level h, say, into its children at level $h+1$ is essentially a constant dimensional problem, as the number of subclusters of a cluster is constant by our assumption. To exploit this observation, we introduce the notion of **subgraph subspaces**: theses are constant dimensional subspaces of $\mathbb{R}^{\kappa_h}$ that suffices to perform the refinement operation. We outline the main ideas here next, and then present our algorithm.

Subgraph subspaces and fast refinement of approximate partitions. Our RefinePartition procedure (Algorithm 4 in Section 4.1) is able to construct an approximate partition $\boldsymbol{P}^h$ with misclassification rate $O(\varphi_{h-1})$ per cluster in time polynomial in the size κ_h of $\boldsymbol{P}^h$, with an $n^{1/2+O(\gamma)}$ overhead for estimating various collision probabilities to access the corresponding spectral embedding. The polynomial dependence on the size of $\boldsymbol{P}^h$ is achieved by refining every cluster $S^* \in \boldsymbol{P}^{h-1}$ individually. Our main innovation here is to not work directly with the κ_h-dimensional embedding of vertices in S^*, but instead to define an appropriate constant dimensional subspace Π of a κ_h dimensional space that is (approximately) spanned by the cluster means of children of S^*, and perform ball-carving there. Intuitively, this idea lets us simulate access to the subgraph of G induced by S^*, and perform clustering there efficiently as a result. The fact that this can be done is somewhat surprising, since **(a)** we are not given the ability to run random walks on subgraphs and **(b)** the outer conductance of the corresponding cut is merely a small constant, so walks of logarithmic length mostly leave the cluster that they start with.

Our algorithm. We now give the full description of our hierarchical clustering oracle. First, linearity of the spectral gap (Lemma 3) allows us to define κ_h-dimensional spectral embeddings for $h \in [H] \cup \{0\}$. Letting $L_G = U\Lambda U^T$ denote the eigendecomposition of the normalized Laplacian of G with $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$, for any $1 \leq \kappa \leq n$ we write $U_{[\kappa]} \in \mathbb{R}^{n\times\kappa}$ for the matrix whose columns are the first κ columns of U. We now introduce

Definition 11. (**κ-dimensional spectral embedding**) For every vertex x we let $f_x^\kappa = U_{[\kappa]}^T \mathbb{1}_x$ be the κ-dimensional spectral embedding of vertex x.

This embedding is indeed well defined because of Lemma 3, as we argue now. Indeed, we have κ_h clusters at level h such that for every cluster $S \in \mathcal{P}^h$ we have $\chi_2(S) \geq \Omega(\varphi_h)$ and $\phi_{\text{out}}(S) \leq O(\varphi_{h-1})$. Thus, by Lemma 4 we have $\lambda_{\kappa_h} \leq 2 \cdot \varphi_{h-1}$ and $\lambda_{\kappa_h+1} \geq \min_{S\in\mathcal{P}^h} \chi_2(S) \geq \Omega(\varphi_h)$. Hence, $\frac{\lambda_{\kappa_h}}{\lambda_{\kappa_h+1}} = O\left(\frac{\varphi_{h-1}}{\varphi_h}\right) = O(\gamma) \ll 1$, and therefore the subspace spanned by the bottom κ_h eigenvectors of L_G is well-defined. For simplicity we set $\kappa = \kappa_h$. Recall that $f_x^\kappa \in \mathbb{R}^\kappa$ is the κ dimensional spectral embedding of vertex x (Definition 11).

The notion of κ-dimensional spectral embedding above lets us apply the recent result of [GKL+21] to estimate $\left\langle f_x^\kappa, f_y^\kappa \right\rangle$ in sublinear time (see Theorem 8). In other words, our problem now reduces to designing a hierarchically clustering oracle assuming dot product access to the spectral embedding above. We now define

Definition 12. (Subgraph projection Π) Let $r, \kappa \in [n]$ and $S \subseteq V$. Let $A \in \mathbb{R}^{\kappa\times|S|}$ be a matrix whose columns are f_x^κ for all $x \in S$. Let $A = Y\Gamma Z^T$ be the SVD of A, where Γ refers to the diagonal matrix of the singular values in non-increasing order. We define the subgraph projection matrix $\Pi \in \mathbb{R}^{\kappa\times\kappa}$ of S with respect to κ and r as the orthogonal projection matrix onto the left r singular vector of A i.e., $\Pi = Y_{[r]}Y_{[r]}^T$.

For intuition, let $S^* \in \mathcal{P}$ be a cluster at level $h-1$ and for $S \in \textsc{Children}_{\mathcal{P}}(S^*)$ let $\mu_S \in \mathbb{R}^\kappa$ denote the κ-dimensional center of S, where $\kappa = |\mathcal{P}^h|$ is the number of clusters in $\mathcal{P}^h$. Then if r is the number of children of S^*, as we show below (see Lemma 14 in Section 4.6), the subgraph projection matrix Π of S^* with respect to κ and r[8] satisfies

$$\Pi \approx \sum_{S\in\textsc{Children}_{\mathcal{P}}(S^*)} |S| \cdot \mu_S \mu_S^T. \tag{20}$$

This is very useful in refining (our approximation to) the level $h-1$ partition $\mathcal{P}^{h-1}$ to $\mathcal{P}^h$. Indeed, when refining (our approximation to) $\mathcal{P}^{h-1}$ that $x \in \boldsymbol{S}^*$, to recover the clusters at level h it suffices to decide which of the children S of S^* the vertex x belongs to. If we had access to the subgraph induced by S^*, we would solve this problem by ball-carving in the spectral embedding of S^*, in particular by checking which of the cluster means μ_S the vertex x is closest to. We do not have direct access to the subgraph, but the subgraph projection matrix Π allows us to simulate this access! In Section 4.9 (Theorem 5) we show that ℓ_2-norm distance between

[8] Note that here we need to know the number of children of S^* in $\mathcal{P}$. This, however, can be estimated efficiently by computing the approximate rank of the Gram matrix of a few vertices sampled from S^* – see Algorithm 11 and its analysis in Section E.

the embedding of any pair of vertices on the projected subspace (i.e., $||\Pi f_x^\kappa - \Pi f_y^\kappa||_2$) can be approximated with high accuracy.

We now have all the ingredients necessary to define our REFINEPARTITION primitive. Suppose for some $h \geq 1$, that (an approximation of) the level $h-1$ partition $\mathcal{P}^{h-1}$ has been constructed, and we are now refining it. Let $S^* \in \mathcal{P}$ be a cluster at level $h-1$, and suppose by induction that we can recover a candidate cluster $\boldsymbol{S}^*$ such that $|\boldsymbol{S}^* \triangle S^*| \leq |S^*| \cdot O(\varphi_{h-2})$. Then our goal is to recover approximations $\boldsymbol{S}$ to children S of S^* in $\mathcal{P}$ such that $|\boldsymbol{S} \triangle S| \leq O(\varphi_{h-1}) \cdot |S|$. Let Π denote the subgraph projection matrix of S^*. Let $\ell = \frac{1}{10^3 \cdot |S^*|}$. For the child $S \in \text{CHILDREN}(S^*)$ we define the cylinder around the mean of cluster S as the set of vertices that are close to $\mu_S \in \mathbb{R}^\kappa$ (Definition 10) in the projected subspace:

$$\text{cyl}(\mu_S, \ell | \boldsymbol{S}^*) = \left\{ x \in \boldsymbol{S}^* : ||\Pi f_x^\kappa - \Pi \mu_S||_2^2 \leq \ell \right\}. \tag{21}$$

We show that for children $S \neq S'$ of S^* one has $\text{cyl}(\mu_S, \ell | \boldsymbol{S}^*) \cap \text{cyl}(\mu_{S'}, \ell | \boldsymbol{S}^*) = \emptyset$. This is basically because by (20) the matrix Π is close to a projection onto the space spanned by the means μ_S of children of S^*, and these means are sufficiently close to orthogonal (see Lemma 12) so that $\Pi \mu_S \approx \mu_S$. Therefore, one can verify that cluster centers are quite far from each other even in the projected space (see Lemma 18 in Section 4.7) and we have $\|\Pi\mu - \Pi\mu'\|_2^2 \geq \frac{1}{|S^*|} = 10^3 \cdot \ell$. Thus every vertex $x \in \boldsymbol{S}^*$ belong to at most one of the candidate clusters $\text{cyl}(\mu_S, \ell | \boldsymbol{S}^*)$.

Showing cyl$(\mu_S, \ell | \boldsymbol{S}^*)$ approximates S well. Now, to complete the proof we need to show that $\text{cyl}(\mu_S, \ell | \boldsymbol{S}^*)$ approximates S up to an $O(\varphi_{h-1})$ error. In other words, we have to prove that for all but an $O(\varphi_{h-1})$ fraction of the vertices $x \in S$ we have $||\Pi f_x^\kappa - \Pi \mu_S||_2^2 \leq \frac{1}{10^3 \cdot |S^*|} = \ell$, where μ_S is the κ-dimensional center of cluster S. Let $x \in S$, and $C \in \mathcal{P}^H$ be the bottom-most descendant of cluster S that contains vertex x. We define $\mu_x = \mu_C$ as the κ-dimensional center of the cluster C. Therefore, by triangle inequality we have

$$||\Pi f_x^\kappa - \Pi \mu_S||_2 \leq ||\Pi f_x^\kappa - \Pi \mu_C||_2 + ||\Pi \mu_C - \Pi \mu_S||_2 \leq ||\Pi f_x^\kappa - \Pi \mu_C||_2 + ||\mu_C - \mu_S||_2 \tag{22}$$

By (3) we have $||\mu_C - \mu||_2^2 \leq \frac{\gamma^{1/4}}{|S|} \leq \frac{1}{4 \cdot 10^3 \cdot |S^*|}$. We will now show for all but an $O(\varphi_{h-1})$ fraction of vertices in S, $||\Pi f_x^\kappa - \Pi \mu_C||_2 \leq \frac{1}{4 \cdot 10^3 \cdot |S^*|}$. Similar to (18) we have

$$\left| \left\{ x \in S : \langle f_x^\kappa - \mu_x, \alpha \rangle^2 > \frac{1}{4 \cdot 10^3 \cdot |S^*|} \right\} \right| \leq |S| \cdot O(\varphi_{h-1}).$$

Summing over all α in an orthonormal basis for the range of Π, we get

$$\sum_{x \in S} ||\Pi f_x^\kappa - \Pi \mu_x||_2^2 \leq \text{rank}(\Pi) \cdot O(\varphi_{h-1}) \leq r \cdot O(\varphi_{h-1}), \tag{23}$$

where, $\text{rank}(\Pi) \leq r = |\text{CHILDREN}(S^*)|$ by (20). Thus, we get that

$$\left| \left\{ x \in S : ||\Pi f_x^\kappa - \Pi \mu_x||_2^2 > \frac{1}{10^4 \cdot |S^*|} \right\} \right| \leq |S^*| \cdot r \cdot O(\varphi_{h-1}) \leq O(\varphi_{h-1}) \cdot |S|, \tag{24}$$

where, the last inequality holds as S^* has constant number of children with comparable size, i.e., $r = O(1)$ and $|S^*| = O(|S|)$. Therefore, we have $|S \setminus \text{cyl}(\mu_S, \ell | \boldsymbol{S}^*)| \leq O(\varphi_{h-1})|S|$ (see Lemma 13 in Section 4.5). Then using $\boldsymbol{S}^* \approx S^*$ and given $\text{cyl}(\mu_S, \ell | \boldsymbol{S}^*) \cap \text{cyl}(\mu_{S'}, \ell | \boldsymbol{S}^*) = \emptyset$ for $S \neq S' \in \text{CHILDREN}(S*)$ one can show $|\text{cyl}(\mu_S, \ell | \boldsymbol{S}^*) \setminus S| \leq O(\varphi_{h-1})|S|$. Therfore,

$$|S \triangle \text{cyl}(\mu_S, \ell | \boldsymbol{S}^*)| \leq O(\varphi_{h-1})|S|.$$

See Theorem 3 in Section 4.8 for the full argument.

Algorithm. For simplicity of technical overview we assume that we know the κ-dimensional center of clusters at level h, i.e., μ_S for all $S \in \mathcal{P}^h$. Under this assumption we present algorithms for recovering the hierarchy of clusters below. For sake of simplicity we explain how to construct the tree $\boldsymbol{T}$ explicitly, even though this would take $\Omega(n)$ time. In Section 4, we will show the actual algorithms with sublinear preprocessing and query time.

Algorithm 1 ConstructTree(G)

```
$\boldsymbol{T} = \emptyset$
for $h = 1$ to $H$ do
    $\kappa = \#$ of clusters at level $h$    ▷ can find $\kappa$ in time $\approx n^{1/2+O(\gamma)}$ (see Appendix F)
    $\boldsymbol{P}^{h-1} \leftarrow$ approximate partition at level $h-1$ (Definition 16)    ▷ not constructed explicitly
    $\boldsymbol{P}^h \leftarrow$ RefinePartition($G, h, \kappa, \boldsymbol{P}^{h-1}, \boldsymbol{T}$).
    generate unique id's from 1 to $\kappa$ for any cluster $\boldsymbol{S} \in \boldsymbol{P}^h$
return $\boldsymbol{T}$
```

Algorithm 2 RefinePartition($G, h, \kappa, \boldsymbol{P}^{h-1}, \boldsymbol{T}$)

```
for $\boldsymbol{S}^* \in \boldsymbol{P}^{h-1}$ do
    $r \leftarrow$ CountChildren($G, h, \kappa, \boldsymbol{S}$).    ▷ Algorithm 11
    $\Pi \leftarrow$ InitializeSubgraphProjMatrix($G, h, \kappa, r, \boldsymbol{S}$)
    for $x \in \boldsymbol{S}^*$ do
        for $\mu_S \in$ centers($\mathcal{P}^h$) do
            $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*) = \left\{x \in \boldsymbol{S}^* : ||\Pi f_x^\kappa - \Pi\mu_S||_2^2 \leq \frac{1}{10^3 \cdot |\boldsymbol{S}^*|}\right\}$
            $\boldsymbol{S} \leftarrow \mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$
            add $\boldsymbol{S}$ as a child of $\boldsymbol{S}^*$ in $\boldsymbol{T}$
            $\boldsymbol{S}^* \leftarrow \boldsymbol{S}^* \setminus \boldsymbol{S}$
            $\boldsymbol{P}^h \leftarrow \boldsymbol{P}^h \cup \{\boldsymbol{S}\}$
return $\boldsymbol{P}^h$
```

Note that to find $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ we need to have access to μ_S, but learning the centers is not straightforward and might require exponential runtime in k as in [GKL$^+$21]. Therefore, to tackle this challenge, we develop a ball carving algorithm that instead of the ball around μ_S, finds a larger ball around a vertex in $\boldsymbol{S}^*$ that contains $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ but is still disjoint from $\mathrm{cyl}(\mu_{S'}, \ell|\boldsymbol{S}^*)$ for $S' \neq S \in \textsc{children}(S^*)$. We present the details of the ball carving algorithm in Section 4.

3 Preliminaries

For $i \in \mathbb{N}$ we use $[i]$ to denote the set $\{1, 2, \ldots, i\}$. Our algorithm and analysis use spectral techniques, and therefore, we setup the following notation. For a symmetric matrix A, we write $\nu_i(A)$ (resp. $\nu_{\max}(A), \nu_{\min}(A)$) to denote the i^{th} largest (resp. maximum, minimum) eigenvalue of A.

We also denote with A_G the adjacency matrix of G and with L_G the *normalized Laplacian* of G where $L_G = I - \frac{A_G}{d}$. We denote the eigenvalues of L_G by $0 \leq \lambda_1 \leq \ldots \leq \lambda_n \leq 2$ and we write Λ to refer to the diagonal matrix of these eigenvalues in non-decreasing order. We also denote by $(u_1, \ldots, u_n)$ an orthonormal basis of eigenvectors of L_G and with $U \in \mathbb{R}^{n \times n}$ the matrix whose columns are the orthonormal eigenvectors of L_G arranged in non-decreasing order of eigenvalues. Therefore the eigendecomposition of L_G is $L_G = U \Lambda U^T$. For any $1 \leq \kappa \leq n$ we write $U_{[\kappa]} \in \mathbb{R}^{n \times \kappa}$ for the matrix whose columns are the first κ columns of U. Now, we introduce a central definition to this work, which is the notion of a spectral embedding.

Definition 11. (**κ-dimensional spectral embedding**) For every vertex x we let $f_x^\kappa = U_{[\kappa]}^T \mathbb{1}_x$ be the κ-dimensional spectral embedding of vertex x.

The spectral embeddings of vertices in a graph provide rich geometric information which has been shown to be useful in graph clustering [LGT14, CPS15, CKK+18, GKL+21].

In this paper, we are interested in (k, γ)-hierarchically-clusterable graphs (Definition 6). Our algorithms often require to estimate the inner product between spectral embeddings of vertices x and y denoted $f_x^{\kappa_h}$ and $f_y^{\kappa_h}$ (here κ_h denotes the dimensionality of the embeddings for the level h partition). The following remark asserts that the inner products between $f_x^{\kappa_h}$ and $f_y^{\kappa_h}$ are well defined even though the choice for these vectors may not be basis free.

Remark 1. *We note that for any $h \in [H]$ G is κ_h-clusterable and $\lambda_{\kappa_h}/\lambda_{\kappa_h+1}$ is smaller than a constant. Thus, the space spanned by the bottom κ_h eigenvectors of the normalized Laplacian of G is uniquely defined, i.e. the choice of $U_{[\kappa_h]}$ is unique up to multiplication by an orthonormal matrix $R \in \mathbb{R}^{\kappa_h \times \kappa_h}$ on the right. Indeed, by Lemma 9 in Section 4 we show $\lambda_{\kappa_h} \leq O(\varphi_{h-1})$ and by Lemma 3 in Section 4 one has $\lambda_{\kappa_h+1} \geq \left(\frac{\beta^3 \cdot \varphi^2}{300}\right) \cdot \varphi_h$. Thus, since we assume that $\frac{\gamma}{\beta^{30} \cdot \varphi^{20}}$ is smaller than an absolute constant, we have $\lambda_{\kappa_h}/\lambda_{\kappa_h+1}$ is smaller than a constant, hence, the subspace spanned by the bottom κ_h eigenvectors of the Laplacian, i.e. the space of $U_{[\kappa_h]}$, is uniquely defined, as required. We note that while the choice of $f_x^{\kappa_h}$ for $x \in V$ is not unique, but the dot product between the spectral embedding of $x \in V$ and $y \in V$ is well defined, since for every orthonormal $R \in \mathbb{R}^{\kappa_h \times \kappa_h}$ one has*

$$\langle R f_x^{\kappa_h}, R f_y^{\kappa_h} \rangle = (R f_x^{\kappa_h})^T (R f_y^{\kappa_h}) = (f_x^{\kappa_h})^T \left(R^T R\right) \left(f_y^{\kappa_h}\right) = (f_x^{\kappa_h})^T \left(f_y^{\kappa_h}\right).$$

For pairs of vertices $x, y \in V$ we use the notation $\langle f_x^\kappa, f_y^\kappa \rangle := (f_x^\kappa)^T (f_y^\kappa)$ to denote the dot product in the embedded domain.

For a multiset $I_S = \{x_1, \ldots, x_s\}$ of vertices from V we abuse notation and also denote by S the $n \times s$ matrix whose i^{th} column is $\mathbb{1}_{x_i}$. For a vertex $x \in V$, we say that $\mathbb{1}_x \in \mathbb{R}^n$ is the indicator of x, that is, the vector which is 1 at index x and 0 elsewhere. For a set $S \subseteq V$, let $\mathbb{1}_S$ we say that $\mathbb{1}_S \in \mathbb{R}^n$ is the indicator of set S, where $\mathbb{1}_S(x) = 1$ if $x \in S$ and $\mathbb{1}_S(x) = 0$ otherwise.

We denote the transition matrix of the *random walk associated with* G by $M = \frac{1}{2} \cdot \left(I + \frac{A_G}{d}\right)$. From any vertex v, this random walk takes every edge incident on v with probability $\frac{1}{2d}$, and stays on v with the remaining probability which is at least $\frac{1}{2}$. Note that $M = I - \frac{L_G}{2}$. Observe that for all i, u_i is also an eigenvector of M, with eigenvalue $1 - \frac{\lambda_i}{2}$. We denote with Σ the diagonal matrix of the eigenvalues of M in descending order. Therefore the eigendecomposition of M is $M = U \Sigma U^T$. We write $\Sigma_{[k]} \in \mathbb{R}^{k \times k}$ for the matrix whose columns are the first k rows and columns of Σ. Furthermore, for any t, M^t is a transition matrix of random walks of length t. For any vertex x, we denote the probability distribution of a t-step random walk started from x by $m_x = M^t \mathbb{1}_x$. For a multiset $I_S = \{x_1, \ldots, x_s\}$ of vertices from V, let matrix $M^t S \in \mathbb{R}^{n \times s}$ is a matrix whose column are probability distribution of t-step random walks started from vertices in I_S. Therefore, the i-th column of $M^t S$ is m_{x_i}.

We also define the Laplacian of an induced subgraph as follows:

Definition 13. (Laplacian of an induced subgraph) Let $G = (V, E)$ be a d-regular graph and let $S \subseteq V$. Let $G[S]$ be the graph obtained by adding $d - d_S(x)$ self-loops to each vertex $x \in S$, where $d_S(x) = |\{y \in S : \{x, y\} \in E\}$. Let L_S denote the normalized Laplacian of $G[S]$. For any $i \in [S]$ we write $\chi_i(S)$ to denote the i-th smallest eigenvalue of L_S.

We will use the following standard result on eigenvalues presented in [GKL$^+$21] whose proof is given in [LGT14] and [CKK$^+$18].

Lemma 4 ([GKL$^+$21])**.** *Let $G = (V, E)$ be a d-regular graph that admits a (k, φ, ϵ)-clustering $C_1, \ldots, C_k$. Then we have $\lambda_k \leq 2\epsilon$ and $\lambda_{k+1} \geq \min_{i \in [k]} \chi_2(C_i)$. In particular, $\lambda_{k+1} \geq \frac{\varphi^2}{2}$.*

Finally, we define the parent, children and the lowest common ancestor of clusters in a hierarchical-clustering as follows:

Definition 14. Let $G = (V, E)$ be a graph that admits a (k, γ)-hierarchical-clustering (Definition 6) and let T be the tree representation of this hierarchical clustering. For $h \in [H]$, and any cluster $S \in \mathcal{P}^{h-1}$, we let $\textsc{children}(S)$ denote the set of children of S in T at level h. If S is not partitioned at level $h - 1$ then $\textsc{children}(S) = \{S\}$. We let $\textsc{parent}(S)$ denote the parent of S in T. For any two internal clusters S_1 and S_2 we define $\mathrm{LCA}(S_1, S_2)$ as the lowest common ancestor of S_1 and S_2 in T.

4 Recovering Hierarchically-Clusterable Graphs

In this section we assume that the graph G admits a (k, γ) hierarchical-clustering $(\mathcal{P}^i)_{i=0}^H$ (Definition 6). The goal of this Section is to design a local computation algorithm that recovers a hierarchical clustering $(\boldsymbol{P})_{i=0}^H$ that is a D-approximation of $(\mathcal{P}^i)_{i=0}^H$ (Definition 7). The main result of this section is Theorem 2.

Theorem 1. *[Informal version of Theorem 2] For sufficiently small constant $\gamma \in (0, 1)$ there exists a* **hierarchical clustering** *oracle with $\approx k^{O(1)} n^{1/2+O(\gamma)}$ preprocessing time and $\approx k^{O(1)} n^{1/2+O(\gamma)}$ query time that achieves a constant factor approximation to Dasgupta cost on (k, γ)-hierarchically clusterable graphs.*

4.1 Algorithm

Assume that the graph G admits a (k, γ) hierarchical-clustering $(\mathcal{P}^h)_{h=0}^H$ represented by the tree T. The goal of the local computation algorithm is to recover $(\boldsymbol{P}^h)_{h=0}^H$ represented by the tree $\boldsymbol{T}$ which is an D-approximation of $(\mathcal{P}^h)_{h=0}^H$. Our algorithm has two phases: the preprocessing phase and the query phase. In the preprocessing phase, the algorithm $\textsc{ConstructTree}(G)$ (Algorithm 3) constructs a sublinear space data structure $\mathcal{D}$ and obtains a small tree $\widetilde{T}$ which is a sketch of $(\boldsymbol{P}^h)_{h=0}^H$. For any hierarchical clustering $\mathcal{P} = (\mathcal{P})_{h=0}^H$, and any $h \leq [H]$, we now define the corresponding subsampled clustering at level h:

Definition 15. (Subsampled clustering) Let $G = (V, E)$ be a graph and let $(\boldsymbol{P}^h)_{h=0}^H$ be a hierarchical clustering. Let $\widetilde{V} \subseteq V$ be a set of vertices sampled independently and uniformly at random from V. For any level $h \in [H]$ and any cluster $\boldsymbol{S} \in \boldsymbol{P}^h$ we define the subsampled cluster $\widetilde{S}$ as $\widetilde{S} = \widetilde{V} \cap \boldsymbol{S}$. For any level $h \in [H]$, we denote the subsampled clustering at level h by $\widetilde{P}^h$ which is the collection of all subsampled clusters at level h. Formally, we have $\widetilde{P}^h$ by $\widetilde{P}^h = \boldsymbol{P}^h \cap \widetilde{V}$.

In the preprocessing phase, the algorithm ConstructTree(G) (Algorithm 3) computes $\{\widetilde{P}^h\}_{h=0}^H$. Then in the query phase algorithm Oracle$(G, z, \widetilde{T}, \mathcal{D})$ (Algorithm 5) receives this information (stored in $\mathcal{D}$) and for any vertex $z \in V$ identifies the cluster this vertex belong to in sublinear time. Therefore, Algorithm 5 determines the path from the root of the tree $\boldsymbol{T}$ to the last cluster containing vertex z.

4.1.1 Preprocessing Phase: Constructing the Sketch of the Tree

We first start by explaning the algorithm ConstructTree(G). This algorithm has H iterations and at iteration h it refines the partition $\widetilde{P}^{h-1}$ to obtain the partition $\widetilde{P}^h$. Let $\boldsymbol{P}^{h-1}$ be the approximate partition constructed implictly by our local computation algorithm for level $h-1$ (Definition 16).

Definition 16. (Approximate clustering) Let G be a graph that admits a (k, γ) hierarchical-clustering $(\mathcal{P}^h)_{h=0}^H$. Let $0 \leq h \leq H$, and $\kappa_h = |\mathcal{P}^h|$ denote the number of clusters at level h. Then for every $i \in [\kappa_h]$ we define the approximate cluster $\boldsymbol{S}_i \subseteq V$ as follows:

$$\boldsymbol{S}_i = \left\{ z \in V : \textsc{Oracle}(G, z, \widetilde{T}, \mathcal{D}) = i \right\}$$

We define the approximate clustering $\boldsymbol{P}^h = \{\boldsymbol{S}_i\}_{i=1}^{\kappa_h}$ as the collection of all approximate clusters at level h

Consider the iteration h of algorithm ConstructTree(G). Recall that at iteration h we want to refine the approximate partition constructed at level $h-1$. Note that our algorithm does not construct $\boldsymbol{P}^{h-1}$ explicitly. However, using algorithm Oracle we have query access to $\boldsymbol{P}^{h-1}$, so for any vertex $z \in V$ we can find the index $i \in [\kappa_h]$ such that $z \in \boldsymbol{S}_i$. Therefore, for a small set of vertices $\widetilde{V}$ we can compute the subsampled partition $\widetilde{P}^{h-1} = \boldsymbol{P}^{h-1} \cap \widetilde{V}$. Then we use the RefinePartition (Algorithm 4) to obtain $\widetilde{P}^h$ from $\widetilde{P}^{h-1}$.

Algorithm 3 ConstructTree(G)

```
$\widetilde{T} = \emptyset$ and $\widetilde{D} = \emptyset$
$\xi = 10^{-3}$
for $h = 1$ to $H$ do
    $\kappa = \#$ of clusters at level $h$        ▷ can find $\kappa$ in time $\approx n^{1/2+O(\gamma)}$ (see Appendix F)
    $\boldsymbol{P}^{h-1} \leftarrow$ approximate partition at level $h-1$ (Definition 16)        ▷ not constructed explicitly
    $\widetilde{V} \leftarrow$ a set of size $\left(\frac{k \cdot n^{\gamma/\varphi} \cdot \log n}{\xi}\right)^{O(1)}$ sampled independently and uniformly at random from $V$
    $\widetilde{P}^{h-1} \leftarrow \boldsymbol{P}^{h-1} \cap \widetilde{V}$
    $\widetilde{P}^h \leftarrow$ RefinePartition$(G, h, \kappa, \widetilde{P}^{h-1}, \widetilde{V}, \xi, \widetilde{T}, \mathcal{D})$.
    generate unique id's from 1 to $\kappa$ for any sampled approximate cluster $\widetilde{S} \in \widetilde{P}^h$
return $\mathcal{D}, \widetilde{T}$
```

Now we explain the RefinePartition algorithm. This algorithm receives $\widetilde{P}^{h-1}$ as input and the goal of the algorithm is to refine every subsampled cluster $\widetilde{S}^*$ to its children. Let $S^* \in \mathcal{P}^{h-1}$ be a cluster at level $h-1$ that has r children at level h. Let $\kappa = |\mathcal{P}^h|$ denote the number of clusters at level h. Let Π and $\widetilde{\Pi}$ be the subgraph projection matrix of S^* and $\boldsymbol{S}^*$ for κ and r respectively (Definition 12).

Our goal is to recover set $\boldsymbol{S}$ that approximates cluster $S \in \textsc{children}(S^*)$ up to $O(\varphi_{h-1})$ misclassification error. In the technical overview, we proved that $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ provides $O(\varphi_{h-1})$ misclassification error for cluster S, and for every $S \neq S' \in \textsc{children}(S^*)$, $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*) \cap \mathrm{cyl}(\mu_{S'}, \ell|\boldsymbol{S}^*) = \emptyset$

However, to find $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ we need to have access to μ_S, but learning the centers is not straightforward. Therefore, to tackle this challenge, we develop a ball carving algorithm that instead of the ball around μ_S, finds a large ball that contains $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ that is still disjoint from $\mathrm{cyl}(\mu_{S'}, \ell|\boldsymbol{S}^*)$. To that end, we first find a representative vertex $x \in \mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$, and then we consider a large enough ball around x denoted by $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*)$ as follows:

$$\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*) = \left\{ y \in \boldsymbol{S}^* : ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 \leq 25\ell \right\},$$

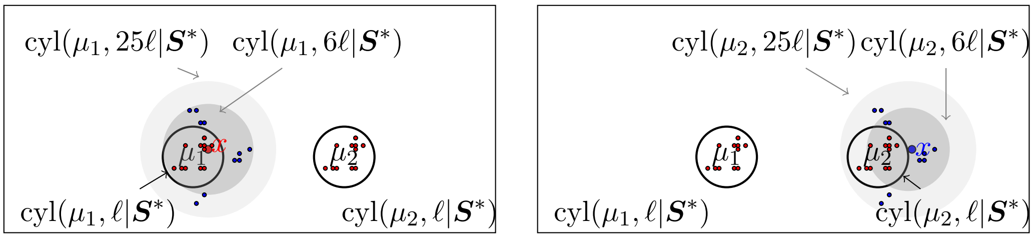
Since $25\ell \gg \ell$ we have $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*) \subseteq \mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*)$, therefore, $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*)$ approximates S up to $O(\varphi_{h-1})$ misclassification error (see Theorem 4 item 1).

After finding a good representative for cluster S, we remove the set $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*)$ from $\boldsymbol{S}^*$ and we proceed to find other children $S' \in \textsc{children}(S^*)$. Since $25\ell \ll 10^3 \cdot \ell$ we have $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*) \cap \mathrm{cyl}(\mu_{S'}, \ell|\boldsymbol{S}^*) = \emptyset$, hence, even by removing $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*)$ we don't touch $\mathrm{cyl}(\mu_{S'}, \ell|\boldsymbol{S}^*)$ of other children (see Theorem 4 item 2).

Finally, we need to explain how one can find a good representative for every $S \in \textsc{children}(S^*)$. Note that any vertex x such that $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*) \supseteq \mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ and is disjoint from $\mathrm{cyl}(\mu_{S'}, \ell|\boldsymbol{S}^*)$ for any $S' \neq S \in \textsc{children}(S^*)$, can be a good representative for cluster S. Therefore, during the ball-carving algorithm we first need to test if vertex $x \in \boldsymbol{S}^*$ can be a good representative for some cluster $S \in \textsc{children}(S^*)$ and then we recover $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*)$. To test this fact we consider a ball $\mathrm{cyl}(f_x^\kappa, 6\ell|\boldsymbol{S}^*)$ around vertex x.

If $\mathrm{cyl}(f_x^\kappa, 6\ell|\boldsymbol{S}^*)$ overlaps with $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$, then $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*) \supseteq \mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$. So if $\mathrm{cyl}(f_x^\kappa, 6\ell|\boldsymbol{S}^*)$ overlaps with one of the $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ for some $S \in \textsc{children}(S^*)$, then x is a good representative for the corresponding child (see Figure 2).

Now note for every cluster $S \in \textsc{children}(S^*)$, $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$ is $O(\varphi_{h-1})$ approximation of S, therefore, the total number of outliers outside of all the balls is at most $O(\varphi_{h-1}) \cdot |S^*|$ (see Lemma 20). Hence, if $|\mathrm{cyl}(f_x^\kappa, 6\ell|\boldsymbol{S}^*)|$ is larger than the number of outliers, then it means that $\mathrm{cyl}(f_x^\kappa, 6\ell|\boldsymbol{S}^*)$ overlaps with one of the balls (i.e., $\mathrm{cyl}(\mu_S, \ell|\boldsymbol{S}^*)$) and hence x is a good representative. In conclusion, to test if vertex x is a good representative it suffices to test if the size of $\mathrm{cyl}(f_x^\kappa, 6\ell|\boldsymbol{S}^*)$ is large enough (see line (8) of Algorithm 4).


cyl(μ_1, 25ℓ|$\boldsymbol{S}^*$)
cyl(μ_1, 6ℓ|$\boldsymbol{S}^*$)
x
μ_1
μ_2
cyl(μ_1, ℓ|$\boldsymbol{S}^*$)
cyl(μ_2, ℓ|$\boldsymbol{S}^*$)
cyl(μ_2, 25ℓ|$\boldsymbol{S}^*$)
cyl(μ_2, 6ℓ|$\boldsymbol{S}^*$)
μ_1
μ_2
x
cyl(μ_1, ℓ|$\boldsymbol{S}^*$)
cyl(μ_2, ℓ|$\boldsymbol{S}^*$)

(a) Carving the first child, S_1. If $x \in S_1$, then $\mathrm{cyl}(\mu_1, 6\ell|\boldsymbol{S}^*)$ contains S_1.

(b) Carving the second child, S_2. If $\mathrm{cyl}(\mu_2, 6\ell|\boldsymbol{S}^*)$ overlaps with S_2, then $\mathrm{cyl}(\mu_2, 25\ell|\boldsymbol{S}^*)$ contains S_2.

Figure 2: Ball Carving

Let $\ell = \frac{1}{10^3 \cdot |S^*|}$. As we explained above, if the size of the ball around vertex x, i.e., $\mathrm{cyl}(f_x^\kappa, 6\ell|\boldsymbol{S}^*)$ is large enough (which means that vertex x passes the test in line (8) of Algorithm 4), then vertex x is a good representative for a unique cluster $S \in \textsc{children}(S^*)$. Therefore, $\mathrm{cyl}(f_x^\kappa, 25\ell|\boldsymbol{S}^*)$ approximates cluster S very well, and we can recover it (see line (9) of Algorithm 4).

As $\boldsymbol{S}^*$ is *close* to S^*, in our algorithms we can use $\ell_{\mathrm{apx}} = 1/10^3 \cdot |\boldsymbol{S}^*|$ as the proxy for radius of the ball i.e., $\ell = 1/10^3 \cdot |S^*|$. We also estimate $\left\|\Pi f_x^\kappa - \Pi f_y^\kappa\right\|_{apx}^2$ by $||\widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa||^2$ with the help of Algorithm 8, and we define the approximate ball around vertex x as:

$$\mathrm{cyl}_{\mathrm{apx}}(f_x^\kappa, \ell_{\mathrm{apx}}|\boldsymbol{S}^*) = \left\{ y \in \boldsymbol{S}^* : \left\|\widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa\right\|_{apx}^2 \leq \ell_{\mathrm{apx}} \right\}.$$

In section 4.10, we prove $\mathrm{cyl}_{\mathrm{apx}}(f_x^\kappa, \ell_{\mathrm{apx}} | \boldsymbol{S}^*)$ approximates cluster S very well (see Lemma 23 and Lemma 30).

Algorithm 4 RefinePartition$(G, h, \kappa, \widetilde{P}^{h-1}, \widetilde{V}, \xi, \widetilde{T}, \mathcal{D})$

```
for $\widetilde{S}^* \in \widetilde{P}^{h-1}$ do
    $s = \frac{|\widetilde{S}^*| \cdot n}{|\widetilde{V}|}$
    $\ell_{\mathrm{apx}} = \frac{1}{1000 \cdot s}$
    $r \leftarrow$ CountChildren$(G, h, \kappa, \widetilde{S}^*, s)$.                 ▷ Algorithm 11
    $\mathcal{D}_{\widetilde{S}^*} \leftarrow$ InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{S}^*, s, \xi)$     ▷ Remark 2
    for $x \in \widetilde{S}^*$ do
        $\widetilde{\mathcal{B}}_x = \left\{ y \in \widetilde{S}^* : \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 \leq 5 \cdot \ell_{\mathrm{apx}} \right\}$     ▷ Remark 2
        if $|\widetilde{\mathcal{B}}_x| > 0.9 \cdot \beta \cdot |\widetilde{S}^*|$ then
            $\widetilde{S} = \left\{ y \in \widetilde{S}^* : \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 \leq 25 \cdot \ell_{\mathrm{apx}} \right\}$     ▷ Remark 2
            rep$(\widetilde{S}) \leftarrow x$                      ▷ $x$ is the representative of $\widetilde{S}$
            add $\widetilde{S}$ as a child of $\widetilde{S}^*$ in $\widetilde{T}$
            $\widetilde{S}^* \leftarrow \widetilde{S}^* \setminus \widetilde{S}$
            $\widetilde{P}^h \leftarrow \widetilde{P}^h \cup \{\widetilde{S}\}$
$\mathcal{D} \leftarrow \mathcal{D} \cup \{\widetilde{P}^h\} \cup \{\mathcal{D}_{\widetilde{S}^*}\}_{\widetilde{S}^* \in \widetilde{P}^{h-1}} \cup \{\mathrm{rep}(\widetilde{S})\}_{\widetilde{S} \in \widetilde{P}^h}$
return $\{\widetilde{P}^h\}$
```

Remark 2. *For computing* $\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2$ *we use Algorithm 8 (Section 4.9.2) as follows:*

$$\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx} = \textsc{ProjectedDistance}(G, x, y, \xi, \mathcal{D}_{\widetilde{S}^*}),$$

where $\mathcal{D}_{\widetilde{S}^*} \leftarrow$ InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{S}^*, s, \xi)$ *(Algorithm 6).*

4.1.2 Query Phase: Hierarchical Clustering Oracle

In the preprocessing phase using algorithm ConstructTree(G) we learn the structure of the tree $\widetilde{T}$ and we find a good representative for every cluster. Furthermore, for every cluster S^* we construct a data structure $\mathcal{D}_{S^*}$ that is to access the subgraph projection matrix and computing $\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2$ during refinement of S^*. We store all this information in a small space data structure $\mathcal{D}$. In the query phase, algorithm Oracle receives the data structure $\mathcal{D}$ and a vertex $z \in V$ as input and assigns z to corresponding candidate cluster $\boldsymbol{S}$. The algorithm proceeds in at most H iteration. Suppose that at iteration $h-1$, algorithm Oracle assigns vertex z to the cluster $\boldsymbol{S}^*$ where $\boldsymbol{S}^*$ is the candidate cluster corresponding to $S^* \in \mathcal{P}^{h-1}$. Recall that for any cluster $S \in$ children(S^*) the representative of cluster S has computed in the preprocessing phase. Let x denote the representative of one of the children, say $S \in$ children(S^*). If $\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_z^\kappa \right\|_{apx}^2 \leq 25 \cdot \ell_{\mathrm{apx}}$, then by definition of $\boldsymbol{S}_x$ we know that $z \in \boldsymbol{S}_x$, hence the algorithm assigns vertex z to $\boldsymbol{S}_x$ and recurs on this child (see line 13 of Algorithm 5). Otherwise, if vertex z does not belong to any of the candidate clusters for children of $\boldsymbol{S}^*$, then vertex z is an outlier. In that case, the algorithm assigns z as a direct child of $\boldsymbol{S}^*$ and stop the algorithm (see line 9 of Algorithm 5).

Algorithm 5 Oracle$(G, z, \widetilde{T}, \mathcal{D})$

```
$\widetilde{S}^* = \mathrm{root}(\widetilde{T})$
for $h = 1$ to $\mathrm{height}(\widetilde{T})$ do
    $S_{\mathrm{child}} = \emptyset$
    for $\widetilde{S}$ that is a child of $\widetilde{S}^*$ in $\widetilde{T}$ do
        $x \leftarrow \mathrm{rep}(\widetilde{S})$
        if $\left\|\widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_z^\kappa\right\|_{apx}^2 \le 25 \cdot \ell_{\mathrm{apx}}$ then        ▷ Remark 2
            $S_{child} \leftarrow \widetilde{S}$
            break from the inner-loop
    if $S_{\mathrm{child}} = \emptyset$ then
        add $z$ as the direct child of $\boldsymbol{S} = \sigma(\widetilde{S}^*)$ in $\boldsymbol{T}$.
        return id of the cluster corresponding to $\boldsymbol{S}$
    else
        $\widetilde{S}^* \leftarrow S_{child}$
add $z$ as the direct child of $\boldsymbol{S} = \sigma(\widetilde{S}^*)$ in $\boldsymbol{T}$.
return id of the cluster corresponding to $\boldsymbol{S}$
```

Finally, we give a roadmap for the rest of Section 4. Section 4.2 develops properties of hierarchically clusterable graphs and presents concentration results which show that spectral embeddings of most vertices in a cluster are close to the center of the cluster. Next, Section 4.3 uses the strucutre of hierarchically clusterable graphs to prove a crucial property which shows that the centers of descendant clusters concentrate around their parent. This property is then used in Section 4.4 to develop one of the key features of hierarchically clusterable graphs. Namely, it is used to show that for any cluster at level h, the spectral gap of the normalized Laplacian induced by the cluster grows only linearly with the inner conductance of the cluster. Building up on these results, Section 4.5 proves the core bounds which show that for any cluster $S \in \mathcal{P}^h$ and for any projection Π onto some space of small dimensionality, at most $O(\varphi_{h-1})|S|$ vertices end up far from the projection of the cluster center. Thereafter, Section 4.6 uses and induction argument to show that the subgraph projection matrix of cluster S is close in operator norm to the subspace spanned by the means of its children. This section also shows that this closeness in operator norm continues to hold even if we replace the subgraph projection matrix of S with the subgraph projection matrix of a set $Q \subseteq V$ that is a good approximation of set S (more formally, Q is *D-hierachically close* to S as per Definition 19). Section 4.7 uses all of this machinery to conclude that centers of subclusters of a cluster remain far from each other. Section 4.8 builds tools which we later use to analyze our algorithms and in particular it uses an inductive argument to show that if a set Q is *D-hierachically close* to a cluster $S \in \mathcal{P}^h$, then the children of Q produced by the ball carving procedure remain D-hierarchically close to corresponding chidren of S at the next level. In Section 4.9 we prove that ℓ_2-norm distance between embedding of any pair of vertices on the projected subspace (i.e., $||\Pi f_x^\kappa - \Pi f_y^\kappa||_2$) can be approximated accurately with high probability. Finally, Section 4.10 puts all this together and proves correctness of our procedures RefinePartition (Algorithm 4) and Oracle (Algorithm 5).

4.2 Properties of Hierarchically Clusterable Graphs

In this section, we fundamental important properties of hierarchically-clusterable-graphs which we use later. We begin with a variant of a *variance bound* which was shown in [GKL+21] (Lemma 6): Suppose $G = (V, E)$ admits a partitioning into m subsets of vertices each of which induce expanders. Then, the sum of squared distances of κ-dimensional embedding of vertices from their corresponding centers along any direction $\alpha \in \mathbb{R}^\kappa$ can be bounded by the

κ-th eigenvalue of the normalized Laplacian. This is a generalization of the variance bound from [GKL$^+$21] as the dimension of embeddings (i.e., κ) could be different from the number of clusters in the partition (i.e., m), which helps us to derive stronger concentration bounds comparing to standard flat clustering techniques. The proof of the following lemma is given in Appendix B.

Lemma 5. *(Variance bounds) Let $\kappa \in [n]$ and $m \geq 2$ be integers. Let $G = (V, E)$ be a d-regular graph. Suppose that V is partitioned into m disjoint subsets $V = S_1 \cup \ldots \cup S_m$. Then for any $\alpha \in \mathbb{R}^\kappa$ with $\|\alpha\| = 1$ we have*

$$\sum_{i=1}^{m} \sum_{x \in S_i} \langle f_x^\kappa - \mu_i, \alpha \rangle^2 \leq \frac{\lambda_\kappa}{\min_{i \in m} \chi_2(S_i)},$$

where $\mu_i \in \mathbb{R}^\kappa$ is the κ-dimensional center of set S_i (Definition 10), $\chi_2(S_i)$ is the second smallest eigenvalue of L_{S_i} (Definition 13), and λ_κ denote the κ-th smallest eigenvalue of L_G.

Next, we state Lemma 6 that is a variant of Lemma 9 from [GKL$^+$21]. The lemma notes the following: Take any partitioning of G into m subsets $S_1, S_2, \cdots, S_m$ each of which induce expanders. Consider κ-dimensional embeddings for all vertices in V and take any subgraph $Q \subseteq V$. Then the sum of outer products $\sum_{x \in Q} f_x^\kappa f_x^{\kappa T}$ is well approximated spectrally by an appropriately weighted sum of $\mu_i \mu_i^T$ (where μ_i is the κ-dimensional center of S_i). This generalizes Lemma 9 from [GKL$^+$21] to **subgraphs** and is a central tool which is used to develop several important properties of hierarchically-clusterable graphs (e.g., Lemma 3 and Lemma 15). The proof of the following lemma is given in Appendix B.

Lemma 6. *Let $\kappa \in [n]$ and $m \geq 2$ be integers. Let $G = (V, E)$ be a d-regular graph. Suppose that V is partitioned into m disjoint subsets $V = S_1 \cup \ldots \cup S_m$. Let $Q \subseteq V$. Then for any $\alpha \in \mathbb{R}^\kappa$ with $\|\alpha\| = 1$ we have*

$$\left| \alpha^T \left(\sum_{i=1}^{m} |Q \cap S_i| \mu_i \mu_i^T - \sum_{x \in Q} f_x^\kappa f_x^{\kappa T} \right) \alpha \right| \leq 2 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{i \in m} \chi_2(S_i)}},$$

where $\mu_i \in \mathbb{R}^\kappa$ is the κ-dimensional center of set S_i (Definition 10), $\chi_2(S_i)$ is the second smallest eigenvalue of L_{S_i} (Definition 13), and λ_κ denote the κ-th smallest eigenvalue of L_G.

The next lemma is an immediate corollary.

Lemma 7. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Then for any $h \in [H]$ we have*

$$\left\| \sum_{S \in \mathcal{P}^h} |S| \mu_S \mu_S^T - I_{\kappa \times \kappa} \right\|_2 \leq 2 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}},$$

where $\mu_S \in \mathbb{R}^\kappa$ is the κ-dimensional center of cluster S (Definition 10), and λ_κ is the κ-th smallest eigenvalue of L_G.

Proof. Note that $\sum_{x \in V} f_x^\kappa f_x^{\kappa T} = U_{[\kappa]}^T U_{[\kappa]} = I_{\kappa \times \kappa}$. Thus the proof follows by Lemma 6 and by choice of $Q = V$. □

Next we state the following properties of cluster means developed in [GKL$^+$21]. It states that the Euclidean length squared of the mean of cluster S is roughly $1/|S|$. Also, for any two different clusters $S, S' \in \mathcal{P}^h$, the cluster means are almost orthogonal.

Lemma 8. *(Cluster means)[GKL$^+$21] Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$ and $\kappa = |\mathcal{P}^h|$ denote the number of clusters at level h. Then we have*

1. *For all $S \in \mathcal{P}^h$, $\left| ||\mu_S||_2^2 - \frac{1}{|S|} \right| \le 4 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}} \cdot \frac{1}{|S|}$*

2. *For all $S \neq S' \in \mathcal{P}^h$, $|\langle \mu_S, \mu_{S'} \rangle| \le 8 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}} \cdot \frac{1}{\sqrt{|S| \cdot |S'|}}$*

where $\mu_i \in \mathbb{R}^\kappa$ is the κ-dimensional center of set S_i (Definition 10), $\chi_2(S_i)$ is the second smallest eigenvalue of L_{S_i} (Definition 13), and λ_κ denote the κ-th smallest eigenvalue of L_G.

We also need the following lemma which follows from multiway Cheeger inequalities [LGT14].

Lemma 9. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$ and let $\mathcal{P}^h$ be the partition at level h. Let $\kappa = |\mathcal{P}^h|$ denote the number of clusters at level h. Then, we have*

$$\lambda_\kappa \le O(\varphi_{h-1}).$$

Proof. By Lemma 4 we have $\lambda_\kappa \le 2 \cdot \max_{S \in \mathcal{P}^h} \phi_{\text{out}}(S)$. By Definition 6 for any cluster $S \in \mathcal{P}^h$ we have $\phi_{\text{out}}(S) \le O(\varphi_{h-1})$. Thus we have $\lambda_\kappa \le O(\varphi_{h-1})$. □

4.3 Concentration of descendant's centers around the ancestor centers

The main result of this section is Lemma 10, in which we prove a fundamental structural property of (k, γ) hierarchically-clusterable graphs. We show that in such graphs the center of clusters at every level are concentrated around the center of their ancestors. In Lemma 10 we prove this property at a given level h^* assuming the spectral gap is linear (see Definition 17) for all the clusters in the levels below (i.e, $h \ge h^*$). Later in Section 4.4 we show the linearity of the spectral gap.

Definition 17. (Linearity of the spectral gap) Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6) and let $\alpha \in [0, 2]$. We say that the spectral gap is α-close to linear at level h if for every cluster $S \in \mathcal{P}^h$

$$\alpha \cdot \phi_{\text{in}}^G(S) \le \chi_2(S) \le 2 \cdot \phi_{\text{in}}^G(S)$$

where $\chi_2(S)$ is the second smallest eigenvalue of L_S (Definition 13).

We outline the major ideas that go inside the proof of Lemma 10 (this was also done in tech overview). We first show that the κ-dimensional center of a node $S \in \mathcal{P}^h$ (with $\kappa = |\mathcal{P}^h|$) and the κ dimensional center of its parent. We then use triangle inequality to bound the distance between the κ-dimensional centers of an ancestor/descendant pair. We first recall the assumption of the model.

Assumption 1. *We assume that the conductance of the base at level H clusters satisfies varphi$_H$ $= \varphi \ge \Omega(\gamma^{1/20})$, and that clusters in the ground truth hierarchical clustering $\mathcal{P}$ get partitioned into constant number of subclusters of comparable size, i.e., for every $S^* \in \mathcal{P}$ and every child S of S^* we have $|S| \ge \beta |S^*|$ for some $\beta \in (0, 1)$. We assume $\beta \ge \Omega(\gamma^{1/30})$.*

We now present the proof of Lemma 10.

Lemma 10. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). For every $h^* \in [H]$ such that the spectral gap is $\frac{\beta^3 \cdot \varphi^2}{300}$-close to linear for all $h \geq h^*$ the following condition holds: For every $S^* \in \mathcal{P}^{h^*}$, every $h > h^*$ and every cluster $S \in \mathcal{P}^h$ that is a descendant of the cluster S^* (i.e, $S \subseteq S^*$) we have*

$$||\mu_S - \mu_{S^*}||_2^2 \leq \frac{\gamma^{1/4}}{|S^*|}$$

where, $\kappa = |\mathcal{P}^{h^*}|$ is the number of clusters at level h^* and $\mu_S, \mu_{S^*} \in \mathbb{R}^\kappa$ are the κ-dimensional center of the cluster S and cluster S^* respectively (Definition 10).

Proof. For any $h^* \leq h \leq H$, and any cluster $S \in \mathcal{P}^h$ let $\mu_S = \frac{\sum_{x \in S} f_x^\kappa}{|S|}$ be the κ-dimensional center of cluster S. Note that by Lemma 7 for any h and we have

$$\left|\left|\sum_{S \in \mathcal{P}^h} |S| \mu_S \mu_S^T - I_{\kappa \times \kappa}\right|\right|_2 \leq 2 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}} \tag{25}$$

By the assumption of the lemma (linearity of the spectral gap) for any $h \geq h^*$ and any cluster $S \in \mathcal{P}^h$ we have $\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}^G(S)$. By Assumption 1 we have $\varphi \geq \gamma^{1/20}$ and $\beta \geq \gamma^{1/30}$. Hence,

$$\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}^G(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \varphi_h \geq \frac{\gamma^{1/5}}{300} \cdot \varphi_h \tag{26}$$

Moreover since $\kappa = |\mathcal{P}^{h^*}|$ by Lemma 9, we have

$$\lambda_\kappa \leq O(\varphi_{h^*-1}) \tag{27}$$

Recall that by Definition 6 for any h' we have $\varphi_{h^*-1} = \varphi_h \cdot \gamma^{h-h^*+1}$. Thus, by putting (25), (26) and (27) together, for every $h \geq h^*$ we get

$$\left|\left|\sum_{S \in \mathcal{P}^h} |S| \mu_S \mu_S^T - I_{\kappa \times \kappa}\right|\right|_2 \leq 2 \cdot \sqrt{\frac{O(\varphi_{h^*-1})}{\frac{\gamma^{1/5}}{300} \cdot \varphi_h}} \leq \sqrt{\frac{\gamma^{h-h^*+1}}{\gamma^{0.3}}}, \tag{28}$$

where the last step follows on choosing a sufficiently small γ to cancel the hidden constant in $O(.)$. Similarly, for every $h \geq h^* + 1$, we get

$$\left|\left|\sum_{S \in \mathcal{P}^{h-1}} |S| \mu_S \mu_S^T - I_{\kappa \times \kappa}\right|\right|_2 \leq \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}} \tag{29}$$

Therefore for every $h \geq h^* + 1$, by (28), (29) and triangle inequality we have

$$\left|\left|\sum_{S \in \mathcal{P}^h} |S| \mu_S \mu_S^T - \sum_{S \in \mathcal{P}^{h-1}} |S| \mu_S \mu_S^T\right|\right|_2 \leq \sqrt{\frac{\gamma^{h-h^*+1}}{\gamma^{0.3}}} + \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}} \leq 2 \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}} \tag{30}$$

For any cluster $S \in \mathcal{P}^h$ at level h we define $\Delta_S \in \mathbb{R}^\kappa$ as follows:

$$\Delta_S = \mu_S - \mu_{\text{PARENT}(S)} \tag{31}$$

Note that for any cluster S' we have

$$\mu_{S'} = \frac{1}{|S'|} \cdot \sum_{S \in \text{CHILDREN}(S')} |S| \cdot \mu_S \tag{32}$$

Thus by (32), (31) and since $|S'| = \sum_{S \in \text{CHILDREN}(S')} |S|$ for any $S' \in \mathcal{P}^{h-1}$ we have

$$\sum_{S \in \text{CHILDREN}(S')} |S| \cdot \Delta_S = 0 \tag{33}$$

Therefore we can write

$$\begin{aligned}
&\sum_{S \in \mathcal{P}^h} |S| \left(\mu_S\right) \left(\mu_S\right)^T \\
&= \sum_{S' \in \mathcal{P}^{h-1}} \sum_{S \in \text{CHILDREN}(S')} |S| \left(\mu_{S'} + \Delta_S\right) \left(\mu_{S'} + \Delta_S\right)^T && \text{By (31)} \\
&= \sum_{S' \in \mathcal{P}^{h-1}} \left(\sum_{S \in \text{CHILDREN}(S')} |S| \Delta_S \Delta_S^T + |S| \mu_{S'} {\mu_{S'}}^T \right) \\
&+ \sum_{S' \in \mathcal{P}^{h-1}} \left(\mu_{S'} \left(\sum_{S \in \text{CHILDREN}(S')} |S| \Delta_S \right) + \left(\sum_{S \in \text{CHILDREN}(S')} |S| \Delta_S \right) \mu_{S'}^T \right) \\
&= \sum_{S' \in \mathcal{P}^{h-1}} \left(\sum_{S \in \text{CHILDREN}(S')} |S| \Delta_S \Delta_S^T + |S| \mu_{S'} {\mu_{S'}}^T \right) && \text{By (33) cross terms are 0} \\
&= \sum_{S \in \mathcal{P}^h} |S| \Delta_S \Delta_S^T + \sum_{S \in \mathcal{P}^{h-1}} |S'| \mu_{S'} \mu_{S'}^T && \text{Since } |S'| = \sum_{S \in \text{CHILDREN}(S')} |S|
\end{aligned} \tag{34}$$

Therefore for any $h \geq h^* + 1$ by (30), and (34) we can write

$$\left|\left| \sum_{S \in \mathcal{P}^h} |S| \mu_S \mu_S^T - \sum_{S \in \mathcal{P}^{h-1}} |S| \mu_S \mu_S^T \right|\right|_2 = \left|\left| \sum_{S \in \mathcal{P}^h} |S| \Delta_S \Delta_S^T \right|\right|_2 \leq 2 \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}} \tag{35}$$

Therefore for every $h \geq h^* + 1$ and for any $\alpha \in \mathbb{R}^\kappa$ with $||\alpha||_2 = 1$ we have

$$\sum_{S \in \mathcal{P}^h} |S| \left\langle \Delta_S, \alpha \right\rangle^2 = \left| \alpha^T \left(\sum_{S \in \mathcal{P}^h} |S| \Delta_S \Delta_S^T \right) \alpha \right| \leq 2 \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}}$$

Thus for any $h \geq h^* + 1$ and for any cluster $S \in \mathcal{P}^h$ at level h, by choice of $\alpha = \frac{\Delta_S}{||\Delta_S||_2}$ we get

$$||\Delta_S||_2^2 \leq \frac{2}{|S|} \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.1}}} \tag{36}$$

Let $S^* \in \mathcal{P}^{h^*}$ and let $S \in \mathcal{P}^{h'}$ be a cluster at level $h' > h^*$ such that S is a descendent of S^* (i.e., $S \subseteq S^*$). Let $S^* = S^{h^*}, S^{h^*+1}, \ldots, S^{h'}$ denote the path from S^* to the cluster S in the underlying tree T. For any $h \geq h^* + 1$ we define $\Delta_{S^h} = \mu_{S^h} - \mu_{S^{h-1}}$. Note that since S^h is the child of S^{h-1} we have $|S^h| \geq \beta \cdot |S^{h-1}|$. Note that $\beta \geq \gamma^{1/30}$ by Assumption 1, thus for any $h^* + 1 \leq h$ we have

$$|S^h| \geq \beta^{h-h^*} \cdot |S^*| \geq \gamma^{(h-h^*)/30} \cdot |S^*| \tag{37}$$

Putting (36) and (37) together for any $h \geq h^* + 1$ we get

$$||\Delta_{S^h}||_2^2 \leq \frac{2}{|S^h|} \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}} \leq \frac{2}{|S^*|} \cdot \frac{1}{\gamma^{(h-h^*)/30}} \cdot \sqrt{\frac{\gamma^{h-h^*}}{\gamma^{0.3}}} \leq \frac{2}{|S^*|} \cdot \frac{\gamma^{7/15 \cdot (h-h^*)}}{\gamma^{3/20}} \tag{38}$$

Note that

$$\mu_S - \mu_{S^*} = \sum_{h=h^*+1}^{h'} \mu_{S^h} - \mu_{S^{h-1}} = \sum_{h=h^*+1}^{h'} \Delta_{S^h}$$

Thus we have

$$\begin{aligned}
||\mu_S - \mu_{S^*}||_2 &= \left|\left|\sum_{h=h^*+1}^{h'} \Delta_{S^h}\right|\right|_2 \\
&\le \sum_{h=h^*+1}^{h'} ||\Delta_{S^h}||_2 && \text{By triangle inequality} \\
&\le \sqrt{\frac{2}{|S^*|}\frac{\gamma^{-7/15\cdot h^*}}{\gamma^{3/20}}} \sum_{h=h^*+1}^{h'} \gamma^{7/30\cdot h} && \text{By (38)} \\
&\le 2\cdot\sqrt{\frac{2\cdot\gamma^{19/60}}{|S^*|}} \\
&\le \sqrt{\frac{\gamma^{1/4}}{|S^*|}} && \text{for small enough } \gamma
\end{aligned}$$

Therefore for any cluster S that is a descendant of the cluster $S^* \in \mathcal{P}^{h^*}$ we have

$$||\mu_S - \mu_{S^*}||_2^2 \le \frac{\gamma^{1/4}}{|S^*|}$$

□

Using Lemma 10 we prove the following lemma (Lemma 11) that we will use later in Section 4.4 to prove the linearity of the spectral gap. Let u_2 denote the second eigenvector of the Laplacian. For $S \subseteq V$, let $\nu_S = \frac{\sum_{x\in S} u_2(x)}{|S|}$. This lemma shows that the second eigenvector of Laplacian in a (k,γ)-hierarchically-clusterable graph inherits some information about the hierarchical cluster structure in the following sense. Take $S^* \in \mathcal{P}^{h^*}$. Then the children of S^* can be partitioned into two collection of sets M, N such that for any cluster $C \in M$ and $C' \in N$ ν_C is far from $\nu_{C'}$. The key to this argument is the following: first, we show that $\exists S \in \textsc{children}(S^*)$ for which $\nu_S \ge \sqrt{1/2|S^*|}$. One can sort the vector $\{\nu_S\}_{S\in\textsc{children}(S^*)}$ and use averaging arguments to find a pair of succesive children $S, T \in \textsc{children}(S^*)$ for which $\nu_S - \nu_T \ge \frac{1}{r\sqrt{2|S^*|}} \ge \frac{\beta}{\sqrt{2|S^*|}}$.

Lemma 11. *Let $G = (V, E)$ be a (k,γ)-hierarchically-clusterable graph (Definition 6). Let $h^* \in [H]$ and $S^* \in \mathcal{P}^{h^*}$. Suppose that the spectral gap is $\frac{\beta^3\cdot\varphi^2}{300}$-close to linear for all h $\ge h^*$. Let u_2 be the second eigenvector of L_{S^*} (Definition 13). For every set $S \subseteq S^*$ let $\nu_S = \frac{\sum_{x\in S} u_2(x)}{|S|}$ be the mean of u_2 in set S. Then there exists a partition of clusters in* $\textsc{children}(S^*)$ *into two collection of sets M, N such that for all $S \in M$ and all $T \in N$ we have*

$$|\nu_S - \nu_T| \ge \frac{\beta}{\sqrt{2|S^*|}}$$

Also, for any $h > h^*$ and $C, C' \in \mathcal{P}^h$ such that $C \subseteq S$ and $C' \subseteq T$ we have

$$|\nu_C - \nu_{C'}| \ge \frac{\beta}{2\cdot\sqrt{|S^*|}}.$$

Proof. Let $G[S^*]$ be the graph obtained by adding $d - d_{S^*}(x)$ self-loops to each vertex $x \in S^*$, where $d_{S^*}(x) = |\{y \in S^* : \{x,y\} \in E\}|$. Therefore, for any cluster S that is a descendant of S^* in G we have $\phi_{\text{in}}^{G[S^*]}(S) = \phi_{\text{in}}^{G}(S)$ and $\phi_{\text{out}}^{G[S^*]}(S) \leq \phi_{\text{out}}^{G}(S)$. Therefore, $G[S^*]$ is (k^*, γ) hierarchically-clusterable (Definition 6) where $k^* = \left|\{C \in \mathcal{P}^H : C \subseteq S^*\}\right|$.

Let $r = |\textsc{children}(S^*)|$, and for every set S let $\mu_S \in \mathbb{R}^r$ be the r-dimensional center of S in graph $G[S^*]$ (Definition 10). By Lemma 7 for any $\alpha \in \mathbb{R}^r$ with $\|\alpha\| = 1$, we have

$$\left|\alpha^T \left(\sum_{S \in \textsc{children}(S^*)} |S| \mu_S \mu_S^T - I_{r \times r}\right) \alpha\right| \leq 2 \cdot \sqrt{\frac{\chi_r}{\min_{S \in \textsc{children}(S^*)} (\chi_2(S))}} \tag{39}$$

Here, χ_r denotes the r-th smallest eigenvalue of $L_{G[S^*]}$, the normalized Laplacian of $G[S^*]$. Note that by assumption of the lemma (linearity of the spectral gap) for any $h \geq h^*$ and any cluster $S \in \mathcal{P}^h$ we have $\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}^{G}(S) = \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}^{G[S^*]}(S)$. Since $\beta \geq \gamma^{1/30}$ and $\varphi \geq \gamma^{1/20}$, we have

$$\min_{S \in \textsc{children}(S^*)} (\chi_2(S)) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \varphi_{h^*+1} = \frac{\gamma^{1/5}}{300} \cdot \varphi_{h^*+1} \tag{40}$$

Moreover since $|\textsc{children}(S^*)| = r$ by Lemma 9 we have

$$\chi_r \leq O(\varphi_{h^*}) \tag{41}$$

Putting (39), (40) and (41) together we have

$$\left|\alpha^T \left(\sum_{S \in \textsc{children}(S^*)} |S| \mu_S \mu_S^T - I_{r \times r}\right) \alpha\right| \leq 2\sqrt{\frac{O(\varphi_{h^*})}{\frac{\gamma^{1/5}}{300} \cdot \varphi_{h^*+1}}} \leq \gamma^{1/4}, \tag{42}$$

where the last step holds because $\varphi_{h^*} = \varphi_{h^*+1} \cdot \gamma$, and by choice of sufficiently small γ.

Let $\alpha = (0,1,0,0,\ldots,0) \in \mathbb{R}^r$ be a r-dimensional vector whose second coordinate is 1 and the rest is zero. Thus for every set S we have $\alpha^T \mu_S = \nu_S$. Therefore by (42) we have

$$\left|\sum_{S \in \textsc{children}(S^*)} |S| \cdot \nu_S^2 - 1\right| \leq \gamma^{1/4} \tag{43}$$

First we show that $\max_{S \in \textsc{children}(S^*)} \nu_S^2 \geq \frac{1}{2|S^*|}$. Suppose by contradiction that $\max_{S \in \textsc{children}(S^*)} \nu_S^2 < \frac{1}{2|S^*|}$. Thus by (43) we get

$$\left|\sum_{S \in \textsc{children}(S^*)} |S| \cdot \nu_S^2 - 1\right| \geq \left||S^*| \cdot \frac{1}{2|S^*|} - 1\right| \geq \frac{1}{2}$$

Thus by (43) we get $1/2 < \gamma^{1/4}$ and given γ is a sufficiently small constant we have the contradiction. Therefore,

$$\max_{S \in \textsc{children}(S^*)} \nu_S^2 \geq \frac{1}{2|S^*|}$$

Suppose that we sort the clusters in $\textsc{children}(S^*)$ based on the value of ν_S. Without loss of generality suppose that $\nu_{S_1} \geq \nu_{S_2} \geq \ldots \geq \nu_{S_r}$ and suppose that $\nu_{S_1} = \text{arg}_{\max} \nu_S^2$. Also note that u_2 is ortogonal to $u_1 = \mathbb{1}$. Thus we have

$$\sum_{S \in \textsc{children}(S^*)} |S| \cdot \nu_S = \sum_{x \in V} u_2(x) = 0 \tag{44}$$

Therefore, $\nu_{S_1} > 0$ and $\nu_{S_r} < 0$. Recall that $\max_{S \in \textsc{children}(S^*)} \nu_S = \nu_{S_1} \geq \sqrt{\frac{1}{2|S^*|}}$. Therefore, there exists an index $2 \leq i \leq r$ such that

$$\nu_{S_i} - \nu_{S_{i-1}} \geq \frac{1}{r \cdot \sqrt{2|S^*|}}$$

By Assumption 1 for every $S \in \textsc{children}(S^*)$ we have $\beta \cdot |S^*| \leq |S| \leq |S^*|$, hence, $r \leq \frac{1}{\beta}$. Thus we have

$$\nu_{S_i} - \nu_{S_{i-1}} \geq \frac{\beta}{\sqrt{2|S^*|}}$$

Now define $M = \{S_j \in \textsc{children}(S^*) : j \geq i\}$ and $N = \{S_j \in \textsc{children}(S^*) : j < i\}$. Hence, for any $S \in M$ and $T \in N$ we have

$$\nu_S - \nu_T \geq \frac{\beta}{\sqrt{2|S^*|}} \tag{45}$$

By Lemma 10 for any cluster $C \in \mathcal{P}^H$ that is a descendant of a cluster $S \in \textsc{children}(S^*)$ we have

$$\begin{aligned}
(\nu_C - \nu_S)^2 &\leq ||\mu_C - \mu_S||_2^2 && \text{Since } \nu_C - \nu_S \text{ is second coordinate of } \mu_C - \mu_S \\
&\leq \frac{\gamma^{1/4}}{|S|} && \text{By Lemma 10} \\
&\leq \frac{\gamma^{1/4}}{\beta \cdot |S^*|} && \text{Since } |S| \geq \beta \cdot |S^*| \\
&\leq \frac{\beta^2}{100 \cdot |S^*|} && \text{By Assumption 1, } \frac{\gamma^{1/4}}{\beta^3} \text{ is a sufficiently small constant}
\end{aligned} \tag{46}$$

Similarly for any cluster $C' \in \mathcal{P}^H$ that is a descendant of a cluster $T \in \textsc{children}(S^*)$ we have

$$(\nu_{C'} - \nu_T)^2 \leq \frac{\beta^2}{100 \cdot |S^*|} \tag{47}$$

Putting (45), (46) and (47) together for any $S \in M$, $T \in N$, $C \subseteq S$, $C' \subseteq T$ we get

$$|\nu_C - \nu_{C'}| \geq |\nu_S - \nu_T| - |\nu_C - \nu_S| - |\nu_{C'} - \nu_T| \geq \frac{\beta}{2\sqrt{|S^*|}}$$

□

4.4 Linearity of the Spectral Gap

The main result of this subsection is Lemma 3 which presents one of the key properties of (k, γ)-hierarchically-clusterable graphs. Namely, for any cluster $S \in \mathcal{P}^h$ the spectral gap of the normalized Laplacian of induced subgraph on S is linear i.e., $\chi_2(S) = \Theta(\phi^G_{\text{in}}(S))$. This is a strong property on hierarchically-clusterable graphs as opposed to general graphs that might suffer from a quadratic gap between the second eigenvalue and inner-conductance.

To prove Lemma 3 we use the fact that the center of base-clusters at level H are concentrated around the center of their ancestor. Using this we show that there exists a small set of bad vertices that are far from the center of their base-cluster. This implies that the number of edges connected to the bad vertices is bounded $E^{\text{bad}(S)} \leq O(|S| \cdot d \cdot \chi_2(S))$. Next, we show that most of the edges crossing the cut between children of S^* are connected to the bad vertices. This implies that $E^{\text{bad}(S)} \geq O(|S| \cdot d \cdot \phi^G_{\text{in}}(S))$. By putting together the lower bound and the upper bound on $E^{\text{bad}(S)}$ we obtain $\chi_2(S) = \Omega(\phi^G_{\text{in}}(S)))$

Lemma 3. *[Linearity of the spectral gap] Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, and $S \in \mathcal{P}^h$ be a cluster at level h. Let $\chi_2(S)$ be the second smallest eigenvalue of L_S (Definition 13). Then we have*

$$\frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi^G_{in}(S) \leq \chi_2(S) \leq 2 \cdot \phi^G_{in}(S).$$

Proof. The upper bound immediately follows from the easy direction of Cheeger's inequality.

We prove the lower bound by induction on h.

Base: Let $h = H$. Note that G is (k, γ)-hierarchically-clusterable. Thus it admits a clustering $C_1, \ldots, C_k$ such that for any cluster C_i we have $\phi^G_{\text{in}}(C_i) \geq \varphi$. Note that $h = H$, thus by Cheeger's inequality we have

$$\chi_2(C_i) \geq \frac{\varphi^2}{2} \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi^G_{\text{in}}(C_i)$$

where the last inequality holds since $\beta < 1$ and $\phi^G_{\text{in}}(C_i) < 1$.

Inductive step: By induction hypothesis we suppose that for any $h' \geq h + 1$ and for any $S' \in \mathcal{P}^{h'}$ we have $\chi_2(S') \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi^G_{\text{in}}(S')$. Then we want to prove that for any cluster $S \in \mathcal{P}^h$ at level h we have $\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi^G_{\text{in}}(S)$.

Let $G[S]$ be the graph obtained by adding $d - d_S(x)$ self-loops to each vertex $x \in S$, where $d_S(x) = |\{y \in S : \{x, y\} \in E\}|$. Let $u_2 \in \mathbb{R}^{|S|}$ and $\chi_2(S) \in \mathbb{R}$ be the second eigenvector and the second eigenvalue of L_S respectively. For any $C \subseteq V$ let $\nu_C = \frac{\sum_{x \in C} u_2(x)}{|C|}$. An application of variance bounds on the graph $G[S]$ (Lemma 5) and the partition defined by clusters $C \in \mathcal{P}^H$ with the choice of $\kappa = 2$ and $\alpha = (0, 1)$ gives

$$\sum_{\substack{C \in \mathcal{P}^H \\ C \subseteq S}} \sum_{x \in C} (u_2(x) - \nu_C)^2 \leq \frac{\chi_2(S)}{\min_{\substack{C \in \mathcal{P}^H \\ C \subseteq S}} \chi_2(C)} \leq \frac{\chi_2(S)}{\frac{\varphi^2}{2}} \tag{48}$$

where the last inequality holds since for any cluster $C \in \mathcal{P}^H$ we have $\phi^G_{\text{in}}(C) \geq \varphi$ and hence, by Cheeger's inequality we have $\chi_2(C) \geq \frac{\varphi^2}{2}$. We define $\nu(x) = \nu_C$ if $x \in C$.

We now define set B as the set of *bad* vertices whose embedding i.e., $u_2(x)$ is from the center of their cluster i.e., $\nu(x) = \nu_C$:

$$B = \left\{ x \in S : (u_2(x) - \nu(x))^2 \geq \frac{\beta^2}{100 \cdot |S|} \right\}$$

By (48), we have

$$|B| \leq \frac{200 \cdot |S| \cdot \chi_2(S)}{\beta^2 \cdot \varphi^2}$$

Let E_S denote the set of edges in $G[S]$. Let E^{bad}_S denote the set of edges in E_S which are adjacent to at least one bad vertex. We have

$$|E^{\text{bad}}_S| = |\{\{x, y\} \in E_S : x \in B \text{ or } y \in B\}| \leq \frac{200 \cdot |S| \cdot d \cdot \chi_2(S)}{\beta^2 \cdot \varphi^2} \tag{49}$$

In the remaining argument, we derive a lower bound on $|E^{\text{bad}}_S|$ which depends on $\phi^G_{\text{in}}(S)$ and we put it together with (49) to complete the proof.

First, we show that the induced subgraph $G[S]$ is hierarchically-clusterable and thus, by the inductive assumption for any $h' \geq h+1$ the spectral gap is linear. To this end, note that for any cluster S' that is a descendant of S in G we have $\phi^{G[S]}_{\text{in}}(S') = \phi^G_{\text{in}}(S')$ and $\phi^{G[S]}_{\text{out}}(S') \leq \phi^G_{\text{out}}(S')$.

Therefore, $G[S]$ is (k', γ) hierarchically-clusterable (Definition 6) with $k' = |\{C \in \mathcal{P}^H : C \subseteq S\}|$. This means the graph $G[S]$ satisfies the assumption of the Lemma 11. Let $S_1, \ldots, S_r$ denote the children of cluster S and let M, N denote the partition of these children into two collections of sets that are constructed as per Lemma 11. Therefore, by Lemma 11 for any $S_i, S_j \in \textsc{children}(S)$, $S_i \in M, S_j \in N$ and $C \subseteq S_i$, $C' \subseteq S_j$ such that $C, C' \in \mathcal{P}^H$ we have

$$|\nu_C - \nu_{C'}| \geq \frac{\beta}{2\sqrt{|S|}} \tag{50}$$

Let $V_M = \bigcup_{S \in M} S$ and $V_N = \bigcup_{T \in N} T$. Recall that $\nu(x) = \nu_C$ if $x \in C$. Thus for any $x \in V_M$ and $y \in V_N$ we have

$$|\nu(x) - \nu(y)| \geq \frac{\beta}{2\sqrt{|S|}} \tag{51}$$

Next we prove for any edge $e = (x, y)$ such that $x \in V_M \setminus B$, and $y \in V_N \setminus B$ we have

$$|u_2(x) - u_2(y)| \geq \frac{\beta}{10\sqrt{|S|}}.$$

We read this as saying that the edge $e = (x, y)$ is *long*. Now, we will show that edges $e = (x, y)$ with $x \in V_M \setminus B$ and $y \in V_N \setminus B$ are long. Note that

$$u_2(x) - u_2(y) = (u_2(x) - \nu(x)) + (\nu(x) - \nu(y)) + (\nu(y) - u_2(y))$$

By definition of B we have $(u_2(x) - \nu(x))^2 \leq \frac{\beta^2}{100 \cdot |S|}$ and $(u_2(y) - \nu(y))^2 \leq \frac{\beta^2}{100 \cdot |S|}$. Thus by triangle inequality we have

$$|u_2(x) - u_2(y)| \geq |\nu(x) - \nu(y)| - \frac{2 \cdot \beta}{10\sqrt{|S|}}$$

Thus by (51) for any edge $e = (x, y)$ such that $x \in V_M \setminus B$, and $y \in V_N \setminus B$ we have

$$|u_2(x) - u_2(y)| \geq \frac{\beta}{2\sqrt{|S|}} - \frac{2 \cdot \beta}{10\sqrt{|S|}} \geq \frac{\beta}{10\sqrt{|S|}} \tag{52}$$

Furthermore we have

$$\sum_{(x,y) \in E_S} (u_2(x) - u_2(y))^2 = d \cdot u_2^T (L_S) u_2 = d \cdot \chi_2(S) \tag{53}$$

Therefore by (52) and (53) we get an upperbound on the number of long edges as follows.

$$|\{\{x, y\} \in E_S : x \in V_M \setminus B \text{ and } y \in V_N \setminus B\}| \leq \frac{100 \cdot |S| \cdot d \cdot \chi_2(S)}{\beta^2} \tag{54}$$

We lowerbound $|E_S^{\text{bad}}|$ by the number of bad edges which are long. This gives.

$$\begin{aligned}
&|\{\{x, y\} \in E_S : x \in B \text{ or } y \in B\}| && (55)\\
&\geq |\{\{x, y\} \in E(V_M, V_N) : x \in V_M \cap B \text{ or } y \in V_N \cap B\}| \\
&= |E(V_M, V_N)| - |\{\{x, y\} \in E_S : x \in V_M \setminus B \text{ and } y \in V_N \setminus B\}| \\
&\geq \phi_{\text{in}}^G(S) \cdot \min(|V_M|, |V_N|) \cdot d - \frac{100 \cdot |S| \cdot d \cdot \chi_2(S)}{\beta^2} && \text{By (54)} \\
&\geq |S| \cdot d \left(\phi_{\text{in}}^G(S) \cdot \beta - \frac{100 \cdot \chi_2(S)}{\beta^2} \right) && \text{By Definition 6 and Assumption 1}
\end{aligned}$$

By putting (49) and (55) together we get

$$|S| \cdot d \cdot \frac{200 \cdot \chi_2(S)}{\beta^2 \cdot \varphi^2} \geq |\{\{x,y\} \in E_S : x \in B \text{ or } y \in B\}| \geq |S| \cdot d \left(\phi_{\text{in}}^G(S) \cdot \beta - \frac{100 \cdot \chi_2(S)}{\beta^2} \right)$$

Hence, we have

$$\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}^G(S).$$

□

The following result extends a lemma from [GKL+21] and shows that for any cluster $S \in \mathcal{P}^h$ the Euclidean length squared of the κ-dimensional center of S is roughly $1/|S|$ where $\kappa = |\mathcal{P}^h|$. Also for every $S \neq S' \in \mathcal{P}^h$, the center of S and S' are almost orthogonal.

Lemma 12. *(Cluster means)[[GKL+21]] Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$ and $\kappa = |\mathcal{P}^h|$ denote the number of clusters at level h. For any cluster $S \in \mathcal{P}^h$ let $\mu_S \in \mathbb{R}^\kappa$ be the κ-dimensional center of cluster S (Definition 10). Then we have*

1. *For all $S \in \mathcal{P}^h$, $\left| ||\mu_S||_2^2 - \frac{1}{|S|} \right| \leq \frac{4 \cdot \gamma^{1/4}}{|S|}$*

2. *For all $S \neq S' \in \mathcal{P}^h$, $|\langle \mu_S, \mu_{S'} \rangle| \leq \frac{8 \cdot \gamma^{1/4}}{\sqrt{|S| \cdot |S'|}}$*

Proof. By Assumption 1 we have $\beta \geq \gamma^{1/30}$ and $\varphi \geq \gamma^{1/20}$ and for any cluster $S \in \mathcal{P}^h$ we have $\phi_{\text{in}}(S) \geq \varphi_h$. Thus by Lemma 3 for any cluster $S \in \mathcal{P}^h$ we have

$$\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \varphi_h \geq \frac{\gamma^{1/5}}{300} \cdot \varphi_h$$

Also by Lemma 9 we have $\lambda_\kappa \leq O(\varphi_{h-1})$. Note that by Definition 6 we have $\varphi_{h-1} = \gamma \cdot \varphi_h$. Therefore, we have

$$\sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}} \leq \sqrt{\frac{O(\varphi_{h-1})}{\frac{\gamma^{1/5}}{300} \cdot \varphi_h}} \leq \gamma^{1/4}$$

where the last step follows by taking γ to be sufficiently small. Therefore by Lemma 8 for any $S \in \mathcal{P}^h$ we have

$$\left| ||\mu_S||_2^2 - \frac{1}{|S|} \right| \leq 4 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}} \cdot \frac{1}{|S|} \leq \frac{4 \cdot \gamma^{1/4}}{|S|}$$

and for any $S \neq S' \in \mathcal{P}^h$ we have

$$|\langle \mu_S, \mu_{S'} \rangle| \leq 8 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}} \cdot \frac{1}{\sqrt{|S| \cdot |S'|}} \leq \frac{8 \cdot \gamma^{1/4}}{\sqrt{|S| \cdot |S'|}}$$

□

4.5 Concentration of Vertices Around the Center of Their Ancestors

The main result of this subsection is Lemma 13. This lemma shows that for every cluster $S \in \mathcal{P}^h$, and any d'-dimensional subspace, at most $O(\varphi_{h-1} \cdot d')$ fraction of vertices in S are far from their cluster center μ_S in the projected subspace. This is a key step in showing that the ball carving procedure correctly classifies at least $1 - O(\varphi_{h-1})$ fraction of vertices in every cluster (Theorem 3). To prove this we first show that for at least $1 - O(\varphi_{h-1})$ fraction of vertices the spectral embedding of vertex x i.e., f_x^κ is is close to the the center of the base cluster C containing x i.e., μ_C in the projected subspace. Next, we bound the distance between μ_C and μ_S by Lemma 10.

Lemma 13. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $\Pi \in \mathbb{R}^{\kappa\times\kappa}$ be an arbitrary orthogonal projection matrix onto a subspace of dimension d'.Then for every $\delta \geq 4 \cdot \gamma^{1/4}$ and every cluster $S \in \mathcal{P}^h$ at level h we have*

$$\left|\left\{x \in S : ||\Pi f_x^\kappa - \Pi\mu||_2^2 \geq \frac{\delta}{|S|}\right\}\right| \leq \left(\frac{d'}{\delta \cdot \varphi^2}\right) \cdot O(\varphi_{h-1}) \cdot |S|$$

where $\kappa = |\mathcal{P}^h|$ is the number of clusters at level h, $\mu \in \mathbb{R}^\kappa$ is the κ-dimensional center of cluster S (Definition 10), and $\varphi = \varphi_H$.

Proof. For any cluster $C \in \mathcal{P}^H$ let $\mu_C \in \mathbb{R}^\kappa$ be the κ-dimensional center of cluster C. For any $x \in C$, let $\mu_x = \mu_C$. Note that by triangle inequality we have

$$||\Pi f_x^\kappa - \Pi\mu||_2 \leq ||\Pi f_x^\kappa - \Pi\mu_x||_2 + ||\Pi\mu_x - \Pi\mu||_2$$

We bound the two terms on the Right Hand Side separately.

Bounding $||\Pi f_x^\kappa - \Pi\mu_x||_2$: By Lemma 5 applied to the partition $\mathcal{P}^H = \{C_1, \ldots, C_k\}$, for any $\alpha \in \mathbb{R}^\kappa$ with $\|\alpha\| = 1$ we have

$$\sum_{x\in V} \langle f_x^\kappa - \mu_x, \alpha\rangle^2 = \sum_{C\in\mathcal{P}^H}\sum_{x\in C} \langle f_x^\kappa - \mu_C, \alpha\rangle^2 \leq \frac{\lambda_\kappa}{\min_{C\in\mathcal{P}^H} \chi_2(C)}.$$

By Cheeger's inequality for any cluster $C \in \mathcal{P}^H$ we have $\chi_2(C) \geq \frac{\varphi^2}{2}$. Also by Lemma 9 we have $\lambda_\kappa \leq O(\varphi_{h-1})$. Thus for any $\alpha \in \mathbb{R}^\kappa$ with $\|\alpha\| = 1$ we have

$$\sum_{x\in V} \langle f_x^\kappa - \mu_x, \alpha\rangle^2 \leq \frac{\lambda_\kappa}{\min_{C\in\mathcal{P}^H} \chi_2(C)} \leq \frac{O(\varphi_{h-1})}{\varphi^2}$$

Let $\alpha_1, \ldots, \alpha_{d'}$ be an orthonormal basis for the columnspace Π. Thus, on applying Lemma 5 to directions $\{\alpha_i\}_{i\in[d']}$, we get

$$\sum_{x\in V} ||\Pi f_x^\kappa - \Pi\mu_x||_2^2 = \sum_{i=1}^{d'}\sum_{x\in V} \langle f_x^\kappa - \mu_x, \alpha_i\rangle^2 \leq d' \cdot \frac{1}{\varphi^2} \cdot O(\varphi_{h-1}),$$

and

$$\left|\left\{x \in S : ||\Pi f_x^\kappa - \Pi\mu_x||_2^2 \geq \frac{\delta}{4} \cdot \frac{1}{|S|}\right\}\right| \leq |S| \cdot O(\varphi_{h-1}) \cdot \left(\frac{d'}{\delta \cdot \varphi^2}\right) \tag{56}$$

Bounding $||\Pi\mu_x - \Pi\mu||_2$: Fix $x \in V$ and let $C \in \mathcal{P}^H$ denote the cluster which contains x. Write $\mu_x = \mu_C$ as before. By Lemma 3 for any $h \in [H]$ and any cluster $S \in \mathcal{P}^h$ at level h we have $\chi_2(S) \geq \frac{\beta^3\cdot\varphi^2}{300} \cdot \phi_{\text{in}}^G(S)$. Therefore by Lemma 10 for any cluster C that is a descendant of the cluster S we have

$$||\mu_C - \mu||_2^2 \leq \frac{\gamma^{1/4}}{|S|} \leq \frac{\delta}{4} \cdot \frac{1}{|S|}.$$

The last inequality holds since $\delta \geq 4\gamma^{1/4}$. Note that $||\Pi||_2 = 1$, thus for any $C \subseteq S$ we have

$$||\Pi\mu_C - \Pi\mu||_2^2 \leq ||\mu_C - \mu||_2^2 \leq \frac{\delta}{4} \cdot \frac{1}{|S|} \tag{57}$$

Thus, by (57), we have

$$||\Pi\mu_x - \Pi\mu||_2^2 = ||\Pi\mu_C - \Pi\mu||_2^2 \leq \frac{\delta}{4} \cdot \frac{1}{|S|} \tag{58}$$

Putting it together: By (56) for at least $|S| \cdot \left(1 - \frac{d' \cdot O(\varphi_{h-1})}{\delta \cdot \varphi^2}\right)$ vertices in S we have

$$\begin{aligned}
||\Pi f_x^\kappa - \Pi\mu||_2 &\leq ||\Pi f_x^\kappa - \Pi\mu_x||_2 + ||\Pi\mu_x - \Pi\mu||_2 && \text{By triangle inequality} \\
&\leq \sqrt{\frac{\delta}{4} \cdot \frac{1}{|S|}} + \sqrt{\frac{\delta}{4} \cdot \frac{1}{|S|}} && \text{By (58) and (56)} \\
&\leq \sqrt{\frac{\delta}{|S|}}
\end{aligned}$$

Therefore, we have $||\Pi f_x^\kappa - \Pi\mu||_2^2 \leq \frac{\delta}{|S|}$. □

4.6 Span of the Embedding of Vertices in the Parent Is Close to a Matrix of Small Rank

The main result of this subsection is Lemma 14 which shows that for cluster $S^* \in \mathcal{P}^{h-1}$, the subgraph projection matrix of S^* (with respect to $r = \text{CHILDREN}(S^*)$ and $\kappa = |\mathcal{P}^h|$) is a good approximation of the subspace spanned by the means of children of S^* in the operator norm. Further, this closeness in operator norm also holds for any set Q that is *D-hierarchically-close* to S as defined below:

Definition 18. Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph and let T be the tree corresponding to the nested partitions. For cluster $S^* \in \mathcal{P}^{h^*}$ at level h^*, and for any level $h < h^*$, we define the set of cousins of the cluster S^* for level h as follows:

$$\text{Cousins}_h(S^*) = \left\{ S \in \mathcal{P}^{h^*} : \text{LCA}(S^*, S) \text{ is at level } h \right\}.$$

We denote the set of vertices in $\text{Cousins}_h(S^*)$ by $B_h(S^*)$:

$$B_h(S^*) = \{x \in \text{Cousins}_h(S^*)\}.$$

Definition 19. (D-hierarchically-close sets) Let $h^* \in [H]$, $S^* \in \mathcal{P}^{h^*}$ be a cluster at level h^* and $Q \subseteq V$. We say that the set Q is D-hierarchically-close to the cluster S^* if

1. $|S^* \setminus Q| \leq D \cdot \varphi_{h^*-1} \cdot |S^*|$
2. For any $0 \leq h \leq h^* - 1$, $|Q \cap B_h(S^*)| \leq |S^*| \cdot \left(\frac{D \cdot \varphi_h}{\beta^{h^*-1-h}}\right)$ (Definition 18).

Note that the definition of D-hierarchically-close depends on the level of the cluster S^*, i.e., h^*.

Lemma 14. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Fix $h \in [H]$, and take a cluster $S^* \in \mathcal{P}^{h-1}$. Let $r = |\text{CHILDREN}(S^*)|$. Then for every $D \geq r$ and every set $Q \subseteq V$ that is D-hierarchically-close to the cluster S^* (Definition 19) we have*

$$\left|\left| \sum_{S \in \text{CHILDREN}(S^*)} |S| \mu_S \mu_S^T - \Pi \right|\right|_2 \leq 30 \cdot D \cdot \gamma^{1/4}$$

where $\kappa = |\mathcal{P}^h|$ and $\Pi \in \mathbb{R}^{\kappa \times \kappa}$ is the subgraph projection matrix of Q with respect to κ and r (Definition 12). Also for any cluster S, $\mu_S \in \mathbb{R}^\kappa$ is the κ-dimensional center of S (Definition 10).

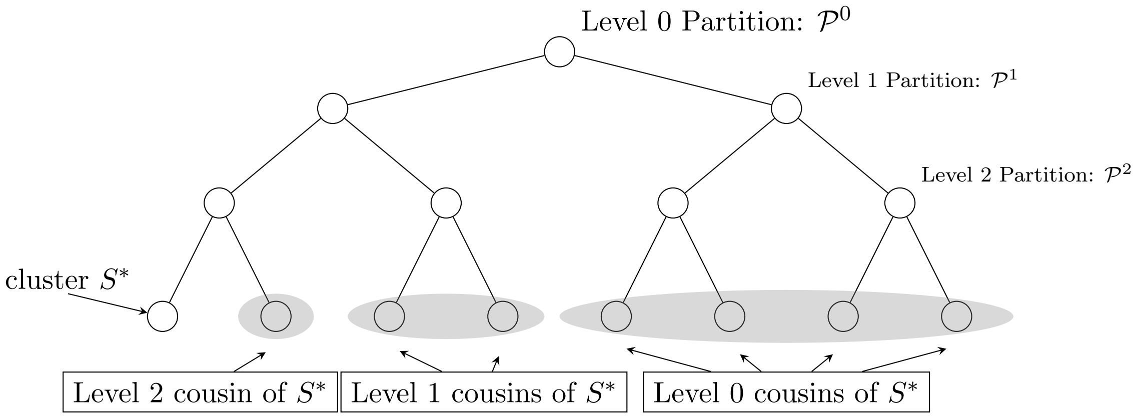

Level 0 Partition: $\mathcal{P}^0$
Level 1 Partition: $\mathcal{P}^1$
Level 2 Partition: $\mathcal{P}^2$
cluster S^*
Level 2 cousin of S^*
Level 1 cousins of S^*
Level 0 cousins of S^*

Figure 3: An illustration of cousins of a level 2 cluster S^* at various levels h, denoted by $\textsc{Cousins}_h(S^*)$ (Definition 18). Level h cousins contain clusters whose LCA with S^* is at level h in the tree.

To prove Lemma 14 we first need to prove Lemma 15. This lemma proves an intuitive intermediate result and shows that for cluster $S^* \in \mathcal{P}^{h-1}$ and any set Q that is D-hierarchically-close to S^*, the projection matrix onto the subspace spanned by the means of $S \in \textsc{children}(S^*)$ is close to the projection matrix onto the subspace spanned by κ-dimensional embeddings of vertices in Q (where $\kappa = |\mathcal{P}^h|$). The following simple proposition would be helpful.

Proposition 1. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6) and let T be the tree representation of the associated ground truth clustering $\mathcal{P}$ of G. Let $S_1 \neq S_2 \in \mathcal{P}^{h^*}$ be two clusters at level h^*. Suppose that $\mathrm{LCA}(S_1, S_2)$ is at level $h < h^*$. Then we have*

$$\beta^{(h^*-h)} \leq \frac{|S_1|}{|S_2|} \leq \left(\frac{1}{\beta}\right)^{(h^*-h)}$$

Proof. Let $S = \mathrm{LCA}(S_1, S_2)$. By Assumption 1 we have $\beta^{(h^*-h)} \cdot |S| \leq |S_1| \leq |S|$ and $\beta^{(h^*-h)} \cdot |S| \leq |S_2| \leq |S|$. Thus we have $\beta^{(h^*-h)} \leq \frac{|S_1|}{|S_2|} \leq \left(\frac{1}{\beta}\right)^{(h^*-h)}$. □

Lemma 15. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Then for every level $h \in [H]$, every cluster $S^* \in \mathcal{P}^{h-1}$, every $D \geq 1$ and every set $Q \subseteq V$ that is D-hierarchically-close to the cluster S^* (Definition 19) we have*

$$\left\|\sum_{S \in \textsc{children}(S^*)} |S| \mu_S \mu_S^T - \sum_{x \in Q} f_x^\kappa f_x^{\kappa T}\right\|_2 \leq 5 \cdot D \cdot \gamma^{1/4}$$

where $\kappa = |\mathcal{P}^h|$ is the number of clusters at level h, and for any cluster S, $\mu_S \in \mathbb{R}^\kappa$ is the κ-dimensional center of S (Definition 10).

Proof. By Lemma 6 we have

$$\left\|\sum_{S \in \mathcal{P}^h} |Q \cap S| \mu_S \mu_S^T - \sum_{x \in Q} f_x^\kappa f_x^{\kappa T}\right\|_2 \leq 2 \cdot \sqrt{\frac{\lambda_\kappa}{\min_{S \in \mathcal{P}^h} \chi_2(S)}}.$$

By Lemma 3 for any cluster $S \in \mathcal{P}^h$ we have $\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}^G(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \varphi_h \geq \frac{\gamma^{1/5}}{300} \cdot \varphi_h$. By Lemma 9 we have $\lambda_\kappa \leq O(\varphi_{h-1})$. Thus we have

$$\left\|\sum_{S \in \mathcal{P}^h} |Q \cap S| \mu_S \mu_S^T - \sum_{x \in Q} f_x^\kappa f_x^{\kappa T}\right\|_2 \leq 2 \cdot \sqrt{\frac{O(\varphi_{h-1})}{\frac{\gamma^{1/5}}{300} \cdot \varphi_h}} \leq \gamma^{1/4}, \tag{59}$$

where the last inequality holds as $\varphi_{h-1} = \gamma \cdot \varphi_h$ and by choice of γ to be sufficiently small to cancel hidden constants in $O(.)$. Note that

$$\begin{aligned}\sum_{S\in\mathcal{P}^h} |Q\cap S|\mu_S\mu_S^T &= \sum_{S\in\textsc{children}(S^*)} |Q\cap S|\mu_S\mu_S^T + \sum_{S\in\mathcal{P}^h\setminus\textsc{children}(S^*)} |Q\cap S|\mu_S\mu_S^T \\ &= \sum_{S\in\textsc{children}(S^*)} (|S| - |S\setminus Q|)\,\mu_S\mu_S^T + \sum_{S\in\mathcal{P}^h\setminus\textsc{children}(S^*)} |Q\cap S|\mu_S\mu_S^T\end{aligned}$$

By triangle inequality we have

$$\begin{aligned}&\left|\left|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - \sum_{S\in\mathcal{P}^h} |Q\cap S|\mu_S\mu_S^T\right|\right|_2 \\ &\leq \left|\left|\sum_{S\in\textsc{children}(S^*)} (|S\setminus Q|)\,\mu_S\mu_S^T\right|\right|_2 + \left|\left|\sum_{S\in\mathcal{P}^h\setminus\textsc{children}(S^*)} |Q\cap S|\mu_S\mu_S^T\right|\right| \end{aligned} \tag{60}$$

Putting (59) and (60) together and by triangle inequality we have

$$\begin{aligned}&\left|\left|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - \sum_{x\in Q} f_x^\kappa f_x^{\kappa T}\right|\right|_2 \leq \\ &\gamma^{1/4} + \left|\left|\sum_{S\in\textsc{children}(S^*)} |S\setminus Q|\mu_S\mu_S^T\right|\right|_2 + \left|\left|\sum_{S\in\mathcal{P}^h\setminus\textsc{children}(S^*)} |Q\cap S|\mu_S\mu_S^T\right|\right| \end{aligned} \tag{61}$$

In the rest of the proof, we will upper bound the second and the third term of (61).

Step 1: First we prove an upper bound on $\left|\left|\sum_{S\in\textsc{children}(S^*)} |S\setminus Q|\mu_S\mu_S^T\right|\right|_2$. Note that by triangle inequality

$$\left|\left|\sum_{S\in\textsc{children}(S^*)} |S\setminus Q|\mu_S{\mu_S}^T\right|\right|_2 \leq \sum_{S\in\textsc{children}(S^*)} |S\setminus Q|\cdot||\mu_S||_2^2 \tag{62}$$

By Lemma 12 and for small enough γ we have

$$||\mu_S||_2^2 \leq \left(1 + 4\cdot\gamma^{1/4}\right)\cdot\frac{1}{|S|} \leq \frac{2}{|S|} \tag{63}$$

Note that S^* is a cluster at level $h-1$, and $Q\subseteq V$ is D-hierarchically-close to cluster S^*. Thus by Definition 19 we have

$$|S^*\setminus Q| \leq D\cdot\varphi_{h-2}\cdot|S^*| \tag{64}$$

Thus by (62) and (63) we have

$$\begin{aligned}
\left\|\sum_{S\in\text{CHILDREN}(S^*)}|S\setminus Q|\mu_S\mu_S^T\right\|_2 &\leq \sum_{S\in\text{CHILDREN}(S^*)}\frac{2\cdot|S\setminus Q|}{|S|} \\
&\leq \sum_{S\in\text{CHILDREN}(S^*)}\frac{2\cdot|S\setminus Q|}{|S^*|\cdot\beta} && \text{By Definition 6, } |S|\geq\beta\cdot|S^*| \\
&= \frac{2}{\beta\cdot|S^*|}\cdot|S^*\setminus Q| && \text{Since } S^*=\bigcup_{S\in\text{CHILDREN}(S^*)} \\
&\leq \frac{2}{\beta\cdot|S^*|}\cdot D\cdot\varphi_{h-2}\cdot|S^*| && \text{By (64)} \\
&\leq \frac{2\cdot D\cdot\gamma\cdot\varphi_{h-1}}{\beta} && \text{Since } \varphi_{h-2}=\varphi_{h-1}\cdot\gamma \\
&\leq 2\cdot D\cdot\gamma^{1/4} && \text{Since } \gamma^{3/4}<\beta \text{ and } \varphi_{h-1}<1 \quad (65)
\end{aligned}$$

Note that $\gamma^{3/4}<\beta$ holds since $\frac{\gamma^{1/30}}{\beta}$ is smaller than a sufficiently small constant by Definition 6.

Step 2: Finally, we prove an upper bound on $\left\|\sum_{S\in\mathcal{P}^h\setminus\text{CHILDREN}(S^*)}|Q\cap S|\left(\mu_S\right)\left(\mu_S\right)^T\right\|$. Note that by Definition 18 we have

$$\bigcup_{h'=0}^{h-2}B_{h'}(S^*)=V\setminus S^*=\bigcup_{S\in\mathcal{P}^h\setminus\text{CHILDREN}(S^*)}S$$

Thus we have

$$\left\|\sum_{S\in\mathcal{P}^h\setminus\text{CHILDREN}(S^*)}|Q\cap S|\mu_S\mu_S^T\right\|=\left\|\sum_{h'=0}^{h-2}\sum_{\substack{S\in\mathcal{P}^h\text{s.t.}\\ S\subseteq B_{h'}(S^*)}}|Q\cap S|\mu_S\mu_S^T\right\| \tag{66}$$

Also note that S^* is a cluster at level $h-1$, and $Q\subseteq V$ is D-hierarchically-close to cluster S^*. Thus by Definition 19 for any $0\leq h'\leq h-2$ we have

$$|Q\cap B_{h'}(S^*)|\leq|S^*|\cdot\left(\frac{D\cdot\varphi_{h'}}{\beta^{h-2-h'}}\right) \tag{67}$$

By triangle inequality we have

$$\left\|\sum_{h'=0}^{h-2} \sum_{\substack{S\in\mathcal{P}^h \text{s.t.}\\ S\subseteq B_{h'}(S^*)}} |Q\cap S|\mu_S\mu_S^T\right\|$$

$$\begin{aligned}
&\leq \sum_{h'=0}^{h-2} \sum_{\substack{S\in\mathcal{P}^h \text{s.t.}\\ S\subseteq B_{h'}(S^*)}} |Q\cap S|\cdot ||\mu_S||_2^2 && \text{By triangle inequality}\\
&\leq \sum_{h'=0}^{h-2} \sum_{\substack{S\in\mathcal{P}^h \text{s.t.}\\ S\subseteq B_{h'}(S^*)}} \frac{2\cdot|Q\cap S|}{|S|} && \text{By (63)}\\
&\leq \sum_{h'=0}^{h-2} \sum_{\substack{S\in\mathcal{P}^h \text{s.t.}\\ S\subseteq B_{h'}(S^*)}} \frac{2\cdot|Q\cap S|}{|S^*|\cdot\beta^{h-h'}} && \text{By Proposition (1) and definition of } B_{h'}(S^*)\\
&= \sum_{h'=0}^{h-2} \frac{2\cdot|Q\cap B_{h'}(S^*)|}{|S^*|\cdot\beta^{h-h'}}
\end{aligned}$$

Therefore, we have

$$\left\|\sum_{h'=0}^{h-2} \sum_{S\in\textsc{Cousins}_{h'}(S^*)} |Q\cap S|\mu_S\mu_S^T\right\|$$

$$\begin{aligned}
&\leq \sum_{h'=0}^{h-2} \frac{2\cdot|Q\cap B_{h'}(S^*)|}{|S^*|\cdot\beta^{h-h'}}\\
&\leq \sum_{h'=0}^{h-2} \frac{2\cdot D\cdot\varphi_{h'}\cdot|S^*|}{|S^*|\cdot\beta^{(2h-2-2h')}} && \text{Since } |Q\cap B_{h'}(S^*)|\leq|S^*|\cdot\left(\frac{D\cdot\varphi_{h'}}{\beta^{(h-2-h')}}\right) \text{ by (67)}\\
&= 2\sum_{h'=0}^{h-2} \frac{D\cdot\varphi_h\cdot\gamma^{h-h'}}{\beta^{2h-2-2h'}}
\end{aligned}$$

Since $\frac{\gamma}{\beta^2}\leq 1/2$, by a geometric sum we get

$$\left\|\sum_{h'=0}^{h-2} \sum_{S\in\textsc{Cousins}_{h'}(S^*)} |Q\cap S|\mu_S\mu_S^T\right\| \leq 2D\cdot\varphi_h\cdot\beta^2\cdot\sum_{h'=0}^{h-2}\left(\frac{\gamma}{\beta^2}\right)^{h-h'} \leq 2D\cdot\varphi_h\cdot\beta^2\cdot\frac{2\cdot\gamma^2}{\beta^4} \leq 2\cdot D\cdot\gamma^{1/4} \tag{68}$$

The last inequality holds since $\frac{\gamma^{1/30}}{\beta}$ is smaller than a sufficiently small constant by Definition 6. Putting (61), (65) and (68) together we get

$$\begin{aligned}
&\left\|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - \sum_{x\in Q} f_x^\kappa f_x^{\kappa T}\right\|_2 \leq\\
&\gamma^{1/4} + \left\|\sum_{S\in\textsc{children}(S^*)} |S\setminus Q|\mu_S\mu_S^T\right\|_2 + \left\|\sum_{h'=0}^{h-2}\sum_{S\in\textsc{Cousins}_{h'}(S^*)} |Q\cap S|\mu_S\mu_S^T\right\| && \text{By (61)}\\
&\leq \gamma^{1/4} + 2\cdot D\cdot\gamma^{1/4} + 2\cdot D\cdot\gamma^{1/4} && \text{By (65) and (68)}\\
&\leq 5\cdot D\cdot\gamma^{1/4} && \text{Since } D\geq 1
\end{aligned}$$

□

We need the next lemma to prove Lemma 17. Recall that for a symmetric matrix H, we write $\nu_i(H)$ (resp. $\nu_{\max}(H), \nu_{\min}(H)$) to denote the i^{th} largest (resp. maximum, minimum) eigenvalue of H.

Lemma 16 (Weyl's Inequality). *Let $H, P \in \mathbb{R}^{n\times n}$ be two symmetric matrices. Then we have for all $i \in \{1, \ldots, n\}$:*

$$\nu_i(H) + \nu_{\min}(P) \leq \nu_i(H+P) \leq \nu_i(H) + \nu_{\max}(P),$$

where for a symmetric matrix $H \in \mathbb{R}^{n\times n}$ $\nu_i(H)$ denotes its i^{th} largest eigenvalue and $\nu_{\min}(H)$ and $\nu_{\max}(H)$ refer to the smallest and largest eigenvalues of H.

To prove Lemma 14 we need the following ingredient.

Lemma 17. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$ and take $S^* \in \mathcal{P}^{h-1}$. Let $r = |\textsc{children}(S^*)|$. Then, for every $D \geq r$ and every set $Q \subseteq V$ that is D-hierarchically-close to the cluster S^* (Definition 19) the following holds:*

1. $\nu_{r+1}\left(AA^T\right) \leq 5 \cdot D \cdot \gamma^{1/4}$
2. $\nu_r\left(AA^T\right) \geq 1 - 13 \cdot D \cdot \gamma^{1/4}$

where $\kappa = |\mathcal{P}^h|$, and $A \in \mathbb{R}^{\kappa\times|Q|}$ is a matrix whose columns are f_x^κ for all $x \in Q$.

Proof. By the definition of A above, we have $AA^T = \sum_{x\in Q} f_x^\kappa f_x^{\kappa T}$. Let $H = \sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T$. We define P as follows:

$$P = AA^T - H = \sum_{x\in Q} f_x^\kappa f_x^{\kappa T} - \sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T,$$

by Lemma 15 we have

$$||P||_2 = \left|\left|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - \sum_{x\in Q} f_x^\kappa f_x^{\kappa T}\right|\right|_2 \leq 5 \cdot D \cdot \gamma^{1/4}.$$

Proof of item (1): Weyl's Inequality (Lemma 16) gives

$$\nu_{r+1}(H+P) \leq \nu_{r+1}(H) + \nu_{\max}(P),$$

and since $\nu_{\max}(P) = ||P||_2 \leq 5 \cdot D \cdot \gamma^{1/4}$, and $H + P = AA^T$. Thus we get

$$\nu_{r+1}\left(AA^T\right) \leq \nu_{r+1}\left(\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T\right) + 5 \cdot D \cdot \gamma^{1/4}$$

Note that S has r children i.e, ($|\textsc{children}(S^*)| = r$), thus the matrix $H = \sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T$ is of rank at most r. Thus we have $\nu_{r+1}(H) = 0$ and therefore,

$$\nu_{r+1}\left(AA^T\right) \leq 5 \cdot D \cdot \gamma^{1/4}$$

Proof of item (2)**:** By Weyl's Inequality (Lemma 16) we have

$$\nu_r(H+P) \geq \nu_r(H) + \nu_{\min}(P) \geq \nu_r(H) - ||P||_2$$

Note that $\nu_{\max}(P) = ||P||_2 \leq 5 \cdot D \cdot \gamma^{1/4}$, and $H + P = AA^T$. Therefore, we have

$$\nu_r\left(AA^T\right) \geq \nu_r\left(\sum_{S \in \text{CHILDREN}(S^*)} |S|\mu_S\mu_S^T\right) - 5 \cdot D \cdot \gamma^{1/4} \tag{69}$$

Recall that $|\text{CHILDREN}(S^*)| = r$. Let $Y \in \mathbb{R}^{r \times r}$ be a matrix whose columns are indexed by the set $\text{CHILDREN}(S^*)$. In particular, for $S \in \text{CHILDREN}(S^*)$, the corresponding column in the vector $\sqrt{|S|} \cdot \mu_S$. Thus we have

$$\nu_r\left(\sum_{S \in \text{CHILDREN}(S^*)} |S|\mu_S\mu_S^T\right) = \nu_{\min}\left(YY^T\right) = \nu_{\min}(Y^TY) \tag{70}$$

Note that $Y^TY \in \mathbb{R}^{r \times r}$ is a matrix where $Y^TY(S,S)$ equals $|S| \cdot ||\mu_S||_2^2$ for every $S \in \text{CHILDREN}(S^*)$ and its off-diagonals are $\left\langle \sqrt{|S|} \cdot \mu_S, \sqrt{|S'|} \cdot \mu_{S'} \right\rangle$ for every $S \neq S' \in \text{CHILDREN}(S^*)$. Note that by Lemma 12 item (1) we have $|S| \cdot ||\mu_S||_2^2 \leq 1 + 4 \cdot \gamma^{1/4}$, and by (2) we have $\left\langle \sqrt{|S|} \cdot \mu_S, \sqrt{|S'|} \cdot \mu_{S'} \right\rangle \leq 8 \cdot \gamma^{1/4}$. Therefore,

$$||Y^TY - I||_2 \leq ||Y^TY - I||_F \leq \sqrt{r^2 \cdot (8 \cdot \gamma^{1/4})^2} \leq 8 \cdot r \cdot \gamma^{1/4}$$

Thus by Weyl's Inequality (Lemma 16) we have

$$\nu_{\min}(Y^TY) \geq \nu_{\min}(I) - ||Y^TY - I||_2 \geq 1 - 8 \cdot r \cdot \gamma^{1/4} \tag{71}$$

Thus by (70) we have

$$\nu_r\left(\sum_{S \in \text{CHILDREN}(S^*)} |S|\mu_S\mu_S^T\right) = \nu_{\min}(Y^TY) \geq 1 - 8 \cdot r \cdot \gamma^{1/4}$$

Note that $D \geq r$, therefore, we have

$$\nu_r\left(\sum_{S \in \text{CHILDREN}(S^*)} |S|\mu_S\mu_S^T\right) \geq 1 - 8 \cdot r \cdot \gamma^{1/4} \geq 1 - 8 \cdot D \cdot \gamma^{1/4} \tag{72}$$

Therefore by (69), and (72) we have

$$\begin{aligned}&\nu_r\left(AA^T\right) \geq \nu_r\left(\sum_{S' \in \text{CHILDREN}(S^*)} |S|\mu_S\mu_S^T\right) - 5 \cdot D \cdot \gamma^{1/4}\\ &\geq 1 - 8 \cdot D \cdot \gamma^{1/4} - 5 \cdot D \cdot \gamma^{1/4}\\ &\geq 1 - 13 \cdot D \cdot \gamma^{1/4}\end{aligned}$$

□

Now we prove the main lemma of this subsection, Lemma 14 below.

Lemma 14. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Fix $h \in [H]$, and take a cluster $S^* \in \mathcal{P}^{h-1}$. Let $r = |\textsc{children}(S^*)|$. Then for every $D \geq r$ and every set $Q \subseteq V$ that is D-hierarchically-close to the cluster S^* (Definition 19) we have*

$$\left|\left|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - \Pi\right|\right|_2 \leq 30 \cdot D \cdot \gamma^{1/4}$$

where $\kappa = |\mathcal{P}^h|$ and $\Pi \in \mathbb{R}^{\kappa\times\kappa}$ is the subgraph projection matrix of Q with respect to κ and r (Definition 12). Also for any cluster S, $\mu_S \in \mathbb{R}^\kappa$ is the κ-dimensional center of S (Definition 10).

Proof. Let $A \in \mathbb{R}^{\kappa\times|Q|}$ be a matrix whose columns are f_x^κ for all $x \in Q$. Thus we have $AA^T = \sum_{x\in Q} f_x^\kappa f_x^{\kappa T}$. Note that by Lemma 15 we have

$$\left|\left|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - AA^T\right|\right|_2 \leq 5 \cdot D \cdot \gamma^{1/4}$$

Thus by triangle inequality we have

$$\left|\left|\sum_{S\in\textsc{children}(S)} |S|\mu_S\mu_S^T - \Pi\right|\right|_2 \leq 5 \cdot D \cdot \gamma^{1/4} + ||AA^T - \Pi||_2 \tag{73}$$

Let $AA^T = Y\Gamma Y^T$ be the eigendecomposition of AA^T. Therefore, by Definition 12 we have $\Pi = Y_{[r]}Y_{[r]}^T$. Thus we need to upper bound $||AA^T - Y_{[r]}Y_{[r]}^T||_2$. Note that

$$AA^T = Y\Gamma Y^T = Y_{[r]}\Gamma_{[r]}Y_{[r]}^T + Y_{[-r]}\Gamma_{[-r]}Y_{[-r]}^T$$

Thus we have

$$\begin{aligned}
||AA^T - Y_{[r]}Y_{[r]}^T||_2 &= ||Y_{[r]}\Gamma_{[r]}Y_{[r]}^T + Y_{[-r]}\Gamma_{[-r]}Y_{[-r]}^T - Y_{[r]}Y_{[r]}^T||_2 \\
&\leq ||Y_{[r]}\Gamma_{[r]}Y_{[r]}^T - Y_{[r]}Y_{[r]}^T||_2 + ||Y_{[-r]}\Gamma_{[-r]}Y_{[-r]}^T||_2 && \text{By triangle inequality} \\
&= ||Y_{[r]}(I - \Gamma_{[r]})Y_{[r]}^T||_2 + ||Y_{[-r]}\Gamma_{[-r]}Y_{[-r]}^T||_2 \\
&= \left(1 - \nu_r(AA^T)\right) + \nu_{r+1}\left(AA^T\right) \\
&\leq 13 \cdot D \cdot \gamma^{1/4} + 5 \cdot D \cdot \gamma^{1/4} && \text{By Lemma 17} \\
&\leq 18 \cdot D \cdot \gamma^{1/4}
\end{aligned}$$

Recall that $\Pi = Y_{[r]}Y_{[r]}^T$. Thus by (73) we have

$$\left|\left|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - \Pi\right|\right|_2 \leq 5 \cdot D \cdot \gamma^{1/4} + ||AA^T - Y_{[r]}Y_{[r]}^T||_2 \leq 30 \cdot D \cdot \gamma^{1/4}.$$

□

4.7 Centers of Subclusters Remain Far in the Projected Subspace

The main result of this subsection is Lemma 18. This lemma uses the properties of (k, γ)-hierarchically clusterable instances developed in Section 4.4 and Section 4.6 to understand the geometric structure of spectral embeddings in such instances better. In particular, the lemma below shows that for any node $S^* \in \mathcal{P}^{h-1}$, it holds that the κ-dimensional means of its children are pairwise far from each other.

Lemma 18. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). For every level $h \geq 1$, every cluster $S^* \in \mathcal{P}^{h-1}$ with $r = |\text{CHILDREN}(S^*)|$, and every D such that $r \leq D$ and $D \cdot \gamma^{1/4}$ is less than a sufficiently small constant, the following holds: For every set $Q^* \subseteq V$ that is D-hierarchically-close to the cluster S^* (Definition 19) and for every $S_1 \neq S_2 \in \text{CHILDREN}(S^*)$ we have that*

$$||\Pi\mu_1 - \Pi\mu_2||_2^2 \geq \frac{1}{|S^*|}$$

where $\kappa = |\mathcal{P}^h|$, $\mu_1, \mu_2 \in \mathbb{R}^\kappa$ is the κ-dimensional center of clusters S_1 and S_2 respectively (Definition 10), and $\Pi \in \mathbb{R}^{\kappa \times \kappa}$ is the subgraph projection matrix of Q^* with respect to κ and r (Definition 12).

Proof. Note that

$$||\Pi\mu_1 - \Pi\mu_2||_2^2 = ||\Pi\mu_1||_2^2 + ||\Pi\mu_2||_2^2 - 2\left\langle \Pi\mu_1, \Pi\mu_2 \right\rangle \tag{74}$$

In the rest of the proof we will upper bound $||\Pi\mu_1||_2^2$, $||\Pi\mu_2||_2^2$ and lower bound $\langle \Pi\mu_1, \Pi\mu_2 \rangle$.

Step 1: We first prove the upper bound for $||\Pi\mu_1||_2^2$ and $||\Pi\mu_2||_2^2$. Let

$$\Pi^* = \sum_{S \in \text{CHILDREN}(S^*)} |S| \mu_S \mu_S^T \tag{75}$$

Note that Π is a projection matrix, thus we have $\Pi^T \Pi = \Pi$. Therefore, we have

$$\begin{aligned}
||\Pi\mu_1||_2^2 &= \mu_1^T \Pi \mu_1 \\
&= \mu_1^T \left(\sum_{S \in \text{CHILDREN}(S^*)} |S| \mu_S \mu_S^T + \Pi - \Pi^* \right) \mu_1 && \text{By (75)} \\
&= \mu_1^T \left(|S_1| \mu_1 \mu_1^T + \left(\sum_{\substack{S \in \text{CHILDREN}(S^*) \\ S \neq S_1}} |S| \mu_S \mu_S^T \right) + \Pi - \Pi^* \right) \mu_1 \\
&\geq |S_1| \cdot \|\mu_1\|_2^4 - \left(\sum_{\substack{S \in \text{CHILDREN}(S^*) \\ S \neq S_1}} |S| \left\langle \mu_S, \mu_1 \right\rangle^2 \right) - \|\Pi - \Pi^*\|_2 \|\mu_1\|_2^2 && (76)
\end{aligned}$$

Consider the three terms above separately. For the first term, we get

$$\begin{aligned}
|S_1| \cdot \|\mu_1\|_2^4 &\geq |S_1| \cdot \left(\frac{1}{|S_1|} - \frac{4 \cdot \gamma^{1/4}}{|S_1|} \right)^2 && \text{By Lemma 12 item (1)} \\
&\geq \frac{0.9}{|S_1|} && \text{As } \gamma \text{ is sufficiently small} \quad (77)
\end{aligned}$$

Now consider the second term on the rhs of (76). We have

$$\begin{aligned}
&\sum_{S\neq S_1\in \textsc{children}(S^*)} |S|\left\langle \mu_S,\mu_1\right\rangle^2 \\
&\leq \sum_{S\neq S_1\in \textsc{children}(S^*)} |S|\left(\frac{8\cdot\gamma^{1/4}}{\sqrt{|S|}\cdot\sqrt{|S_1|}}\right)^2 && \text{By Lemma 12 item (2)} \\
&\leq \sum_{S\neq S_1\in \textsc{children}(S^*)} \frac{64\cdot\sqrt{\gamma}}{|S_1|} \\
&\leq \frac{64\cdot(r-1)\cdot\sqrt{\gamma}}{|S_1|} && \text{Since } |\textsc{children}(S^*)| = r \\
&\leq \frac{64\cdot\sqrt{\gamma}}{\beta\cdot|S_1|} && \text{By Definition 6, } r = |\textsc{children}(S^*)| \leq \frac{1}{\beta} \\
&\leq \frac{0.1}{|S_1|} && \text{By Definition 6, } \frac{\sqrt{\gamma}}{\beta} \text{ is sufficiently smalls} \qquad (78)
\end{aligned}$$

Now consider the last term on the rhs of (76). Note that by Lemma 14 we have

$$||\Pi^*-\Pi||_2 = \left|\left|\sum_{S\in\textsc{children}(S^*)} |S|\mu_S\mu_S^T - \Pi\right|\right|_2 \leq 30\cdot D\cdot\gamma^{1/4} \tag{79}$$

Thus by (79) we have

$$\begin{aligned}
\|\Pi-\Pi^*\|_2\|\mu\|_2^2 &\leq \left(30\cdot D\cdot\gamma^{1/4}\right)\cdot\frac{1}{|S_1|}\cdot\left(1+4\cdot\gamma^{1/4}\right) && \text{By Lemma 12 item (1)} \\
&\leq \frac{0.1}{|S_1|} && \text{As } D\cdot\gamma^{1/4} \text{ is sufficiently small} \qquad (80)
\end{aligned}$$

Putting together (76), (77),(78) and (80) we have

$$||\Pi\mu_1||_2^2 \geq |S_1|\cdot\|\mu_1\|_2^4 - \left(\sum_{S\neq S_1\in\textsc{children}(S^*)} |S|\left\langle\mu_S,\mu_1\right\rangle^2\right) - \|\Pi-\Pi^*\|_2\|\mu_1\|_2^2 \geq \frac{0.7}{|S_1|} \tag{81}$$

Similarly we have

$$||\Pi\mu_2||_2^2 \geq \frac{0.7}{|S_2|} \tag{82}$$

Step 2: It remains to upper bound $\langle\Pi\mu_1,\Pi\mu_2\rangle$. Recall that $\Pi^* = \sum_{S\in\textsc{children}(S^*)}|S|\left(\mu_S\right)\left(\mu_S\right)^T$. Thus we have

$$\begin{aligned}
\langle\Pi\mu_1,\Pi\mu_2\rangle &= \mu_1^T(\Pi)\mu_2 \\
&= \mu_1^T(\Pi^*+\Pi-\Pi^*)\mu_2 \\
&= \mu_1^T\left(|S_1|\mu_1\mu_1^T + |S_2|\mu_2\mu_2^T + \left(\sum_{S\in\textsc{children}(S^*),S\neq S_1,S_2}|S|\mu_S\mu_S^T\right) + (\Pi-\Pi^*)\right)\mu_2 \\
&\leq |S_1|\cdot\|\mu_1\|_2^2\langle\mu_1,\mu_2\rangle + |S_2|\cdot\|\mu_2\|_2^2\langle\mu_1,\mu_2\rangle \\
&+ \left(\sum_{S\in\textsc{children}(S^*),S\neq S_1,S_2}|S|\left\langle\mu_1,\mu_S\right\rangle\cdot\left\langle\mu_2,\mu_S\right\rangle\right) + \|\Pi-\Pi^*\|_2\langle\mu_1,\mu_2\rangle \qquad (83)
\end{aligned}$$

We consider the four terms on rhs of(83). For the first term we have

$$\begin{aligned}
|S_1| \cdot \|\mu_1\|_2^2 \langle \mu_1, \mu_2 \rangle &\leq |S_1| \cdot \left(\frac{1}{|S_1|} + \frac{4 \cdot \gamma^{1/4}}{|S_1|} \right) \left(\frac{8 \cdot \gamma^{1/4}}{\sqrt{|S_1| \cdot |S_2|}} \right) && \text{By Lemma 12} \\
&\leq 2 \cdot \left(\frac{8 \cdot \gamma^{1/4}}{\sqrt{|S_1| \cdot |S_2|}} \right) && \text{As } 4 \cdot \gamma^{1/4} < 1 \\
&\leq \frac{1}{|S^*|} \cdot \frac{16}{\beta} \cdot \gamma^{1/4} && \text{As } \min(|S_1|, |S_2|) \geq \beta \cdot |S^*| \\
&\leq \frac{0.05}{|S^*|} && \text{As } \frac{\sqrt{\gamma}}{\beta} \text{ is sufficiently small by Definition 6}
\end{aligned} \tag{84}$$

And similarly, we upper bound the second term in (83) by $\frac{0.05}{|S^*|}$. For the third term we have

$$\begin{aligned}
&\sum_{S \in \text{CHILDREN}(S^*), S \neq S_1, S_2} |S| \left\langle \mu_1, \mu_S \right\rangle \cdot \left\langle \mu_2, \mu_S \right\rangle \\
&\leq \sum_{S \in \text{CHILDREN}(S^*), S \neq S_1, S_2} |S| \left(\frac{8 \cdot \gamma^{1/4}}{\sqrt{|S_1| \cdot |S|}} \right) \cdot \left(\frac{8 \cdot \gamma^{1/4}}{\sqrt{|S_2| \cdot |S|}} \right) && \text{By Lemma 12, item 1} \\
&\leq \sum_{S \in \text{CHILDREN}(S^*), S \neq S_1, S_2} \frac{64 \cdot \sqrt{\gamma}}{\sqrt{|S_1| \cdot |S_2|}} \\
&\leq (r-2) \cdot \frac{64 \cdot \sqrt{\gamma}}{\sqrt{|S_1| \cdot |S_2|}} && \text{Since } |\text{CHILDREN}(S^*)| = r \\
&\leq \frac{1}{\beta^2} \cdot \frac{64 \cdot \sqrt{\gamma}}{|S^*|} && \text{By Definition 6, } r \leq \frac{1}{\beta} \text{ and } \min(|S_1|, |S_2|) \geq \beta \cdot |S \\
&\leq \frac{0.05}{|S^*|} && \text{As } \frac{\gamma}{\beta} \text{ is sufficiently small by Definition 6}
\end{aligned} \tag{85}$$

For the last term on the rhs of (83) by (79) and by Lemma 12 item 1 we have

$$\begin{aligned}
\|\Pi - \Pi^*\|_2 \langle \mu_1, \mu_2 \rangle &\leq \left(30 \cdot D \cdot \gamma^{1/4} \right) \cdot \left(\frac{8 \cdot \gamma^{1/4}}{\sqrt{|S_1| \cdot |S_2|}} \right) \\
&\leq \frac{8 \cdot \gamma^{1/4}}{\sqrt{|S_1| \cdot |S_2|}} && \text{As } D \cdot \gamma^{1/4} \text{ is sufficiently small} \\
&\leq \left(\frac{8 \cdot \sqrt{\gamma}}{\beta \cdot |S^*|} \right) && \text{By Definition 6, } \min(|S_1|, |S_2|) \geq \beta \cdot |S^*| \\
&\leq \frac{0.05}{|S^*|} && \text{As } \frac{\sqrt{\gamma}}{\beta} \text{ is sufficiently small by Definition 6}
\end{aligned} \tag{86}$$

Thus by (83), (84),(85),(86) we have

$$\begin{aligned}\langle \Pi\mu_1, \Pi\mu_2\rangle &\le |S| \cdot \|\mu_1\|_2^2 \langle \mu_1, \mu_2\rangle + |S_2| \cdot \|\mu_2\|_2^2 \langle \mu_1, \mu_2\rangle \\ &\quad + \left(\sum_{S \in \textsc{children}(S^*), S \neq S_1, S_2} |S| \langle \mu_1, \mu_S\rangle \cdot \langle \mu_2, \mu_S\rangle \right) + \|\Pi - \Pi^*\|_2 \langle \mu_1, \mu_2\rangle \\ &\le \frac{0.05}{|S^*|} + \frac{0.05}{|S^*|} + \frac{0.05}{|S^*|} + \frac{0.05}{|S^*|} && (87) \\ &\le \frac{0.2}{|S^*|} && (88)\end{aligned}$$

Putting it together Thus by (74), (81), (82) and (87) we have

$$\begin{aligned}||\Pi\mu_1 - \Pi\mu_2\|_2^2 &= ||\Pi\mu_1||_2^2 + ||\Pi\mu_2||_2^2 - 2\langle \Pi\mu_1, \Pi\mu_2\rangle \\ &\ge \frac{0.7}{|S_1|} + \frac{0.7}{|S_2|} - 2 \cdot \frac{0.2}{|S^*|} \\ &\ge \frac{1}{|S^*|}. && \text{Since } |S_1| \le |S^*| \text{ and } |S_2| \le |S^*|\end{aligned}$$

□

4.8 Bounding the Intersection of Candidate Clusters with True Clusters

The main result of this subsection is Theorem 3 that shows the inductive step for the proof of Ball-Carving. First we state the definition of the cylinder of clusters:

Definition 20. (Cylinder) Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$ and $Q^* \subseteq V$ be a set that is D-hierarchically-close to cluster $S^* \in \mathcal{P}^{h-1}$. For any center $\alpha \in \mathbb{R}^\kappa$ and any radius $\ell \in \mathbb{R}$ we define the cylinder of radius ℓ around the center α as follows:

$$\text{cyl}(\alpha, \ell | Q^*) = \left\{ y \in Q^* : ||\Pi\alpha - \Pi f_y^\kappa||_2^2 \le \ell \right\}$$

where, $\kappa = |\mathcal{P}^h|$, $r = |\textsc{children}(S^*)|$, and Π is the subgraph projection matrix of Q^* with respect to κ and r (Definition 12).

Theorem 3. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, $S^* \in \mathcal{P}^{h-1}$, $\ell = \frac{1}{10^3 \cdot |S^*|}$, and $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ where D_0 is a sufficiently large constant. For every set Q^* that is D-hierarchically-close to S^* (Definition 19) the following holds:*

1. *For every $S_1 \neq S_2 \in \textsc{children}(S^*)$, $cyl(\mu_{S_1}, \ell | Q^*) \bigcap cyl(\mu_{S_2}, \ell | Q^*) = \emptyset$*
2. *For every $S \in \textsc{children}(S^*)$, $\left| S \triangle cyl(\mu_S, \ell | Q^*) \right| \le \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S|$*
3. *For every $S \in \textsc{children}(S^*)$, $cyl(\mu_S, \ell | Q^*)$ is D-hierarchically-close to S*

Proof. **Proof of** (1) : Let $r = |\textsc{children}(S^*)|$, $\kappa = |\mathcal{P}^h|$ and $\Pi \in \mathbb{R}^{\kappa \times \kappa}$ be the subgraph projection matrix of Q^* with respect to κ and r (Definition 12). Suppose that $r \ge 2$ and let $S_1 \neq S_2 \in \textsc{children}(S^*)$ be two of the children of S^*. For simplicity of notation we let $\mu_1 = \mu_{S_1}$, $\mu_2 = \mu_{S_1}$. We now define

$$S_1' = \left\{ x \in V : ||\Pi f_x^\kappa - \Pi\mu_1||_2^2 \le \frac{1}{10^3 \cdot |S^*|} \right\}, \text{ and } S_2' = \left\{ x \in V : ||\Pi f_x^\kappa - \Pi\mu_2||_2^2 \le \frac{1}{10^3 \cdot |S^*|} \right\}$$

We first show that $S_1' \cap S_2' = \emptyset$. By contradiction suppose that $S_1' \cap S_2' \neq \emptyset$. Therefore, there exists a vertex $x \in V$ such that $||\Pi f_x^\kappa - \Pi\mu_1\|_2^2 \leq \frac{1}{10^3 \cdot |S^*|}$ and $||\Pi f_x^\kappa - \Pi\mu_2\|_2^2 \leq \frac{1}{10^3 \cdot |S^*|}$. Thus we have

$$\begin{aligned} ||\Pi\mu_1 - \Pi\mu_2\|_2 &\leq ||\Pi\mu_1 - \Pi f_x^\kappa||_2 + ||\Pi f_x^\kappa - \Pi\mu_2||_2 && \text{By triangle inequality} \\ &\leq \frac{1}{30 \cdot \sqrt{|S^*|}} + \frac{1}{30 \cdot \sqrt{|S^*|}} && \text{By definition of } S_1' \text{ and } S_2' \\ &\leq \frac{1}{15 \cdot \sqrt{|S^*|}} && (89) \end{aligned}$$

Note that by Definition 6, for every $S \in \text{CHILDREN}(S^*)$ we have $|S| \geq \beta \cdot |S^*|$, hence,

$$|\text{CHILDREN}(S^*)| = r \leq \frac{1}{\beta} \tag{90}$$

Also note that $\max(\frac{\gamma}{\beta^{30}}, \frac{\gamma}{\varphi^{20}})$ is smaller than a sufficiently small constant by Definition 6. Therefore, by choice of $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ we have $r \leq D \ll \frac{1}{\gamma^{1/4}}$. Thus by Lemma 18 we have

$$||\Pi\mu_1 - \Pi\mu_2\|_2^2 \geq \frac{1}{|S^*|}$$

and this contradicts with (89). Therefore, we have

$$S_1' \cap S_2' = \emptyset \tag{91}$$

Note that $\ell = \frac{1}{10^3 \cdot |S^*|}$, hence, $\text{cyl}(\mu_1, \ell|Q^*) = Q^* \cap S_1'$ and $\text{cyl}(\mu_2, \ell|Q^*) = Q^* \cap S_2'$. Therefore by (91) we have

$$\text{cyl}(\mu_1, \ell|Q^*) \cap \text{cyl}(\mu_2, \ell|Q^*) = Q^* \cap S_1' \cap S_2' = \emptyset$$

Proof of (2)**:** We now prove $\left|\text{cyl}(\mu_S, \ell|Q^*) \triangle S\right| \leq \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S|$ for every $S \in \text{CHILDREN}(S^*)$. Let $\delta = \frac{\beta}{10^3}$. We now define

$$S' = \left\{ x \in V : ||\Pi f_x^\kappa - \Pi\mu_S||_2^2 \leq \frac{1}{10^3 \cdot |S^*|} \right\}, \text{ and } S'' = \left\{ x \in S : ||\Pi f_x^\kappa - \Pi\mu_S||_2^2 \leq \frac{\delta}{|S|} \right\}$$

For any $S \in \text{CHILDREN}(S^*)$ we have $|S| \geq \beta \cdot |S^*|$, thus for any $x \in S''$ we get

$$||\Pi f_x^\kappa - \Pi\mu_S||_2^2 \leq \frac{\delta}{|S|} \leq \frac{\delta}{\beta \cdot |S^*|} = \frac{\beta}{10^3 \cdot \beta \cdot |S^*|} \leq \frac{1}{10^3 \cdot |S^*|},$$

Therefore, we have $S'' \subseteq S'$, also by definition of S'' we have $S'' \subseteq S$. Therefore, we have $S'' \subset S' \cap S$, hence, we get

$$|S' \cap S| \geq |S''| \tag{92}$$

By Definition 6, $\frac{\gamma}{\beta^{30}}$ is a sufficiently small constant, hence, $\delta = \frac{\beta}{10^3} \geq 4 \cdot \gamma^{1/4}$, thus we can apply Lemma 13, and we get

$$|S''| \geq |S| \left(1 - \frac{\text{rank}(\Pi) \cdot O(\varphi_{h-1})}{\delta \cdot \varphi^2}\right) \geq |S| \left(1 - \frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2}\right), \tag{93}$$

where the last inequality holds since $\text{rank}(\Pi) = r \leq \frac{1}{\beta}$ by (90), and $\delta = \frac{\beta}{10^3}$. Putting (93) and (92) together for every $S \in \text{CHILDREN}(S^*)$ we get

$$|S' \cap S| \geq |S''| \geq |S| \cdot \left(1 - \frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2}\right) \tag{94}$$

Note that

$$|\text{cyl}(\mu_S, \ell|Q^*) \triangle S| = |S \setminus \text{cyl}(\mu_S, \ell|Q^*)| + |\text{cyl}(\mu_S, \ell|Q^*) \setminus S| \tag{95}$$

Therefore, we need to upper bound $|S \setminus \text{cyl}(\mu_S, \ell|Q^*)|$ and $|\text{cyl}(\mu_S, \ell|Q^*) \setminus S|$.

Step 1: First we upper bound $|S \setminus \mathrm{cyl}(\mu_S, \ell|Q^*)|$.

$$\begin{aligned}
&|S \setminus \mathrm{cyl}(\mu_S, \ell|Q^*)| \\
&= |S \setminus (Q^* \cap S')| && \text{By definition of } \mathrm{cyl}(\mu_S, \ell|Q^*) \text{ and } S' \\
&\leq |S \setminus Q^*| + |S \setminus S'| \\
&\leq |S^* \setminus Q^*| + |S| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2} \right) && \text{As } S \subseteq S^* \text{ and by (94)} \\
&\leq D \cdot \varphi_{h-2} \cdot |S^*| + |S| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2} \right) && \text{As } S^* \in \mathcal{P}^{h-1} \text{ is } D\text{-hierarchically-close to } Q^*\text{, so } |S^* \setminus Q^*| \leq D \cdot \varphi_{h-2} \cdot \\
&\leq D \cdot \varphi_{h-2} \cdot \frac{|S|}{\beta} + |S| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2} \right) && \text{As } |S| \geq \beta \cdot |S^*| \\
&\leq |S| \cdot \varphi_{h-1} \left(\frac{D \cdot \gamma}{\beta} + \frac{D_1}{\beta^2 \cdot \varphi^2} \right) && \text{As } \varphi_{h-2} = \varphi_{h-1} \cdot \gamma \qquad (96)
\end{aligned}$$

In the last step, D_1 is the constant hidden in $O(.)$ above.

Step 2: Next we upper bound $|\mathrm{cyl}(\mu_S, \ell|Q^*) \setminus S|$.

$$\begin{aligned}
|\mathrm{cyl}(\mu_S, \ell|Q^*) \setminus S| &= |(Q^* \cap S') \setminus S| && \text{By definition of } \mathrm{cyl}(\mu_S, \ell|Q^*) \text{ and } S' \\
&= |((Q^* \cap S^*) \cap S') \setminus S| + |((Q^* \setminus S^*) \cap S') \setminus S| \\
&\leq |S' \cap (S^* \setminus S)| + |Q^* \setminus S^*| \\
&= \left(\sum_{\substack{S_2 \in \textsc{children}(S^*) \\ \text{s.t.} S_2 \neq S}} |S' \cap S_2| \right) + |Q^* \setminus S^*| && (97)
\end{aligned}$$

Thus we need to upper bound $|Q^* \setminus S^*|$ and $|S' \cap S_2|$ for $S_2 \neq S \in \textsc{children}(S^*)$. Note that

$$\begin{aligned}
|S' \cap S_2| &= |S' \cap \big((S_2 \cap S_2') \cup (S_2 \setminus S_2')) \big) | \\
&= |S' \cap S_2 \cap S_2'| + |S' \cap (S_2 \setminus S_2')| \\
&= |S' \cap (S_2 \setminus S_2')| && \text{By (91), } S' \cap S_2' = \emptyset \\
&\leq |S_2 \setminus S_2'| \\
&\leq |S_2| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2} \right) && \text{By (94)}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\sum_{\substack{S_2 \in \textsc{children}(S^*) \\ \text{s.t.} S_2 \neq S}} |S' \cap S_2| &\leq \sum_{\substack{S_2 \in \textsc{children}(S^*) \\ \text{s.t.} S_2 \neq S}} |S_2| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2} \right) \\
&\leq |S^*| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^2 \cdot \varphi^2} \right) \\
&\leq |S| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^3 \cdot \varphi^2} \right) && \text{As } |S| \geq \beta \cdot |S^*| \qquad (98)
\end{aligned}$$

Next we will bound $|Q^* \setminus S^*|$. Note that since Q^* is D-hierarchically-close to S^* by Lemme 19 we have $|Q^* \setminus S^*| \leq 2 \cdot D \cdot \varphi_{h-2} \cdot |S^*|$. Since $|S| \geq \beta \cdot |S^*|$ and $\varphi_{h-2} = \gamma \cdot \varphi_{h-1}$ we get

$$|Q^* \setminus S^*| \leq 2 \cdot D \cdot \varphi_{h-2} \cdot |S^*| \leq 2 \cdot D \cdot \varphi_{h-1} \cdot \gamma \cdot \frac{|S|}{\beta} \qquad (99)$$

By putting (97), (98) and (99) together we get

$$\begin{aligned}
|\text{cyl}(\mu_S, \ell|Q^*) \setminus S| &\leq \left(\sum_{\substack{S_2 \in \text{CHILDREN}(S^*) \\ \text{s.t.} S_2 \neq S}} |S' \cap S_2| \right) + |Q^* \setminus S^*| \\
&\leq |S| \cdot \left(\frac{O(\varphi_{h-1})}{\beta^3 \cdot \varphi^2} \right) + |S| \cdot \frac{2 \cdot D \cdot \gamma}{\beta} \cdot \varphi_{h-1} && \text{By (98) and (99)} \\
&\leq |S| \cdot \varphi_{h-1} \left(\frac{D_2}{\beta^3 \cdot \varphi^2} + \frac{2 \cdot D \cdot \gamma}{\beta} \right) && \text{where } D_2 \text{ is the constant hidden in big-Oh abov}
\end{aligned} \tag{100}$$

Putting it together: By (95), (96) and (100) we have

$$\begin{aligned}
|\text{cyl}(\mu_S, \ell|Q^*) \triangle S| &= |S \setminus \text{cyl}(\mu_S, \ell|Q^*)| + |\text{cyl}(\mu_S, \ell|Q^*) \setminus S| \\
&\leq |S| \cdot \varphi_{h-1} \left(\frac{D \cdot \gamma}{\beta} + \frac{D_1}{\beta^2 \cdot \varphi^2} + \frac{D_2}{\beta^3 \cdot \varphi^2} + \frac{2 \cdot D \cdot \gamma}{\beta} \right) \\
&\leq |S| \cdot \varphi_{h-1} \left(\frac{3 \cdot D \cdot \gamma}{\beta} + \frac{D_1 + D_2}{\beta^3 \cdot \varphi^2} \right) \\
&\leq \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S|
\end{aligned}$$

The last inequality holds since $\frac{\gamma}{\beta^2} < \frac{1}{60}$ and on choosing $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, where $D_0 = 20(D_1 + D_2)$.

Proof of (3): So far we proved property (1) of definition D-hierarchically-close. To complete the proof, we need to show that $\text{cyl}(\mu_S, \ell|Q^*)$ is D-hierarchically-close to S by verifying property (2) of Definition 19. Thus we need to show that for any $h' \in [h-1]$, $|\text{cyl}(\mu_S, \ell|Q^*) \cap B_{h'}(S)| \leq |S| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{(h-1-h')}} \right)$. Since S is a child of S^* thus for for any $h' \leq h-2$ we have $B_{h'}(S) = B_{h'}(S^*)$. Therefore, we have

$$\begin{aligned}
|\text{cyl}(\mu_S, \ell|Q^*) \cap B_{h'}(S)| &\leq |Q^* \cap B_{h'}(S^*)| && \text{Since } B_{h'}(S) = B_{h'}(S^*), \text{ and } \text{cyl}(\mu_S, \ell|Q^*) \subseteq Q^* \\
&\leq |S^*| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{h-2-h'}} \right) && \text{Since } Q^* \text{ is } D\text{-hierarchically-close to } S^* \in \mathcal{P}^{h-1} \\
&\leq |S| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{h-1-h'}} \right) && \text{Since } |S| \geq \beta \cdot |S^*|
\end{aligned} \tag{101}$$

Note that $B_{h-1}(S) = S^* \setminus S$. Thus we have

$$\begin{aligned}
|\text{cyl}(\mu_S, \ell|Q^*) \cap B_{h-1}(S)| &= |\text{cyl}(\mu_S, \ell|Q^*) \cap (S^* \setminus S)| \\
&\leq |\text{cyl}(\mu_S, \ell|Q^*) \setminus S| \\
&\leq D \cdot \varphi_{h-1} \cdot |S| && \text{As } |\text{cyl}(\mu_S, \ell|Q^*) \triangle S| \leq \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S|
\end{aligned}$$

Therfore, for any $0 \leq h' \leq h-1$ we have

$$|\text{cyl}(\mu_S, \ell|Q^*) \cap B_{h'}(S)| \leq |S| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{h-1-h'}} \right).$$

□

Lemma 19. *Let $h^* \in [H]$, $S \in \mathcal{P}^{h^*}$ be a cluster at level h^* and let $Q \subseteq V$ be a set that is D-hierarchically-close to S^*. Then we have*

$$|Q \setminus S^*| \leq 2 \cdot D \cdot \varphi_{h^*-1} \cdot |S^*|$$

Proof. Note that since Q is D-hierarchically-close to S^* we have

$$\begin{aligned}
|Q \setminus S^*| &= \sum_{h=0}^{h^*-1} |B_h(S^*) \cap S^*| \\
&\leq \sum_{h=0}^{h^*-1} |S^*| \cdot \left(\frac{D \cdot \varphi_h}{\beta^{(h^*-1-h)}} \right) && \text{By item 2 of Definition 19} \\
&= D \cdot |S^*| \cdot \sum_{h=0}^{h^*-1} \left(\frac{\varphi_{h^*-1} \cdot \gamma^{h^*-1-h}}{\beta^{h^*-1-h}} \right) && \text{As } \varphi_h = \varphi_{h^*-1} \cdot \gamma^{h^*-1-h} \\
&= D \cdot |S^*| \cdot \varphi_{h^*-1} \cdot \sum_{h=0}^{h^*-1} \left(\frac{\gamma}{\beta} \right)^{h^*-1-h} \\
&\leq 2 \cdot D \cdot \varphi_{h^*-1} \cdot |S^*| && \text{It is a geometric sum where } \gamma/\beta < 1/2
\end{aligned}$$

□

Lemma 20. *(Bounded outliers) Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, $|\mathcal{P}^h| = \kappa$, $S^* \in \mathcal{P}^{h-1}$, $\ell = \frac{1}{10^3 \cdot |S^*|}$, and $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ where D_0 is a sufficiently large constant. Let Q^* be a set that is D-hierarchically-close to S^* (Definition 19). Let $O = Q^* \setminus \bigcup_{S \in \text{CHILDREN}(S^*)} cyl(\mu_S, \ell | Q^*)$ where $\mu_S \in \mathbb{R}^\kappa$ is the κ-dimensional center of cluster S (Definition 10). Then we have*

$$|O| \leq \frac{D \cdot \beta}{5} \cdot \varphi_{h-1} \cdot |S^*|.$$

Proof. By Theorem 3, for every $S_1 \neq S_2 \in \text{CHILDREN}(S^*)$ we have $\text{cyl}(\mu_{S_1}, \ell | Q^*) \cap \text{cyl}(\mu_{S_2}, \ell | Q^*) = \emptyset$. We now define

$$O = Q^* \setminus \bigcup_{S \in \text{CHILDREN}(S^*)} \text{cyl}(\mu_S, \ell | Q^*) \tag{102}$$

Therefore, we have

$$\begin{aligned}
|O| &= |O \setminus S^*| + |O \cap S^*| \\
&= |O \setminus S^*| + \sum_{S \in \text{CHILDREN}(S^*)} |S \cap O| && \text{As } S^* = \bigcup_{S \in \text{CHILDREN}(S^*)} S \\
&\leq |Q^* \setminus S^*| + \sum_{S \in \text{CHILDREN}(S^*)} \left| S \cap \left(Q^* \setminus \bigcup_{S \in \text{CHILDREN}(S^*)} \text{cyl}(\mu_S, \ell | Q^*) \right) \right| && \text{Since } O \subseteq Q^* \text{ and by (102)} \\
&\leq |Q^* \setminus S^*| + \sum_{S \in \text{CHILDREN}(S^*)} |S \setminus \text{cyl}(\mu_S, \ell | Q^*)| && (103)
\end{aligned}$$

Note that Q^* is D-hierarchically-close to $S^* \in \mathcal{P}^{h-1}$. Thus by Lemma 19 we have

$$|Q^* \setminus S^*| \leq 2 \cdot D \cdot \varphi_{h-2} \cdot |S^*| \tag{104}$$

Also by Theorem 3, item (2) for every $S \in \textsc{children}(S^*)$ we have

$$|\mathrm{cyl}(\mu_S, \ell | Q^*) \triangle S| \leq \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S| \tag{105}$$

Putting (103), (104) and (105) together we get

$$\begin{aligned}
|O| &\leq |Q^* \setminus S^*| + \sum_{S \in \textsc{children}(S^*)} |S \setminus \mathrm{cyl}(\mu_S, \ell | Q^*)| && \text{By (103)} \\
&\leq 2 \cdot D \cdot \varphi_{h-2} \cdot |S^*| + \sum_{S_i \in \textsc{children}(S^*)} \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S| && \text{By (104) and (105)} \\
&\leq |S^*| \left(2 \cdot D \cdot \varphi_{h-2} + \frac{D \cdot \beta}{10} \cdot \varphi_{h-1}\right) && \text{As } S^* = \sum_{S \in \textsc{children}(S^*)} |S| \\
&= D \cdot \varphi_{h-1} \cdot |S^*| \cdot \left(2 \cdot \gamma + \frac{\beta}{10}\right) && \text{As } \varphi_{h-2} = \gamma \cdot \varphi_{h-1} \\
&\leq \frac{D \cdot \beta}{5} \cdot \varphi_{h-1} \cdot |S^*| && \text{As } \frac{\gamma}{\beta} \text{ is sufficiently small}
\end{aligned} \tag{106}$$

□

Theorem 4. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, $|\mathcal{P}^h| = \kappa$, $S^* \in \mathcal{P}^{h-1}$, $\ell = \frac{1}{10^3 \cdot |S^*|}$, and $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ where D_0 is a sufficiently large constant. Let Q^* be a set that is D-hierarchically-close to S^* (Definition 19). Let $x \in Q^*$ be a vertex such that $|cyl(f_x^\kappa, 6\ell | Q^*)| \geq 0.85 \cdot \beta \cdot |S^*|$. Then for every set Q satisfying*

$$cyl(f_x^\kappa, 20\ell | Q^*) \subseteq Q \subseteq cyl(f_x^\kappa, 30\ell | Q^*),$$

there exists a unique cluster $S \in \textsc{children}(S^*)$ such that:

1. *Q is D-hierarchically-close to S*
2. *for every $S' \neq S \in \textsc{children}(S^*)$, $cyl(\mu_{S'}, \ell | Q^*) \cap Q = \emptyset$*

where $\mu_{S'} \in \mathbb{R}^\kappa$ is the κ-dimensional center of set S' (Definition 10).

Proof. We first prove that there exists a cluster $S \in \textsc{children}(S^*)$ such that $\mathrm{cyl}(f_x^\kappa, 6\ell | Q^*) \cap \mathrm{cyl}(\mu_S, \ell | Q^*) \neq \emptyset$ and $\mathrm{cyl}(\mu_S, \ell | Q^*) \subseteq Q$. Then we use these two facts to show items (1) and (2). By Theorem 3, for every $S_1 \neq S_2 \in \textsc{children}(S^*)$ we have

$$\mathrm{cyl}(\mu_{S_1}, \ell | Q^*) \cap \mathrm{cyl}(\mu_{S_2}, \ell | Q^*) = \emptyset$$

We now define

$$O = Q^* \setminus \bigcup_{S \in \textsc{children}(S^*)} \mathrm{cyl}(\mu_S, \ell | Q^*) \tag{107}$$

Note that by the assumptions of the lemma we have $|\mathrm{cyl}(f_x^\kappa, 6\ell | Q^*)| \geq 0.85 \cdot \beta \cdot |S^*|$. Therefore, we get

$$\begin{aligned}
|\mathrm{cyl}(f_x^\kappa, 6\ell | Q^*)| &\geq 0.85 \cdot \beta \cdot |S^*| \\
&> \frac{\beta}{5} \cdot \frac{D_0}{\beta^4 \cdot \varphi^2} \cdot \gamma \cdot |S^*| && \text{Since } \frac{\gamma}{\beta^4 \cdot \varphi^2} \text{ is sufficiently small by Definition 6} \\
&\geq \frac{D \cdot \beta}{5} \cdot \varphi_{h-1} \cdot |S^*| && \text{Since } \varphi_{h-1} \leq \varphi_{H-1} \leq \gamma \text{ and } D = \frac{D_0}{\beta^4 \cdot \varphi^2} \\
&\geq |O| && \text{By Lemma 20}
\end{aligned}$$

Recall that $Q^* = O \cup \Big(\bigcup_{S \in \text{CHILDREN}(S^*)} \text{cyl}(\mu_S, \ell | Q^*) \Big)$ and $\text{cyl}\,(f_x^\kappa, 6\ell | Q^*) \subseteq Q^*$. Since $|\text{cyl}\,(f_x^\kappa, 6\ell | Q^*)\,| > |O|$, therefore, there exists a cluster S such that

$$\text{cyl}\,(f_x^\kappa, 6\ell | Q^*) \cap \text{cyl}\,(\mu_S, \ell | Q^*) \neq \emptyset \tag{108}$$

Since $\text{cyl}\,(f_x^\kappa, 6\ell | Q^*) \cap \text{cyl}\,(\mu_S, \ell | Q^*) \neq \emptyset$, there exists a vertex $z \in \text{cyl}\,(f_x^\kappa, 6\ell | Q^*) \cap \text{cyl}\,(\mu_S, \ell | Q^*)$. Let $r = |\text{CHILDREN}(S^*)|$, $\kappa = |\mathcal{P}^h|$ and $\Pi \in \mathbb{R}^{\kappa \times \kappa}$ be the subgraph projection matrix of Q^* with respect to κ and r (Definition 12). Since $z \in \text{cyl}\,(f_x^\kappa, 6\ell | Q^*)$, by Definition 20 we have

$$||\Pi f_x^\kappa - \Pi f_z^\kappa||_2^2 \leq 6\ell \tag{109}$$

Since $z \in \text{cyl}\,(\mu_S, \ell | Q^*)$, we have $||\Pi f_z^\kappa - \Pi \mu_S||_2^2 \leq \ell$. Also for every $y \in (\mu_S, \ell | Q^*)$, we have $||\Pi f_y^\kappa - \Pi \mu_S||_2^2 \leq \ell$. Thus by triangle inequality for every $y \in Q_i$ we get

$$||\Pi f_z^\kappa - \Pi f_y^\kappa||_2^2 \leq 4\ell \tag{110}$$

Putting (109) and (110) together and by triangle inequality for every $y \in \text{cyl}\,(\mu_S, \ell | Q^*)$ we have

$$||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 \leq 20\ell \tag{111}$$

Therfore, we have

$$\text{cyl}\,(\mu_S, \ell | Q^*) \subseteq \text{cyl}\,(f_x^\kappa, 20\ell | Q^*) \subseteq Q \tag{112}$$

Proof of (2)**:** We want to show that for every $S' \neq S \in \text{CHILDREN}(S^*)$, $\text{cyl}\,(\mu_{S'}, \ell | Q^*) \cap Q = \emptyset$. Since $Q \subseteq \text{cyl}\,(f_x^\kappa, 30\ell | Q^*)$, it suffices to show that $\text{cyl}\,(\mu_{S'}, \ell | Q^*) \cap \text{cyl}\,(f_x^\kappa, 30\ell | Q^*) = \emptyset$, or equivalently, for every $w \in \text{cyl}\,(\mu_{S'}, \ell | Q^*)$ we need to show that

$$||\Pi f_x^\kappa - \Pi f_w^\kappa||^2 > 30\ell.$$

Let $y \in \text{cyl}\,(\mu_S, \ell | Q^*)$. By triangle inequality we have

$$||\Pi f_x^\kappa - \Pi f_w^\kappa||_2 \geq ||\Pi f_y^\kappa - \Pi f_w^\kappa||_2 - ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2 \tag{113}$$

and

$$\begin{aligned}
&||\Pi f_y^\kappa - \Pi f_w^\kappa||_2 \\
&\geq ||\Pi \mu_S - \Pi \mu_{S'}||_2 - ||\Pi \mu_S - \Pi f_y^\kappa||_2 - ||\Pi \mu_{S'} - \Pi f_w^\kappa||_2 && \text{By triangle inequality} \\
&\geq ||\Pi \mu_S - \Pi \mu_{S'}||_2 - 2\sqrt{\ell} && \text{As } y \in \text{cyl}\,(\mu_S, \ell | Q^*) \text{ and } w \in \text{cyl}\,(\mu_{S'}, \ell | Q^*)
\end{aligned} \tag{114}$$

Note that by Definition 6, for every $S \in \text{CHILDREN}(S^*)$ we have $|S| \geq \beta \cdot |S^*|$, hence, $|\text{CHILDREN}(S^*)| = r \leq \frac{1}{\beta}$. Also recall that $\frac{\gamma}{\beta^{30}}$ and $\frac{\gamma}{\varphi^{20}}$ are sufficiently small. Therefore, by choice of $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ we have $r \leq D \leq \frac{1}{\gamma^{1/4}}$. Thus we can apply Lemma 18 and we get

$$||\Pi \mu_S - \Pi \mu_{S'}||_2 \geq \frac{1}{\sqrt{|S^*|}} = \sqrt{10^3 \cdot \ell} \tag{115}$$

Putting (114) and (115) together we get

$$||\Pi f_y^\kappa - \Pi f_w^\kappa||_2 \geq 29 \cdot \sqrt{\ell} \tag{116}$$

Recall that $y \in \text{cyl}\,(\mu_S, \ell | Q^*)$ and by (112) we have $\text{cyl}\,(\mu_S, \ell | Q^*) \subseteq \text{cyl}\,(f_x^\kappa, 20\ell | Q^*)$. Thus, $y \in \text{cyl}\,(f_x^\kappa, 20\ell | Q^*)$. Therefore, we have

$$||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 \leq 20 \cdot \ell \tag{117}$$

Putting (113), (116), and (117) together we get

$$||\Pi f_x^\kappa - \Pi f_w^\kappa||_2 \geq ||\Pi f_y^\kappa - \Pi f_w^\kappa||_2 - ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2 \geq 29\sqrt{\ell} - \sqrt{20 \cdot \ell} \geq 24 \cdot \sqrt{\ell}$$

Thus, for every $w \in \text{cyl}\,(\mu_{S'}, \ell | Q^*)$ we have $||\Pi f_x^\kappa - \Pi f_w^\kappa||^2 > 30 \cdot \ell$. Thus, $\text{cyl}\,(f_x^\kappa, 30\ell | Q^*) \cap \text{cyl}\,(\mu_{S'}, \ell | Q^*) = \emptyset$. Since $Q \subseteq \text{cyl}\,(f_x^\kappa, 30\ell | Q^*)$, for every $S' \neq S \in \text{CHILDREN}(S^*)$ we have

$$Q \cap \text{cyl}\,(\mu_{S'}, \ell | Q^*) = \emptyset \tag{118}$$

Proof of (1)**:** Let Q be a set such that $\text{cyl}\left(f_x^\kappa, 20\ell|Q^*\right) \subseteq Q \subseteq \text{cyl}\left(f_x^\kappa, 30\ell|Q^*\right)$. To show that Q is D-hierarchically-close to S we need to verify property (1) and property (2) of Definition 19. So we need to prove that $|S \setminus Q| \leq D \cdot \varphi_{h-1} \cdot |S|$ and for any $0 \leq h' \leq h-1$, $|Q \cap B_{h'}(S)| \leq |S| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{h-1-h'}}\right)$. We first bound $|S \setminus Q|$. For property (1) we have

$$\begin{aligned} |S \setminus Q| &\leq |S \setminus \text{cyl}(\mu_S, \ell|Q^*)| && \text{By (112) } \text{cyl}(\mu_S, \ell|Q^*) \subseteq \text{cyl}\left(f_x^\kappa, 20\ell|Q^*\right) \subseteq Q \\ &\leq \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S| && \text{By Theorem 3} \\ &\leq D \cdot \varphi_{h-1} \cdot |S| && (119) \end{aligned}$$

Now, we verify property (2) of Definition 19 by proving that for any $0 \leq h' \leq h-1$, $|Q \cap B_{h'}(S)| \leq |S| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{(h-1-h')}}\right)$. For $h' = h-1$, since $B_{h-1}(S) = S^* \setminus S$ we have

$$|Q \cap B_{h-1}(S)| \leq |Q \cap (S^* \setminus S)| \leq |Q \setminus S|$$

By (118) for every $S' \neq S \in \text{CHILDREN}(S^*)$ we have $Q \cap \text{cyl}\left(\mu_{S'}, \ell|Q^*\right) = \emptyset$. By (107) we have $O = Q^* \setminus \bigcup_{S \in \text{CHILDREN}(S^*)} \text{cyl}(\mu_S, \ell|Q^*)$. Therefore we have

$$Q \subseteq O \cup \text{cyl}\left(\mu_S, \ell|Q^*\right)$$

Hence,

$$|Q \cap B_{h-1}(S)| \leq |Q \setminus S| \leq |O| + |\text{cyl}\left(\mu_S, \ell|Q^*\right) \setminus S|$$

By Lemma (20) we have $|O| \leq \frac{D \cdot \beta}{5} \cdot \varphi_{h-1} \cdot |S^*|$, also we know $|S| \geq \beta \cdot |S^*|$. Therefore, we get

$$|O| \leq \frac{D \cdot \beta}{5} \cdot \varphi_{h-1} \cdot |S^*| \leq \frac{D \cdot \beta}{5} \cdot \varphi_{h-1} \cdot \frac{|S|}{\beta} \leq \frac{D}{5} \cdot \varphi_{h-1} \cdot |S| \tag{120}$$

Moreover, by Theorem 3, item (2) we have $|\text{cyl}\left(\mu_S, \ell|Q^*\right) \triangle S| \leq \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S|$. Therefore, we get

$$\begin{aligned} |Q \cap B_{h-1}(S)| &\leq |\text{cyl}\left(\mu_S, \ell|Q^*\right) \setminus S| + |O| \\ &\leq \frac{D \cdot \beta}{10} \cdot \varphi_{h-1} \cdot |S| + \frac{D}{5} \cdot \varphi_{h-1} \cdot |S| \\ &\leq D \cdot \varphi_{h-1} \cdot |S| && (121) \end{aligned}$$

Also note that for any $h' \leq h-2$ we have

$$\begin{aligned} |Q \cap B_{h'}(S)| &\leq |Q^* \cap B_{h'}(S^*)| && \text{Since } B_{h'}(S) = B_{h'}(S^*) \text{ and } Q \subseteq Q^* \\ &\leq |S^*| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{h-2-h'}}\right) && \text{Since } Q^* \text{ is } D\text{-hierarchically-close to } S^* \in \mathcal{P}^{h-1} \\ &\leq |S| \cdot \left(\frac{D \cdot \varphi_{h'}}{\beta^{h-1-h'}}\right) && \text{Since } |S| \geq \beta \cdot |S^*| \quad (122) \end{aligned}$$

Putting (119), (121) and (122) together we get that Q is D-hierarchically-close to S. □

4.9 Dot Product Oracle on the Projected Subspace

The main result of this Section is Theorem 5.

Theorem 5. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$, $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ where D_0 is some large constant. Let $h \in [H]$, $\kappa = |\mathcal{P}^h|$, $S^* \in \mathcal{P}^{h-1}$ and $r = |\text{CHILDREN}(S^*)|$. Let $\boldsymbol{S}^* \subseteq V$ be a set that is D-hierarchically-close to S^* (Definition 19). Let $\Pi \in \mathbb{R}^{\kappa \times \kappa}$ denote the subgraph projection matrix of $\boldsymbol{S}^*$ for κ and r. Let $A_0, c > 1$*

be large constants and let $\widetilde{S}^*$ be a set of size $\widetilde{s} \geq \frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}$ sampled independently and uniformly at random from Q. Let $\boldsymbol{s}^*$ be an estimation of $|\boldsymbol{S}^*|$ such that $|\boldsymbol{s}^* - |\boldsymbol{S}^*|| \leq \frac{|\boldsymbol{S}^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$. Then InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{S}^*, \boldsymbol{s}^*, \xi)$ *(Algorithm 6) computes a sublinear space data structure $\mathcal{D}$ such that with probability at least $1 - n^{-96}$ the following property is satisfied:*

1. *For every pair of vertices $x, y \in V$,* ProjectedDotProduct$(G, x, y, \xi, \mathcal{D})$ *(Algorithm 7) runs in time $n^{1/2 + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \xi \cdot \varphi_h}\right)^{O(1)}$ and computes an output value $\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx}$ such that with probability at least $1 - n^{-96}$,*
$$\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle \right| \leq \frac{\xi}{n}$$

2. *For every pair of vertices $x, y \in V$,* ProjectedDistance$(G, x, y, \xi, \mathcal{D})$ *(Algorithm 8) runs in time $n^{1/2 + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \xi \cdot \varphi_h}\right)^{O(1)}$ and computes an output value $\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2$ such that with probability at least $1 - n^{-96}$,*
$$\left| \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 - ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 \right| \leq \frac{\xi}{n}$$

Notice that by Remark 3 we can achieve a tradeoff in the preprocessing/ query runtime.

Remark 3. *For any $\omega \in [0, 1/2]$, one can obtain the following trade-offs between preprocessing time and query time: Algorithm* ProjectedDotProduct$(G, x, y, \xi, \mathcal{D})$ *requires $n^{\omega + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \xi \cdot \varphi_h}\right)^{O(1)}$ per query when the prepressing time of Algorithm* InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{Q},$ *is increased to $n^{1 - \omega + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \xi \cdot \varphi_h}\right)^{O(1)}$.*

We first set up notations and then state the algorithms below.

Let $m \leq n$ be integers. For any matrix $A \in \mathbb{R}^{n \times m}$ with singular value decomposition (SVD) $A = Y \Gamma Z^T$ we assume $Y \in \mathbb{R}^{n \times n}$ and $Z \in \mathbb{R}^{m \times n}$ are orthogonal matrices and $\Gamma \in \mathbb{R}^{n \times n}$ is a diagonal matrix of singular values. Since Y and Z are orthogonal matrices, their columns form an orthonormal basis. For any integer $q \in [m]$ we denote $Y_{[q]} \in \mathbb{R}^{n \times q}$ as the first q columns of Y and $Y_{-[q]}$ to denote the matrix of the remaining columns of Y. We also denote $Z_{[q]}^T \in \mathbb{R}^{q \times n}$ as the first q rows of Z^T and $Z_{-[q]}^T$ to denote the matrix of the remaining rows of Z. Finally we denote $\Gamma_{[q]}^T \in \mathbb{R}^{q \times q}$ as the first q rows and columns of Γ and we use $\Gamma_{-[q]}$ as the last $n - q$ rows and columns of Γ. So for any $q \in [m]$ the span of $Y_{-[q]}$ is the orthogonal complement of the span of $Y_{[q]}$, also the span of $Z_{-[q]}$ is the orthogonal complement of the span of $Z_{[q]}$. Thus we can write $A = Y_{[q]} \Gamma_{[q]} Z_{[q]}^T + Y_{-[q]} \Gamma_{-[q]} Z_{-[q]}^T$.

Algorithm 6 InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{Q}, s, \xi)$

```
$\mathcal{D}_h$ = InitializeDotProductOracle$(G, \omega, \xi, h, \kappa)$          ▷ Remark 5
$\widetilde{s} \leftarrow |\widetilde{Q}|$
$\widehat{\Upsilon} \in \mathbb{R}^{\widetilde{s} \times \widetilde{s}} \leftarrow$ gram-matrix of $\langle f_{z_1}^\kappa, f_{z_2}^\kappa \rangle_{apx}$ for $z_1, z_2 \in \widetilde{Q}$,          ▷ Remark 5
Let $\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) = \widehat{Z} \widehat{\Gamma} \widehat{Z}^T$ be the eigendecomposition of $\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right)$
$\widehat{\Psi} = \frac{s}{\widetilde{s}} \cdot \widehat{Z}_{[r]} \widehat{\Gamma}_{[r]}^{-1} \widehat{Z}_{[r]}^T$
return $\mathcal{D} = \{\mathcal{D}_h, \widehat{\Psi}, \widetilde{Q}\}$
```

```
Algorithm 7 ProjectedDotProduct$(G, x, y, \xi, \mathcal{D})$        ▷ $\mathcal{D} = \{\mathcal{D}_h, \widehat{\Psi}, \widetilde{Q}\}$
1: Let $\alpha_x \in \mathbb{R}^{|\widetilde{Q}|}$ be a vector such that for any $z \in \widetilde{Q}$ we have $\alpha_x(z) = \langle f_x^\kappa, f_z^\kappa \rangle_{apx}$ ▷ Remark 5
2: Let $\alpha_y \in \mathbb{R}^{|\widetilde{Q}|}$ be a vector such that for any $z \in \widetilde{Q}$, we have $\alpha_y(z) = \left\langle f_y^\kappa, f_z^\kappa \right\rangle_{apx}$ ▷ Remark 5
3: return $\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} = \alpha_x^T \widehat{\Psi} \alpha_y$
```

Remark 4. *Note that as per line 3 of Algorithm 7 we have* $\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx}$ *is symmetric:*

$$\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} = \left\langle \widetilde{\Pi} f_x^\kappa, f_y^\kappa \right\rangle_{apx}$$

Remark 5. *For computing* $\langle f_a^\kappa, f_b^\kappa \rangle_{apx}$ *we use Algorithm 10 given in Appendix C:*

$$\langle f_a^\kappa, f_b^\kappa \rangle_{apx} = \textsc{SpectralDotProductOracle}(G, a, b, \omega, \xi, \mathcal{D}_h)$$

where $\mathcal{D}_h = \textsc{InitializeDotProductOracle}(G, \omega, \xi, h, \kappa)$ *(see Algorithm 9).*

```
Algorithm 8 ProjectedDistance$(G, x, y, \xi, \mathcal{D})$        ▷ $\mathcal{D} = \{\mathcal{D}_h, \widehat{\Psi}, \widetilde{Q}\}$
$\xi' = \frac{\xi}{4}$
$\left\langle f_x^\kappa, \widetilde{\Pi} f_x^\kappa \right\rangle_{apx} =$ ProjectedDotProduct$(G, x, x, \xi', \mathcal{D})$        ▷ Algorithm 7
$\left\langle f_y^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} =$ ProjectedDotProduct$(G, y, y, \xi', \mathcal{D})$        ▷ Algorithm 7
$\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} =$ ProjectedDotProduct$(G, x, y, \xi', \mathcal{D})$        ▷ Algorithm 7
$\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 = \left\langle f_x^\kappa, \widetilde{\Pi} f_x^\kappa \right\rangle_{apx} + \left\langle f_y^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - 2 \cdot \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx}$
return $\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2$
```

4.9.1 Proof of Theorem 6 (Correctness of Algorithm 7)

To prove Theorem 5 we first present a more general result (Theorem 6) with the help of Definition 21.

Definition 21. (**δ-close to r-clusterable**) Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6) and let $Q \subseteq V$. We say that set Q is δ-close to be r-clusterable if there exists $\kappa \in [k]$, $r \in [\kappa]$ such that $\nu_{r+1}(AA^T) \leq \delta$ and $\nu_r(AA^T) \geq 1 - \delta$, where $A \in \mathbb{R}^{\kappa \times |Q|}$ is a matrix whose columns are f_x^κ for all $x \in Q$.

Theorem 6. *Let* $G = (V, E)$ *be a* (k, γ)*-hierarchically-clusterable graph (Definition 6). Let* $\kappa \in [k], r \in [\kappa]$, $\delta \in [0, \frac{1}{1000})$, $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$ *and* $A_0, c > 1$ *be large enough constants. Let* $Q \subseteq V$ *be a set that is* δ*-close to* r*-clusterable (Definition 21) and let* $\Pi \in \mathbb{R}^{\kappa \times \kappa}$ *be the subgraph projection matrix of* Q *for* κ *and* r *(Definition 12). Let* $\widetilde{Q}$ *be a set of size* $\widetilde{s} \geq \frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}$ *sampled independently and uniformly at random from* Q*. Let* s *be an estimation of* $|Q|$ *such that* $|s - |Q|| \leq \frac{|Q| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$*. Then* $\textsc{InitializeSubgraphProjMatrix}(G, ., \kappa, r, \widetilde{Q}, s, \xi)$ *(Algorithm 6) computes a sublinear space data structure* $\mathcal{D}$ *such that with probability at least* $1 - n^{-96}$ *the following property is satisfied:*

For every pair of vertices $x, y \in V$, $\textsc{ProjectedDotProduct}(G, x, y, \xi, \mathcal{D})$ *(Algorithm 7) computes an output value* $\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx}$ *such that with probability at least* $1 - n^{-96}$

$$\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle \right| \leq \frac{\xi}{n}$$

The running time of $\textsc{InitializeSubgraphProjMatrix}(G, h, \kappa, r, \widetilde{Q}, s, \xi)$ *is* $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$. *The running time of* $\textsc{ProjectedDotProduct}(G, x, y, \xi, \mathcal{D})$ *is* $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$. *Moreover, the size of the data structure* $\mathcal{D}$ *is* $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi}\right)^{O(1)}$.

To prove Theorem 6 we need the following two lemmas whose proofs are deferred to Appendix D. Lemma 21 shows that if set S is almost r-clusterable (Definition 21), then the projection subgraph matrix $\widetilde{\Pi}$ of a subsampled set $\widetilde{S}$ can be used as a proxy to the projection subgraph matrix Π of the set S.

Lemma 21. *Let* $G = (V, E)$ *be a* (k, γ)*-hierarchically-clusterable graph (Definition 6). Let* $Q \subseteq V$ *be a set that is* δ*-close to* r*-clusterable (Definition 21) and let* $\widetilde{Q}$ *be a set of size* $\widetilde{s}$ *that is sampled independently and uniformly at random from* Q*. Let* $\Pi, \widetilde{\Pi} \in \mathbb{R}^{\kappa\times\kappa}$ *denote the subgraph projection matrix of* Q *and* $\widetilde{Q}$ *for* κ *and* r *respectively (Definition 12). Then with probabaility at least* $1 - n^{-100}$ *for every* $x, y \in V$ *we have*

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle\right| \leq \frac{\xi}{n},$$

where, $\kappa \in [k], r \in [\kappa]$, $\delta \in [0, \frac{1}{1000})$, $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$, $\widetilde{s} \geq \frac{k^c\cdot n^{160A_0\cdot\gamma/\varphi}}{\xi^2}$ *and* $A_0, c > 1$ *are large enough constants.*

Next, Lemma 22 asserts that the approximate inner products between f_x^κ, f_y^κ in the projected space obtained in Line 3 of $\textsc{ProjectedDotProduct}$ using $\widetilde{\Pi}$ are very close to the actual inner products computed using $\widetilde{\Pi}$ when we know f_x^κ, f_y^κ.

Lemma 22. *Let* $G = (V, E)$ *be a* (k, γ)*-hierarchically-clusterable graph (Definition 6). Let* $Q \subseteq V$ *be a set that is* δ*-close to* r*-clusterable (Definition 21) and let* $\Pi \in \mathbb{R}^{\kappa\times\kappa}$ *be the subgraph projection matrix of* Q *for* κ *and* r *(Definition 12). Then* $\textsc{InitializeSubgraphProjMatrix}(G, ., \kappa, r, \widetilde{Q}, s, \xi)$ *(Algorithm 6) computes a data structure* $\mathcal{D}$ *such that with probability at least* $1 - n^{-97}$ *the following property is satisfied: With probability at least* $1-n^{-97}$*, for every pair of vertices* $x, y \in V$*,* $\textsc{ProjectedDotProduct}(G, x, y, \xi, \mathcal{D})$ *(Algorithm 7) computes an output value* $\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx}$ *such that*

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle\right| \leq \frac{\xi}{n},$$

where, $\kappa \in [k], r \in [\kappa]$, $\delta \in (0, \frac{1}{1000})$, $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$, *and* $A_0, c > 1$ *are large enough constants. Also,* $\widetilde{Q}$ *is a set of size* $\widetilde{s} \geq \frac{k^c\cdot n^{560A_0\cdot\gamma/\varphi}}{\xi^6}$ *sampled independently and uniformly at random from* Q*, and* s *is an estimation of* $|Q|$ *such that* $|s - |Q|| \leq \frac{|Q|\cdot\xi^3}{k^c\cdot n^{280A_0\cdot\gamma/\varphi}}$.

Now we are ready to prove Theorem 6.

Proof. (Of Theorem 6) **Correctness:** Let $\xi' = \frac{\xi}{2}$. Note that by choice of $\widetilde{s}$ for large enough constant c we have $\widetilde{s} \geq \frac{k^c\cdot n^{560A_0\cdot\gamma/\varphi}}{\xi^6} \geq \frac{k^{c'}\cdot n^{160A_0\cdot\frac{\gamma}{\varphi}}}{\xi'^2}$ where c' is the constant from Lemma 21. Therefore, by Lemma 21 with probabaility at least $1 - n^{-100}$ for any $x, y \in V$ we have

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle\right| \leq \frac{\xi}{2\cdot n}$$

Also note that by choice of $s, \widetilde{s}$ for large enough constant c we have $\widetilde{s} \geq \frac{k^c\cdot n^{560A_0\cdot\gamma/\varphi}}{\xi^6} \geq \frac{k^{c''}\cdot n^{560A_0\cdot\frac{\gamma}{\varphi}}}{\xi'^6}$ and $|s - |Q|| \leq \frac{|Q|\cdot\xi^3}{k^c\cdot n^{280A_0\cdot\gamma/\varphi}} \leq \frac{|Q|\cdot\xi'^3}{k^{c''}\cdot n^{280A_0\cdot\gamma/\varphi}}$ where c'' is the constant from Lemma 22. Therefore, by Lemma 22 with probabaility at least $1 - n^{-97}$ for any $x, y \in V$ we have

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle\right| \leq \frac{\xi}{2\cdot n}.$$

Therefore, by triangle inequality with probabaility at least $1-n^{-96}$ for any $x, y \in V$ we have we have

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa\right\rangle_{apx} - \left\langle f_x^\kappa, \Pi f_y^\kappa\right\rangle\right| \le \left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa\right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa\right\rangle\right| + \left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa\right\rangle - \left\langle f_x^\kappa, \Pi f_y^\kappa\right\rangle\right| \le \frac{\xi}{n}$$

Runtime: We first bound the running time of InitializeSubgraphProjMatrix (Algorithm 6). We consider individual steps of this procedure and consider the running time for each of these. By Theorem 8, line 1 takes time $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$. Line 3 computes $\widetilde{s}^2$ dot products. For $z_1, z_2 \in \widetilde{Q}$, by Theorem 8 computing $\left\langle f_{z_1}^\kappa, f_{z_2}^\kappa\right\rangle_{apx}$ takes time $t_{z_1,z_2} \le n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$. Thus, overall time taken by Line 3 is

$$\sum_{z_1,z_2\in\widetilde{Q}} t_{z_1,z_2} \le \frac{k^{2c}\cdot n^{1120A_0\cdot\gamma/\varphi}}{\xi^{12}} \cdot \max_{z_1,z_2\in\widetilde{Q}} t_{z_1,z_2} \le n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$$

The time taken by lines 4 and 5 both is $\widetilde{s}^3$. Thus, the overall time taken by InitializeSubgraphProjMatrix (Algorithm 6) is at most $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$.

Now we bound the running time of ProjectedDotProduct (Algorithm 7). Lines 1 and 2 find vectors $\alpha_x, \alpha_y \in \mathbb{R}^{\widetilde{s}}$ one coordinate at a time. By Theorem 8, finding a single coordinate $\alpha_x(z), \alpha_y(z)$ takes time $t_z \le n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$. So, the overall time taken to construct these vectors is at most $\widetilde{s}\cdot t_z \le \frac{k^c\cdot n^{560A_0\gamma/\varphi}}{\xi^6}\cdot t_z \le n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$.

Line 3 computes the approximate inner product between every pair of vectors in $\widetilde{Q}$. According to Theorem 8, each of these $\widetilde{s}^2$ computations takes time $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$. Thus, asymptotically the overall time is $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$.

Space Used: Finally, we bound the size of the data structure $\mathcal{D}$ computed by the procedure InitializeSubgraphProjMatrix. Recall $\mathcal{D} = \{\mathcal{D}_h, \widehat{\Psi}, \widetilde{Q}\}$. Here, the data structure $\mathcal{D}_h$ is obtained by calling InitializeDotProductOracle in Line 1 of InitializeSubgraphProjMatrix. By Theorem 8, the size of $\mathcal{D}_h$ is $n^{1/2+o(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi}\right)^{o(1)}$. Further, $\widehat{\psi}$ is just a $\widetilde{s}$-by-$\widetilde{s}$ matrix and the set $\widetilde{Q}$ contains $\widetilde{s}$ vertices. In all, the overall size is dominated by the size of $\mathcal{D}_h$ which can be upperbounded as $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi}\right)^{O(1)}$

□

4.9.2 Proof of Theorem 5 (Correctness of Algorithm 8)

Theorem 5. *Let $G=(V,E)$ be a (k,γ)-hierarchically-clusterable graph (Definition 6). Let $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$, $D = \frac{D_0}{\beta^4\cdot\varphi^2}$ where D_0 is some large constant. Let $h \in [H]$, $\kappa = |\mathcal{P}^h|$, $S^* \in \mathcal{P}^{h-1}$ and $r = |$children$(S^*)|$. Let $\boldsymbol{S}^* \subseteq V$ be a set that is D-hierarchically-close to S^* (Definition 19). Let $\Pi \in \mathbb{R}^{\kappa\times\kappa}$ denote the subgraph projection matrix of $\boldsymbol{S}^*$ for κ and r. Let $A_0, c > 1$ be large constants and let $\widetilde{S}^*$ be a set of size $\widetilde{s} \ge \frac{k^c\cdot n^{560A_0\cdot\gamma/\varphi}}{\xi^6}$ sampled independently and uniformly at random from Q. Let $\boldsymbol{s}^*$ be an estimation of $|\boldsymbol{S}^*|$ such that $|\boldsymbol{s}^* - |\boldsymbol{S}^*|| \le \frac{|\boldsymbol{S}^*|\cdot\xi^3}{k^c\cdot n^{280A_0\cdot\gamma/\varphi}}$. Then* InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{S}^*, \boldsymbol{s}^*, \xi)$ *(Algorithm 6) computes a sublinear space data structure $\mathcal{D}$ such that with probability at least $1-n^{-96}$ the following property is satisfied:*

1. *For every pair of vertices $x, y \in V$,* $\textsc{ProjectedDotProduct}(G, x, y, \xi, \mathcal{D})$ *(Algorithm 7) runs in time* $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \xi \cdot \varphi_h}\right)^{O(1)}$ *and computes an output value* $\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx}$ *such that with probability at least* $1 - n^{-96}$*,*
$$\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle \right| \leq \frac{\xi}{n}$$

2. *For every pair of vertices $x, y \in V$,* $\textsc{ProjectedDistance}(G, x, y, \xi, \mathcal{D})$ *(Algorithm 8) runs in time* $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \xi \cdot \varphi_h}\right)^{O(1)}$ *and computes an output value* $\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2$ *such that with probability at least* $1 - n^{-96}$*,*
$$\left| \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 - ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 \right| \leq \frac{\xi}{n}$$

Proof. Let $A \in \mathbb{R}^{\kappa \times |\boldsymbol{S}^*|}$ be matrix whose columns are f_x^κ for all $x \in \boldsymbol{S}^*$. Note that $r = \textsc{children}(S^*) \leq \frac{1}{\beta}$ since for every $S \in \textsc{children}(S^*)$ we have $|S| \geq \beta \cdot |S^*|$ by Definition 6. Therefore, we have
$$r \leq \frac{1}{\beta} \leq \frac{D_0}{\beta^4 \cdot \varphi^2} = D$$
Let $\delta = 13 \cdot D \cdot \gamma^{1/4}$. Note that by Definition 6, $\min\left(\frac{\gamma}{\beta^{30}}, \frac{\gamma}{\varphi^{20}}\right)$ is a sufficiently small constant. Therefore, by choice of $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ we have $\delta < 0.01$, we can apply Lemma 17 and we get
$$\nu_{r+1}\left(AA^T\right) \leq 5 \cdot D \cdot \gamma^{1/4} \leq \delta$$
and
$$\nu_r\left(AA^T\right) \geq 1 - 13 \cdot D \cdot \gamma^{1/4} = 1 - \delta$$
Therefore, set Q is δ-close to be r-clusterable (Definition 21). Let $\xi' = \frac{\xi}{4}$ and c' be the constant from Theorem 6. Note that by choice of $s, \widetilde{s}$ for large enough constant c we have $\widetilde{s} \geq \frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6} \geq \frac{k^{c'} \cdot n^{560 A_0 \cdot \frac{\gamma}{\varphi}}}{\xi'^6}$ and $|s - |\boldsymbol{S}^*|| \leq \frac{|\boldsymbol{S}^*| \cdot \xi^3}{k^c \cdot\ n^{280 A_0 \cdot \gamma/\varphi}} \leq \frac{|\boldsymbol{S}^*| \cdot \xi'^3}{k^{c'} \cdot n^{280 A_0 \cdot \gamma/\varphi}}$. Therefore by Theorem 6 with probabaility at least $1 - n^{-96}$ for any $x, y \in V$ we have
$$\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle \right| \leq \frac{\xi'}{n} \leq \frac{\xi}{n}$$
Also as per line (5) of Algorithm 8 we define
$$\left\| \Pi f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 = \langle f_x^\kappa, \Pi f_x^\kappa \rangle_{apx} + \left\langle f_y^\kappa, \Pi f_y^\kappa \right\rangle_{apx} - 2 \cdot \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle_{apx}$$
Note that
$$||\Pi f_x^\kappa - \widetilde{\Pi} f_y^\kappa||_2^2 = \langle f_x^\kappa, \Pi f_x^\kappa \rangle + \left\langle f_y^\kappa, \Pi f_y^\kappa \right\rangle - 2 \cdot \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle$$
Therefore, by triangle inequality and by item (1) with probabaility at least $1 - n^{-96}$ we have
$$\begin{aligned}
&\left| \left\| \Pi f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 - ||\Pi f_x^\kappa - \widetilde{\Pi} f_y^\kappa||_2^2 \right| \\
&\leq \left| \langle f_x^\kappa, \Pi f_x^\kappa \rangle_{apx} - \langle f_x^\kappa, \Pi f_x^\kappa \rangle \right| + \left| \left\langle f_y^\kappa, \Pi f_y^\kappa \right\rangle_{apx} - \left\langle f_y^\kappa, \Pi f_y^\kappa \right\rangle \right| + 2 \cdot \left| \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle \right| \\
&\leq \frac{\xi'}{n} + \frac{\xi'}{n} + 2 \cdot \frac{\xi'}{n} \\
&= \frac{\xi}{n}
\end{aligned}$$

The last equality holds by choice of $\xi' = \frac{\xi}{4}$.

Finally, we bound the running time of ProjectedDistance. Note that the running time is dominated by ProjectedDotProduct (Algorithm 7) that is at most $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\xi\cdot\varphi_h}\right)^{O(1)}$.

□

4.10 Correctness of hierarchical-clustering oracle

4.10.1 Quality of approximated cylinders

The main result of this Section is Lemma 23. that shows if a cylinder around vertex x has large enough size then it overlaps with a unique cluster S, hence, a bigger cylinder around x can be used to recover (a good approximation to) S.

Definition 22. (Approximate cylinder) Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, $S^* \in \mathcal{P}^{h-1}$, $\kappa = |\mathcal{P}^h|$, $r = |\textsc{children}(S^*)|$ and $A_0, c > 1$ be large constants. For vertex $x \in Q^*$ we define

$$\text{cyl}_{\text{apx}}(f_x^\kappa, \ell|Q^*) = \left\{ y \in Q^* : \left\|\widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa\right\|_{apx}^2 \leq \ell \right\}$$

where, $\left\|\widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa\right\|_{apx}^2 = \textsc{ProjectedDistance}(G, x, y, \xi, \mathcal{D})$ (Algorithm 8). Note that $\mathcal{D}$ is the data structure computed by InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{Q}, s, \xi)$ (Algorithm 6), where $\xi = 10^{-3}$, $\widetilde{Q}$ is a set of size $\widetilde{s} \geq \frac{k^c\cdot n^{560A_0\cdot\gamma/\varphi}}{\xi^6}$ sampled independently and uniformly at ranodm from Q^*, and s is an estimation of $|Q^*|$ such that $|s - |Q^*|| \leq \frac{|Q^*|\cdot\xi^3}{k^c\cdot n^{280A_0\cdot\gamma/\varphi}}$.

Lemma 23. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, and Q^* be a set that is D-hierarchically-close to $S^* \in \mathcal{P}^{h-1}$ (Definition 19). Let $x \in Q^*$ be a vertex such that $\left|cyl_{apx}(f_x^\kappa, 5\ell_{apx}|Q^*)\right| \geq 0.85 \cdot \beta \cdot |S^*|$. Then with probability at least $1 - n^{-96}$ there exists a unique cluster $S \in \textsc{children}(S^*)$ such that:*

1. *$cyl_{apx}(f_x^\kappa, 25\ell_{apx}|Q^*)$ is D-hierarchically-close to S,*
2. *for every $S' \neq S \in \textsc{children}(S^*)$, $cyl(\mu_{S'}, \ell|Q^*) \cap cyl_{apx}(f_x^\kappa,\ 25\ell_{apx}|Q^*) = \emptyset$,*

where, $|\mathcal{P}^h| = \kappa$, $D = \frac{D_0}{\beta^4\cdot\varphi^2}$, $\ell = \frac{1}{10^3\cdot|S^*|}$, and $\ell_{apx} = \frac{1}{1000\cdot s}$, where, s is an estimation of $|Q^*|$ such that $|s - |Q^*|| \leq \frac{|Q^*|\cdot\xi^3}{k^c\cdot n^{280A_0\cdot\gamma/\varphi}}$ and $A_0, D_0, c > 1$ are large constants. Also, $\mu_{S'} \in \mathbb{R}^\kappa$ is the κ-dimensional center of set S' (Definition 10).

To prove Lemma 23 we need Claim 1 that we defer its proof to Appendix G.

Claim 1. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). For some large constant D_0 and let $D = \frac{D_0}{\beta^4\cdot\varphi^2}$, $\xi \leq 10^{-3}$, $h \in [H]$, $S^* \in \mathcal{P}^{h-1}$, and Q^* be a set that is D-hierarchically-close to S^* (Definition 19). Let s be an estimation of $|Q^*|$ such that $|s - |Q^*|| \leq \frac{|Q^*|\cdot\xi^3}{k^c\cdot n^{280A_0\cdot\gamma/\varphi}}$ where $A_0, c > 1$ are large enough constants. Then we have $|s - |S^*|| \leq \frac{|S^*|}{10^3}$.*

Now we are ready to prove Lemma 23.

Proof. Let $r = |\textsc{children}(S^*)|$, $\kappa = |\mathcal{P}^h|$ and $\Pi \in \mathbb{R}^{\kappa\times\kappa}$ be the subgraph projection matrix of Q^* with respect to κ and r (Definition 12). Let $\xi = \frac{1}{10^3}$. By Definition 22, we have $\widetilde{Q}$ is a set of size $\widetilde{s} \geq \frac{k^c\cdot n^{560A_0\cdot\gamma/\varphi}}{\xi^6}$ sampled independently and uniformly at random from Q^*, and s is an estimation of $|Q^*|$ such that $|s - |Q^*|| \leq \frac{|Q^*|\cdot\xi^3}{k^c \cdot n^{280A_0\cdot\gamma/\varphi}}$. Therfore, by Theorem 5, with probability at least $1 - n^{-96}$ for every $y \in V$ we have

$$||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 - \frac{1}{10^3 \cdot n} \leq \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 \leq ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 + \frac{1}{10^3 \cdot n}$$

Let $\ell = \frac{1}{10^3 \cdot |S^*|}$. Therefore, for every $y \in \mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 5\ell_{\mathrm{apx}}|Q^*\right)$ we have

$$\begin{aligned}
||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 &\leq \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 + \frac{1}{10^3 \cdot n} && \text{By Theorem 5} \\
&\leq 5 \cdot \ell_{\mathrm{apx}} + \frac{1}{10^3 \cdot n} && \text{As } y \in \mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 5\ell_{\mathrm{apx}}|Q^*\right) \\
&= \frac{1}{10^3 \cdot s} + \frac{1}{10^3 \cdot n} && \text{As } \ell_{\mathrm{apx}} = \frac{1}{10^3 \cdot s} \\
&\leq \frac{6}{10^3 \cdot |S^*|} && \text{By Claim 1, } s \in (1 \pm 10^{-3})|S^*| \\
&= 6 \cdot \ell && \text{As } \ell = \frac{1}{10^3 \cdot |S^*|}
\end{aligned}$$

Hence, for every $y \in \mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 5\ell_{\mathrm{apx}}|Q^*\right)$, we have $y \in \mathrm{cyl}\left(f_x^\kappa, 6\ell|Q^*\right)$. Thus $\mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 5\ell_{\mathrm{apx}}|Q^*\right) \subseteq \left(f_x^\kappa, 6\ell|Q^*\right)$. Therefore, by the assumption of the lemma we have

$$|\mathrm{cyl}\left(f_x^\kappa, 6\ell|Q^*\right)| \geq |\mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 5\ell_{\mathrm{apx}}|Q^*\right)| \geq 0.85 \cdot \beta \cdot |S^*|$$

Therefore, by Theorem 4 for every set Q satisfying $\mathrm{cyl}\left(f_x^\kappa, 20\ell|Q^*\right) \subseteq Q \subseteq \mathrm{cyl}\left(f_x^\kappa, 30\ell|Q^*\right)$, there exists a unique cluster $S \in \textsc{children}(S^*)$ such that Q is D-hierarchically-close to S, and for every $S' \neq S \in \textsc{children}(S^*)$, $\mathrm{cyl}_{\mathrm{apx}}\left(\mu_{S'}, \ell|Q^*\right) \cap Q = \emptyset$. Therefore, to complete the proof it suffices to show that

$$\mathrm{cyl}\left(f_x^\kappa, 20\ell|Q^*\right) \subseteq \mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 25\ell_{\mathrm{apx}}|Q^*\right) \subseteq \mathrm{cyl}\left(f_x^\kappa, 30\ell|Q^*\right)$$

Note that for every $y \in \mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 25\ell_{\mathrm{apx}}|Q^*\right)$ we have

$$\begin{aligned}
||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 &\leq \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 + \frac{1}{10^3 \cdot n} && \text{By Theorem 5} \\
&\leq 5 \cdot \ell_{\mathrm{apx}} + \frac{1}{10^3 \cdot n} && \text{As } y \in \mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 25\ell_{\mathrm{apx}}|Q^*\right) \\
&= \frac{25}{10^3 \cdot s} + \frac{1}{10^3 \cdot n} && \text{As } \ell_{\mathrm{apx}} = \frac{1}{10^3 \cdot s} \\
&\leq \frac{30}{10^3 \cdot |S^*|} && \text{By Claim 1, } s \in (1 \pm 10^{-3})|S^*| \\
&= 30 \cdot \ell && \text{As } \ell = \frac{1}{10^3 \cdot |S^*|}
\end{aligned}$$

Hence, for every $y \in \mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 25\ell_{\mathrm{apx}}|Q^*\right)$, we have $y \in \mathrm{cyl}\left(f_x^\kappa, 30\ell|Q^*\right)$. Thus $\mathrm{cyl}_{\mathrm{apx}}\left(f_x^\kappa, 25\ell_{\mathrm{apx}}|Q^*\right) \subseteq \left(f_x^\kappa, 30\ell|Q^*\right)$.

Also note that for every $y \in \mathrm{cyl}\left(f_x^\kappa, 20\ell|Q^*\right)$ we have

$$\begin{aligned}
\left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 &\leq ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 + \frac{1}{10^3 \cdot n} && \text{By Theorem 5} \\
&\leq 20 \cdot \ell + \frac{1}{10^3 \cdot n} && \text{As } y \in \mathrm{cyl}\left(f_x^\kappa, 20\ell|Q^*\right) \\
&= \frac{20}{10^3 \cdot |S^*|} + \frac{1}{10^3 \cdot n} && \text{As } \ell = \frac{1}{10^3 \cdot |S^*|} \\
&\leq \frac{25}{10^3 \cdot s} && \text{By Claim 1, } s \in (1 \pm 10^{-3})|S^*| \\
&= 25 \cdot \ell_{\mathrm{apx}} && \text{As } \ell_{\mathrm{apx}} = \frac{1}{10^3 \cdot s}
\end{aligned}$$

Therefore, with probability at least $1 - n^{-96}$ we have $\text{cyl}\left(f_x^\kappa, 20\ell|Q^*\right) \subseteq \text{cyl}_{\text{apx}}\left(f_x^\kappa, 25\ell_{\text{apx}}|Q^*\right) \subseteq \text{cyl}\left(f_x^\kappa, 30\ell|Q^*\right)$. Thus, by Theorem 4, $\text{cyl}_{\text{apx}}\left(f_x^\kappa, 25\ell_{\text{apx}}|Q^*\right)$ satisfies the required guarantees. □

Lemma 24. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, and Q^* be a set that is D-hierarchically-close to $S^* \in \mathcal{P}^{h-1}$ (Definition 19). Then for every $S \in \text{CHILDREN}(S^*)$ and $x \in cyl\left(\mu_S, \ell|Q^*\right)$ with probability at least $1 - n^{-96}$, we have*

$$cyl\left(\mu_S, \ell|Q^*\right) \subseteq cyl_{apx}\left(f_x^\kappa, 5\ell_{apx}|Q^*\right),$$

where, $|\mathcal{P}^h| = \kappa$, $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, $\ell = \frac{1}{10^3 \cdot |S^*|}$ and $\ell_{apx} = \frac{1}{1000 \cdot s}$, where s is an estimation of $|Q^*|$ such that $|s - |Q^*|| \le \frac{|Q^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$ and $A_0, D_0, c > 1$ are large enough constants. Also, $\mu_S \in \mathbb{R}^\kappa$ is the κ-dimensional center of S (Definition 10).

Proof. Let $r = |\text{CHILDREN}(S^*)|$, $\kappa = |\mathcal{P}^h|$ and $\Pi \in \mathbb{R}^{\kappa \times \kappa}$ be the subgraph projection matrix of Q^* with respect to κ and r (Definition 12). Note that $x \in \text{cyl}\left(\mu_S, \ell|Q^*\right)$. Thus by Definition 20 we have $||\Pi\mu_S - \Pi f_x^\kappa||_2^2 \le \ell$. Also for every $y \in \text{cyl}\left(\mu_S, \ell|Q^*\right)$ we have $||\Pi\mu_S - \Pi f_y^\kappa||_2^2 \le \ell$. Therefore, by triangle inequality for every $y \in \text{cyl}\left(\mu_S, \ell|Q^*\right)$ we have

$$||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 \le 4\ell \tag{123}$$

Let $\xi = \frac{1}{10^3}$. By Definition 22, we have $\widetilde{Q}$ is a set of size $\widetilde{s} \ge \frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}$ sampled independently and uniformly at random from Q^*, and s is an estimation of $|Q^*|$ such that $|s - |Q^*|| \le \frac{|Q^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$. Therfore, by Theorem 5, with probability at least $1 - n^{-96}$ for every $y \in V$ we have

$$||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 \le \left\|\widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa\right\|_{apx}^2 + \frac{1}{10^3 \cdot n}$$

Thus, for every $y \in \text{cyl}\left(\mu_S, \ell|Q^*\right)$ we have

$$\begin{aligned}
\left\|\widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa\right\|_{apx}^2 &\le ||\Pi f_x^\kappa - \Pi f_y^\kappa||_2^2 + \frac{1}{10^3 \cdot n} && \text{By Theorem 5} \\
&\le 4\ell + \frac{1}{10^3 \cdot n} && \text{By (123)} \\
&= \frac{4}{10^3 \cdot |S^*|} + \frac{1}{10^3 \cdot n} && \text{As } \ell = \frac{1}{10^3 \cdot |S^*|} \\
&\le \frac{5}{10^3 \cdot s} && \text{By Claim 1, } s \in (1 \pm 10^{-3})|S^*| \\
&= 5 \cdot \ell_{\text{apx}} && \text{As } \ell_{\text{apx}} = \frac{1}{10^3 \cdot s}
\end{aligned}$$

Hence, for every $y \in \text{cyl}\left(\mu_S, \ell|Q^*\right)$, we have $y \in \text{cyl}_{\text{apx}}\left(f_x^\kappa, 5\ell_{\text{apx}}|Q^*\right)$. Thus, with probability at least $1 - n^{-96}$ we have

$$\text{cyl}\left(\mu_S, \ell|Q^*\right) \subseteq \text{cyl}_{\text{apx}}\left(f_x^\kappa, 5\ell_{\text{apx}}|Q^*\right)$$

□

4.10.2 Quality of subsampled cylinders

Lemma 25. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, and Q^* be a set that is D-hierarchically-close to $S^* \in \mathcal{P}^{h-1}$ (Definition 19). Let $\widetilde{Q}$ be* *a set of size* $|\widetilde{Q}| \geq \frac{10^7 \cdot \log n}{\beta}$ *sampled independently and uniformly at random from* Q^**. Then for every* $S \in$ CHILDREN(S^*)*, with probability at least* $1 - n^{-100}$*, we have*

$$|\widetilde{Q} \cap cyl\,(\mu_S, \ell | Q^*)\,| \geq 0.95 \cdot \beta \cdot |\widetilde{Q}|,$$

where, $|\mathcal{P}^h| = \kappa$*,* $S^* \in \mathcal{P}^{h-1}$*,* $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$*,* $\ell = \frac{1}{10^3 \cdot |S^*|}$*,* $D_0 > 1$ *is a large enough constant and* $\mu_S \in \mathbb{R}^\kappa$ *is the* κ*-dimensional center of* S *(Definition 10).*

Proof. By Theorem 3, for every $S \in$ CHILDREN(S^*) we have $\mathrm{cyl}\,(\mu_S, \ell|Q^*)$ is D-hierarchically-close to sets S. Therefore, by Definition 19 and by Lemma 19 we have

$$\begin{aligned} ||\mathrm{cyl}\,(\mu_S, \ell|Q^*)\,| - |S|| &\leq 3 \cdot D \cdot \varphi_{h-1} \cdot |S| \\ &\leq 3 \cdot \left(\frac{D_0}{\beta^4 \cdot \varphi^2}\right) \cdot \gamma \cdot |S| && \text{As } D = \frac{D_0}{\beta^4 \cdot \varphi^2} \text{ and } \varphi_{h-1} \leq \varphi_{H-1} = \gamma \cdot \varphi \leq \gamma \\ &\leq \frac{|S|}{100} && \text{By Definition 6, } \frac{\gamma}{\beta^{30}} \text{ and } \frac{\gamma}{\varphi^{20}} \text{ are sufficiently small} \end{aligned} \tag{124}$$

Let X_i be a a random variable which is 1 if the i-th sampled vertex is in $\mathrm{cyl}\,(\mu_S, \ell|Q^*)$, and 0 otherwise. Thus $\mathbb{E}[X_i] = \frac{\mathrm{cyl}(\mu_S, \ell|Q^*)}{Q^*}$. Observe that $|\widetilde{Q} \cap \mathrm{cyl}\,(\mu_S, \ell|Q^*)\,|$ is a random variable defined as $\sum_{i=1}^{|\widetilde{Q}|} X_i$, where its expectation is given by

$$\begin{aligned} |\widetilde{Q} \cap \mathrm{cyl}\,(\mu_S, \ell|Q^*)\,| &= |\widetilde{Q}| \cdot \frac{\mathrm{cyl}\,(\mu_S, \ell|Q^*)}{|Q^*|} \\ &\geq |\widetilde{Q}| \cdot \frac{0.99 \cdot |S|}{|Q^*|} && \text{By (124)} \\ &\geq 0.99 \cdot |\widetilde{Q}| \cdot \frac{\beta \cdot |S^*|}{|Q^*|} && \text{By Definition 6, } |S| \geq \beta \cdot |S^*| \\ &\geq 0.99 \cdot \beta \cdot |\widetilde{Q}| \cdot \frac{0.99 \cdot |Q^*|}{|Q^*|} && \text{By Claim 4, } |S^*| \geq 0.99 \cdot |Q^*| \\ &\geq 0.98 \cdot \beta \cdot |\widetilde{Q}| \end{aligned} \tag{125}$$

Therefore, by Chernoff bound,

$$\Pr\left[|\widetilde{Q} \cap \mathrm{cyl}\,(\mu_S, \ell|Q^*)\,| < 0.95 \cdot \beta \cdot |\widetilde{Q}|\right] \leq \exp\left(-\frac{0.98 \cdot \beta \cdot |\widetilde{Q}|}{2 \cdot 10^4}\right) \leq n^{-100},$$

where the last inequality holds since $|\widetilde{Q}| \geq \frac{10^7 \cdot \log n}{\beta}$. Thus, for every $S \in S^*$ with probability at least $1 - n^{-100}$, we have

$$|\widetilde{Q} \cap \mathrm{cyl}\,(\mu_S, \ell|Q^*)\,| \geq 0.95 \cdot \beta \cdot |\widetilde{Q}|,$$

□

We defer the proof of the Lemma 26 and Lemma 27 to Appendix G.

Lemma 26. *Let* $G = (V, E)$ *be a* (k, γ)*-hierarchically-clusterable graph (Definition 6). Let* $h \in [H]$*, and suppose* $(\boldsymbol{P}^i)_{i=0}^{h-1}$ *is a* D*-approximation of* $(\mathcal{P}^i)_{i=0}^{h-1}$ *(Definition 7). Let* $\widetilde{V}$ *be a set sampled independently and uniformly at ranodm from* V*. Then for every* $\boldsymbol{S}^* \in \boldsymbol{P}^{h-1}$ *with probability at least* $1 - n^{-100}$ *we have*

1. $|\widetilde{V} \cap \boldsymbol{S}^*| \geq \max\left(\frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}, \frac{10^7 \cdot \log n}{\beta}\right)$

2. $\left| |\boldsymbol{S}^*| - \frac{n \cdot |\widetilde{V} \cap \boldsymbol{S}^*|}{|\widetilde{V}|} \right| \leq \frac{|\boldsymbol{S}^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$,

where, $A_0, c > 1$ are constants, $\xi = 10^{-3}$, $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, $D_0, c' > 1$ are large constants, and $|\widetilde{V}| \geq \frac{k^{c'} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6}$.

Lemma 27. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, where D_0 is a large constant. Let $h \in [H]$, $S^* \in \mathcal{P}^{h-1}$ and $\boldsymbol{S}^*$ be a set that is D-hierarchically-close to $S^* \in \mathcal{P}^{h-1}$ (Definition 19). Let $\widetilde{S}^*$ be a set of size $|\widetilde{S}^*| \geq \frac{10^7 \cdot \log n}{\beta}$ sampled independently and uniformly at ranodm from $\boldsymbol{S}^*$. Let $\mathcal{B} \subseteq \boldsymbol{S}^*$ and $\widetilde{\mathcal{B}} = \widetilde{S}^* \cap \mathcal{B}$. If $|\widetilde{\mathcal{B}}| \geq 0.9 \cdot \beta \cdot |\widetilde{S}^*|$, then with probability at least $1 - n^{-100}$ we have*

$$|\mathcal{B}| \geq 0.85 \cdot \beta \cdot |S^*|$$

4.10.3 Correctness of Algorithm 3 and 5

In this section, we prove the correctness of the RefinePartition (Lemma 30) and Oracle (Lemma 31). Intuitively, a good representative for a cluster S is a vertex such that the cylender around the vertex is D-hierarchically-close to the cluster S. The formal definition follows:

Definition 23 (Good Representative)**.** Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, $|\mathcal{P}^h| = \kappa$, and $S^* \in \mathcal{P}^{h-1}$. Let Q^* be a set that is D-hierarchically-close to S^* (Definition 19). Let $\ell_{\text{apx}} = \frac{1}{1000 \cdot s}$, where, s is an estimation of $|Q^*|$ such that $|s - |Q^*|| \leq \frac{|Q^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$, and $A_0, c > 1$ are large constants. We say that vertex $x \in Q^*$ is a good representative for cluster S if $\text{cyl}_{\text{apx}}(f_x^\kappa, 25 \ell_{\text{apx}} |Q^*)$ (Definition 20) is D-hierarchically-close to S.

To prove Lemma 30, we need Lemma 28 to count the number of clusters at every level. We defer the proof of Lemma 28 to Appendix F.

Lemma 28. *Let $k \in \mathbb{N}$ and let $\gamma > 0$ be a sufficiently small constant. There exists an algorithm which on input a (k, γ)-hierarchically clusterable graph G (Definition 6) and a parameter $h \leq H$ (where the associated hierarchical clustering is denoted $\mathcal{P} = (\mathcal{P}^0, \dots, \mathcal{P}^h)$) runs in time $(dn)^{1/2 + O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot \log n}{\gamma}\right)^{O(1)}$ and computes a number κ where $\kappa = |\mathcal{P}^h|$ holds with probability at least $1 - n^{-100}$.*

To prove Lemma 30, we also need Lemma 29 to count the number of children of every cluster. We defer the proof of Lemma 29 to Appendix E.

Lemma 29. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). For some large enough constant $D_0 > 1$, let $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, $h \in [H]$, $\kappa = |\mathcal{P}^h|$, $S^* \in \mathcal{P}^{h-1}$, $r = |\text{children}(S^*)|$ and $c > 1$ be a large enough constant. Let $\boldsymbol{S}^* \subseteq V$ be a set that is D-hierarchically-close to S^* (Definition 19). Let $\widetilde{S}^*$ be a set of size $\widetilde{s} \geq k^c \cdot n^{80 A_0 \cdot \gamma/\varphi}$ sampled independently and uniformly at random from S^*. Let s be an estimation of $|\boldsymbol{S}^*|$ such that $|s - |\boldsymbol{S}^*|| \leq \frac{|\boldsymbol{S}^*|}{k^c \cdot n^{40 A_0 \cdot \gamma/\varphi}}$. Then CountChildren$(G, \kappa, \widetilde{S}^*, s)$ runs in time $n^{1/2 + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$ and with probability at least $1 - n^{-97}$ returns r.*

Now we are ready to prove Lemma 30.

Lemma 30. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, $\kappa = |\mathcal{P}^h|$, $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ where $D_0 > 1$ is some large constant. Suppose that $(\boldsymbol{P}^i)_{i=0}^{h-1}$ is a $3D$-approximation of $(\mathcal{P}^i)_{i=0}^{h-1}$ (Definition 7) such that for every $S \in \mathcal{P}$, $\sigma(S)$ is D-hierarchically close to S (Definition 19). Let $\widetilde{P}^{h-1}$ be the subsampled clustering of $\boldsymbol{P}^{h-1}$ (Definition 15). Then with probability at least $1 - \kappa \cdot n^{-95}$, RefinePartition$(G, h, \kappa, \widetilde{P}^{h-1}, \widetilde{V}, \xi, \widetilde{T}, \mathcal{D})$ (Algorithm 4) finds a good representative for every $S \in \mathcal{P}^h$.*

Proof. By Lemma 28 and as per line 4 of the ConstructTree (Algorithm 3), we find $\kappa = |\mathcal{P}^h|$ correctly with probability at least $1 - n^{-100}$. Then as per line 6 of Algorithm 3, let $\widetilde{V}$ be a set of size $|\widetilde{V}| \geq \frac{k^{c'} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6}$ sampled independently and uniformly at random from V where c' is the constant from Lemma 26 and $\xi = 10^{-3}$ as per line 2. Let $\widetilde{P}^{h-1} = \widetilde{V} \cap \boldsymbol{P}^{h-1}$ denote the clustering subsampled from $\boldsymbol{P}^{h-1}$ (Definition 15). Then in Line 8 we call RefinePartition (Algorithm 4) on the subsampled clustering $\widetilde{P}^{h-1}$. Now, to argue correctness of the ConstructTree, it suffices to show that RefinePartition (Algorithm 4) finds a good representative for every $S \in \mathcal{P}^h$.

Let $S^* \in \mathcal{P}^{h-1}$ and let $\boldsymbol{S}^* = \sigma(S^*)$ be the corresponding set in $\boldsymbol{P}^{h-1}$ (Definition 7). Let $\widetilde{S}^* = \boldsymbol{S}^* \cap \widetilde{V}$, and $s = \frac{n \cdot |\widetilde{S}^*|}{|\widetilde{V}|}$ be an estimation of $|\widetilde{S}^*|$. Note that by choice of $|\widetilde{V}| \geq \frac{k^{c'} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6}$, and by Lemma 26 we have

$$|\widetilde{S}^*| \geq \max\left(\frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}, \frac{10^7 \cdot \log n}{\beta}\right) \tag{126}$$

and

$$||\boldsymbol{S}^*| - s| \leq \frac{|\boldsymbol{S}^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}} \tag{127}$$

Let $r = |\text{children}(S^*)|$. By (126), we have $|\widetilde{S}^*| \geq k^c \cdot n^{80 A_0 \cdot \gamma/\varphi}$, and by (127) we have $|s - |\boldsymbol{S}^*|| \leq \frac{|\boldsymbol{S}^*|}{k^c \cdot n^{40 A_0 \cdot \gamma/\varphi}}$. Therefore, by Lemma 29, as per line 4 of Algorithm 4, we find $r = |\text{children}(S^*)|$ correctly with probability at least $1 - n^{-97}$. Next we will show that for every $S \in \text{children}(S^*)$ we will find a good representative. As per line 3 of Algorithm 4, let $\ell_{\text{apx}} = \frac{1}{1000 \cdot s}$. For every $x \in \boldsymbol{S}^*$ we define

$$\mathcal{B}_x = \text{cyl}_{\text{apx}}\left(f_x^\kappa, 5\ell_{\text{apx}} | \boldsymbol{S}^*\right) = \left\{ y \in \boldsymbol{S}^* : \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 \leq 5 \cdot \ell_{\text{apx}} \right\}$$

and

$$\boldsymbol{S}_x = \text{cyl}_{\text{apx}}\left(f_x^\kappa, 25\ell_{\text{apx}} | \boldsymbol{S}^*\right) = \left\{ y \in \boldsymbol{S}^* : \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 \leq 25 \cdot \ell_{\text{apx}} \right\}.$$

Note that as per line 7 of Algorithm 4 we have

$$\widetilde{\mathcal{B}}_x = \left\{ y \in \widetilde{S}^* : \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 \leq 5 \cdot \ell_{\text{apx}} \right\} = \mathcal{B}_x \cap \widetilde{S}^*$$

In line 6 of Algorithm 4, we iterate over every vertex $x \in \widetilde{S}^*$. If for some vertex $x \in \widetilde{S}^*$, the condition in line 8 passes (i.e., $|\widetilde{\mathcal{B}}_x| \geq 0.9 \cdot \beta \cdot |\widetilde{S}^*|$), then, in line 10 we set vertex x as the representative of $\widetilde{S}$, where as per line 9, $\widetilde{S}$ is defined as follows:

$$\widetilde{S} = \left\{ y \in \widetilde{S}^* : \left\| \widetilde{\Pi} f_x^\kappa - \widetilde{\Pi} f_y^\kappa \right\|_{apx}^2 \leq 25 \cdot \ell_{\text{apx}} \right\} = \boldsymbol{S}_x \cap \widetilde{S}^*$$

We will show that for every cluster $S \in \text{children}(S^*)$, there exists a good representative x in $\widetilde{S}^*$ that passes the test (i.e., $|\widetilde{\mathcal{B}}_x| \geq 0.9 \cdot \beta \cdot |\widetilde{S}^*|$), and this vertex does not get removed while ball-carving other children $S' \in \text{children}(S^*)$.

For every $S \in$ children(S^*) there exists a good representative in $\widetilde{S}^*$ that passes the test : Note that $\widetilde{S}^*$ is uniformly distributed within $\boldsymbol{S}^*$ and $|\widetilde{S}^*| \geq \frac{10^7 \cdot \log n}{\beta}$, thus by Lemma 25, with probability at least $1 - n^{-100}$ we have

$$|\widetilde{S}^* \cap \text{cyl}\left(\mu_S, \ell | \boldsymbol{S}^*\right)| \geq 0.95 \cdot \beta \cdot |\widetilde{S}^*|$$

By a union bound over children of S^*, with probability at least $1 - r \cdot n^{-100}$ for all $S \in \text{children}(S^*)$ we have $|\widetilde{S}^* \cap \text{cyl}\left(\mu_S, \ell | \boldsymbol{S}^*\right)| \geq 0.95 \cdot \beta \cdot |\widetilde{S}^*|$.

We will show that for each $S \in \textsc{children}(S^*)$, there exists $x \in \widetilde{S}^*$ for which $|\widetilde{\mathcal{B}}_x| \geq 0.9{\cdot}\beta{\cdot}|\widetilde{S}^*|$ holds. Pick $x \in \widetilde{S}^* \cap \text{cyl}\left(\mu_S, \ell|\boldsymbol{S}^*\right)$. By Lemma 24, with probability at least $1 - n^{-96}$ we have $\text{cyl}\left(\mu_S, \ell|\boldsymbol{S}^*\right) \subseteq \mathcal{B}_x$. Therefore, we get

$$\widetilde{S}^* \cap \text{cyl}\left(\mu_S, \ell|\boldsymbol{S}^*\right) \subseteq \widetilde{S}^* \cap \mathcal{B}_x = \widetilde{\mathcal{B}}_x$$

Thus, we have $|\widetilde{\mathcal{B}}_x| \geq |\widetilde{S}^* \cap \text{cyl}\left(\mu_S, \ell|\boldsymbol{S}^*\right)| \geq 0.95 \cdot \beta \cdot |\widetilde{S}^*|$. Therefore, the test in line 8 will be passed for every $x \in \widetilde{S}^* \cap \text{cyl}\left(\mu_S, \ell|\boldsymbol{S}^*\right)$.

If vertex x passes the size test (i.e., $|\widetilde{\mathcal{B}}_x| \geq 0.9{\cdot}\beta{\cdot}|\widetilde{S}^*|$), then x is a good representative for a unique cluster $S \in$ children(S^*): Note that vertices in $\widetilde{S}^*$ are uniformly distributed in $\boldsymbol{S}^*$. Therefore, by Lemma 27 if $|\widetilde{\mathcal{B}}_x| \geq 0.9 \cdot \beta \cdot |\widetilde{S}^*|$, then with probability at least $1 - n^{-100}$ we have

$$|\mathcal{B}_x| \geq 0.85 \cdot \beta \cdot |S^*| \tag{128}$$

Therefore, by Lemma 23, with probability at least $1 - n^{-96}$ there exists a unique cluster $S \in \textsc{children}(S^*)$ such that $\boldsymbol{S}_x = \text{cyl}_{\text{apx}}\left(f_x^{\kappa}, 25\ell_{\text{apx}}|\boldsymbol{S}^*\right)$ is D-hierarchically-close to S, and for every $S' \neq S \in \textsc{children}(S^*)$,

$$\text{cyl}\left(\mu_{S'}, \ell|\boldsymbol{S}^*\right) \cap \boldsymbol{S}_x = \text{cyl}\left(\mu_{S'}, \ell|\boldsymbol{S}^*\right) \cap \text{cyl}_{\text{apx}}\left(f_x^{\kappa}, 25\ell_{\text{apx}}|\boldsymbol{S}^*\right) = \emptyset \tag{129}$$

For each $S \in$ children(S^*), a good representative survives: Finally, note that in line 12, we remove vertices from $\widetilde{S}^*$. We show that despite these removals, for every $S \in \textsc{children}(S^*)$, there still exists some x for which $|\widetilde{\mathcal{B}}_x| \geq 0.9\beta|\widetilde{S}^*|$ holds. By Lemma 23, for any x such that $|\mathcal{B}_x| \geq 0.9\beta|\boldsymbol{S}^*|$, and for every $S' \neq S \in \textsc{children}(S^*)$ we have $\text{cyl}\left(\mu_{S'}, \ell|\boldsymbol{S}^*\right) \cap \boldsymbol{S}_x = \emptyset$. Therefore, a good representative for S survives with probability $1 - n^{-96}$.

Overall, by a union bound, for a fixed $S^* \in \mathcal{P}^{h-1}$, we pick good representatives for all $S \in \textsc{children}(S^*)$ with probability at least $1 - r \cdot n^{-96} - n^{-99}$. Thus, by a union bound over all $S^* \in \mathcal{P}^{h-1}$, we pick a correct representatives for all $S \in \mathcal{P}^h$ with probability at least $1 - \kappa \cdot n^{-95}$. Finally, in line 14 of RefinePartition (Algorithm 4) stores the representative of clusters in data structure $\mathcal{D}$. □

Now we prove the correctness of ORACLE (Algorithm 5) in Lemma 31.

Lemma 31. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, $\kappa = |\mathcal{P}^h|$, and $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$ where $D_0 > 1$ is a large constant. Suppose that $(\boldsymbol{P}^i)_{i=0}^{h-1}$ is a $3D$-approximation of $(\mathcal{P}^i)_{i=0}^{h-1}$ (Definition 7), such that for every $S \in \mathcal{P}$, $\sigma(S)$ is D-hierarchically close to S (Definition 19). For every $i \in [\kappa]$ and $z \in V$, let*

$$\boldsymbol{S}_i = \left\{ z \in V : \textsc{Oracle}(G, z, \widetilde{T}, \mathcal{D}) = i \right\},$$

where $\mathcal{D}$ and $\widetilde{T}$ are constructed by ConstructTree(G) *(Algorithm 3) until iteration $h - 1$. Let $(\boldsymbol{P})^h = \{\boldsymbol{S}_i\}_{i=1}^{\kappa}$. Then with probability at least $1 - \kappa \cdot n^{-95}$ we have $(\boldsymbol{P}^i)_{i=0}^{h}$ is a $3D$-approximation of $(\mathcal{P}^i)_{i=0}^{h}$.*

Proof. Note that by Lemma 30, RefinePartition$(G, h, \kappa, \widetilde{P}^{h-1}, \widetilde{V}, \xi, \widetilde{T}, \mathcal{D})$ (Algorithm 4) finds a good representative for every $S \in \mathcal{P}^h$ with probability at least $1 - \kappa \cdot n^{-95}$. Let $x_1, \ldots, x_\kappa$ denote a set of representatives found by Algorithm 4. We then define

$$\boldsymbol{S}_{x_i} = \text{cyl}_{\text{apx}}\left(f_{x_i}^{\kappa}, 25\ell_{\text{apx}}|\boldsymbol{S}^*\right) = \left\{ y \in \boldsymbol{S}^* : \left\| \widetilde{\Pi} f_{x_i}^{\kappa} - \widetilde{\Pi} f_y^{\kappa} \right\|_{apx}^2 \leq 25 \cdot \ell_{\text{apx}} \right\}.$$

Since x_i is a good representative for cluster $\boldsymbol{S}_{x_i}$, then by Definition 23, $\boldsymbol{S}_{x_i}$ is D-hierarchically-close to a unique cluster $S \in \mathcal{P}^h$. Since $\boldsymbol{S}_{x_i}$ is D-hierarchically-close to S, by Definition 19 and Lemma 19 we have $|\boldsymbol{S}_{x_i} \triangle S| \leq 3 \cdot D \cdot \varphi_{h-1} \cdot |S|$. Therefore, with probability at least $1 - \kappa \cdot n^{-95}$ we have $(\boldsymbol{P}^i)_{i=0}^{h}$ is a $3D$-approximation of $(\mathcal{P}^i)_{i=0}^{h}$. □

4.10.4 Proof of Theorem 2

Now, we prove the main theorem.

Theorem 1. *[Informal version of Theorem 2] For sufficiently small constant $\gamma \in (0,1)$ there exists a* **hierarchical clustering** *oracle with $\approx k^{O(1)} n^{1/2+O(\gamma)}$ preprocessing time and $\approx k^{O(1)} n^{1/2+O(\gamma)}$ query time that achieves a constant factor approximation to Dasgupta cost on (k,γ)-hierarchically clusterable graphs.*

Proof. **Correctness:** By Lemma 31, and by union bound over all $h \in [H]$, with probability at least $1 - \sum_{h=1}^{H} \kappa_h \cdot n^{-95} \geq 1 - k^2 \cdot n^{-95} \geq 1 - n^{-93}$, we have that $(\boldsymbol{P}^h)_{h=0}^{H}$ is a $3D$-approximation of $(\mathcal{P}^h)_{h=0}^{H}$.

Running time in the Preprocessing phase: Line 3 of ConstructTree (Algorithm 3) makes H iterations in all. Line 4 of ConstructTree finds the number of clusters at level h in $\mathcal{P}^h$ (where $h \leq H$). By Lemma 28, each of these calls takes time at most $t_{\kappa_h} \leq n^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot d \cdot \log n}{\gamma}\right)^{O(1)}$. Now we consider the calls to RefinePartition (Algorithm 4) and the contribution to running time from these calls. By Lemma 29 line 4 of the RefinePartition procedure (Algorithm 4) takes time at most $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$. Line 5 in RefinePartition calls InitializeSubgraphProjMatrix with $\xi = 0.001$. By Theorem thm:cluster-pi-apx, the call to InitializeSubgraphProjMatrix (Algorithm 6) takes time at most $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$. The last major step in Algorithm 4, line 7 finds the cylinder $\mathcal{B}_x$ which by Remark 2 which computes $\left\|\widetilde{\Pi} f_x^{\kappa} - \widetilde{\Pi} f_y^{\kappa}\right\|_{apx}^2$ at most s^2 times for each pair of vertices $x, y \in \widetilde{S}$ and thus takes total time at most $s^2 \cdot n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$ which is asymptotically dominated by $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$. In all, therefore the total time taken by a single iteration of ConstructTree procedure (Algorithm 3) is dominated by line 4 which is at most $n^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot d \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$. For any $h \geq 1$, recall $\varphi_h \geq \varphi_1 \geq 1/\gamma^H$ and since $H = O(\log k)$, we get, $\frac{1}{\varphi_h} \leq k^{O(1)}$. So, the overall running time over all the H calls is asymptotically at most $n^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot d \cdot \log n}{\gamma}\right)^{O(1)}$.

Running time in the Query phase: Now we bound the running time of the Oracle$(G, z, \widetilde{T}, \mathcal{D})$ procedure (Algorithm 5). Note that this procedure takes the data structure $\mathcal{D}$ as an argument which stores all the representatives associated with each node in $\widetilde{T}$. Also, note that $\mathcal{D}$ also contains the subgraph projection matrices for each node in $\widetilde{T}$. Thus, in each iteration of the loop in line 2 comes from line 6. By Remark 2, since $\xi = 0.001$, each computation in line 6 takes time $t_{\text{hc}} = n^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot d \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$. In total, line 6 is executed at most H times per vertex. And thus the total time taken by Oracle$(G, z, \widetilde{T}, \mathcal{D})$ procedure (Algorithm 5) is at most $H \cdot t_{\text{hc}}$ which asymptotically, as seen with the preprocessing phase, is at most $n^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot d \cdot \log n}{\gamma \cdot \varphi_0}\right)^{O(1)}$.

Number of Random bits: Finally, we bound the number of random bits used in the preprocessing phase (ConstructTree, Algorithm 3) and the query phase (Oracle, Algorithm 5). First, consider the preprocessing phase. Fix $h \in [H]$. Line 6 samples $s_0 = \left(\frac{k \cdot n^{\gamma/\varphi} \cdot \log n}{\xi}\right)^{O(1)}$ vertices. The number of random bits used by this step is at most $s_0 \cdot \log n$. Line 8 calls RefinePartition$(G, h, \kappa, \widetilde{P}^{h-1}, \widetilde{V}, \xi, \widetilde{T}, \mathcal{D})$ (Algorithm 4). We now bound the number of bits used by this procedure. Line 4 calls the CountChildren$(G, h, \kappa, r, \widetilde{S}^*, \boldsymbol{s}^*)$ procedure (Algorithm 11). For $x, y \in \widetilde{S}^*$, CountChildren computes $\left\langle f_x^{\kappa}, f_y^{\kappa} \right\rangle_{apx}$. We compute this inner product by estimating collision probabilities between walks from x and y. Using 4-wise independent hash functions, the random walks from x can be implemented with $\log d \cdot \frac{1}{\varphi} \cdot O(\log n)$ random bits. In total, this step takes $B \le |\widetilde{S}^*|^2 \cdot \log d \cdot \frac{1}{\varphi} \cdot O(\log n)$ bits. Line 5 of RefinePartition calls InitializeSubgraphProjMatrix$(G, h, \kappa, r, \widetilde{S}^*, \boldsymbol{s}^*, \xi)$ (Algorithm 6). It can be implemented using the same number of random bits as the previous step. In line 7, number of random bits used to find $\mathcal{B}_x$ for each $x \in \widetilde{S}^*$ is at most $\log d \cdot \frac{1}{\varphi} \cdot O(\log n)$. So in total line 8 takes at most B bits. So the total number of bits used across the $H = \mathit{height}(\widetilde{T})$ levels in the preprocessing phase is at most $O(HB \cdot s_0 \log n)$ bits which is at most $\frac{k^{O(1)} \cdot n^{O(\gamma/\varphi)} \cdot (\log n)^{O(1)}}{\xi^{O(1)}}$. Recalling that Line 2 of ConstructTree sets $\xi = 0.001$, this is at most $\widetilde{O}\left(k^{O(1)} \cdot n^{O_{\beta,\varphi}(\gamma)}\right)$.

Now we consider the Oracle$(G, z, \widetilde{T}, \mathcal{D})$. For any vertex $x \in V$, line 2 makes H iterations. Number of random bits used in line 6 is at most $H \cdot \log d \cdot \frac{1}{\varphi} \cdot \log n \le \widetilde{O}\left(k^{O(1)} \cdot n^{O_{\beta,\varphi}(\gamma)}\right)$.

□

4.11 Bounding Dasgupta cost (Proof of Lemma 1 and Lemma 2)

In this section we first show that in a (k, γ)-hierarchically clusterable graph, any hierarchical clustering $\boldsymbol{P}$ which is a D-approximation of the ground truth hierarchical clustering $\mathcal{P}$ provides an $O\left(\frac{D}{\beta}\right)$ approximation to Dasgupta's cost of the hierarchical clustering $\mathcal{P}$. Next, in Lemma 2 we show that in a (k, γ)-hierarchically clusterable graph, the cost of the ground truth hierarchical clustering $\mathcal{P}$ is an $O_\beta(1)$ approximation of the optimum Dasgupta's cost. Finally, we put these results together in Corollary 1 to construct a hierarchical clustering with $O_{\beta,\varphi}(1)$ approximation to the optimum Dasgupta cost of G.

Lemma 1. *Let $G = (V, E)$ be a (k, γ)-hierarchical clusterable graph and let $\mathcal{P}$ be the hierarchical-clustering. If $\boldsymbol{P}$ is a D-approximation of $\mathcal{P}$, then $COST(\boldsymbol{P}) \le O\left(\frac{D}{\beta}\right) COST(\mathcal{P})$.*

Proof. Let T and $\boldsymbol{T}$ be the tree representation of a hierarchical clustering $\mathcal{P}$ and $\boldsymbol{P}$ respectively (Definition 3). We can write $\mathrm{COST}(\mathcal{P})$ and $\mathrm{COST}(\boldsymbol{P})$ as

$$\mathrm{COST}(\mathcal{P}) = \sum_{(x,y)\in E} |\textsc{leaves}(T[\mathrm{LCA}(x,y)])| = \sum_{h=0}^{H} \sum_{S^* \in \mathcal{P}^h} \sum_{\substack{(x,y)\in E \text{ s.t.} \\ T[\mathrm{LCA}(x,y)]=S^*}} |S^*| \tag{130}$$

and

$$\mathrm{COST}(\boldsymbol{P}) = \sum_{(x,y)\in E} |\textsc{leaves}(T[\mathrm{LCA}(x,y)])| = \sum_{h=0}^{H} \sum_{\boldsymbol{S}^* \in \boldsymbol{P}^h} \sum_{\substack{(x,y)\in E \text{ s.t.} \\ \boldsymbol{T}[\mathrm{LCA}(x,y)]=\boldsymbol{S}^*}} |\boldsymbol{S}^*|, \tag{131}$$

Recall that every cluster $S \in \mathcal{P}$, the corresponding approximation cluster in $\boldsymbol{P}$ is denoted by $\boldsymbol{S} = \sigma(S)$. For any $\boldsymbol{S}^* = \sigma(S^*)$ we want to estimate the number of edges with $T[\mathrm{LCA}(x, y)] = S^*$ by the number of edges with $\boldsymbol{T}[\mathrm{LCA}(x, y)] = \boldsymbol{S}^*$. Note that $\boldsymbol{S}^*$ is an approximation of S^*, hence, during refinement of $\boldsymbol{S}^*$ to its children some outliers migh have been generated. Let

$$O(\boldsymbol{S}^*) = \boldsymbol{S}^* \setminus \left(\bigcup_{\boldsymbol{S} \in \textsc{children}_{\boldsymbol{T}}(\boldsymbol{S}^*)} \boldsymbol{S} \right)$$

denote the set of outliers generated during refinement of $\boldsymbol{S}^*$ in $\boldsymbol{T}$. We have

$$\sum_{\substack{(x,y)\in E \text{ s.t.} \\ \boldsymbol{T}[\mathrm{LCA}(x,y)]=\boldsymbol{S}^*}} |\boldsymbol{S}^*| = |\boldsymbol{S}^*| \cdot |E(O(\boldsymbol{S}^*), \boldsymbol{S}^*)| + |\boldsymbol{S}^*| \cdot \sum_{\boldsymbol{S} \neq \boldsymbol{S}' \in \textsc{children}_{\boldsymbol{T}}(\boldsymbol{S}^*)} |E(\boldsymbol{S}, \boldsymbol{S}')| \tag{132}$$

Next, we will bound $|E(O(\boldsymbol{S}^*), \boldsymbol{S}^*)|$, and $\sum_{\boldsymbol{S} \neq \boldsymbol{S}' \in \textsc{children}_{\boldsymbol{T}}(\boldsymbol{S}^*)} |E(\boldsymbol{S}, \boldsymbol{S}')|$. Note that we have

$$\begin{aligned}
&|O(\boldsymbol{S}^*)| \\
&= |\boldsymbol{S}^*| - \sum_{\boldsymbol{S} \in \textsc{children}_{\boldsymbol{T}}(\boldsymbol{S}^*)} |\boldsymbol{S}| \\
&\leq |S^*| \cdot (1 + D \cdot \varphi_{h-1}) - \sum_{S \in \textsc{children}(S^*)} (1 - D \cdot \varphi_h)|S| && \text{As } \boldsymbol{S}^* \text{ (resp. } \boldsymbol{S}\text{) is } D\text{-hierarchically-close to } S^* \text{ (resp.} \\
&\leq D \cdot |S^*| \cdot (\varphi_{h-1} + \varphi_h) && \text{As } |S^*| = \sum_{S \in \textsc{children}(S^*)} |S| \\
&\leq 2 \cdot D \cdot \varphi_h \cdot |S^*| && (133)
\end{aligned}$$

Since $\boldsymbol{S}^*$ is D-hierarchically-close to S^* we have $|\boldsymbol{S}^*| \leq 2 \cdot |S^*|$. Thus, by (133) we have

$$|\boldsymbol{S}^*| \cdot |E(O(\boldsymbol{S}^*), \boldsymbol{S}^*)| \leq |\boldsymbol{S}^*| \cdot d \cdot |O(\boldsymbol{S}^*)| \leq 4 \cdot d \cdot D \cdot \varphi_h \cdot |S^*|^2 \tag{134}$$

Next, we bound $\sum_{\boldsymbol{S} \neq \boldsymbol{S}' \in \textsc{children}_{\boldsymbol{T}}(\boldsymbol{S}^*)} |E(\boldsymbol{S}, \boldsymbol{S}')|$ recalling that $\boldsymbol{S} = \sigma(S)$.

$$\begin{aligned}
&\sum_{\boldsymbol{S} \neq \boldsymbol{S}' \in \textsc{children}_{\boldsymbol{T}}(\boldsymbol{S}^*)} |E(\boldsymbol{S}, \boldsymbol{S}')| \\
&\leq \sum_{S \neq S' \in \textsc{children}(S^*)} |E(\sigma(S) \cap S, \sigma(S') \cap S')| + \sum_{S \in \textsc{children}(S^*)} |E(\sigma(S) \setminus S, \boldsymbol{S}^*)| \\
&\leq \sum_{S \neq S' \in \textsc{children}(S^*)} |E(S, S')| + \sum_{S \in \textsc{children}(S^*)} d \cdot |S \triangle \sigma(S)| \\
&\leq \sum_{S \neq S' \in \textsc{children}(S^*)} |E(S, S')| + \sum_{S \in \textsc{children}(S^*)} d \cdot D \cdot \varphi_h \cdot |S| \\
&= \sum_{S \neq S' \in \textsc{children}(S^*)} |E(S, S')| + d \cdot D \cdot \varphi_h \cdot |S^*| && \text{As } |S^*| = \sum_{S \in \textsc{children}(S^*)} |S|
\end{aligned}$$

Since $\boldsymbol{S}^*$ is D-approximation of S^* we have $|\boldsymbol{S}^*| \leq 2 \cdot |S^*|$. Therefore,

$$\begin{aligned}
|\boldsymbol{S}^*| \cdot \sum_{\boldsymbol{S} \neq \boldsymbol{S}' \in \textsc{children}_{\boldsymbol{T}}(\boldsymbol{S}^*)} |E(\boldsymbol{S}, \boldsymbol{S}')| &\leq 2 \cdot |S^*| \cdot \left(\sum_{S \neq S' \in \textsc{children}(S^*)} |E(S, S')| + d \cdot D \cdot \varphi_h \cdot |S^*| \right) \\
&\leq 2 \sum_{\substack{(x,y) \in E \text{ s.t.} \\ T[\mathrm{LCA}(x,y)] = S^*}} |S^*| + 2 \cdot d \cdot D \cdot \varphi_h \cdot |S^*|^2 \qquad (135)
\end{aligned}$$

Thus, by (130), (131), (132), (134) and (135) we have

$$\begin{aligned}
\mathrm{COST}(\boldsymbol{P}) &= \sum_{h=0}^{H} \sum_{\boldsymbol{S}^* \in \boldsymbol{P}^h} \sum_{\substack{(x,y)\in E \text{ s.t.} \\ \boldsymbol{T}[\mathrm{LCA}(x,y)]=\boldsymbol{S}^*}} |\boldsymbol{S}^*| && \text{By (131)} \\
&\le \sum_{h=0}^{H} \sum_{S^* \in \mathcal{P}^h} \left(4 \cdot d \cdot D \cdot \varphi_h \cdot |S^*|^2 + 2 \cdot d \cdot D \cdot \varphi_h \cdot |S^*|^2 + 2 \cdot \sum_{\substack{(x,y)\in E \text{ s.t.} \\ T[\mathrm{LCA}(x,y)]=S^*}} |S^*| \right) && \text{By (132), (134), (1} \\
&= 2 \cdot \mathrm{COST}(\mathcal{P}) + 6 \cdot \sum_{h=0}^{H} \sum_{S^* \in \mathcal{P}^h} d \cdot D \cdot \varphi_h \cdot |S^*|^2 && \text{By (130)}
\end{aligned} \tag{136}$$

Note that we have

$$\begin{aligned}
\mathrm{COST}(\mathcal{P}) &= \sum_{h=0}^{H} \sum_{S^* \in \mathcal{P}^h} \sum_{\substack{(x,y)\in E \text{ s.t.} \\ T[\mathrm{LCA}(x,y)]=S^*}} |S^*| \\
&= \frac{1}{2} \cdot \sum_{h=0}^{H} \sum_{S^* \in \mathcal{P}^h} \sum_{S \in \text{CHILDREN}(S^*)} |S^*| \cdot |E(S, S^* \setminus S)| \\
&\ge \frac{1}{2} \cdot \sum_{h=0}^{H} \sum_{S^* \in \mathcal{P}^h} |S^*| \cdot |S^*| \cdot d \cdot \varphi_h \cdot \beta && \text{By Definition 6 and as } |S| \ge \beta \cdot |S^*|
\end{aligned} \tag{137}$$

Therefore, by (136) and (137) we have

$$\mathrm{COST}(\boldsymbol{P}) \le \left(2 + \frac{12D}{\beta}\right) \cdot \mathrm{COST}(\mathcal{P}) \le \frac{14 \cdot D}{\beta} \cdot \mathrm{COST}(\mathcal{P}).$$

□

The second main result of this section is Lemma 2 whose proof is a modification of Theorem 2.3 of [CC17]. This lemma essentially asserts that in a (k, γ)-hierarchically clusterable graph, the cost of the ground truth hierarchical clustering $\mathcal{P}$ is an $O_\beta(\frac{1}{\gamma})$ approximation of the optimum Dasgupta's cost. To prove Lemma 2 we first need the following definition from [CC17].

Definition 24 (Maximal clusters induced by a tree)**.** Let $G = (V, E)$ be a graph. Let T be a tree with n leaves on vertices of G. Let $T(s)$ denote the clusters of size at most s induced by T. We refer to these clusters as maximal clusters of size at most s. We denote by $E_T(s)$ the edges that are cut in $T(s)$, i.e. edges with end points in different clusters in $T(s)$. For convenience, we also define $E_T(0) = E$. We remark that $T(s)$ is a partition of V.

Lemma 2. *Let $G = (V, E)$ be a (k, γ)-hierarchical clusterable graph and let $\mathcal{P}$ be the hierarchical-clustering. Suppose that $\phi_{in}(G) \ge \varphi_0$. Let $\mathcal{P}^*$ be a hierarchical clustering of the graph G that minimizes Dasgupta cost, then $COST(\mathcal{P}) \le O\left(\frac{1}{\beta^2}\right) \cdot COST(\mathcal{P}^*)$.*

Proof. Let T and T^* be the tree representation of a hierarchical clustering $\mathcal{P}$ and $\mathcal{P}^*$ respectively. Let T^* be the optimal solution for Dasgupta's cost. Let $T^*(t)$ be the maximal clusters in T^* of size at most t (Definition 24). Recall that $T^*(t)$ is a partition of V. We denote $E^*(t)$ the edges that are cut in $T^*(t)$, i.e. edges with end points in different clusters in $T^*(t)$. For convenience, we also define $E^*(0) = E$. By Claim 2.1 of [CC17] we have

$$\mathrm{COST}(T^*) = \sum_{t=0}^{n-1} |E^*(t)| \tag{138}$$

It will be convenient to use the following bound that is directly implied by the above claim.

$$2 \cdot \mathrm{COST}(T^*) = 2 \cdot \sum_{t=0}^{n-1} |E^*(t)| \geq \sum_{t=0}^{n} |E^*\left(\lfloor t/2 \rfloor\right)| \tag{139}$$

For convenience, we define $\varphi_{H+1} = 1$. Let $0 \leq h \leq H$, and let's look at a cluster $S^* \in \mathcal{P}^h$ with size $|S^*| = s$ in the solution produced by T. Suppose that S^* has r children $S_1, \ldots, S_r$. Note that by Definition 6 we have $\beta \cdot |S^*| \leq |S_i| \leq (1-\beta) \cdot |S^*|$. For every $S_i \neq S_j \in \textsc{children}(S^*)$, the contribution of the edges $E(S_i, S_j)$ to the hierarchical clustering objective function is $s \cdot |E(S_i, S_j)|$. We want to charge this cost to $T^*(\lfloor s/2 \rfloor)$ and for that we first observe that the edges cut in $T^*(\lfloor s/2 \rfloor)$satisfy the following:

$$s \cdot |E^*(\lfloor s/2 \rfloor)| = \frac{1}{\beta} \cdot (\beta \cdot s \cdot |E^*(\lfloor s/2 \rfloor)|) \leq \frac{1}{\beta} \cdot \sum_{t=(1-\beta)\cdot s+1}^{s} |E^*(\lfloor t/2 \rfloor)| \tag{140}$$

This follows easily from the fact that $|E^*(t) \cap S^*| \leq |E^*(t-1) \cap S^*|$. For any partition of S^* into disjoint sets $Q_1, \ldots, Q_\ell$ we define the value of the partition as follows:

$$\mathrm{VAL}(Q_1, \ldots, Q_\ell) = \frac{\sum_{i=1}^{\ell} |E(Q_i, S^* \setminus Q_i)|}{\sum_{i=1}^{\ell} |Q_i| \cdot |S^* \setminus Q_i|} \tag{141}$$

Now in order to explain our charging scheme, let's look at the partition $O_1, \ldots, O_m$ induced inside the cluster S^* by $T^*(\lfloor s/2 \rfloor)$, where by design the size of each $|O_i| = \gamma_i \cdot |S^*|$, $\gamma_i \leq \frac{1}{2}$. We have:

$$\mathrm{VAL}(O_1, \ldots, O_m) = \frac{\sum_{i=1}^{m} |E(O_i, S^* \setminus O_i)|}{\sum_{i=1}^{m} \gamma_i (1-\gamma_i) \cdot s^2} \leq 2 \cdot \frac{|E^*(\lfloor s/2 \rfloor)|}{s^2/2} = 4 \cdot \frac{|E^*(\lfloor s/2 \rfloor)|}{s^2} \tag{142}$$

The first inequality holds because $\sum_{i=1}^{m} \gamma_i = 1$ and $\sum_{i=1}^{m} \gamma_i^2 \leq 1/2$, and the factor of 2 is introduced since we double counted every edge. Let $D_1, \ldots, D_b$ be a partition of S^* into at least two parts that minimizes $\mathrm{VAL}(D_1, \ldots, D_b)$. Therfore, by the definition of minimum we have

$$\mathrm{VAL}(D_1, \ldots, D_b) \leq \mathrm{VAL}(O_1, \ldots, O_m) \leq 4 \cdot \frac{|E^*(\lfloor s/2 \rfloor)|}{s^2} \tag{143}$$

Note that since $S^* \in \mathcal{P}^h$ we have $\phi_{\mathrm{in}}^G(S^*) \geq \varphi_h$. Therefore, we have

$$\begin{aligned}
\mathrm{VAL}(D_1, \ldots, D_b) &= \frac{\sum_{i=1}^{b} |E(D_i, S^* \setminus D_i)|}{\sum_{i=1}^{b} |D_i| \cdot |S^* \setminus D_i|} \\
&\geq \min_{i \in [b]} \frac{|E(D_i, S^* \setminus D_i)|}{|D_i| \cdot |S^* \setminus D_i|} \\
&\geq \min_{i \in [b]} \frac{\varphi_h \cdot d \cdot \min(|D_i|, |S^* \setminus D_i|)}{|D_i| \cdot |S^* \setminus D_i|} && \text{As } \phi_{\mathrm{in}}^G(S^*) \geq \varphi_h \\
&\geq \frac{\varphi_h \cdot d}{|S^*|}
\end{aligned} \tag{144}$$

Also, note that for every $S_i \in \textsc{children}(S^*)$ we have $S_i \in \mathcal{P}^{h+1}$, therefore, we have

$$\begin{aligned}
\text{VAL}(S_1, \ldots, S_r) &= \frac{\sum_{i=1}^r |E(S_i, S^* \setminus S_i)|}{\sum_{i=1}^r |S_i| \cdot |S^* \setminus S_i|} \\
&\leq \max_{i \in [r]} \frac{|E(S_i, S^* \setminus S_i)|}{|S_i| \cdot |S^* \setminus S_i|} \\
&\leq \max_{i \in [r]} \frac{\varphi_h \cdot d \cdot |S_i|}{|S_i| \cdot |S^* \setminus S_i|} && \text{As } \phi^G_{\text{out}}(S_i) \leq O(\varphi_h) \\
&\leq O\left(\frac{\varphi_h \cdot d}{\beta \cdot |S^*|}\right) && \text{As } |S^* \setminus S_i| \geq \beta \cdot |S^*| \qquad (145)
\end{aligned}$$

Putting (143), (144), and (145) together we get

$$\text{VAL}(S_1, \ldots, S_r) \leq O\left(\frac{1}{\beta}\right) \cdot \text{VAL}(D_1, \ldots, D_b) \leq O\left(\frac{1}{\beta}\right) \cdot \frac{|E^*(\lfloor s/2 \rfloor)|}{s^2} \qquad (146)$$

The contribution of this step to the hierarchical clustering objective function is:

$$\begin{aligned}
s \cdot \frac{1}{2} \cdot \sum_{i=1}^r |E(S_i, S^* \setminus S_i)| &= \frac{s}{2} \cdot \text{VAL}(S_1, \ldots, S_r) \cdot \sum_{i=1}^r |S_i| \cdot |S^* \setminus S_i| \\
&\leq \frac{s}{2} \cdot \text{VAL}(S_1, \ldots, S_r) \cdot s^2 \cdot \sum_{i=1}^r z_i(1 - z_i) && \text{As } z_i = \frac{|S_i|}{|S^*|} \\
&\leq \frac{s}{2} \cdot O\left(\frac{1}{\beta}\right) \cdot \frac{|E^*(\lfloor s/2 \rfloor)|}{s^2} \cdot s^2 && \text{By (146), and } \sum_{i=1}^r z_i(1 - zi) \leq 1 \\
&\leq O\left(\frac{s}{\beta}\right) \cdot |E^*(\lfloor s/2 \rfloor)| && (147)
\end{aligned}$$

Therefore, the total cost of T is

$$\begin{aligned}
\text{COST}(T) &= \sum_{S^* \in T} |S^*| \cdot \frac{1}{2} \cdot \sum_{S \in \textsc{children}(S^*)} |E(S, S^* \setminus S)| \\
&\leq \sum_{S^* \in T} O\left(\frac{|S^*|}{\beta}\right) \cdot |E^*(\lfloor s/2 \rfloor)| && \text{By (147)} \\
&= O\left(\frac{1}{\beta}\right) \cdot \sum_{S^* \in T} \frac{1}{\beta} \sum_{t=(1-\beta)\cdot|S^*|+1}^{|S^*|} |E^*(\lfloor t/2 \rfloor)| && \text{By (140)} \\
&\leq O\left(\frac{1}{\beta^2}\right) \cdot \sum_{t=1}^n |E^*(\lfloor t/2 \rfloor)| && \text{explained below} \\
&\leq O\left(\frac{1}{\beta^2}\right) \cdot 2 \cdot \text{COST}(T^*) && \text{By (139)} \qquad (148)
\end{aligned}$$

The fourth inequality holds because

$$\sum_{S^* \in T} \sum_{t=(1-\beta)|S^*|+1}^{|S^*|} |E^*(\lfloor t/2 \rfloor) \cap S^*| \leq \sum_{t=1}^n |E^*(\lfloor t/2 \rfloor)|,$$

as for a fixed value of t and S^*, the LHS is: $|E^*(\lfloor t/2 \rfloor) \cap S^*|$. Consider which clusters S^* contribute such a term to the LHS. From the fact that $(1 - \beta)|S^*| + 1 \leq t \leq |S^*|$ we need to have that $|S^*| \geq t$ and $\max_{S \in \textsc{children}(S^*)} |S| \leq (1 - \beta) \cdot |S^*| \leq t$. We deduce that S^* is a

minimal cluster of size $|S^*| \geq t > \max_{S \in \text{CHILDREN}(S^*)} |S|$. Thus, if all of the children of S^* are of size less than t, then this cluster S^* contributes such a term. The set of all such S^* form a disjoint partition of V because of the definition for minimality (in order for them to overlap in the hierarchical clustering, one of them needs to be ancestor of the other and this cannot happen because of minimality). Since $|E^*(\lfloor t/2 \rfloor) \cap S^*|$ for all such S^* forms a disjoint partition of $E^*(\lfloor t/2 \rfloor)$, the claim follows by summing up over all t. Thus we have

$$\text{COST}(T) \leq O\left(\frac{1}{\beta^2}\right) \cdot \text{COST}(T^*).$$

□

Finally, we state the following corollary of Theorem 2 that bounds the Dasgupta's cost of the approximate hierarchical clustering.

Corollary 1. *For every integer $k \geq 2$, every $H \in O(\log k)$, every $\beta, \varphi \in (0,1)$, every $\gamma \leq O(\min(\varphi^{20}, \beta^{30}))$ and every graph $G = (V, E)$ that $\phi_{in}(G) \geq \varphi_0$ and admits a (k,γ)-hierarchical clustering $\mathcal{P}$, there exists a D-approximate hierarchical clustering $\boldsymbol{P}$ (Definition 8) with $D = O\left(\frac{1}{\beta^4 \cdot \varphi^2}\right)$ such that*

$$COST(\boldsymbol{P}) \leq O\left(\frac{1}{\beta^7 \cdot \varphi^2}\right) \cdot COST(\mathcal{P}^*),$$

where, $\mathcal{P}^*$ is the hierachical clustering with optimum Dasgupta cost.

Proof. Let $\boldsymbol{P}$ be the approximate hierarchical clustering that is obtained by Theorem 2. With high probability, $\boldsymbol{P}$ is a D-approximation of $\mathcal{P}$. Thus, by Lemma 1, we have

$$\text{COST}(\boldsymbol{P}) \leq O\left(\frac{D}{\beta}\right) \cdot \text{COST}(\mathcal{P}). \tag{149}$$

Let $\mathcal{P}^*$ denote a hierarchical clustering of G with the optimum Dasgupta cost. Thus, by Lemma 2 we have

$$\text{COST}(\mathcal{P}) \leq O\left(\frac{1}{\beta^2}\right) \cdot \text{COST}(\mathcal{P}^*). \tag{150}$$

Thus, by (149), (150) and as $D = \frac{D_0}{\beta^4 \varphi^2}$ we get

$$\text{COST}(\boldsymbol{P}) \leq O\left(\frac{D}{\beta^3}\right) \cdot \text{COST}(\mathcal{P}^*) \leq O\left(\frac{1}{\beta^7 \cdot \varphi^2}\right) \cdot \text{COST}(\mathcal{P}^*).$$

□

A Graphs in Hierarchical Stochastic Block Models are (k, γ)-clusterable

In this section we present a natural family of graphs that is (k, γ)-hierarchically-clusterable showing that our definition is at the same time natural and well-founded. In particular, the main result of this section is Theorem 7. We briefly sketch the proof of Theorem 7. Fix $H \in \mathbb{N}$, $\varphi \in (0, 1)$ (which is some large constant bounded away from zero) and some sufficiently small γ. We use a standard random graph model to sample 2^H expanders, $C_1, C_2, \ldots, C_{2^H}$, each containing $t = n/2^H$ vertices and each with inner conductance at least $\varphi \geq \Omega(1)$. The parameters of our model are chosen so that it is possible to (recursively) merge these expanders, two at a time which produces a collection of 2^{H-1} sets each containing $2t$ vertices such that each set has inner conductance $\varphi \cdot \gamma$. We recursively merge these clusters to obtain a collection $\mathcal{P}^h$ of sets for all $0 \leq h \leq H$. Finally, we prove that at any level h, the collection $\mathcal{P}^h$ at that level satisfies the properties needed in Definition 6. The following lemma is a useful primitive which we use to show that a cluster obtained after merging two clusters (according to our random model) has nice expansion properties.

Lemma 32. *Let $\varphi > 0$ be a constant bounded away from 0 and let $d \in \mathbb{N}$. Let $\epsilon \in (1/d, \varphi/16)$. For sufficiently large n, let $G = (V, E)$ be a $d + \epsilon d$-regular graph with two clusters C_1 and C_2 where*

- $|C_1| = |C_2| = n/2$
- *$G[C_1], G[C_2]$ are d-regular.*
- $\min(\varphi_{in}^G(C_1), \varphi_{in}^G(C_2)) \geq \varphi$.
- *For each $u \in C_1$, $v \in C_2$, we have $|E(u, C_2)| = \epsilon d = |E(v, C_1)| > 1$.*

Then $\varphi(G) \geq \epsilon/16$.

Proof. Let $S \subseteq C_1 \cup C_2$ be such that $|S| \leq n/2$. We will show that

$$|E(S, V \setminus S)| \geq \frac{\epsilon \cdot d|S|}{16}.$$

Let $S_1 = S \cap C_1, T_1 = C_1 \setminus S_1$ and $S_2 = S \cap C_2, T_2 = C_2 \setminus S_2$. Consider the following cases

- $|S_1| \leq |T_1|$ and $|S_2| \leq |T_2|$. In this case, note that

$$|E(S, V \setminus S)| \geq |E(S_1, T_1)| + |E(S_2, T_2)| \geq \varphi d|S_1| + \varphi d|S_2| \geq \varphi d|S|.$$

- $|S_1| \geq |T_1|, |S_2| \leq |T_2|$. We now split into two more cases as below.

  1. $|S_1| - |S_2| \geq n/8$. In this case, note

  $$|E(S, V \setminus S)| \geq |E(S_1, T_2)| \geq \epsilon d \cdot (|S_1| - |S_2|) \geq \epsilon d \cdot n/8 \geq \frac{\epsilon d \cdot n}{16}.$$

  2. $|S_1| - |S_2| < n/8$. In this case, note that $|S_1|, |T_1|, |S_2|, |T_2|$ all have comparable sizes. In particular, $|S_1| \in [n/4, n/4 + n/8]$ and $|S_2| \in [n/4 - n/8, n/4]$. Thus, both $|T_1|$ and $|S_2|$ contain at least $n/8$ vertices. Therefore,

  $$|E(S, V \setminus S)| \geq |E(S_1, T_1)| + |E(S_2, T_2)| \geq \varphi d|T_1| + \varphi d|S_2| \geq \varphi \cdot \frac{dn}{4} \geq \varphi \cdot \frac{d|S|}{12}.$$

- $|S_2| \geq |T_2|, |S_1| \leq |T_1|$. Identical to the case above.

□

The key to our argument lies in the precise description of the random model our graphs come from. We first setup the stage to describe our random model.

Fact 1. *Let n be a sufficiently large positive integer. For $p \geq 10 \log n/n$, sample a graph $G \sim \mathcal{G}(n, p)$. Then with probability at least $1 - 1/n^2$, G has expansion at least $\Omega(1)$.*

Claim 2. *Let n be a sufficiently large positive integer. For $c > 30$ and $p = c \log n/n$, sample a graph $G \sim \mathcal{G}(n, p)$. Then with probability at least $1 - 1/n^2$, all vertices in G have degree between $2/3 \cdot np$ and $4/3 \cdot np$ is at least $1 - 1/n^2$.*

Proof. The proof is a direct application of Chernoff Bounds. The average degree is $np = c \log n$. By a Chernoff bound, the probability that for a fixed vertex the degree is between $2/3np$ and $4/3np$ is at most $\exp(-1/9 \cdot 30 \log n) \leq 1/n^3$. By a union bound over all the vertices, the probability that some vertex has degree outside this interval is at most $1/n^2$. □

We are now ready to describe our base cluster.

Construction of Erdos Renyi Clusters: Fix $H \in \mathbb{N}$ and let $\varphi \in (0, 1)$ denote a constant so that for large enough integer n, it holds that $G \sim \mathcal{G}(n/2^H, \frac{30 \log n}{n/2^H})$ has inner conductance $\varphi \geq \Omega(1)$ (by Fact 1). Now, let

$$\varphi_0 \geq \frac{1}{2^{O(H)}}.$$

Set $k = 2^H$ and for some large enough n, sample graphs

$$C'_1, C'_2, \dots C'_k \sim \mathcal{G}\left(\frac{n}{k}, \frac{30 \log n}{n/k}\right).$$

By Claim 2 w.h.p. all vertices in each C_i have degree between $20k \cdot \log n$ and $40k \cdot \log n$. Add enough self loops at each vertex to obtain a d' regular graph where $d' = 40k \cdot \log n$. Call the resulting clusters $C_1, C_2, \dots C_k$. We refer to these $k = 2^H$ clusters as Erdos-Renyi collection of clusters with parameter H.

Claim 3. *Fix $H \in \mathbb{N}$, let $k = 2^H$ and let $C_1, C_2, \dots C_k$ denote the Erdos Renyi Clusters with paramter H as defined above. Then each of these clusters has conductance at least $\Omega(1)$ with probability at least $1 - 2/n$.*

Proof. Follows from Fact 1 and Claim 2 on taking the union bound. □

Hierarchy of (H, φ_0, γ)-clusters: Fix $H \in \mathbb{N}$, $\varphi_0 \geq \frac{1}{2^{O(H)}}$. Let $k = 2^H$ and for sufficiently large n, let $C_1, C_2, \dots C_k$ denote the Erdos Renyi collection of clusters with paramter H where each cluster has inner conductance at least $2 \cdot \varphi \geq \Omega(1)$.

Recall each cluster is d'-regular with $d' = 40k \log n$. Let $\gamma \leq \varphi^{20}$ and set $\varphi_h = \varphi_0/\gamma^h$. Add $\epsilon d'$ half-edges at each vertex where $\epsilon = 16(\varphi_0 + \varphi_1 + \ldots + \varphi_{H-1})$. Write $d = d' + \epsilon d'$ and note that the resulting graph is a collection of k disjoint components each of which is d-regular. We say that these half-edges come in H different colors and for $h \in [H]$, we have $16\varphi_{h-1}d'$ half-edges colored h at each vertex. We now describe a tree with base clusters $C_1, C_2, \dots C_k$ at level H. The definition of our tree is recursive. Suppose we already have a collection of sets (all of which are actually parition V) $\mathcal{P}_H, \mathcal{P}_{H-1}, \cdots \mathcal{P}_h$ where $0 < h \leq H$. We now define the partition $\mathcal{P}_{h-1}$. This is done in the following steps. Denote the clusters at level h as $A_1, A_2, \dots A_{2^h}$.

- For an odd $i \leq 2^h$, take the clusters A_i, A_{i+1}. Take all the half-edges colored h between A_i and A_{i+1}. Add a perfect matching between the all the half-edges. If the matching pairs up a half-edge colored h on u with a half-edge colored h on v, we add in the edge (u, v). Note that this results in a graph which has parallel edges.

- Drop all the half-edges colored h from A_i and A_{i+1}.

This gives level $h-1$. Note that this procedure preserves d-regularity while going from level h to level $h-1$. Repeat this process till it terminates at level 0 and return the final graph obtained. We denote this model as $\mathcal{G}(n, H, \varphi_0, \gamma)$.

Theorem 7. *Fix $H \in \mathbb{N}$, $\varphi_0 \geq \frac{1}{2^{O(H)}}$ and $\varphi \in (0,1)$. Let $\gamma \leq \varphi^{20}$ and for $h \in [H]$ set $\varphi_h = \varphi_0/\gamma^h$ and let $k = 2^H$. For sufficiently large n, let $G \sim \mathcal{G}(n, H, \varphi_0, \gamma)$. Then, with high probability G is (k, γ)-hierarchically clusterable (Definition 6).*

Proof. We need to verify that with high probability, a graph G sampled from $\mathcal{G}(n, H, \varphi_0, \gamma)$ satisfies the conditions in Definition 6. As noted in Claim 3, with probability at least $1 - 2/n$, all the clusters $C_1, C_2, \ldots C_k$ are d'-regular for $d' = 40k \cdot \log n$ and have inner conductance at least 2φ. As per the definition of the hierarchy of (H, φ_0, γ)-clusters, we add $\epsilon d'$ half-edges at each vertex where $\epsilon = 16(\varphi_0 + \varphi_1 + \ldots + \varphi_{H-1})$ (and thus each C_i still has inner conductance at least $\varphi \geq \Omega(1)$). Write $d = d' + \epsilon d'$ and note that each vertex has degree d. We recall the half-edges come in H different colors and we have $16\varphi_{h-1}d'$ half-edges colored h. Now, we proceed to verify the conditions of Definition 6.

To do this, take any $h \in [H]$ and note that for each $S^* \in \mathcal{P}^{h-1}$ and any $S \in \textsc{children}(S^*)$ we have $|S| = |S^*|/2$ and thus the size condition in Definition 6 is met. For the first condition, we will show the following. Take any pair of sibling clusters at level h. For convenience denote them as A_1 and A_2. Suppose for some $0 < \delta < \alpha \ll 1$, both A_1 and A_2 have inner conductance α and the sparsity of the cut between A_1 and A_2 is δ.

By Lemma 32, the inner conductance of $A = A_1 \cup A_2$ is at least $\delta/16$. Now, we apply the above to the bottom level collection $\mathcal{P}^H$ of G by setting $\delta = 16\varphi_{H-1}$. This means the inner conductance of any cluster at level $H-1$ is at least φ_{H-1}. Inductively, this means all clusters at level $h \geq 1$ have inner conductance at least φ_h and outer conductance at most $16\varphi_{h-1}$. □

B Properties of Hierarchically Clusterable Graphs

Lemma 5. *(Variance bounds) Let $\kappa \in [n]$ and $m \geq 2$ be integers. Let $G = (V, E)$ be a d-regular graph. Suppose that V is partitioned into m disjoint subsets $V = S_1 \cup \ldots \cup S_m$. Then for any $\alpha \in \mathbb{R}^\kappa$ with $\|\alpha\| = 1$ we have*

$$\sum_{i=1}^{m} \sum_{x \in S_i} \langle f_x^\kappa - \mu_i, \alpha \rangle^2 \leq \frac{\lambda_\kappa}{\min_{i \in m} \chi_2(S_i)},$$

where $\mu_i \in \mathbb{R}^\kappa$ is the κ-dimensional center of set S_i (Definition 10), $\chi_2(S_i)$ is the second smallest eigenvalue of L_{S_i} (Definition 13), and λ_κ denote the κ-th smallest eigenvalue of L_G.

Proof. Let $z = U_{[\kappa]}\alpha$. Note that

$$\langle z, L_G z \rangle \leq \lambda_\kappa \tag{151}$$

Fix some $i \in [k]$, let $z' \in \mathbb{R}^n$ be a vector such that $z'(x) := z(x) - \langle \mu_i, \alpha \rangle$. For any $S \subseteq V$, we define $z'_S \in \mathbb{R}^n$ to be a vector such that for all $x \in V$ $z'_S(x) = z'(x)$ if $x \in S$ and $z'_S(x) = 0$ otherwise. Note that $z(x) = \langle f_x^\kappa, \alpha \rangle$, thus we have

$$\sum_{x \in V} z'_{S_i}(x) = \sum_{x \in S_i} z'(x) = \sum_{x \in S_i} z(x) - \langle \mu_i, \alpha \rangle = \sum_{x \in S_i} \langle f_x^\kappa - \mu_i^\kappa, \alpha \rangle = 0$$

Thus we have $z'|_{S_i} \perp \mathbb{1}$, so by properties of Rayleigh quotient we get

$$\chi_2(G[S_i]) \leq \frac{\langle z'_{S_i}, L_i z'_{S_i} \rangle}{\langle z'_{S_i}, z'_{S_i} \rangle} = \frac{1}{d} \frac{\sum_{x,y \in S_i, (x,y) \in E} (z'(x) - z'(y))^2}{\sum_{x \in S_i} (z'(x))^2} = \frac{1}{d} \frac{\sum_{x,y \in S_i, (x,y) \in E} (z(x) - z(y))^2}{\sum_{x \in S_i} (z(x) - \langle \mu_i, \alpha \rangle)^2} \tag{152}$$

Thus we have

$$\begin{aligned}
\lambda_\kappa &\geq \langle z, L_G z\rangle && \text{By (151)}\\
&= \frac{1}{d}\cdot \sum_{(x,y)\in E} (z(x)-z(y))^2\\
&\geq \frac{1}{d}\cdot \sum_{i=1}^{m} \sum_{x,y\in S_i,(x,y)\in E} (z(x)-z(y))^2\\
&\geq \min_{i\in m} \chi_2(G[S_i])\cdot \sum_{i=1}^{m}\sum_{x\in S_i} (z(x)-\langle \mu_i,\alpha\rangle)^2 && \text{By (152)}
\end{aligned}$$

Recall that for all $x \in V$, $z(x) = \langle f_x^\kappa, \alpha\rangle$. Therefore we have

$$\sum_{i=1}^{m}\sum_{x\in S_i} \langle f_x^\kappa - \mu_i, \alpha\rangle^2 \leq \frac{\lambda_\kappa}{\min_{i\in m}\chi_2(G[S_i])}$$

□

Lemma 6. *Let $\kappa \in [n]$ and $m \geq 2$ be integers. Let $G = (V, E)$ be a d-regular graph. Suppose that V is partitioned into m disjoint subsets $V = S_1 \cup \ldots \cup S_m$. Let $Q \subseteq V$. Then for any $\alpha \in \mathbb{R}^\kappa$ with $\|\alpha\| = 1$ we have*

$$\left|\alpha^T \left(\sum_{i=1}^{m} |Q\cap S_i|\mu_i\mu_i^T - \sum_{x\in Q} f_x^\kappa f_x^{\kappa T}\right)\alpha\right| \leq 2\cdot\sqrt{\frac{\lambda_\kappa}{\min_{i\in m}\chi_2(S_i)}},$$

where $\mu_i \in \mathbb{R}^\kappa$ is the κ-dimensional center of set S_i (Definition 10), $\chi_2(S_i)$ is the second smallest eigenvalue of L_{S_i} (Definition 13), and λ_κ denote the κ-th smallest eigenvalue of L_G.

Proof. Let $\Upsilon \in \mathbb{R}^{\kappa\times|Q|}$ denote a matrix whose x-th column is μ_i if $x \in S_i$. Note that

$$\Upsilon\Upsilon^T = \sum_{i=1}^{m} |Q\cap S_i|\,(\mu_i)\,(\mu_i)^T.$$

We define $z, \widetilde{z} \in \mathbb{R}^{|Q|}$ as follows: $\widetilde{z} := \Upsilon^T\alpha$, and for any $x \in Q$, $z(x) := \langle f_x^\kappa, \alpha\rangle$. Therefore we have

$$\begin{aligned}
\left|\alpha^T\left(\Upsilon\Upsilon^T - \sum_{x\in Q}(f_x^\kappa)(f_x^\kappa)^T\right)\alpha\right| &= \left|\sum_{x\in Q}\widetilde{z}(x)^2 - z(x)^2\right| && \text{From definition of } z(x) \text{ and } \widetilde{z}(x)\\
&\leq \sum_{x\in Q} |(z(x)-\widetilde{z}(x))\,(z(x)+\widetilde{z}(x))|\\
&\leq \sqrt{\sum_{x\in Q}(z(x)-\widetilde{z}(x))^2 \sum_{x\in Q}(\widetilde{z}(x)+z(x))^2} && \text{By Cauchy-Schwarz inequality}
\end{aligned} \tag{153}$$

Note that for any $x \in Q$, we have $z(x) = \langle f_x^\kappa, \alpha\rangle$ and $\widetilde{z}(x) = \langle \mu_x, \alpha\rangle$. Therefore by Lemma 5 we have

$$\sqrt{\sum_{x\in Q}(z(x)-\widetilde{z}(x))^2} \leq \sqrt{\sum_{x\in V}\langle f_x^\kappa - \mu_x^\kappa, \alpha\rangle^2} \leq \sqrt{\frac{\lambda_\kappa}{\min_{i\in m}\chi_2(G[S_i])}} \tag{154}$$

To complete the proof it suffices to show that $\sum_{x\in Q}(\tilde{z}(x)+z(x))^2 \leq 2$. Note that

$$\begin{aligned}
\sum_{x\in Q} \tilde{z}(x)^2 &\leq \sum_{x\in V} \langle \alpha, \mu_x\rangle^2 \\
&= \sum_{i=1}^{m} |S_i| \left\langle \alpha, \frac{\sum_{x\in S_i} f_x^\kappa}{|S_i|} \right\rangle^2 \\
&= \sum_{i=1}^{m} |S_i| \left(\frac{\sum_{x\in S_i} \langle \alpha, f_x^\kappa\rangle}{|S_i|} \right)^2 \\
&\leq \sum_{i=1}^{m} \sum_{x\in S_i} \langle \alpha, f_x^\kappa\rangle^2 && \text{By Jensen's inequality} \\
&= ||U_{[\kappa]}\alpha||_2^2 \\
&= 1
\end{aligned}$$

Thus we have

$$\sum_{x\in Q} (\tilde{z}(x)+z(x))^2 \leq \sum_{x\in Q} \left(2\cdot \tilde{z}(x)^2 + 2\cdot z(x)^2\right) \leq 2 + 2\sum_{x\in Q} \langle \alpha, f_x^\kappa\rangle^2 \leq 4 \tag{155}$$

In the first inequality we used the fact that $(\tilde{z}(x)-z(x))^2 \geq 0$ and for the second inequality we used the fact that $\sum_{x\in Q} \langle \alpha, f_x^\kappa\rangle^2 \leq \sum_{x\in V} \langle \alpha, f_x^\kappa\rangle^2 = ||U_{[\kappa]}\alpha||_2^2 = 1$. Putting (155), (154), and (153) together we get

$$\left| \alpha^T \left(\sum_{i=1}^{m} |Q\cap S_i| \left(\mu_i\right)\left(\mu_i\right)^T - \sum_{x\in Q} (f_x^\kappa)(f_x^\kappa)^T \right) \alpha \right| \leq 2\cdot \sqrt{\frac{\lambda_\kappa}{\min_{i\in m} \chi_2(G[S_i])}}$$

□

C Dot Product Oracle of κ-dimensional Spectral Embeddings

The main result of this section is Theorem 8. This is a variant of Theorem 2 in [GKL$^+$21] and asserts the following: In a (k,γ)-hierarchically-clusterable graph, for every level h, and every pair of vertices $x,y\in V$, one can estimate $\left\langle f_x^\kappa, f_y^\kappa \right\rangle$ in time $\approx n^{1/2+O(\gamma/\varphi)}$, where $\kappa = |\mathcal{P}^h|$. Note that whereas the result in [GKL$^+$21] requires the clusters to have sizes within constant factor of each other, here the clusters in level h can have sizes which are within a $O(\frac{1}{\beta^h}) \approx k^{O(1)}$ factor of each other. However, the result continues to hold at the expense of extra $k^{O(1)}$ factors in the running time.

Moreover, the proof of Theorem 2 in [GKL$^+$21] assumes that the input instance admits a k-clustering where each cluster has outer conductance at most ϵ and relies on the fact that $||f_x^k||_2^2 \leq \frac{k^{O(1)}\cdot n^{O(\gamma/\varphi)}}{n}$. In a (k,γ)-hierarchically-clusterable graph, for every level h we have $||f_x^\kappa||_2^2 \leq ||f_x^k||_2^2$, where $\kappa = |\mathcal{P}^h|$, and $k = |\mathcal{P}^H|$. Moreover as shown in Lemma 36, we have $||f_x^\kappa||_2^2 \leq \frac{k^{O(1)}\cdot n^{O(\gamma/\varphi)}}{n}$, hence, the proof of Theorem 2 in [GKL$^+$21] can be used for our setting as well (and we do **not** require to assume $\varphi_{h-1}/\varphi_h^2 \ll 1$).

Theorem 8. *[Spectral Dot Product Oracle[GKL$^+$21]] Let $G=(V,E)$ be a (k,γ)-hierarchically-clusterable graph (Definition 6). Let $\xi \in (\frac{1}{n^5},1)$ and let $\kappa = |\mathcal{P}^h|$ denote the number of clusters at level h. Then* InitializeDotProductOracle$(G,1/2,\xi,h,\kappa)$ *(Algorithm 9) computes in time $n^{1/2+O(\gamma/\varphi)}\cdot\left(\frac{k\cdot\log n}{\xi\cdot\gamma\cdot\varphi_h}\right)^{O(1)}$ a sublinear space data structure $\mathcal{D}_h$ of size $n^{1/2+O(\gamma/\varphi)}\cdot\left(\frac{k\cdot\log n}{\gamma\cdot\xi}\right)^{O(1)}$ such that with probability at least $1-n^{-100}$ the following property is satisfied:*

For every pair of vertices $x, y \in V$, $\textsc{SpectralDotProduct}(G, x, y, 1/2, \xi, \mathcal{D}_h)$ *(Algorithm 10) computes an output value $\left\langle f_x^\kappa, f_y^\kappa \right\rangle_{apx}$ such that with probability at least $1 - n^{-100}$*

$$\left| \left\langle f_x^\kappa, f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, f_y^\kappa \right\rangle \right| \leq \frac{\xi}{n}.$$

The running time of $\textsc{SpectralDotProduct}(G, x, y, 1/2, \xi, \mathcal{D}_h)$ *is $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\xi \cdot \gamma \cdot \varphi_h}\right)^{O(1)}$ and the space used by this procedure is $\left(\frac{k \log n}{\xi \gamma \varphi_h}\right)^{O(1)} \cdot n^{O(\gamma/\varphi)}$.*

Furthermore, for any $0 \leq \omega \leq 1/2$, one can obtain the following trade-offs between preprocessing time and query time: Algorithm $\textsc{SpectralDotProduct}(G, x, y, \omega, \xi, \mathcal{D}_h)$ *requires $n^{\omega+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\xi \cdot \gamma \cdot \varphi_h}\right)^{O(1)}$ per query when the preprocessing time of Algorithm* $\textsc{InitializeDotProductOracle}(G$ *is increased to $n^{1-\omega+O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\xi \cdot \gamma \cdot \varphi_h}\right)^{O(1)}$.*

We first set up notations and then state the algorithms below.

Let $m \leq n$ be integers. For any matrix $A \in \mathbb{R}^{n \times m}$ with singular value decomposition (SVD) $A = Y \Gamma Z^T$ we assume $Y \in \mathbb{R}^{n \times n}$ and $Z \in \mathbb{R}^{m \times n}$ are orthogonal matrices and $\Gamma \in \mathbb{R}^{n \times n}$ is a diagonal matrix of singular values. Since Y and Z are orthogonal matrices, their columns form an orthonormal basis. For any integer $q \in [m]$ we denote $Y_{[q]} \in \mathbb{R}^{n \times q}$ as the first q columns of Y and $Y_{-[q]}$ to denote the matrix of the remaining columns of Y. We also denote $Z_{[q]}^T \in \mathbb{R}^{q \times n}$ as the first q rows of Z^T and $Z_{-[q]}^T$ to denote the matrix of the remaining rows of Z. Finally we denote $\Gamma_{[q]}^T \in \mathbb{R}^{q \times q}$ as the first q rows and columns of Γ and we use $\Gamma_{-[q]}$ as the last $n - q$ rows and columns of Γ. So for any $q \in [m]$ the span of $Y_{-[q]}$ is the orthogonal complement of the span of $Y_{[q]}$, also the span of $Z_{-[q]}$ is the orthogonal complement of the span of $Z_{[q]}$. Thus we can write $A = Y_{[q]} \Gamma_{[q]} Z_{[q]}^T + Y_{-[q]} \Gamma_{-[q]} Z_{-[q]}^T$.

Algorithm 9 $\textsc{InitializeDotProductOracle}(G, \omega, \xi, h, \kappa)$

```
$t := \frac{6000 \cdot \log n}{\varphi_h \cdot \beta^3 \cdot \varphi^2}$
$R_{\text{init}} := n^{1-\omega+O(\gamma/\varphi)} \cdot \left(\frac{k}{\xi}\right)^{O(1)}$
$s := n^{O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\xi}\right)^{O(1)}$
Let $I_S$ be the multiset of $s$ indices chosen independently and uniformly at random from $\{1, \ldots, n\}$
for $i = 1$ to $O(\log n)$ do
    $\widehat{Q}_i := \textsc{EstimateTransitionMatrix}(G, I_S, R_{\text{init}}, t)$ ▷ $\widehat{Q}_i$ has at most $R_{\text{init}} \cdot s$ non-zeros
$\mathcal{G} := \textsc{EstimateCollisionProbabilities}(G, I_S, R_{\text{init}}, t)$
Let $\frac{n}{s} \cdot \mathcal{G} := \widehat{W} \widehat{\Sigma} \widehat{W}^T$ be the eigendecomposition of $\frac{n}{s} \cdot \mathcal{G}$    ▷ $\mathcal{G} \in \mathbb{R}^{s \times s}$
$\Psi := \frac{n}{s} \cdot \widehat{W}_{[\kappa]} \widehat{\Sigma}_{[\kappa]}^{-2} \widehat{W}_{[\kappa]}^T$    ▷ $\Psi \in \mathbb{R}^{s \times s}$
return $\mathcal{D}_h := \{\Psi, \widehat{Q}_1, \ldots, \widehat{Q}_{O(\log n)}\}$
```

Algorithm 10 SpectralDotProductOracle$(G, x, y, \omega, \xi, \mathcal{D}_h)$

```
$R_{\text{query}} := n^{\omega + O(\gamma/\varphi)} \cdot \left(\frac{k}{\xi}\right)^{O(1)}$
for $i = 1$ to $O(\log n)$ do
    $\widehat{m}_x^i :=$ RunRandomWalks$(G, R_{\text{query}}, t, x)$
    $\widehat{m}_y^i :=$ RunRandomWalks$(G, R_{\text{query}}, t, y)$
Let $\alpha_x$ be a vector obtained by taking the entrywise median of $(\widehat{Q}_i)^T(\widehat{m}_x^i)$ over all runs
Let $\alpha_y$ be a vector obtained by taking the entrywise median of $(\widehat{Q}_i)^T(\widehat{m}_y^i)$ over all runs
return $\left\langle f_x^\kappa, f_y^\kappa \right\rangle_{apx} := \alpha_x^T \Psi \alpha_y$
```

D Projection On Subgraph Projection Matrix

D.1 Stability Bounds under Sampling of Vertices

The main result of this subsection is Lemma 21.

Lemma 21. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $Q \subseteq V$ be a set that is δ-close to r-clusterable (Definition 21) and let $\widetilde{Q}$ be a set of size $\widetilde{s}$ that is sampled independently and uniformly at random from Q. Let $\Pi, \widetilde{\Pi} \in \mathbb{R}^{\kappa \times \kappa}$ denote the subgraph projection matrix of Q and $\widetilde{Q}$ for κ and r respectively (Definition 12). Then with probabaility at least $1 - n^{-100}$ for every $x, y \in V$ we have*

$$\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle - \left\langle f_x^\kappa, \Pi f_y^\kappa \right\rangle \right| \leq \frac{\xi}{n},$$

where, $\kappa \in [k], r \in [\kappa]$, $\delta \in [0, \frac{1}{1000})$, $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$, $\widetilde{s} \geq \frac{k^c \cdot n^{160 A_0 \cdot \gamma/\varphi}}{\xi^2}$ and $A_0, c > 1$ are large enough constants.

To prove Lemma 21 we require the matrix concentration bound, which is a generalization of Bernstein's inequality bound to matrices. Equipped with the Matrix Bernstein bound, we can show that under certain spectral conditions we can approximate a matrix AA^T by $(AS)(AS)^T$, i.e. by sampling columns of A. The idea is to write $AA^T = \sum_{i=1}^n (A\mathbb{1}_i)(A\mathbb{1}_i)^T$ as a sum over the outer products of its columns and make the sample size depend on the spectral norm of the summands (Lemma 34). To prove Lemma 34 we require the following matrix concentration bound, which is a generalization of Bernstein's inequality bound to matrices.

Lemma 33 (Matrix Bernstein [Tro12])**.** *Consider a finite sequence X_i of independent, random matrices with dimensions $d_1 \times d_2$. Assume that each random matrix satisfies $\mathbb{E}[X_i] = 0$ and $\|X_i\|_2 \leq b$ almost surely. Define $\sigma^2 = \max\{\| \sum_i \mathbb{E}[X_i X_i^T] \|_2, \| \sum_i \mathbb{E}[X_i^T X_i] \|_2\}$. Then for all $t \geq 0$,*

$$\mathbb{P}\left[\| \sum_i X_i \|_2 \geq t \right] \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + bt/3} \right).$$

Lemma 34. *Let $p, q \leq n$ and $A \in \mathbb{R}^{p \times q}$ be a matrix. Let $B = \max_{h \in \{1, \dots, q\}} \|(A\mathbb{1}_h)(A\mathbb{1}_h)^T\|_2$. Let $\xi \in (0, 1)$ and $s \geq \frac{40 \cdot q^2 B^2 \log n}{\xi^2}$. Let $I_S = \{i_1, \dots, i_s\}$ be a multiset of s indices chosen independently and uniformly at random from $\{1, \dots, q\}$. Let S be the $q \times s$ matrix whose j-th column equals $\mathbb{1}_{i_j}$. Then we have*

$$\mathbb{P}\left[\|AA^T - \frac{q}{s}(AS)(AS)^T\|_2 \geq \xi \right] \leq n^{-100}.$$

Proof. Observe that

$$A(A)^T = \sum_{\ell \in \{1,\ldots,q\}} (A\mathbb{1}_\ell)(A\mathbb{1}_\ell)^T. \tag{156}$$

and

$$\frac{q}{s}(AS)(AS)^T = \frac{q}{s} \cdot \sum_{i_j \in I_S} (A\mathbb{1}_{i_j})(A\mathbb{1}_{i_j})^T. \tag{157}$$

Let $X_j \in \mathbb{R}^{p\times p}$ be a random variable defined with value $X_j = \frac{q}{s} \cdot (A\mathbb{1}_{i_j})(A\mathbb{1}_{i_j})^T$. Thus we have

$$\mathbb{E}[X_j] = \frac{q}{s} \cdot \mathbb{E}[(A\mathbb{1}_{i_j})(A\mathbb{1}_{i_j})^T] = \frac{q}{s} \cdot \frac{1}{q} \sum_{\ell \in \{1,\ldots,n\}} (A\mathbb{1}_\ell)(A\mathbb{1}_\ell)^T = \frac{1}{s} \cdot A(A)^T \tag{158}$$

By equality (157) we have $\frac{q}{s}(AS)(AS)^T = \sum_{j=1}^s X_j$. Thus by equality (158) we get

$$\|\frac{q}{s}(AS)(AS)^T - A(A)^T\|_2 = \|\sum_{j=1}^s (X_j - \mathbb{E}[X_j])\|_2. \tag{159}$$

Let $Z_j = X_j - \mathbb{E}[X_j]$. We then have $\|Z_j\|_2 = \|X_j - \mathbb{E}[X_j]\|_2 \leq \|X_j\|_2 + \|\mathbb{E}[X_j]\|_2$ Now let $B = \max_{\ell\in\{1,\ldots,q\}} \|(A\mathbb{1}_\ell)(A\mathbb{1}_\ell)^T\|_2^2$. Furthermore, by our assumption we have

$$\|X_j\|_2 = \left\|\frac{q}{s} \cdot (A\mathbb{1}_j)(A\mathbb{1}_j)^T\right\|_2 \leq \frac{q}{s} \cdot B \tag{160}$$

By subadditivity of the spectral norm and (158) we get

$$\|\mathbb{E}[X_j]\|_2 \leq \frac{q}{s} \cdot B \tag{161}$$

Putting (160) and (161) together we get

$$\|Z_j\|_2 = \|X_j - \mathbb{E}[X_j]\|_2 \leq \|X_j\|_2 + \|\mathbb{E}[X_j]\|_2 \leq 2 \cdot \frac{q}{s} \cdot B \tag{162}$$

Now we would like to get a bound for the variance. Since Z_j is symmetric, we have $Z_j^T Z_j = Z_j Z_j^T = Z_j^2$.

$$\|\sum_{j=1}^s \mathbb{E}[Z_j^2]\|_2 \leq s \cdot \|\mathbb{E}[Z_j^2]\|_2 = s \cdot \|\mathbb{E}[X_j^2] - \mathbb{E}[X_j]^2\|_2 \leq s \cdot \|\mathbb{E}[X_j^2]\|_2 + s \cdot \|\mathbb{E}[X_j]^2\|_2$$

By submultiplicativity of the spectral norm we get

$$\|\mathbb{E}[X_j^2]\|_2 = \left\|\frac{1}{q} \cdot \frac{q^2}{s^2} \sum_{\ell\in\{1,\ldots,q\}} ((A\mathbb{1}_\ell)(A\mathbb{1}_\ell)^T)^2\right\|_2 \leq \frac{q^2}{s^2} \cdot B^2 \tag{163}$$

Moreover by submultiplicativity of spectral norm we have $\|\mathbb{E}[X_j]^2\,\|_2 \leq \|\mathbb{E}[X_j]\|_2^2 \leq \frac{q^2}{s^2} \cdot B^2$. Putting things together we obtain

$$\left\|\sum_{j=1}^s \mathbb{E}[Z_j^2]\right\|_2 \leq \frac{2 \cdot q^2 B^2}{s}$$

Now we can apply Lemma 33 and we get with $b = 2 \cdot \frac{q}{s} B$ and $\sigma^2 \leq \frac{2\cdot q^2 B^2}{s}$ using $s \geq \frac{40 \cdot q^2 B^2 \log n}{\xi^2}$

$$\mathbb{P}\left[\|\sum_{j=1}^s Z_j\|_2 > \xi\right] \leq (p+q) \cdot \exp\left(\frac{\frac{-\xi^2}{2}}{\sigma^2 + \frac{b\xi}{3}}\right) \leq n^{-100} \tag{164}$$

□

The following lemma proves a 2-norm bound on the κ-dimensional vectors f_x^κ. In turn, its proof requires the following lemma which was proved in [GKL+21].

Lemma 35. *[[GKL+21]] Let $\varphi \in (0,1)$ and $\epsilon \leq \frac{\varphi^2}{100}$, and let $G = (V, E)$ be a d-regular graph that admits (k, φ, ϵ)-clustering $C_1, \dots, C_k$. Let u be a normalized eigenvector of L with $||u||_2 = 1$ and with eigenvalue at most 2ϵ. Then we have*

$$||u||_\infty \leq n^{20\cdot\epsilon/\varphi^2} \cdot \sqrt{\frac{160}{\min_{i\in[k]} |C_i|}}.$$

Lemma 36. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Then there exists a constant A_0 such that for every $\kappa \in [k]$ and every $x \in V$ we have*

$$||f_x^\kappa||_2 \leq \frac{k^{O(1)} \cdot n^{\left(\frac{A_0\gamma}{\varphi}\right)}}{\sqrt{n}}$$

Proof. Let $C_1, \dots, C_k \in \mathcal{P}^H$ denote the clusters at level H. Note that for any cluster C at level H we have $\phi_{\text{in}}(C) \geq \varphi_H = \varphi$ and $\phi_{\text{out}}(C) \leq O(\varphi_{H-1}) = A_0 \cdot \varphi \cdot \gamma$ for some constant A_0. Therefore, G is $(k, \varphi, A_0\varphi \cdot \gamma)$-clusterable. Also note that $\frac{A_0\gamma\cdot\varphi}{\varphi^2} = \frac{A_0\gamma}{\varphi}$ is smaller than a sufficiently small constant. Thus by Lemma 35 for any $\kappa \leq k$ we have

$$||f_x^\kappa||_\infty \leq n^{20A_0\cdot\frac{\gamma}{\varphi}} \cdot \sqrt{\frac{160}{\min_{i\in[k]} |C_i|}}$$

Also note that by Proposition 1 we have $\min_{i\in[k]} |C_i| \geq n\cdot\beta^H$. By Definition 6, $H = O(\log k)$, thus we get

$$||f_x^\kappa||_2 \leq \sqrt{160} \cdot \sqrt{k} \cdot \left(\frac{1}{\beta}\right)^{(H/2)} \cdot \frac{n^{20A_0\cdot\frac{\gamma}{\varphi}}}{\sqrt{n}} \leq \frac{k^{O(1)} \cdot n^{\left(\frac{20A_0\cdot\gamma}{\varphi}\right)}}{\sqrt{n}}$$

□

Lemma 37. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $\xi \in (\frac{1}{n^5}, 1)$ and $c > 1$, $A_0 > 1$ be sufficiently large constants. Let $Q \subseteq V$ and $\widetilde{Q}$ be a set of size $\widetilde{s} \geq \frac{k^c\cdot n^{80A_0\cdot\frac{\gamma}{\varphi}}}{\xi^2}$ that is sampled independently and uniformly at random from Q. Let s be an estimation of $|Q|$ such that $|s - |Q|| \leq \frac{|Q|\cdot\xi}{k^c\cdot n^{40A_0\cdot\gamma/\varphi}}$. Let $A \in \mathbb{R}^{\kappa\times|Q|}$ and $\widetilde{A} \in \mathbb{R}^{\kappa\times|\widetilde{Q}|}$ be matrices whose columns are f_x^κ for $x \in Q$ and $x \in \widetilde{Q}$ respectively. Then with probabaility at least $1-n^{-100}$ we have:*

$$\left\| AA^T - \frac{s}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T \right\|_2 \leq \xi$$

Proof. Note that $|\widetilde{Q}| = \widetilde{s}$. Recall that $A \in \mathbb{R}^{\kappa\times|Q|}$ and $\widetilde{A} \in \mathbb{R}^{\kappa\times\widetilde{s}}$ are matrices whose columns are f_x^κ for $x \in Q$ and $x \in \widetilde{Q}$ respectively. Let $S \in \mathbb{R}^{|Q|\times\widetilde{s}}$ be a matrix whose x-th column equals $\mathbb{1}_x$ for any $x \in \widetilde{Q}$. Note that $\widetilde{A} = AS$. Let $B = \max_{x\in Q} ||f_x^\kappa||_2^2$. Note that by Lemma 36 we have $B \leq \frac{k^{O(1)}\cdot n^{40A_0\cdot\gamma/\varphi}}{n}$ holds for some large constant A_0. Also, recall that $|Q| \leq n$. Therefore, by choice of $\widetilde{s}$ for large enough c we have

$$\widetilde{s} \geq \frac{k^c \cdot n^{80A_0\cdot\gamma/\varphi}}{\xi^2} \geq \frac{40 \cdot |Q|^2 \cdot B^2}{\left(\frac{\xi}{2}\right)^2}$$

Thus by Lemma 34 with probabaility at least $1 - n^{-100}$ we have

$$\left\| AA^T - \frac{|Q|}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T \right\|_2 \leq \frac{\xi}{2} \tag{165}$$

Also note that

$$\begin{aligned}
&\left\| \frac{|Q|}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T - \frac{s}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T \right\|_2 && (166)\\
&= \left| \frac{|Q| - s}{\widetilde{s}} \right| \cdot ||(\widetilde{A})(\widetilde{A})^T||_2 \\
&\leq \left| \frac{|Q| - s}{\widetilde{s}} \right| \cdot ||\widetilde{A}||_F^2 && \text{Since } ||.||_F \leq ||.||_2 \text{ and } ||(\widetilde{A})(\widetilde{A})^T||_2 = ||\widetilde{A}||_2^2 \\
&= \left| \frac{|Q| - s}{\widetilde{s}} \right| \cdot \sum_{x \in \widetilde{Q}} ||f_x||_2^2 \\
&\leq \left| \frac{|Q| - s}{\widetilde{s}} \right| \cdot |\widetilde{Q}| \cdot \frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma/\varphi}}{n} && \text{By Lemma 36, } ||f_x||_2^2 \leq \frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma/\varphi}}{n} \\
&\leq \left(\frac{|Q| \cdot \xi}{k^c \cdot n^{40 \cdot \gamma/\varphi}} \right) \cdot \frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma/\varphi}}{n} && \text{Since } \widetilde{s} = |\widetilde{Q}| \text{ and } ||Q| - s| \leq \frac{|Q| \cdot \xi}{k^c \cdot n^{40 A_0 \cdot \gamma/\varphi}} \\
&\leq \frac{\xi}{2} && \text{Since } |Q| \leq n
\end{aligned}$$

Note that the last inequality holds since constant c is large enough to cancel the hidden constants in $\frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma/\varphi}}{n}$. Therefore by (165) and (166) with probabaility at least $1 - n^{-100}$ we have

$$\begin{aligned}
\left\| AA^T - \frac{s}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T \right\|_2 &\leq \left\| AA^T - \frac{|Q|}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T \right\|_2 + \left\| \frac{|Q|}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T - \frac{s}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T \right\|_2 && \text{By triangle inequality} \\
&\leq \frac{\xi}{2} + \frac{\xi}{2} && \text{By (165) and (166)} \\
&\leq \xi
\end{aligned}$$

□

To prove Lemma 39 we need Lemma 6 in which we will use the following result from [HJ90] (Theorem 1.3.20 on page 53).

Lemma 38 ([HJ90])**.** *Let j, m, n be integers such that $1 \leq j \leq m \leq n$. For any matrix $A \in \mathbb{R}^{m \times n}$ and any matrix $B \in \mathbb{R}^{n \times m}$, the multisets of nonzero eigenvalues of AB and BA are equal. In particular, if one of AB and BA is positive semidefinite, then $\nu_j(AB) = \nu_j(BA)$.*

Lemma 39. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $\kappa \in [k], r \in [\kappa]$, $\delta \in (0, \frac{1}{2})$, $\xi \in (\frac{1}{n^5}, 1)$ and $c > 1$, $A_0 > 1$ be large enough constants. Let $Q \subseteq V$ be a set that is δ-close to r-clusterable (Definition 21) and let $\widetilde{Q}$ be a set of size $\widetilde{s} \geq \frac{k^c \cdot n^{80 A_0 \cdot \frac{\gamma}{\varphi}}}{\xi^2}$ that is sampled independently and uniformly at random from Q. Let s be an estimation of $|Q|$ such that $|s - |Q|| \leq \frac{|Q| \cdot \xi}{k^c \cdot n^{40 A_0 \cdot \gamma/\varphi}}$. Let $\widetilde{A} \in \mathbb{R}^{\kappa \times \widetilde{s}}$ be a matrix whose columns are f_x^κ for $x \in \widetilde{Q}$ and let $\widetilde{\Upsilon} = \widetilde{A}^T \widetilde{A}$. Then with probabaility at least $1 - n^{-100}$ we have:*

1. $\nu_r \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon} \right) = \nu_r \left(\frac{s}{\widetilde{s}} \cdot \widetilde{A} \widetilde{A}^T \right) \geq 1 - (\delta + \xi)$
2. $\nu_{r+1} \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon} \right) = \nu_{r+1} \left(\frac{s}{\widetilde{s}} \cdot \widetilde{A} \widetilde{A}^T \right) \leq (\delta + \xi)$

Proof. Let $A \in \mathbb{R}^{\kappa \times |Q|}$ be a matrix whose columns are f_x^κ for $x \in Q$. Note that Q is δ-close to r-clusterable. Therefore, by Definition 21 we have

$$\nu_r(AA^T) \geq 1 - \delta \tag{167}$$

and

$$\nu_{r+1}(AA^T) \leq \delta \tag{168}$$

Also note that since $\widetilde{s} \geq \frac{k^c \cdot n^{80A_0 \cdot \frac{\gamma}{\varphi}}}{\xi^2}$ and $|s - |Q|| \leq \frac{|Q| \cdot \xi}{k^c \cdot n^{40A_0 \cdot \gamma/\varphi}}$, by Lemma 37 with probabaility at least $1 - n^{-100}$ we have

$$\left\| AA^T - \left(\frac{s}{\widetilde{s}}\right) \cdot \widetilde{A}\widetilde{A}^T \right\|_2 \leq \xi \tag{169}$$

Therefore, by Weyls inequality (Lemma 16), (167) and (169) we have

$$\nu_r\left(\frac{s}{\widetilde{s}} \cdot \widetilde{A}\widetilde{A}^T\right) \geq \nu_r\left(AA^T\right) - \left\| AA^T - \left(\frac{s}{\widetilde{s}}\right) \cdot \widetilde{A}\widetilde{A}^T \right\|_2 \geq 1 - (\delta + \xi)$$

Also, by Weyls inequality (Lemma 16),(168) and (169) we have

$$\nu_{r+1}\left(\frac{s}{\widetilde{s}} \cdot \widetilde{A}\widetilde{A}^T\right) \leq \nu_{r+1}\left(AA^T\right) + \left(AA^T\right) + \left\| AA^T - \left(\frac{s}{\widetilde{s}}\right) \cdot \widetilde{A}\widetilde{A}^T \right\|_2 \leq (\delta + \xi)$$

Recall that $\widetilde{\Upsilon} \in \mathbb{R}^{\widetilde{s} \times \widetilde{s}}$ is a matrix such that for any $z_1, z_2 \in \widetilde{Q}$ we have $\widetilde{\Upsilon}(z_1, z_2) = \left\langle f_{z_1}^\kappa, f_{z_2}^\kappa \right\rangle$ and $\widetilde{A} \in \mathbb{R}^{\kappa \times \widetilde{s}}$ is a matrix whose columns are f_x^κ for $x \in \widetilde{Q}$. Therefore, we have $\widetilde{\Upsilon} = \widetilde{A}^T\widetilde{A}$. Thus by Lemma 38 we have

$$\nu_r\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) = \nu_r\left(\frac{s}{\widetilde{s}} \cdot \widetilde{A}\widetilde{A}^T\right) \geq 1 - (\delta + \xi)$$

and

$$\nu_{r+1}\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) = \nu_{r+1}\left(\frac{s}{\widetilde{s}} \cdot \widetilde{A}\widetilde{A}^T\right) \leq (\delta + \xi)$$

□

Our main technical tool is the Davis-Kahan $\sin(\theta)$ Theorem [DK70].

Theorem 9 (Davis-Kahan *sin*(θ)-Theorem [DK70])**.** *Let $A = Y_0\Gamma_0Y_0^T + Y_1\Gamma_1Y_1^T$ and $A + E = \widetilde{Y}_0\widetilde{\Gamma}_0\widetilde{Y}_0^T + \widetilde{Y}_1\widetilde{\Gamma}_1\widetilde{Y}_1^T$ be symmetric real-valued matrices with Y_0, Y_1 and $\widetilde{Y}_0, \widetilde{Y}_1$ orthogonal. If the eigenvalues of Γ_0 are contained in an interval (a, b), and the eigenvalues of $\widetilde{\Gamma}_1$ are excluded from the interval $(a - D, b + D)$for some $D > 0$, then for any unitarily invariant norm $\|.\|$.*

$$\|\widetilde{Y}_1^T Y_0\| \leq \frac{\|\widetilde{Y}_1^T E Y_0\|}{D}.$$

Lemma 40 ([GKL$^+$21])**.** *For every symmetric matrix E and every pair of orthogonal projection matrices $P, \widetilde{P}$ one has*

$$||P \cdot E \cdot P - \widetilde{P} \cdot E \cdot \widetilde{P}||_2 \leq 2\|E\|_2 \cdot (\|P \cdot (I - \widetilde{P})\|_2 + \|\widetilde{P} \cdot (I - P)\|_2).$$

Lemma 41. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $\kappa \in [k], r \in [\kappa]$, $\delta \in (0, \frac{1}{1000})$, $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$ and $A_0 > 1, c > 1$ be large enough constants. Let $Q \subseteq V$ be a set that is δ-close to r-clusterable (Definition 21) and let $\widetilde{Q}$ be a set of size $\widetilde{s} \geq \frac{k^c \cdot n^{80A_0 \cdot \frac{\gamma}{\varphi}}}{\xi^2}$ that is sampled independently and uniformly at ranodm from Q. Let $\Pi, \widetilde{\Pi} \in \mathbb{R}^{\kappa \times \kappa}$ denote the subgraph projection matrix of Q and $\widetilde{Q}$ for κ and r respectively (Definition 12). Then with probabaility at least $1 - n^{-100}$ we have*

$$\left\| \Pi - \widetilde{\Pi} \right\|_2 \leq \xi$$

Proof. Let $s = |Q|$ and let $A \in \mathbb{R}^{\kappa\times s}$ and $\widetilde{A} \in \mathbb{R}^{\kappa\times\widetilde{s}}$ be matrices whose columns are f_x^κ for $x \in Q$ and $x \in \widetilde{Q}$ respectively. Let $B = (A)(A)^T$ and $\widetilde{B} = \frac{s}{\widetilde{s}} \cdot (\widetilde{A})(\widetilde{A})^T$. Let $B = Y\Gamma Y^T$ and $\widetilde{B} = \widetilde{Y}\widetilde{\Gamma}\widetilde{Y}^T$ denote the eigendecomposition of B and $\widetilde{B}$ respectively. Therefore by Definition 12 we have $\Pi = Y_{[r]}Y_{[r]}^T$ and $\widetilde{\Pi} = \widetilde{Y}_{[r]}\widetilde{Y}_{[r]}^T$. We define

$$\left\|\Pi - \widetilde{\Pi}\right\|_2 = \left\|Y_{[r]}Y_{[r]}^T - \widetilde{Y}_{[r]}\widetilde{Y}_{[r]}^T\right\|_2$$

We define $P = Y_{[r]}Y_{[r]}^T$, $\widetilde{P} = \widetilde{Y}_{[r]}\widetilde{Y}_{[r]}^T$ and $E = I_{\kappa\times\kappa}$. Note that since P and $\widetilde{P}$ are projection matrices we have $PP = P$ and $\widetilde{P}\widetilde{P} = \widetilde{P}$. Therefore, we have

$$P \cdot E \cdot P - \widetilde{P} \cdot E \cdot \widetilde{P} = P - \widetilde{P} = Y_{[r]}Y_{[r]}^T - \widetilde{Y}_{[r]}\widetilde{Y}_{[r]}^T \tag{170}$$

Therefore, (170) and by Lemma 40 we have

$$\begin{aligned}
\left\|\Pi - \widetilde{\Pi}\right\|_2 &= \left\|Y_{[r]}Y_{[r]}^T - \widetilde{Y}_{[r]}\widetilde{Y}_{[r]}^T\right\|_2 \\
&\le 2\|I\|_2 \cdot (\|P \cdot (I - \widetilde{P})\|_2 + \|\widetilde{P} \cdot (I - P)\|_2) && \text{By Lemma 40} \\
&= 2 \cdot \|(Y_{[r]}Y_{[r]}^T)(\widetilde{Y}_{[-r]}\widetilde{Y}_{[-r]}^T)\|_2 + 2 \cdot \|(\widetilde{Y}_{[r]}\widetilde{Y}_{[r]}^T)(Y_{[-r]}Y_{[-r]}^T)\|_2 \\
&\le 2 \cdot ||Y_{[r]}||_2 \cdot \|Y_{[r]}^T\widetilde{Y}_{[-r]}\|_2 \cdot ||\widetilde{Y}_{[-r]}^T||_2 + 2 \cdot ||\widetilde{Y}_{[r]}||_2 \cdot \|\widetilde{Y}_{[r]}^T Y_{[-r]}\|_2 ||Y_{[-r]}^T||_2 && \text{By submultiplicativity of nor} \\
&= 2 \cdot \left(\|Y_{[r]}^T\widetilde{Y}_{[-r]}\|_2 + \|\widetilde{Y}_{[r]}^T Y_{[-r]}\|_2\right) && (171)
\end{aligned}$$

where the last inequality holds since $||Y_{[r]}||_2 = ||Y_{[-r]}||_2 = ||\widetilde{Y}_{[r]}||_2 = ||\widetilde{Y}_{[-r]}||_2 = 1$. Therefore, we need to upper bound $\|Y_{[r]}^T\widetilde{Y}_{[-r]}\|_2$ and $\|\widetilde{Y}_{[r]}^T Y_{[-r]}\|_2$.

Let $\xi' = \frac{\xi}{8}$. Note that by choice of $\widetilde{s}$ for large enough c we have $\widetilde{s} \ge \frac{k^c \cdot n^{80A_0\cdot\gamma/\varphi}}{\xi^2} \ge \frac{k^{c'} \cdot n^{80A_0\cdot\gamma/\varphi}}{\xi'^2}$ where c' is the constant from Lemma 37. Thus by Lemma 37 with probability $1 - n^{-100}$ we have

$$||B - \widetilde{B}||_2 = \left\|AA^T - \left(\frac{s}{\widetilde{s}}\right) \cdot (\widetilde{A})(\widetilde{A})^T\right\|_2 \le \xi' \tag{172}$$

Bounding $||Y_{[r]}^T\widetilde{Y}_{[-r]}||_2$: Note that since Q is δ-close to r-clusterable we have $\nu_r(B) \ge 1 - \delta$. Moreover, by Lemma 39 with probability $1-n^{-100}$ we have $\nu_{r+1}(\widetilde{B}) \le (\delta+\xi')$. Since $\delta, \xi' < \frac{1}{1000}$, thus we have

$$\nu_r(B) \ge 1/2 + \nu_{r+1}(\widetilde{B}) \tag{173}$$

Thus by Davis-Kahan (Theorem 9) and (173) we have

$$\begin{aligned}
||Y_{[r]}^T\widetilde{Y}_{[-r]}||_2 &\le \frac{\left\|Y_{[r]}^T(B - \widetilde{B})\widetilde{Y}_{[-r]}\right\|_2}{1/2} \\
&\le 2 \cdot ||Y_{[r]}^T||_2 \cdot \left\|B - \widetilde{B}\right\|_2 \cdot ||\widetilde{Y}_{[-r]}||_2 && \text{By submultiplicativity of norm} \\
&\le 2 \cdot \xi' && \text{By (172), and since } ||Y_{[r]}^T||_2 = ||\widetilde{Y}_{[-r]}||_2 = 1
\end{aligned} \tag{174}$$

Bounding $||\widetilde{Y}_{[r]}^T Y_{[-r]}||_2$: Note that since Q is δ-close to r-clusterable we have $\nu_{r+1}(B) \le \delta$. Moreover, by Lemma 39 with probability $1 - n^{-100}$ we have $\nu_r(\widetilde{B}) \ge 1 - (\delta + \xi')$. Since $\delta, \xi' < \frac{1}{100}$, thus we have

$$\nu_r(\widetilde{B}) \ge \nu_r(B) + 1/2 \tag{175}$$

Thus, by Davis-Kahan (Theorem 9), (172) and (175) we have

$$\begin{aligned}
||\widetilde{Y}_{[r]}^T Y_{[-r]}||_2 &\le \frac{\left\|\widetilde{Y}_{[r]}^T (B-\widetilde{B}) Y_{[-r]}\right\|_2}{1/2} \\
&\le 2 \cdot ||\widetilde{Y}_{[r]}^T||_2 \cdot \left\|B-\widetilde{B}\right\|_2 \cdot ||Y_{[-r]}||_2 && \text{By submultiplicativity of norm} \\
&\le 2 \cdot \xi' && \text{By (172), and since } ||\widetilde{Y}_{[r]}^T||_2 = ||Y_{[-r]}||_2 = 1
\end{aligned} \tag{176}$$

Put together: By (171), (174) and (176) with probability $1 - n^{-100}$ we have

$$\left\|\Pi - \widetilde{\Pi}\right\|_2 \le 2 \cdot \|Y_{[r]}^T \widetilde{Y}_{[-r]}\|_2 + 2 \cdot \|\widetilde{Y}_{[r]}^T Y_{[-r]}\|_2 \le 8 \cdot \xi' \le \xi \tag{177}$$

□

Now we are ready to prove Lemma 21.

Lemma 21. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $Q \subseteq V$ be a set that is δ-close to r-clusterable (Definition 21) and let $\widetilde{Q}$ be a set of size $\widetilde{s}$ that is sampled independently and uniformly at random from Q. Let $\Pi, \widetilde{\Pi} \in \mathbb{R}^{\kappa \times \kappa}$ denote the subgraph projection matrix of Q and $\widetilde{Q}$ for κ and r respectively (Definition 12). Then with probabaility at least $1 - n^{-100}$ for every $x, y \in V$ we have*

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle - \langle f_x^\kappa, \Pi f_y^\kappa \rangle\right| \le \frac{\xi}{n},$$

where, $\kappa \in [k], r \in [\kappa]$, $\delta \in [0, \frac{1}{1000})$, $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$, $\widetilde{s} \ge \frac{k^c \cdot n^{160 A_0 \cdot \gamma/\varphi}}{\xi^2}$ and $A_0, c > 1$ are large enough constants.

Proof. Note by submultiplicativity of norm we have

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle - \langle f_x^\kappa, \Pi f_y^\kappa \rangle\right| = \left|(f_x^\kappa)^T (\widetilde{\Pi} - \Pi)(f_y^\kappa)\right| \le ||f_x^\kappa||_2 \cdot ||\widetilde{\Pi} - \Pi||_2 \cdot ||f_y^\kappa||_2 \tag{178}$$

Note that by Lemma 36 for any $x \in V$ we have

$$||f_x^\kappa||_2 \le \frac{k^{O(1)} \cdot n^{20 A_0 \cdot \gamma/\varphi}}{\sqrt{n}} \tag{179}$$

Let $\xi' = \frac{\xi}{k^{c'} \cdot n^{40 A_0 \cdot \gamma/\varphi}}$ where we set c' later. By choice of $\widetilde{s}$ and for large enough constant c we have $\widetilde{s} \ge \frac{k^c \cdot n^{160 A_0 \cdot \frac{\gamma}{\varphi}}}{\xi^2} \ge \frac{k^{c''} \cdot n^{80 A_0 \cdot \frac{\gamma}{\varphi}}}{\xi'^2}$ where c'' is the constant from Lemma 41. Therefore, by Lemma 41 with probability $1 - n^{-100}$ we have

$$||\widetilde{\Pi} - \Pi||_2 \le \xi' \tag{180}$$

Therefore, with probability $1 - n^{-100}$ for every $x, y \in V$ we have

$$\begin{aligned}
\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle - \langle f_x^\kappa, \Pi f_y^\kappa \rangle\right| &\le ||f_x^\kappa||_2 \cdot ||\widetilde{\Pi} - \Pi||_2 \cdot ||f_y^\kappa||_2 && \text{By (178)} \\
&\le \xi' \cdot \left(\frac{k^{O(1)} \cdot n^{20 A_0 \cdot \gamma/\varphi}}{\sqrt{n}}\right)^2 && \text{By (179) and (180)} \\
&= \frac{\xi}{k^{c'} \cdot n^{40 A_0 \cdot \gamma/\varphi}} \cdot \frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma/\varphi}}{n} && \text{By choice of } \xi' = \frac{\xi}{k^{c'} \cdot n^{40 A_0 \cdot \gamma/\varphi}} \\
&\le \frac{\xi}{n}
\end{aligned}$$

The last inequality holds by choice of c' as the constant hidden in O notation. □

D.2 Stability Bounds under Approximations by Dot Product Oracle

The main result of this subsection is Lemma 22.

Lemma 22. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $Q \subseteq V$ be a set that is δ-close to r-clusterable (Definition 21) and let $\Pi \in \mathbb{R}^{\kappa\times\kappa}$ be the subgraph projection matrix of Q for κ and r (Definition 12). Then* InitializeSubgraphProjMatrix$(G, ., \kappa, r, \widetilde{Q}, s, \xi)$ *(Algorithm 6) computes a data structure $\mathcal{D}$ such that with probability at least $1 - n^{-97}$ the following property is satisfied: With probability at least $1-n^{-97}$, for every pair of vertices $x, y \in V$,* ProjectedDotProduct$(G, x, y, \xi, \mathcal{D})$ *(Algorithm 7) computes an output value $\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx}$ such that*

$$\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle \right| \leq \frac{\xi}{n},$$

where, $\kappa \in [k], r \in [\kappa]$, $\delta \in (0, \frac{1}{1000})$, $\xi \in (\frac{1}{n^5}, \frac{1}{1000})$, and $A_0, c > 1$ are large enough constants. Also, $\widetilde{Q}$ is a set of size $\widetilde{s} \geq \frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}$ sampled independently and uniformly at random from Q, and s is an estimation of $|Q|$ such that $|s - |Q|| \leq \frac{|Q|\cdot\xi^3}{k^c \cdot n^{280 A_0\cdot\gamma/\varphi}}$.

To prove Lemma 22 we first need to prove Lemma 42 and Lemma 43.

Lemma 42. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $\kappa \in [k], r \in [\kappa]$, and $s \geq 1$. Let $\widetilde{Q} \subseteq V$ be a set of size $\widetilde{s}$ and let $\widetilde{\Pi} \in \mathbb{R}^{\kappa\times\kappa}$ be the subgraph projection matrix of $\widetilde{Q}$ with respect to κ and r (Definition 12). Let $\widetilde{\Upsilon} \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}$ be a matrix such that for any $z_1, z_2 \in \widetilde{Q}$ we have $\widetilde{\Upsilon}(z_1, z_2) = \langle f_{z_1}^\kappa, f_{z_2}^\kappa\rangle$. Let $\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) = \widetilde{Z}\widetilde{\Gamma}\widetilde{Z}^T$ be the eigendecomposition of $\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)$. Let $\Psi = \frac{s}{\widetilde{s}} \cdot \widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T$. Let $x, y \in V$ and $\mathbf{a}_x, \mathbf{a}_y \in \mathbb{R}^{\widetilde{s}}$ be vectors such that for any $z \in \widetilde{Q}$ we have $\mathbf{a}_x(z) = \langle f_x^\kappa, f_z^\kappa\rangle$ and $\mathbf{a}_y(z) = \langle f_y^\kappa, f_z^\kappa\rangle$. Then we have*

$$\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle = \mathbf{a}_x^T \Psi \mathbf{a}_y$$

Proof. Let $\widetilde{A} \in \mathbb{R}^{\kappa\times\widetilde{s}}$ be a matrix whose columns are f_z^κ for all $z \in \widetilde{Q}$. Note that $\widetilde{\Upsilon} \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}$ is a matrix such that for any $z_1, z_2 \in \widetilde{Q}$ we have $\widetilde{\Upsilon}(z_1, z_2) = \langle f_{z_1}^\kappa, f_{z_2}^\kappa\rangle$. Therefore, we have

$$\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon} = \left(\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A}\right)^T \left(\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A}\right)$$

Therefore, $\widetilde{\Upsilon}$ is a gram matrix, hence $\widetilde{\Gamma} \succcurlyeq 0$. Therefore, we denote the singular value decomposition of $\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A}$ by $\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A} = \widetilde{Y}(\widetilde{\Gamma}^{1/2})\widetilde{Z}^T$. Observe that $\widetilde{Y} \in \mathbb{R}^{\kappa\times s}, \Gamma \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}, \widetilde{Z} \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}$ and we have

$$\left(\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A}\right)^T \left(\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A}\right) = \left(\widetilde{Y}(\widetilde{\Gamma}^{1/2})\widetilde{Z}^T\right)^T \left(\widetilde{Y}(\widetilde{\Gamma}^{1/2})\widetilde{Z}^T\right) = \widetilde{Z}\widetilde{\Gamma}\widetilde{Z}^T = \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)$$

Note that $\mathbf{a}_x, \mathbf{a}_y \in \mathbb{R}^{\widetilde{s}}$ are vectors such that for any $z \in \widetilde{Q}$ we have $\mathbf{a}_x(z) = \langle f_x^\kappa, f_z^\kappa\rangle$ and $\mathbf{a}_y(z) = \langle f_y^\kappa, f_z^\kappa\rangle$. Therefore, we have $\mathbf{a}_x = \widetilde{A}^T f_x$ and $\mathbf{a}_y = \widetilde{A}^T f_y$. Thus we can write

$$\begin{aligned}
&\mathbf{a}_x^T \Psi \mathbf{a}_y \\
&= \mathbf{a}_x^T \left(\frac{s}{\widetilde{s}} \cdot \widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T\right) \mathbf{a}_y && \text{As } \Psi = \frac{s}{\widetilde{s}} \cdot \widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T \\
&= (f_x^\kappa)^T (\widetilde{A}) \left(\frac{s}{\widetilde{s}} \cdot \widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T\right) (\widetilde{A})^T (f_y^\kappa) && \text{As } \mathbf{a}_x = \widetilde{A}^T f_x \text{ and } \mathbf{a}_y = \widetilde{A}^T f_y \\
&= (f_x^\kappa)^T \left(\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A}\right) \left(\widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T\right) \left(\sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A}^T\right) (f_y^\kappa) \\
&= (f_x^\kappa)^T \left(\widetilde{Y}(\widetilde{\Gamma}^{1/2})\widetilde{Z}^T\right) \left(\widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T\right) \left(\widetilde{Z}(\widetilde{\Gamma}^{1/2})\widetilde{Y}^T\right) (f_y^\kappa) && \text{As } \sqrt{\frac{s}{\widetilde{s}}} \cdot \widetilde{A} = \widetilde{Y}(\widetilde{\Gamma}^{1/2})\widetilde{Z}^T \qquad (181)
\end{aligned}$$

We define the padded identity matrix $I_{a\times b} \in \mathbb{R}^{a\times b}$ as a matrix such that for all $i \leq \min(a, b)$, $I_{a,b}(i, i) = 1$ and the rest is zero. Note that $\widetilde{Z}^T \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}$ and $\widetilde{Z}_{[r]} \in \mathbb{R}^{s\times r}$. Also note that since $\widetilde{Z}$ is the right singular vector of $\widetilde{A}$, hence, we have $\widetilde{Z}^T\widetilde{Z} = I_{s\times s}$. Therefore, we have $\widetilde{Z}^T\widetilde{Z}_{[r]} = I_{\widetilde{s}\times r}$ and $\widetilde{Z}^T_{[r]}\widetilde{Z} = I_{r\times\widetilde{s}}$. Therefore, we get

$$
\begin{aligned}
&\mathbf{a}_x^T \Psi \mathbf{a}_y \\
&= (f_x^\kappa)^T \left(\widetilde{Y}(\widetilde{\Gamma}^{1/2})\widetilde{Z}^T\right)\left(\widetilde{Z}_{[r]}\widetilde{\Gamma}^{-1}_{[r]}\widetilde{Z}^T_{[r]}\right)\left(\widetilde{Z}(\widetilde{\Gamma}^{1/2})\widetilde{Y}^T\right)(f_y^\kappa) && \text{By (181)} \\
&= (f_x^\kappa)^T(\widetilde{Y})\left((\widetilde{\Gamma}^{1/2})(I_{\widetilde{s}\times r})(\widetilde{\Gamma}^{-1}_{[r]})(I_{r\times\widetilde{s}})(\widetilde{\Gamma}^{1/2})\right)(\widetilde{Y}^T)(f_y^\kappa) && \text{As } \widetilde{Z}^T\widetilde{Z}_{[r]} = I_{\widetilde{s}\times r} \text{ and } \widetilde{Z}^T_{[r]}\widetilde{Z} = I_{r\times\widetilde{s}} \\
&= (f_x^\kappa)^T\widetilde{Y}(I_{\widetilde{s}\times r})(I_{r\times\widetilde{s}})\widetilde{Y}^T(f_y^\kappa) && \text{As } (\widetilde{\Gamma}^{1/2})(I_{\widetilde{s}\times r})(\widetilde{\Gamma}^{-1}_{[r]})(I_{r\times\widetilde{s}})(\widetilde{\Gamma}^{1/2}) = (I_{\widetilde{s}\times r})(I_{r\times\widetilde{s}}) \\
&= (f_x^\kappa)^T\left(\widetilde{Y}_{[r]}\widetilde{Y}^T_{[r]}\right)(f_y^\kappa) && \text{As } \widetilde{Y}(I_{\widetilde{s}\times r}) = \widetilde{Y}_{[r]} \qquad (182)
\end{aligned}
$$

Recall that $\widetilde{A} \in \mathbb{R}^{\kappa\times\widetilde{s}}$ is a matrix whose columns are f_z^κ for all $z \in \widetilde{Q}$ and the singular value decomposition of $\sqrt{\frac{s}{\widetilde{s}}}\cdot\widetilde{A}$ is $\sqrt{\frac{s}{\widetilde{s}}}\cdot\widetilde{A} = \widetilde{Y}(\widetilde{\Gamma}^{1/2})\widetilde{Z}^T$. Therefore, the eigendecomposition of $\widetilde{A}\widetilde{A}^T$ is

$$\widetilde{A}\widetilde{A}^T = \widetilde{Y}\left(\frac{\widetilde{s}}{s}\cdot\widetilde{\Gamma}\right)\widetilde{Y}^T$$

Note that by Definition 12 we have

$$\widetilde{\Pi} = \widetilde{Y}_{[r]}\widetilde{Y}^T_{[r]} \tag{183}$$

Therefore, by (182) and (183) we have

$$\mathbf{a}_x^T\Psi\mathbf{a}_y = (f_x^\kappa)^T\left(\widetilde{Y}_{[r]}\widetilde{Y}^T_{[r]}\right)(f_y^\kappa) = \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa\right\rangle$$

□

To prove Lemma 22 we also need to use the following lemma from [GKL+21].

Lemma 43. *Let $A, \widehat{A} \in \mathbb{R}^{n\times n}$ be symmetric matrices with eigendecompositions $A = Z\Gamma Z^T$ and $\widehat{A} = \widehat{Z}\widehat{\Gamma}\widehat{Z}^T$. Let the eigenvalues of A be $1 \geq \gamma_1 \geq \cdots \geq \gamma_n \geq 0$. Suppose that $\|A - \widehat{A}\|_2 \leq \frac{\gamma_r}{100}$ and $\gamma_{r+1} < \gamma_r/4$. Then we have*

$$\|Z_{[r]}\Gamma^{-1}_{[r]}Z^T_{[r]} - \widehat{Z}_{[r]}\widehat{\Gamma}^{-1}_{[r]}\widehat{Z}^T_{[r]}\|_2 \leq \frac{32\left(\|A - \widehat{A}\|_2\right)^{1/3}}{\gamma_r^2}$$

Now we are ready to prove Lemma 22.

Proof. **Proof of Lemma 22.** Let $\widehat{\Upsilon} \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}$ be a matrix such that for any $z_1, z_2 \in \widetilde{Q}$ we have $\widehat{\Upsilon}(z_1, z_2) = \left\langle f_{z_1}^\kappa, f_{z_2}^\kappa\right\rangle_{apx}$. Let $\left(\frac{s}{\widetilde{s}}\cdot\widehat{\Upsilon}\right) = \widehat{Z}\widehat{\Gamma}\widehat{Z}^T$ be the eigendecomposition of $\left(\frac{s}{\widetilde{s}}\cdot\widehat{\Upsilon}\right)$. Then as per line (5) of Algorithm 6 we have

$$\widehat{\Psi} = \frac{s}{\widetilde{s}}\cdot\widehat{Z}_{[r]}\widehat{\Gamma}^{-1}_{[r]}\widehat{Z}^T_{[r]} \tag{184}$$

Let $\alpha_x, \alpha_y \in \mathbb{R}^{\widetilde{s}}$ be vectors such that for any $z \in \widetilde{Q}$ we have $\alpha_x(z) = \langle f_x^\kappa, f_z^\kappa\rangle_{apx}$ and $\alpha_y(z) = \left\langle f_y^\kappa, f_z^\kappa\right\rangle_{apx}$. Then as per line 3 of Algorithm 7 we have

$$\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa\right\rangle_{apx} = \alpha_x^T\widehat{\Psi}\alpha_y \tag{185}$$

Let $\widetilde{A} \in \mathbb{R}^{\kappa\times\widetilde{s}}$ be a matrix whose columns are f_z^κ for all $z \in \widetilde{Q}$. Let $\widetilde{\Upsilon} = \widetilde{A}^T\widetilde{A}$. Note that $\widetilde{\Upsilon} \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}$ is a matrix such that for any $z_1, z_2 \in \widetilde{Q}$ we have $\widetilde{\Upsilon}(z_1, z_2) = \langle f_{z_1}^\kappa, f_{z_2}^\kappa\rangle$. Let $\left(\frac{s}{\widetilde{s}}\cdot\widetilde{\Upsilon}\right) = \widetilde{Z}\widetilde{\Gamma}\widetilde{Z}^T$ be the eigendecomposition of $\left(\frac{s}{\widetilde{s}}\cdot\widetilde{\Upsilon}\right)$ and we define

$$\Psi = \frac{s}{\widetilde{s}} \cdot \widetilde{Z}_{[r]} \widetilde{\Gamma}_{[r]}^{-1} \widetilde{Z}_{[r]}^T \tag{186}$$

Let $\mathbf{a}_x, \mathbf{a}_y \in \mathbb{R}^{\widetilde{s}}$ be vectors such that for any $z \in \widetilde{Q}$ we have $\mathbf{a}_x(z) = \langle f_x^\kappa, f_z^\kappa \rangle$ and $\mathbf{a}_y(z) = \left\langle f_y^\kappa, f_z^\kappa \right\rangle$. Therefore, by Lemma 42 we have

$$\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle = \mathbf{a}_x^T \Psi \mathbf{a}_y \tag{187}$$

Therefore, by (185) and (187) we have

$$\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle \right| = \left| \alpha_x^T \widehat{\Psi} \alpha_y - \mathbf{a}_x^T \Psi \mathbf{a}_y \right| \tag{188}$$

In the rest of the proof we bound $\left| \alpha_x^T \widehat{\Psi} \alpha_y - \mathbf{a}_x^T \Psi \mathbf{a}_y \right|$. Let $E = \widehat{\Psi} - \Psi$, $\mathbf{e}_x = \alpha_x - \mathbf{a}_x$ and $\mathbf{e}_y = \alpha_y - \mathbf{a}_y$. Thus we have

$$\begin{aligned}
&\left| \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle \right| \\
&= \left| \alpha_x^T \widehat{\Psi} \alpha_y - \mathbf{a}_x^T \Psi \mathbf{a}_y \right| \\
&= \left| (\mathbf{a}_x + \mathbf{e}_x)^T (\Psi + E)(\mathbf{a}_y + \mathbf{e}_y) - \mathbf{a}_x^T \Psi \mathbf{a}_y \right| \\
&\leq ||\mathbf{a}_x||_2 ||E||_2 ||\mathbf{a}_y||_2 + ||\mathbf{e}_x||_2 ||E||_2 ||\mathbf{a}_y||_2 + ||\mathbf{a}_x||_2 ||E||_2 ||\mathbf{e}_y||_2 + ||\mathbf{e}_x||_2 ||E||_2 ||\mathbf{e}_y||_2 \\
&+ ||\mathbf{e}_x||_2 ||\Psi||_2 ||\mathbf{a}_y||_2 + ||\mathbf{a}_x||_2 ||\Psi||_2 ||\mathbf{e}_y||_2 + ||\mathbf{e}_x||_2 ||\Psi||_2 ||\mathbf{e}_y||_2
\end{aligned} \tag{189}$$

Thus to complete the proof we need to bound $||\mathbf{a}_x||_2$, $||\mathbf{a}_y||_2$, $||\mathbf{e}_x||_2$, $||\mathbf{e}_y||_2$, $||\Psi||_2$ and $||E||_2$.

Bounding $||\mathbf{a}_x||_2$ and $||\mathbf{a}_y||_2$: Note that for any $z \in \widetilde{Q}$ we have $\mathbf{a}_x(z) = \langle f_x^\kappa, f_z^\kappa \rangle$. Thus we have

$$\begin{aligned}
||\mathbf{a}_x||_2 &= \sqrt{\sum_{z \in \widetilde{Q}} \langle f_x^\kappa, f_z^\kappa \rangle^2} \\
&= \sqrt{|\widetilde{Q}|} \cdot ||f_x^\kappa||_2 ||f_z^\kappa||_2 && \text{By Cauchy Schwarz} \\
&\leq \sqrt{\widetilde{s}} \cdot \frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma / \varphi}}{n} && \text{By Lemma 36}
\end{aligned}$$

Thus we have

$$||\mathbf{a}_x||_2 \leq \sqrt{\widetilde{s}} \cdot \frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma / \varphi}}{n} \tag{190}$$

and

$$||\mathbf{a}_y||_2 \leq \sqrt{\widetilde{s}} \cdot \frac{k^{O(1)} \cdot n^{40 A_0 \cdot \gamma / \varphi}}{n} \tag{191}$$

Bounding $||\mathbf{e}_x||_2$ and $||\mathbf{e}_y||_2$: Note that $\mathbf{e}_x = \alpha_x - \mathbf{a}_x$. Recall that for any $z \in \widetilde{Q}$ we have $\alpha_x(z) = \langle f_x^\kappa, f_z^\kappa \rangle_{apx}$ and $\mathbf{a}_x(z) = \langle f_x^\kappa, f_z^\kappa \rangle$. Let $\xi' = \frac{\xi^3}{k^{c'} \cdot n^{240 A_0 \cdot \gamma / \varphi}}$ where we set c' later. By Theorem 8 with probabaility at least $1 - n^2 \cdot n^{-100}$ we have

$$||\mathbf{e}_x||_2 = \sqrt{\sum_{z \in \widetilde{Q}} \mathbf{e}_x(z)^2} = \sqrt{\sum_{z \in \widetilde{Q}} \left(\langle f_x^\kappa, f_z^\kappa \rangle_{apx} - \langle f_x^\kappa, f_z^\kappa \rangle \right)^2} \leq \sqrt{\sum_{z \in \widetilde{Q}} \left(\frac{\xi'}{n} \right)^2} \leq \frac{\sqrt{\widetilde{s}} \cdot \xi'}{n} \tag{192}$$

Similarly we have

$$||\mathbf{e}_y||_2 \leq \frac{\sqrt{\widetilde{s}} \cdot \xi'}{n} \tag{193}$$

Bounding $||\Psi||_2$: Note that $\Psi = \left(\frac{s}{\widetilde{s}} \cdot \widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T\right)$. Since $\widetilde{Z}_{[r]}$ is orthonormal we have $||\widetilde{Z}_{[r]}||_2 = 1$. Therefore, we get

$$||\Psi||_2 \leq \frac{s}{\widetilde{s}} \cdot ||\widetilde{Z}_{[r]}||_2 \cdot ||\widetilde{\Gamma}_{[r]}^{-1}||_2 \cdot ||\widetilde{Z}_{[r]}^T||_2 \leq \frac{s}{\widetilde{s}} \cdot \frac{1}{\nu_r(\widetilde{\Gamma})} \tag{194}$$

Then, we need to bound $\nu_r(\widetilde{\Gamma})$. Recall that $\widetilde{\Upsilon} = \widetilde{A}\widetilde{A}^T$ and the eigendecomposition of $\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)$ is $\widetilde{Z}\widetilde{\Gamma}\widetilde{Z}^T$. Therefore, we have $\nu_r(\widetilde{\Gamma}) = \nu_r\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)$. Recall that $\xi' = \frac{\xi^3}{k^{c'} \cdot n^{240A_0 \cdot \gamma/\varphi}}$ and let c'' be the constant from Lemma 39, hence, for large enough constant c we have $|s - |Q|| \leq \frac{|Q| \cdot \xi^3}{k^c \cdot n^{280A_0 \cdot \gamma/\varphi}} \leq \frac{|Q| \cdot \xi'}{k^{c''} \cdot n^{40A_0 \cdot \gamma/\varphi}}$. Also by choice of $\widetilde{s}$ we have $\widetilde{s} \geq \frac{k^c \cdot n^{560A_0 \cdot \frac{\gamma}{\varphi}}}{\xi^6} \geq \frac{k^{c''} \cdot n^{80A_0 \cdot \frac{\gamma}{\varphi}}}{\xi'^2}$. Therefore, by Lemma 39 item (1) with probabaility at least $1 - n^{-100}$ we have

$$\nu_r(\widetilde{\Gamma}) = \nu_r\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) \geq 1 - \delta - \xi' \tag{195}$$

Therefore, by (194), (195) and since $0 < \delta, \xi' < \frac{1}{1000}$ we have

$$||\Psi||_2 \leq \frac{s}{\widetilde{s}} \cdot \frac{1}{\nu_r(\widetilde{\Gamma})} \leq \frac{s}{\widetilde{s}} \cdot \frac{1}{1 - \delta - \xi'} \leq 2 \cdot \frac{s}{\widetilde{s}} \tag{196}$$

Bounding $||E||_2$: Note that

$$E = \widehat{\Psi} - \Psi = \widehat{Z}_{[r]}\widehat{\Gamma}_{[r]}^{-1}\widehat{Z}_{[r]}^T - \widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T$$

To bound $||E||_2$ we will use Lemma 43. Hence, we first verify that prerequisites of this lemma are satisfied. Recall that $\widehat{\Upsilon} \in \mathbb{R}^{\widetilde{s} \times \widetilde{s}}$ and $\widetilde{\Upsilon} \in \mathbb{R}^{\widetilde{s} \times \widetilde{s}}$ are matrices such that for any $z_1, z_2 \in \widetilde{Q}$ we have $\widehat{\Upsilon}(z_1, z_2) = \left\langle f_{z_1}^{\kappa}, f_{z_2}^{\kappa}\right\rangle_{apx}$ and $\widetilde{\Upsilon}(z_1, z_2) = \left\langle f_{z_1}^{\kappa}, f_{z_2}^{\kappa}\right\rangle$. Therefore, by Theorem 8 with probabaility at least $1 - n^2 \cdot n^{-100}$ we have

$$||\widehat{\Upsilon} - \widetilde{\Upsilon}||_2 \leq ||\widehat{\Upsilon} - \widetilde{\Upsilon}||_F = \sqrt{\sum_{z_1, z_2 \in \widetilde{Q}} \left(\left\langle f_{z_1}^{\kappa}, f_{z_2}^{\kappa}\right\rangle_{apx} - \left\langle f_{z_1}^{\kappa}, f_{z_2}^{\kappa}\right\rangle\right)^2} \leq \widetilde{s} \cdot \frac{\xi'}{n}$$

Thus since $s \leq 2 \cdot |Q| \leq 2 \cdot n$ we have

$$\left|\left|\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) - \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)\right|\right|_2 \leq \frac{s}{\widetilde{s}} \cdot \widetilde{s} \cdot \frac{\xi'}{n} \leq 2 \cdot \xi' \tag{197}$$

Recall that the eigendecomposition of $\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)$ is $\widetilde{Z}\widetilde{\Gamma}\widetilde{Z}^T$ Also recall that $\xi' = \frac{\xi^3}{k^{c'} \cdot n^{240A_0 \cdot \gamma/\varphi}}$ and c'' is the constant from Lemma 39. Thus, for large enough constant c we have $|s - |Q|| \leq \frac{|Q| \cdot \xi^3}{k^c \cdot n^{280A_0 \cdot \gamma/\varphi}} \leq \frac{|Q| \cdot \xi'}{k^{c''} \cdot n^{40A_0 \cdot \gamma/\varphi}}$ and by choice of $\widetilde{s}$ we have $\widetilde{s} \geq \frac{k^c \cdot n^{560A_0 \cdot \frac{\gamma}{\varphi}}}{\xi^6} \geq \frac{k^{c''} \cdot n^{80A_0 \cdot \frac{\gamma}{\varphi}}}{\xi'^2}$. Therefore, by Lemma (39) item 2 we have

$$\nu_{r+1}(\widetilde{\Gamma}) = \nu_{r+1}\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) \leq \delta + \xi'$$

Also by (195) we have $\nu_r(\widetilde{\Gamma}) \geq 1 - \delta - \xi'$. Since $\delta, \xi' < \frac{1}{1000}$, hence, we have

$$\nu_{r+1}(\widetilde{\Gamma}) \leq \frac{\nu_r(\widetilde{\Gamma})}{4} \tag{198}$$

Also by (195), (197) and since $\delta, \xi' < \frac{1}{1000}$ we have

$$\left|\left|\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) - \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)\right|\right|_2 \leq 2 \cdot \xi' \leq \frac{1 - \delta - \xi'}{100} \leq \frac{\nu_r(\widetilde{\Gamma})}{100} \tag{199}$$

Putting (198) and (199) together we can apply Lemma 43 to matrices $\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right)$ and $\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)$. Thus we get

$$\begin{aligned}
||E||_2 &= \|\widehat{Z}_{[r]}\widehat{\Gamma}_{[r]}^{-1}\widehat{Z}_{[r]}^T - \widetilde{Z}_{[r]}\widetilde{\Gamma}_{[r]}^{-1}\widetilde{Z}_{[r]}^T\|_2 \\
&\leq \frac{32 \cdot \left(\left|\left|\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) - \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)\right|\right|_2\right)^{1/3}}{\nu_r(\widetilde{\Gamma})^2} && \text{By Lemma 43} \\
&\leq \frac{32 \cdot (2 \cdot \xi')^{1/3}}{(1 - \delta - \xi')^2} && \text{By (195) and (197)} \\
&\leq 100 \cdot {\xi'}^{1/3} && \text{Since } \delta, \xi' < \frac{1}{1000} \qquad (200)
\end{aligned}$$

Put together: Putting upper bounds of $||\mathbf{a}_x||_2$, $||\mathbf{a}_y||_2$, $||\mathbf{e}_x||_2$, $||\mathbf{e}_y||_2$, $||\Psi||_2$ and $||E||_2$ together by (190), (191), (192), (193), (196), (200) with probabaility at least $1 - n^{-98} - n^{-100}$ we get

$$\begin{aligned}
&\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle\right| \\
&\leq ||\mathbf{a}_x||_2||E||_2||\mathbf{a}_y||_2 + ||\mathbf{e}_x||_2||E||_2||\mathbf{a}_y||_2 + ||\mathbf{a}_x||_2||E||_2||\mathbf{e}_y||_2 + ||\mathbf{e}_x||_2||E||_2||\mathbf{e}_y||_2 \\
&+ ||\mathbf{e}_x||_2||\Psi||_2||\mathbf{a}_y||_2 + ||\mathbf{a}_x||_2||\Psi||_2||\mathbf{e}_y||_2 + ||\mathbf{e}_x||_2||\Psi||_2||\mathbf{e}_y||_2 && \text{By (189)} \\
&\leq \left(\widetilde{s} \cdot \frac{k^{O(1)} \cdot n^{80A_0 \cdot \gamma/\varphi}}{n^2}\right) \cdot \left(100 \cdot {\xi'}^{1/3}\right) \\
&+ 2 \cdot \left(\sqrt{\widetilde{s}} \cdot \frac{k^{O(1)} \cdot n^{40A_0 \cdot \gamma/\varphi}}{n}\right) \cdot \left(100 \cdot {\xi'}^{1/3}\right) \cdot \left(\frac{\sqrt{\widetilde{s}} \cdot \xi'}{n}\right) + \left(100 \cdot {\xi'}^{1/3}\right) \cdot \left(\frac{\widetilde{s} \cdot \xi'^2}{n^2}\right) \\
&+ 2 \cdot \left(\sqrt{\widetilde{s}} \cdot \frac{k^{O(1)} \cdot n^{40A_0 \cdot \gamma/\varphi}}{n}\right) \cdot \left(2 \cdot \frac{s}{\widetilde{s}}\right) \cdot \left(\frac{\sqrt{\widetilde{s}} \cdot \xi'}{n}\right) + \left(\frac{\widetilde{s} \cdot \xi'^2}{n^2}\right)\left(2 \cdot \frac{s}{\widetilde{s}}\right) \\
&\leq O\left(\frac{{\xi'}^{1/3} \cdot k^{O(1)} \cdot n^{80A_0 \cdot \gamma/\varphi}}{n}\right) \\
&\leq \frac{\xi}{n}.
\end{aligned}$$

The last inequaliy holds by choice of $\xi' = \frac{\xi^3}{k^{c'} \cdot n^{240A_0 \cdot \gamma/\varphi}}$ for large enough constant c'. Therefore, with probabaility at least $1 - n^{-97}$ for any $x, y \in V$ we have

$$\left|\left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle_{apx} - \left\langle f_x^\kappa, \widetilde{\Pi} f_y^\kappa \right\rangle\right| \leq \frac{\xi}{n}.$$

□

E Counting the Number of Children

In this section we prove the correctness of Algorithm 11 that counts the number of children of a cluster. The main result of this section is Lemma 29.

Algorithm 11 $\textsc{CountChildren}(G, \kappa, \widetilde{S}^*, s)$

```
1: $\widetilde{s} = |\widetilde{S}^*|$.
2: $\widehat{\Upsilon} \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}} \leftarrow$ Gram-matrix of $\left\langle f_x^\kappa, f_y^\kappa \right\rangle_{apx}$ for $x, y \in \widetilde{S}^*$        ▷ Remark 6
3: $r \leftarrow$ the largest number such that $\nu_r\left(\frac{\boldsymbol{s}^*}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) \geq 0.9$, and $\nu_{r+1}\left(\frac{\boldsymbol{s}^*}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) \leq 0.1$
4: return $r$
```

Remark 6. *$\widehat{\Upsilon}(x,y) = \left\langle f_x^\kappa, f_y^\kappa \right\rangle_{apx}$ and for computing $\left\langle f_x^\kappa, f_y^\kappa \right\rangle_{apx}$ we use Algorithm 10 given in Appendix C.*

$$\left\langle f_x^\kappa, f_y^\kappa \right\rangle_{apx} = \textsc{SpectralDotProductOracle}(G, x, y, \omega, 0.01, \mathcal{D}_h)$$

Lemma 29. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). For some large enough constant $D_0 > 1$, let $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, $h \in [H]$, $\kappa = |\mathcal{P}^h|$, $S^* \in \mathcal{P}^{h-1}$, $r = |\text{CHILDREN}(S^*)|$ and $c > 1$ be a large enough constant. Let $\boldsymbol{S}^* \subseteq V$ be a set that is D-hierarchically-close to S^* (Definition 19). Let $\widetilde{S}^*$ be a set of size $\widetilde{s} \geq k^c \cdot n^{80A_0\cdot\gamma/\varphi}$ sampled independently and uniformly at random from S^*. Let s be an estimation of $|\boldsymbol{S}^*|$ such that $|s - |\boldsymbol{S}^*|| \leq \frac{|\boldsymbol{S}^*|}{k^c \cdot n^{40A_0\cdot\gamma/\varphi}}$. Then $\textsc{CountChildren}(G, \kappa, \widetilde{S}^*, s)$ runs in time $n^{1/2+O(\gamma/\varphi)} \cdot \left(\frac{k\cdot\log n}{\gamma\cdot\varphi_h}\right)^{O(1)}$ and with probability at least $1 - n^{-97}$ returns r.*

Proof. Let $A \in \mathbb{R}^{\kappa\times|\boldsymbol{S}^*|}$ be a matrix whose columns are f_x^κ for all $x \in \boldsymbol{S}^*$. Note that $r = \text{CHILDREN}(S^*) \leq \frac{1}{\beta}$ since for every $S \in \text{CHILDREN}(S^*)$, $|S| \geq \beta \cdot |S^*|$. Recall that $D = \frac{D_0}{\beta^4\cdot\varphi^2}$. Thus we have $r \leq D$. Let $\delta = 13 \cdot D \cdot \gamma^{1/4}$. By Definition 6, $\frac{\gamma}{\beta^{30}}$ and $\frac{\gamma}{\varphi^{20}}$ are sufficiently small. Thus we have $\delta \leq 0.01$, hence, by Lemma 17 we have

$$\nu_{r+1}\left(AA^T\right) \leq 5 \cdot D \cdot \gamma^{1/4} \leq \delta$$

and

$$\nu_r\left(AA^T\right) \geq 1 - 13 \cdot D \cdot \gamma^{1/4} = 1 - \delta$$

Therefore, by Definition 21, we have set $\boldsymbol{S}^*$ is δ-close to r-clusterable. Let $\xi = \frac{1}{100}$ and c' be the constant from Lemma 39. Let $\widetilde{A} \in \mathbb{R}^{\kappa\times\widetilde{s}}$ be a matrix whose columns are f_x^κ for $x \in \widetilde{S}^*$ and let $\widetilde{\Upsilon} = \widetilde{A}^T\widetilde{A}$. Therefore, by choice of $\boldsymbol{s}^*, \widetilde{s}$ for large enough constant c we have $\widetilde{s} \geq k^c \cdot n^{80A_0\cdot\frac{\gamma}{\varphi}} \geq \frac{k^{c'}\cdot n^{80A_0\cdot\frac{\gamma}{\varphi}}}{\xi^2}$ and $|s - |\boldsymbol{S}^*|| \leq \frac{|\boldsymbol{S}^*|}{k^c\cdot\ n^{40A_0\cdot\gamma/\varphi}} \leq \frac{|\boldsymbol{S}^*|\cdot\xi}{k^{c'}\cdot n^{40A_0\cdot\gamma/\varphi}}$. Therefore, by Lemma 39 with probabaility at least $1 - n^{-100}$ we have:

$$\nu_r\left(\frac{\boldsymbol{s}^*}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) \geq 1 - (\delta + \xi) \tag{201}$$

and

$$\nu_{r+1}\left(\frac{\boldsymbol{s}^*}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) \leq (\delta + \xi) \tag{202}$$

Note that $\widetilde{\Upsilon} = \widetilde{A}^T\widetilde{A}$ where $\widetilde{A} \in \mathbb{R}^{\kappa\times\widetilde{s}}$ is a matrix whose columns are f_x^κ for $x \in \widetilde{S}^*$. Therefore, for any $z_1, z_2 \in \widetilde{S}^*$ we have $\widetilde{\Upsilon}(z_1, z_2) = \left\langle f_{z_1}^\kappa, f_{z_2}^\kappa \right\rangle$. Also note that as per line (2) of Algorithm 11,

$\widehat{\Upsilon} \in \mathbb{R}^{\widetilde{s}\times\widetilde{s}}$ is a matrix such that for any $z_1, z_2 \in \widetilde{S}^*$ we have $\widehat{\Upsilon}(z_1, z_2) = \left\langle f^{\kappa}_{z_1}, f^{\kappa}_{z_2}\right\rangle_{apx}$. Therefore, by Therorem 8 with probabaility at least $1 - n^2 \cdot n^{-100}$ we have

$$||\widehat{\Upsilon} - \widetilde{\Upsilon}||_2 \leq ||\widehat{\Upsilon} - \widetilde{\Upsilon}||_F = \sqrt{\sum_{z_1, z_2 \in \widetilde{S}^*} \left(\left\langle f^{\kappa}_{z_1}, f^{\kappa}_{z_2}\right\rangle_{apx} - \left\langle f^{\kappa}_{z_1}, f^{\kappa}_{z_2}\right\rangle\right)^2} \leq \widetilde{s} \cdot \frac{\xi}{n}$$

Thus since $s \leq 2 \cdot |\boldsymbol{S}^*| \leq 2 \cdot n$ we have

$$\left|\left|\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) - \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)\right|\right|_2 \leq \frac{s}{\widetilde{s}} \cdot \widetilde{s} \cdot \frac{\xi}{n} \leq 2 \cdot \xi \tag{203}$$

Therefore, by Weyl's inequality (Lemma 16) and choice of $H = \frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}$ and $P = \left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) - \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)$ we have

$$\begin{aligned}
\nu_r\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) &\geq \nu_r\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) - \left|\left|\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) - \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)\right|\right|_2 \\
&\geq 1 - (\delta + \xi) - (2 \cdot \xi) && \text{By (201), (203)} \\
&\geq 0.9 && \text{As } \xi = \frac{1}{100} \text{ and } \delta \leq \frac{1}{100}
\end{aligned}$$

Also by Weyls inequality (Lemma 16) we have

$$\begin{aligned}
\nu_{r+1}\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) &\leq \nu_{r+1}\left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right) + \left|\left|\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) + \left(\frac{s}{\widetilde{s}} \cdot \widetilde{\Upsilon}\right)\right|\right|_2 \\
&\geq (\delta + \xi) + (2 \cdot \xi) && \text{By (202), (203)} \\
&\leq 0.1 && \text{As } \xi = \frac{1}{100} \text{ and } \delta \leq \frac{1}{100}
\end{aligned}$$

Therefore, with probabaility at least $1 - n^2 \cdot n^{-100} - n^{-100}$ we have $\nu_r\left(\frac{s}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) \geq 0.9$ and $\nu_{r+1}\left(\frac{\boldsymbol{s}^*}{\widetilde{s}} \cdot \widehat{\Upsilon}\right) \leq 0.1$. Thus as per line (3) of Algorithm 11, with probabaility at least $1 - n^{-97}$ it returns r.

Now, we bound the running time of CountChildren (Algorithm 11). Line 2 computes $\widetilde{s}^2$ dot products. As per Remark 6, this is done by calling SpectralDotProductOracle with $\xi = 0.01$. With this choice of ξ, for $x, y \in \widetilde{S}^*$, by Theorem 8, the time taken to compute $\left\langle f^{\kappa}_x, f^{\kappa}_y\right\rangle_{apx}$ is at most $t_{x,y} \leq n^{1/2 + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$.

So, the total time taken by line 2 is at most

$$\widetilde{s}^2 \max_{x,y \in \widetilde{S}^*} t_{x,y} \leq k^{2c} \cdot\ n^{160 A_0 \gamma/\varphi} \cdot \max_{x,y \in \widetilde{S}^*} t_{x,y} \leq n^{1/2 + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}.$$

Line 3 computes the eigenvalues of $\widehat{\Upsilon}$ which takes time at most $\widetilde{s}^3$. So, the overall running time is dominated by line 2 which is at most $n^{1/2 + O(\gamma/\varphi)} \cdot \left(\frac{k \cdot \log n}{\gamma \cdot \varphi_h}\right)^{O(1)}$

□

F Counting the Number of Clusters

Let $G = (V, E)$ be a d-regular graph. The main result of this section is Lemma 28 that shows we can count the number of clusters at every level with high probability.

Lemma 28. *Let $k \in \mathbb{N}$ and let $\gamma > 0$ be a sufficiently small constant. There exists an algorithm which on input a (k, γ)-hierarchically clusterable graph G (Definition 6) and a parameter $h \leq H$ (where the associated hierarchical clustering is denoted $\mathcal{P} = (\mathcal{P}^0, \ldots, \mathcal{P}^h)$) runs in time $(dn)^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot \log n}{\gamma}\right)^{O(1)}$ and computes a number κ where $\kappa = |\mathcal{P}^h|$ holds with probability at least $1 - n^{-100}$.*

To prove lemma 28 we first show Theorem 10. This is a modification of Theorem 1 of [CKK+18], and the goal of this section is to prove guarantees of Theorem 10 as stated below.

Definition 25 (**Graph clusterability**)**.** Let $G = (V, E)$ be a d-regular graph. We say that G is (k, χ_{in})*-clusterable* if V can be partitioned into $S_1, \ldots, S_\ell$ for some $\ell \leq k$ such that for all $i \in [\ell]$, $\chi_2(S_i) \geq \chi_{\text{in}}$ (Definition 13). Graph G is defined to be (k, φ_{out}, τ)*-unclusterable* if V contains $k+1$ pairwise disjoint subsets $S_1, \ldots, S_{k+1}$ such that for all $i \in [k+1]$, $\text{vol}(S_i) \geq \tau \cdot \frac{\text{vol}(V)}{k+1}$, and $\phi^G_{\text{out}}(S_i) \leq \varphi_{\text{out}}$.

Theorem 10. *Suppose $\frac{\varphi_{out}}{\chi_{in}} < 10^{-3}$. For every graph G, integer $k \geq 1$, and $\tau \in (0, 1)$,*

1. *If G is (k, χ_{in})-clusterable (YES case), then* PartitionTest*$(G, k, \chi_{in}, \varphi_{out}, \tau)$ accepts with probability at least $1 - n^{-100}$.*

2. *If G is (k, φ_{out}, τ)-unclusterable (NO case), then* PartitionTest*$(G, k, \chi_{in}, \varphi_{out}, \tau)$ rejects with probability at least $1 - n^{-100}$.*

The algorithm PartitionTest*$(G, k, \chi_{in}, \varphi_{out}, \tau)$ runs in time $(dn)^{1/2+O(\varphi_{out}/\chi_{in})} \cdot \left(\frac{k \cdot \log n}{\chi_{in} \cdot \tau}\right)^{O(1)}$.*

Algorithm PartitionTest calls the procedure Estimate given by Algorithm 12, compares the value returned with a threshold, and then decides whether to accept or reject.

Algorithm 12 Estimate(G, k, s, t, σ, R)

```
Sample $s$ vertices from $V$ independently and uniformly at random and let $S$ be the multiset of sampled vertices.
$r = 192 \cdot s \cdot \sqrt{d \cdot n}$.
for each sample $a \in S$ do
    if $\ell_2^2$-norm tester$(G, a, \sigma, r)$ rejects then return $\infty$.        ▷ High collision probability
for Each sample $a \in S$ do
    Run $2R$ random walks of length $t$ starting from $a$. Let $\mathbf{q}_a$ and $\mathbf{q}'_a$ be the empirical distribution of running $R$ random walks started at $a$.
Let $Q$ and $Q'$ be matrices whose columns are $\left\{\frac{\mathbf{q}_a}{\sqrt{d}} : a \in S\right\}$ and $\left\{\frac{\mathbf{q}'_a}{\sqrt{d}} : a \in S\right\}$ respectively.
Let $\mathcal{G} := \frac{1}{2} \cdot \left(Q^\top Q' + Q'^\top Q\right)$
Return $\nu_{k+1}(\mathcal{G})$.
```

Algorithm 13 PartitionTest$(G, k, \chi_{\text{in}}, \varphi_{\text{out}}, \tau)$ $\triangleright$ Need: $\frac{\varphi_{\text{out}}}{\chi_{\text{in}}} < 10^{-3}$

```
$s := \frac{1600\cdot(k+1)^2\cdot\log(12(k+1))\cdot\log(dn)}{\tau}$.
$c := \frac{10}{\chi_{\text{in}}}$
$t := c \cdot \log(n \cdot d)$
$\sigma := \frac{192\cdot s\cdot k}{n\cdot d}$.
$\Delta_{\text{thres}} := \frac{1}{2} \cdot \frac{8(k+1)\log(12(k+1))}{\tau} \cdot (n \cdot d)^{-1-120\cdot c\cdot\varphi_{\text{out}}}$.
$\Delta_{\text{err}} = \frac{1}{3} \cdot \frac{8(k+1)\log(12(k+1))}{\tau} \cdot (n \cdot d)^{-1-120\cdot c\cdot\varphi_{\text{out}}}$
$R := \max\left(\frac{100\cdot s^2\sigma^{1/2}}{\Delta_{\text{err}}}, \frac{200\cdot s^4\sigma^{3/2}}{\Delta_{\text{err}}^2}\right)$.
if Estimate$(G, k, s, t, \sigma, R) \leq \Delta_{\text{thres}}$ then
    Accept $G$.
else
    Reject $G$.
```

Recall that with the graph G we associated a random walk, and let M be the transition matrix of that random walk. For a vertex a of G, denote by $\mathbf{p}_a^t = M^t \mathbb{1}_a$ the probability distribution of of a t step random walk starting from a. For any vertex b, the fraction of the random walks ending in b is taken as an estimate of $\mathbf{p}_a^t(b) = \mathbb{1}_b^\top M^t \mathbb{1}_a$, the probability that the t-step random walk started from a ends in b. However, for this estimate to have sufficiently small variance, the quantity $\frac{\|\mathbf{p}_a^t\|_2^2}{d}$ needs to be small enough. To check this, Estimate uses the procedure **ℓ_2^2-norm tester**, whose guarantees are formally specified in the following lemma.

Lemma 44 ([CKK$^+$18])**.** *Let $G = (V, E)$. Let $a \in V$, $\sigma > 0$, $0 < \delta < 1$, and $R \geq \frac{16\sqrt{d\cdot n}}{\delta}$. Let $t \geq 1$, and $\mathbf{p}_a^t$ be the probability distribution of the endpoints of a t-step random walk starting from a. There exists an algorithm, denoted by **ℓ_2^2-norm tester**(G, a, σ, R), that outputs accept if $\frac{\|\mathbf{p}_a^t\|_2^2}{d} \leq \frac{\sigma}{4}$, and outputs reject if $\frac{\|\mathbf{p}_a^t\|_2^2}{d} > \sigma$, with probability at least $1 - \delta$. The running time of the tester is $O(R \cdot t)$.*

Definition 26. We say that a vertex $a \in V$, is (σ, t)*-good* if $\frac{\|\mathbf{p}_a^t\|_2^2}{d} \leq \sigma$.

We first claim that for all multisets S containing only (σ, t)-good vertices, with a good probability over the R random walks, the quantity $\mathcal{G}$ that Algorithm 12 returns is a good approximation to $\frac{1}{d} \cdot (M^t S)^\top (M^t S)$ in Frobenius norm.

Lemma 45 ([CKK$^+$18])**.** *Let $G = (V, E)$ be a graph. Let $0 < \sigma \leq 1$, $t > 0$, $\mu_{err} > 0$, k be an integer, and let S be a multiset of s vertices, all whose elements are (σ, t)-good. Let $R = \max\left(\frac{100\cdot s^2\cdot\sigma^{1/2}}{\mu_{err}}, \frac{200\cdot s^4\cdot\sigma^{3/2}}{\mu_{err}^2}\right)$. For each $a \in S$ and each $b \in V_G$, let $\mathbf{q}_a(b)$ and $\mathbf{q}'_a(b)$ be random variables which denote the fraction out of the R random walks starting from a, which end in b. Let Q and Q' be matrices whose columns are $(D^{-\frac{1}{2}}\mathbf{q}_a)_{a\in S}$ and $(D^{-\frac{1}{2}}\mathbf{q}'_a)_{a\in S}$ respectively. Let $\mathcal{G} = \frac{1}{2}\left(Q^\top Q' + Q'^\top Q\right)$. Then with probability at least $49/50$,*

$$\left|\nu_{k+1}(\mathcal{G}) - \nu_{k+1}\left(\frac{1}{d} \cdot (M^t S)^T (M^t S)\right)\right| \leq \mu_{err}$$

We now prove that Algorithm 13 indeed outputs a YES with good probability on a YES instance. For this, we need the following lemma.

Lemma 46. *For all $\alpha \in (0, 1)$, and all $G = (V, E)$ which is (k, χ_{in})-clusterable (Definition 25), there exists $V' \subseteq V$ with $|V'| \geq (1 - \alpha)n$ such that for any $t \geq \frac{\log n}{\chi_{in}}$, every $x \in V'$ is $\left(\frac{2k}{\alpha\cdot d\cdot n}, t\right)$-good.*

Proof. Recall that we say that vertex x is (σ,t)-*good* if $\frac{||\mathbf{p}_x^t||_2^2}{d} \le \sigma$. We have $\mathbf{p}_x^t = M^t \mathbb{1}_x$ Recall that $1-\frac{\lambda_1}{2} \ge \cdots \ge 1-\frac{\lambda_n}{2}$, are eigenvalues of M, and $u_1,\dots,u_n$ are the corresponding orthonormal eigenvectors. We write $\mathbb{1}_x$ in the eigenbasis of M as $\mathbb{1}_x = \sum_{i=1}^n \beta_i(x)\cdot u_i$ where $\beta_i(x) = \mathbb{1}_x^T u_i = u_i(x)$. Therefore we get,

$$\begin{aligned}
||\mathbf{p}_x^t||_2^2 &= ||M^t \mathbb{1}_x||_2^2 \\
&= \sum_{i=1}^n \beta_i(x)^2 \left(1-\frac{\lambda_i}{2}\right)^{2t} \\
&= \sum_{i=1}^k \beta_i(x)^2 \left(1-\frac{\lambda_i}{2}\right)^{2t} + \sum_{i=k+1}^n \beta_i(x)^2 \left(1-\frac{\lambda_i}{2}\right)^{2t} \\
&\le \sum_{i=1}^k \beta_i(x)^2 + \left(1-\frac{\lambda_{k+1}}{2}\right)^{2t} \sum_{i=k+1}^n \beta_i(x)^2 \\
&\le \sum_{i=1}^k \beta_i(x)^2 + \left(1-\frac{\chi_{\text{in}}}{4}\right)^{2t}
\end{aligned}$$

The last inequality follows by an application of Cheeger's inequality, and the fact that $\sum_{i=k+1}^n \beta_i(x)^2 \le ||u_i||_2^2 \le 1$. We now bound $h(x) := \sum_{i=1}^k \beta_i(x)^2$. Observe that

$$\sum_{x\in V} h(x) = \sum_{i=1}^k \sum_{x\in V} u_i(x)^2 = \sum_{i=1}^k ||u_i||_2^2 = \frac{k}{n}$$

Thus by Markov's inequality there exists a set $V' \subseteq V$ with $|V'| \ge (1-\alpha)|V|$ such that for any $x \in V'$, $h(x) \le \frac{1}{\alpha}\cdot\frac{k}{n}$. Thus if $t \ge \frac{\log(n)}{\chi_{\text{in}}}$ for any $x\in V'$ we have

$$||\mathbf{p}_x^t||_2^2 \le \frac{k}{\alpha\cdot n} + \left(1-\frac{\chi_{\text{in}}}{2}\right)^{2t} \le \frac{2k}{\alpha\cdot n},$$

therefore every $x \in V'$ is $\left(\frac{2k}{\alpha\cdot d\cdot n}, t\right)$-good. □

Lemma 47 ([CKK⁺18]). *Let $\varphi_{in} > 0$, integer $k \ge 1$, and $G = (V_G, E_G)$ be a (k,φ_{in})-clusterable graph (Definition 25). Let L be its normalized Laplacian matrix, and M be the transition matrix of the associated random walk. Let S be a multiset of s vertices of G. Then*

$$\nu_{k+1}\left(\frac{1}{d}\cdot(M^tS)^T(M^tS)\right) \le s\cdot\left(1-\frac{\lambda_{k+1}}{2}\right)^{2t},$$

where λ_{k+1} is the $(k+1)$-st smallest eigenvalue of L.

Theorem 11. *Let $\chi_{in} > 0$, and integer $k \ge 1$. Then for every (k,χ_{in})-clusterable graph $G=(V,E)$ (Definition 25), Algorithm 13 accepts G with probability at least $\frac{5}{6}$.*

Proof. If Algorithm 13 outputs a NO one of the following events must happen.

- E_1: Some vertex in S is not $(\frac{\sigma}{4}, t)$-good.
- E_2: All vertices in S are $(\frac{\sigma}{4}, t)$-good, but ℓ_2^2**-norm tester** fails on some vertex.
- E_3: All vertices in S are $(\frac{\sigma}{4}, t)$-good, and ℓ_2^2**-norm tester** succeeds on all vertices, but $|\nu_{k+1}(\mathcal{G}) - \nu_{k+1}\left(\frac{1}{d}\cdot(M^tS)^T(M^tS)\right)| > \Delta_{\text{err}}$.

If none of the above happen then Algorithm 12 returns $\nu_{k+1}(\mathcal{G}) \leq \nu_{k+1}\left(\frac{1}{d}\cdot (M^tS)^T(M^tS)\right) + \Delta_{\text{err}}$. Note that

$$
\begin{aligned}
&\nu_{k+1}(\mathcal{G}) \\
&\leq \nu_{k+1}\left(\frac{1}{d}\cdot (M^tS)^T(M^tS)\right) + \Delta_{\text{err}} \\
&\leq s\cdot\left(1-\frac{\lambda_{k+1}}{2}\right)^{2t} + \Delta_{\text{err}} && \text{By Lemma 47} \\
&\leq s\cdot\left(1-\frac{\chi_{in}}{2}\right)^{2t} + \Delta_{\text{err}} && \text{By Cheeger bound, } \lambda_{k+1} \geq \min_{S_i}\chi_2(S_i) \geq \chi_{\text{i}} \\
&\leq s\cdot\exp\left(-t\cdot\chi_{\text{in}}\right) + \Delta_{\text{err}} \\
&= s\cdot\exp\left(-c\cdot\log(n\cdot d)\cdot\chi_{\text{in}}\right) + \Delta_{\text{err}} && \text{By choice of } t = c\cdot\log(n\cdot d) \\
&= \frac{1600\cdot(k+1)^2\cdot\log\left(12(k+1)\right)\cdot\log(n\cdot d)}{\tau}\cdot(n\cdot d)^{-c\cdot\chi_{\text{in}}} + \Delta_{\text{err}} && \text{By choice of } s \\
&\leq \frac{1}{10}\cdot\frac{8(k+1)\log(12(k+1))}{\tau}\cdot(n\cdot d)^{2-c\cdot\chi_{\text{in}}} + \Delta_{\text{err}} && \text{As } k+1\leq n\cdot d\text{, and } 2\cdot 10^3\log(n\cdot d)\leq n\cdot \\
&\leq \frac{1}{10}\cdot\frac{8(k+1)\log(12(k+1))}{\tau}\cdot(n\cdot d)^{-1-c\cdot\varphi_{\text{out}}} + \Delta_{\text{err}} && \text{As } \frac{\varphi_{out}}{\chi_{\text{in}}} < \frac{1}{1000} \\
&\leq \Delta_{\text{thres}} && \text{By choice of } \Delta_{\text{thres}} \text{ and } \Delta_{\text{err}}
\end{aligned}
$$

Therefore, if none of the events above happen, Algorithm 13 accepts G.

Now we bound the probability of the events. Apply Lemma 46 with $\alpha = \frac{1}{24\cdot s}$. Then by the union bound, with probability at least $1-\alpha = 1-\frac{1}{24}$ all the vertices in S are $\left(\frac{48\cdot s\cdot k}{d\cdot n}, t\right)$-good, that is, $(\frac{\sigma}{4}, t)$-good, where $\sigma = \frac{192\cdot s\cdot k}{d\cdot n}$, as chosen in Algorithm 12. Thus, $\Pr[E_1] \leq \frac{1}{24}$. Given that E_1 doesn't happen, by Lemma 26, on any sample, **ℓ_2^2-norm tester** fails with probability at most $\frac{16\sqrt{d\cdot n}}{r} < \frac{1}{12s}$ for $r = 192\cdot s\cdot\sqrt{d\cdot n}$, as chosen in Algorithm 12. Thus, with probability at least $1-\frac{1}{12}$, **ℓ_2^2-norm tester** succeeds on all the sampled vertices, which implies $\Pr[E_2] \leq \frac{1}{12}$. Given that both E_1 and E_2 don't happen, by Lemma 45, with probability at least $\frac{49}{50}$, Algorithm 12 returns a value that is at most Δ_{err} away from $\nu_{k+1}(\mathcal{G}) \leq \nu_{k+1}\left(\frac{1}{d}\cdot(M^tS)^T(M^tS)\right)$. Thus, $\Pr[E_3] \leq \frac{1}{50}$. By the union bound, the probability that Algorithm 13 rejects is at most $\frac{1}{24}+\frac{1}{12}+\frac{1}{50} < \frac{1}{6}$. □

Theorem 12 ([CKK$^+$18])**.** *Let $\varphi_{out} > 0$, $\tau \in (0,1)$, and integer $k \geq 1$. Then for every (k, φ_{out}, τ)-unclusterable graph $G = (V, E)$ (Definition 25) Algorithm 13 rejects G with probability at least $\frac{4}{7}$.*

Now we are set to prove Theorem 10.

Theorem 10. *Suppose $\frac{\varphi_{out}}{\chi_{in}} < 10^{-3}$. For every graph G, integer $k \geq 1$, and $\tau \in (0,1)$,*

1. *If G is (k, χ_{in})-clusterable (YES case), then* PartitionTest*$(G, k, \chi_{in}, \varphi_{out}, \tau)$ accepts with probability at least $1-n^{-100}$.*

2. *If G is (k, φ_{out}, τ)-unclusterable (NO case), then* PartitionTest*$(G, k, \chi_{in}, \varphi_{out}, \tau)$ rejects with probability at least $1-n^{-100}$.*

The algorithm PartitionTest*$(G, k, \chi_{in}, \varphi_{out}, \tau)$ runs in time $(dn)^{1/2+O(\varphi_{out}/\chi_{in})}\cdot\left(\frac{k\cdot\log n}{\chi_{in}\cdot\tau}\right)^{O(1)}$.*

Proof. The correctness of the algorithm is guaranteed by Theorem 11 and Theorem 12. Since these theorems give correctness probability that is a constant larger than $1/2$, it can be boosted up to $2/3$ using standard techniques (majority of the answers of $O(\log n)$ independent runs). It remains to analyze the query complexity. For each of the s sampled vertices, we run **ℓ_2^2-norm**

tester once, followed by R random walks of t steps each. Each call to the ℓ_2^2**-norm tester** takes $O(rt) = O(st\sqrt{n \cdot d})$ queries, as guaranteed by Lemma 44. The random walks from each vertex take $O(Rt)$ time. Thus, the overall query complexity is $O(srt + sRt + s\sqrt{n \cdot d})$. Substituting the values of s, r, R, and t as defined in Algorithm 13, we get that its runtime is $(dn)^{1/2+O(\varphi_{\text{out}}/\chi_{\text{in}})} \cdot \left(\frac{k \cdot \log n}{\chi_{\text{in}} \cdot \tau}\right)^{O(1)}$. □

Now we are ready to prove Lemma 28.

Lemma 28. *Let $k \in \mathbb{N}$ and let $\gamma > 0$ be a sufficiently small constant. There exists an algorithm which on input a (k, γ)-hierarchically clusterable graph G (Definition 6) and a parameter $h \leq H$ (where the associated hierarchical clustering is denoted $\mathcal{P} = (\mathcal{P}^0, \ldots, \mathcal{P}^h)$) runs in time $(dn)^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot \log n}{\gamma}\right)^{O(1)}$ and computes a number κ where $\kappa = |\mathcal{P}^h|$ holds with probability at least $1 - n^{-100}$.*

Proof. Let $\mathcal{P}^h$ denote the clustering at level h in the tree representation of $(\mathcal{P}^h)_{0 \leq h \leq H}$. Therefore, by Definition 6 for any cluster $S \in \mathcal{P}^h$ we have $\phi_{\text{in}}^G(S) \geq \varphi_h$. Note that by Lemma 3 for any cluster $S \in \mathcal{P}^h$ we have $\chi_2(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \phi_{\text{in}}^G(S) \geq \frac{\beta^3 \cdot \varphi^2}{300} \cdot \varphi_h$. We define $\chi_{\text{in}} = \frac{\beta^3 \cdot \varphi^2}{300} \cdot \varphi_h$. Note that by Definition 25 we have G is $(\kappa, \chi_{\text{in}})$-clusterable.

Also note that by Definition 6 for any cluster $S \in \mathcal{P}^h$ we have $\phi_{\text{out}}^G(S) \leq O(\varphi_{h-1})$. Note that by Proposition 1 we have $\min_{S \in \mathcal{P}^h} |S| \geq n \cdot \beta^h$ thus for any $S \neq S' \in \mathcal{P}^H$ we have $\frac{|S|}{|S'|} \leq \beta^h$. We define $D = \beta^h$ and $\varphi_{\text{out}} = D_0 \cdot \varphi_{h-1}$ where D_0 is a large enough constant. Therefore, by Definition 25 G is $(\kappa - 1, \varphi_{\text{out}}, \tau)$-unclusterable. Note that

$$\begin{aligned}
\frac{\varphi_{\text{out}}}{\chi_{in}} &\leq \frac{D_0 \varphi_{h-1}}{\frac{\beta^3 \cdot \varphi^2}{300} \cdot \varphi_h} && \text{By Lemma 3 and Definition 6} \\
&= \frac{D_0 \cdot \gamma}{\frac{\beta^3 \cdot \varphi^2}{300}} && \text{By Definition 6, } \varphi_{h-1} = \gamma \cdot \varphi_h \\
&\leq \frac{1}{1000} && \text{By Definition 6, } \frac{\gamma}{\beta^{30}} \text{ and } \frac{\gamma}{\varphi^{20}} \text{ is sufficiently small}
\end{aligned}$$

Therefore, we have G is $(\kappa, \chi_{\text{in}})$-clusterable and $(\kappa - 1, \varphi_{\text{out}}, \tau)$-unclusterable. Thus by Theorem 10 with probability at least $1 - n^{-100}$ algorithm PartitionTest$(G, k, \chi_{\text{in}}, \varphi_{\text{out}}, \tau)$ accepts G for all $k \geq \kappa$ and rejects G for all $k \leq \kappa - 1$. Thus to count the number of clusters it suffices to find the smallest k such that PartitionTest$(G, k, \chi_{\text{in}}, \varphi_{\text{out}}, \tau)$ accepts G. Note that by Theorem 10 algorithm PartitionTest runs in time $(dn)^{1/2+O(\varphi_{\text{out}}/\chi_{\text{in}})} \cdot \left(\frac{k \cdot \log n}{\chi_{\text{in}} \cdot \tau}\right)^{O(1)}$ where $\tau = \beta^h$, $\chi_{in} = \frac{\varphi_h}{\frac{\beta^3 \cdot \varphi^2}{300}}$ and $\frac{\varphi_{\text{out}}}{\chi_{in}} \leq \frac{D_0 \cdot \gamma}{\frac{\beta^3 \cdot \varphi^2}{300}} \leq O_{\beta,\varphi}(\gamma)$. Therefore, the runtime of our algorithm is $(dn)^{1/2+O_{\beta,\varphi}(\gamma)} \cdot \left(\frac{k \cdot \log n}{\gamma}\right)^{O(1)}$. □

G Quality of Subsampled and Approximated Cylinders

Claim 4. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). For a large constant $D_0 > 1$, let $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, and let $\xi \leq 10^{-3}$, $h \in [H]$, $S^* \in \mathcal{P}^{h-1}$, and Q^* be a set that is D-hierarchically-close to S^* (Definition 19). Then we have $||Q^*| - |S^*|| \leq \frac{|S^*|}{2 \cdot 10^3}$.*

Proof. Note that Q^* is D-hierarchically-close to S^*. Therefore, by Definition 19, we have $|S^* \setminus Q^*| \leq D \cdot \varphi_{h^*-1} \cdot |S^*|$ and by Lemma 19 we have $|Q^* \setminus S^*| \leq 2 \cdot D \cdot \varphi_{h^*-1} \cdot |S^*|$. Therefore, we have

$$
\begin{aligned}
||Q^*| - |S^*|| &\leq 3 \cdot D \cdot \varphi_{h-1} \cdot |S^*| \\
&\leq 3 \cdot \left(\frac{D_0}{\beta^4 \cdot \varphi^2}\right) \cdot \gamma \cdot |S^*| && \text{As } D = \frac{D_0}{\beta^4 \cdot \varphi^2} \text{ and } \varphi_{h-1} \leq \varphi_{H-1} = \gamma \cdot \varphi \leq \gamma \\
&\leq \frac{|S^*|}{2 \cdot 10^3} && \text{By Definition 6, } \frac{\gamma}{\beta^{30}} \text{ and } \frac{\gamma}{\varphi^{20}} \text{ is sufficiently small}
\end{aligned}
\tag{204}
$$

□

Claim 1. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). For some large constant D_0 and let $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, $\xi \leq 10^{-3}$, $h \in [H]$, $S^* \in \mathcal{P}^{h-1}$, and Q^* be a set that is D-hierarchically-close to S^* (Definition 19). Let s be an estimation of $|Q^*|$ such that $|s - |Q^*|| \leq \frac{|Q^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$ where $A_0, c > 1$ are large enough constants. Then we have $|s - |S^*|| \leq \frac{|S^*|}{10^3}$.*

Proof. Note that $|s - |Q^*|| \leq \frac{|Q^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$. Therefore, we have

$$
\begin{aligned}
|s - |S^*|| &\leq |s - |Q^*|| + ||Q^*| - |S^*|| && \text{By triangle inequality} \\
&\leq \frac{|Q^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}} + \frac{|S^*|}{2 \cdot 10^3} && \text{By Claim 4} \\
&\leq \frac{2 \cdot |S^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}} + \frac{|S^*|}{2 \cdot 10^3} && \text{By Claim 4, } |Q^*| \leq 2 \cdot |S^*| \\
&\leq \frac{|S^*|}{10^3} && \text{As } \xi \leq \frac{1}{10^3}
\end{aligned}
$$

□

Lemma 26. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $h \in [H]$, and suppose $(\boldsymbol{P}^i)_{i=0}^{h-1}$ is a D-approximation of $(\mathcal{P}^i)_{i=0}^{h-1}$ (Definition 7). Let $\widetilde{V}$ be a set sampled independently and uniformly at ranodm from V. Then for every $\boldsymbol{S}^* \in \boldsymbol{P}^{h-1}$ with probability at least $1 - n^{-100}$ we have*

1. $|\widetilde{V} \cap \boldsymbol{S}^*| \geq \max\left(\frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}, \frac{10^7 \cdot \log n}{\beta}\right)$
2. $\left||\boldsymbol{S}^*| - \frac{n \cdot |\widetilde{V} \cap \boldsymbol{S}^*|}{|\widetilde{V}|}\right| \leq \frac{|\boldsymbol{S}^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$,

where, $A_0, c > 1$ are constants, $\xi = 10^{-3}$, $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, $D_0, c' > 1$ are large constants, and $|\widetilde{V}| \geq \frac{k^{c'} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6}$.

Proof. Let $S^* \in \mathcal{P}^{h-1}$ and $\boldsymbol{S}^* = \sigma(S^*)$ be the corresponding set in $\boldsymbol{P}^{h-1}$. For $1 \leq i \leq |\widetilde{V}|$, let X_i be a random variable which is 1 if the i-th sampled vertex is in $\boldsymbol{S}^*$, and 0 otherwise. Thus $\mathbb{E}[X_i] = \frac{|\boldsymbol{S}^*|}{n}$. Observe that $|\widetilde{V} \cap \boldsymbol{S}^*|$ is a random variable defined as $\sum_{i=1}^{|\widetilde{V}|} X_i$, where its expectation is given by

$$
\begin{aligned}
|\widetilde{V} \cap \boldsymbol{S}^*| &= |\widetilde{V}| \cdot \frac{|\boldsymbol{S}^*|}{n} \\
&\geq |\widetilde{V}| \cdot \frac{0.99 \cdot |S^*|}{n} && \text{By Claim 4} \\
&\geq |\widetilde{V}| \cdot \frac{0.99 \cdot \beta^{h-1} \cdot n}{n} && \text{By Definition 6, and as } S^* \in \mathcal{P}^{h-1} \\
&\geq \max\left(\frac{10^3 \cdot k^{2c} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6}, \frac{10^8 \cdot \log n}{\beta}\right),
\end{aligned}
$$

the last inequality holds by choice of $|\widetilde{V}| \geq \frac{k^{c'} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6} \geq \max\left(\frac{10^3 \cdot k^{2c} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6 \cdot \beta^{h-1}}, \frac{10^8 \cdot \log n}{\beta^h}\right)$, where, c' is a large enough constant such that $k^{c'} \geq \frac{10^8 \cdot k^{2c}}{\beta^H} \geq \frac{10^8 \cdot k^{2c}}{\beta^h}$. Therefore, by Chernoff bound,

$$\Pr\left[\left||\widetilde{V} \cap \boldsymbol{S}^*|\right| < \max\left(\frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}, \frac{10^7 \cdot \log n}{\beta}\right)\right] \leq \exp\left(-\frac{10^8 \cdot \log n}{2 \cdot 0.81 \cdot \beta}\right) \leq n^{-100},$$

Thus, with probability at least $1 - n^{-100}$ we have

$$|\widetilde{V} \cap \boldsymbol{S}^*| \geq \max\left(\frac{k^c \cdot n^{560 A_0 \cdot \gamma/\varphi}}{\xi^6}, \frac{10^7 \cdot \log n}{\beta}\right)$$

Also, by Chernoff bound,

$$\begin{aligned} &\Pr\left[\left||\widetilde{V} \cap \boldsymbol{S}^*| - |\widetilde{V}| \cdot \frac{|\boldsymbol{S}^*|}{n}\right| > \frac{\xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}} \cdot |\widetilde{V}| \cdot \frac{|\boldsymbol{S}^*|}{n}\right] \\ &\leq \exp\left(-\frac{1}{3} \cdot \left(\frac{\xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}\right)^2 \cdot \frac{10^3 \cdot k^{2c} \cdot n^{560 A_0 \cdot \gamma/\varphi} \cdot \log n}{\xi^6}\right) \\ &\leq n^{-100} \end{aligned}$$

Thus, with probability at least $1 - n^{-100}$ we have

$$\left||\boldsymbol{S}^*| - \frac{n \cdot |\widetilde{V} \cap \boldsymbol{S}^*|}{|\widetilde{V}|}\right| \leq \frac{|\boldsymbol{S}^*| \cdot \xi^3}{k^c \cdot n^{280 A_0 \cdot \gamma/\varphi}}$$

□

Lemma 27. *Let $G = (V, E)$ be a (k, γ)-hierarchically-clusterable graph (Definition 6). Let $D = \frac{D_0}{\beta^4 \cdot \varphi^2}$, where D_0 is a large constant. Let $h \in [H]$, $S^* \in \mathcal{P}^{h-1}$ and $\boldsymbol{S}^*$ be a set that is D-hierarchically-close to $S^* \in \mathcal{P}^{h-1}$ (Definition 19). Let $\widetilde{S}^*$ be a set of size $|\widetilde{S}^*| \geq \frac{10^7 \cdot \log n}{\beta}$ sampled independently and uniformly at ranodm from $\boldsymbol{S}^*$. Let $\mathcal{B} \subseteq \boldsymbol{S}^*$ and $\widetilde{\mathcal{B}} = \widetilde{S}^* \cap \mathcal{B}$. If $|\widetilde{\mathcal{B}}| \geq 0.9 \cdot \beta \cdot |\widetilde{S}^*|$, then with probability at least $1 - n^{-100}$ we have*

$$|\mathcal{B}| \geq 0.85 \cdot \beta \cdot |S^*|$$

Proof. Suppose $|\mathcal{B}| = \delta |\boldsymbol{S}^*|$ and for convenience write $\alpha = 0.9\beta$. We will show that with probability at least $1 - n^{-100}$, $\delta > 0.99\alpha$. We do this in two steps. First, we show that with high probability $\delta = 0.99\alpha$ cannot hold. Thereafter, we show that with high probability $\delta < 0.99\alpha$ cannot hold either. Consider case 1 above where $\delta = 0.99\alpha$. Let $X = |\widetilde{\mathcal{B}}| = |\widetilde{S}^* \cap \mathcal{B}|$. Note that $\mathbb{E}[X] = 0.99\alpha|\widetilde{S}^*|$. By a Chernoff Bound,

$$Pr(X \geq \alpha|\widetilde{S}^*|) = Pr(X - 0.99\alpha|\widetilde{S}^*| \geq 0.01\alpha|\widetilde{S}^*|) = Pr(X - \mathbb{E}[X] \geq 0.01\alpha|\widetilde{S}^*|) \leq \exp(-0.01^2 \cdot 0.99\alpha|\widetilde{S}^*|) \leq n^{-10}$$

Next, we rule out case 2 where $\delta < 0.99\alpha$. We do this via a coupling argument. Details follow. We define random variables $X_1, X_2, \ldots, X_s$ where each $X_i \sim Ber(\delta)$ is an indicator which takes on the value 1 if the i-th vertex belongs to $\widetilde{\mathcal{B}}$. We also define another (coupled) sequence of Bernoulli random variables $Y_1, Y_2, \ldots, Y_s$ where

- If $X_i = 1$, $Y_i = 1$.
- If $X_i = 0$, then $Y_i = 1$ with probability $\frac{0.99\alpha - \delta}{1 - \delta}$.

Note that each $Y_i \sim Ber(0.99\alpha)$ and that all the Y_i's are independent. Let $X = \sum X_i$ and $Y = \sum Y_i$. Further, by definition of X and Y, for each $t \in \mathbb{N}$, it holds that $Pr(X \geq t) \leq Pr(Y \geq t)$. And this holds in particular for $t = \alpha|\widetilde{S}^*|$. This means $Pr(X \geq \alpha|\widetilde{S}^*|) \leq n^{-100}$ as desired. Putting case 1 and case 2 together, this means that with probability at least $1 - n^{-100}$, it holds that

$$|\mathcal{B}| > 0.99\alpha|\boldsymbol{S}^*| \geq 0.99 \cdot 0.9\beta \cdot |\boldsymbol{S}^*|.$$

By Claim 4, this gives $|\mathcal{B}| \geq 0.85 \cdot \beta \cdot |S^*|$ as S^* is D-close to $\boldsymbol{S}^*$.

□

References

[Abb18] Emmanuel Abbe. Community detection and stochastic block models. *Found. Trends Commun. Inf. Theory*, 14(1-2):1–162, 2018.

[ACL$^+$22] Sepehr Assadi, Vaggos Chatziafratis, Jakub Lacki, Vahab Mirrokni, and Chen Wang. Hierarchical clustering in graph streams: Single-pass algorithms and space lower bounds. In *COLT*, volume 178 of *Proceedings of Machine Learning Research*, pages 4643–4702. PMLR, 2022.

[AGPT16] Reid Andersen, Shayan Oveis Gharan, Yuval Peres, and Luca Trevisan. Almost optimal local graph clustering using evolving sets. *J. ACM*, 63(2):15:1–15:31, 2016.

[AKLP22] Arpit Agarwal, Sanjeev Khanna, Huan Li, and Prathamesh Patil. Sublinear algorithms for hierarchical clustering. *CoRR*, abs/2206.07633, 2022.

[CC17] Moses Charikar and Vaggos Chatziafratis. Approximate hierarchical clustering via sparsest cut and spreading metrics. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 841–854. SIAM, 2017.

[CCN19] Moses Charikar, Vaggos Chatziafratis, and Rad Niazadeh. Hierarchical clustering better than average-linkage. In *SODA*, pages 2291–2304. SIAM, 2019.

[CKK$^+$18] Ashish Chiplunkar, Michael Kapralov, Sanjeev Khanna, Aida Mousavifar, and Yuval Peres. Testing graph clusterability: Algorithms and lower bounds. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 497–508. IEEE, 2018.

[CKM17] Vincent Cohen-Addad, Varun Kanade, and Frederik Mallmann-Trenn. Hierarchical clustering beyond the worst-case. In *NIPS*, pages 6201–6209, 2017.

[CKMM19] Vincent Cohen-Addad, Varun Kanade, Frederik Mallmann-Trenn, and Claire Mathieu. Hierarchical clustering: Objective functions and algorithms. *J. ACM*, 66(4):26:1–26:42, 2019.

[CNC18] Vaggos Chatziafratis, Rad Niazadeh, and Moses Charikar. Hierarchical clustering with structural constraints. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 773–782. PMLR, 2018.

[CPS15] Artur Czumaj, Pan Peng, and Christian Sohler. Testing cluster structure of graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 723–732, 2015.

[CS04] Artur Czumaj and Christian Sohler. Sublinear-time approximation for clustering via random sampling. In Josep Díaz, Juhani Karhumäki, Arto Lepistö, and Donald Sannella, editors, *Automata, Languages and Programming: 31st International Colloquium, ICALP 2004, Turku, Finland, July 12-16, 2004. Proceedings*. Springer, 2004.

[CS07] Artur Czumaj and Christian Sohler. Testing expansion in bounded-degree graphs. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pages 570–578. IEEE Computer Society, 2007.

[Das16] Sanjoy Dasgupta. A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 118–127, 2016.

[DK70] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[GKL$^+$21] Grzegorz Gluch, Michael Kapralov, Silvio Lattanzi, Aida Mousavifar, and Christian Sohler. Spectral clustering oracles in sublinear time. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, 2021.

[GR11] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation - In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, pages 68–75. 2011.

[HJ90] Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, 1990.

[KKLM22] Michael Kapralov, Akash Kumar, Silvio Lattanzi, and Aida Mousavifar. Learning hierarchical structure of clusterable graphs. *CoRR*, abs/2207.02581, 2022.

[KPS08] Satyen Kale, Yuval Peres, and C. Seshadhri. Noise tolerance of expanders and sublinear expander reconstruction. In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 719–728. IEEE Computer Society, 2008.

[KS08] Satyen Kale and C. Seshadhri. An expansion tester for bounded degree graphs. In *Automata, Languages and Programming, 35th International Colloquium, ICALP 2008, Reykjavik, Iceland, July 7-11, 2008, Proceedings, Part I: Tack A: Algorithms, Automata, Complexity, and Games*, pages 527–538, 2008.

[KVV04] Ravi Kannan, Santosh S. Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[LGT14] James R Lee, Shayan Oveis Gharan, and Luca Trevisan. Multiway spectral partitioning and higher-order cheeger inequalities. *Journal of the ACM (JACM)*, 61(6):37, 2014.

[MOP01] Nina Mishra, Daniel Oblinger, and Leonard Pitt. Sublinear time approximate clustering. In S. Rao Kosaraju, editor, *Proceedings of the Twelfth Annual Symposium on Discrete Algorithms, January 7-9, 2001, Washington, DC, USA*, pages 439–447, 2001.

[MS21] Bogdan-Adrian Manghiuc and He Sun. Hierarchical clustering: $o(1)$-approximation for well-clustered graphs. In *NeurIPS*, 2021.

[MW17] Benjamin Moseley and Joshua R. Wang. Approximation bounds for hierarchical clustering: Average linkage, bisecting k-means, and local search. In *NIPS*, pages 3094–3103, 2017.

[NJW02] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[NS10] Asaf Nachmias and Asaf Shapira. Testing the expansion of a graph. *Inf. Comput.*, 208(4):309–314, 2010.

[Pen20] Pan Peng. Robust clustering oracle and local reconstructor of cluster structure of graphs. In Shuchi Chawla, editor, *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, pages 2953–2972. SIAM, 2020.

[SM00] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.

[ST13] Daniel A. Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM J. Comput.*, 42(1):1–26, 2013.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.