Polylogarithmic Inapproximability*

[Extended Abstract]

Eran Halperin
International Computer Science Institute
and Computer Science Division
University of California
Berkeley, CA 94720.
eran@cs.berkeley.edu

Robert Krauthgamer
International Computer Science Institute
and Computer Science Division
University of California
Berkeley, CA 94720.
robi@cs.berkeley.edu

ABSTRACT

We provide the first hardness result of a polylogarithmic approximation ratio for a natural NP-hard optimization problem. We show that for every fixed $\epsilon > 0$, the GROUP-STEINER-TREE problem admits no efficient $\log^{2-\epsilon} k$ approximation, where k denotes the number of groups (or, alternatively, the input size), unless NP has quasi-polynomial Las-Vegas algorithms. This hardness result holds even for input graphs which are *Hierarchically Well-Separated Trees*, introduced by Bartal [FOCS, 1996]. For these trees (and also for general trees), our bound is nearly tight with the log-squared approximation currently known. Our results imply that for every fixed $\epsilon > 0$, the DIRECTED-STEINER-TREE problem admits no $\log^{2-\epsilon} n$–approximation, where n is the number of vertices in the graph, under the same complexity assumption.

Categories and Subject Descriptors

F.2.0 [**Theory of Computation**]: Analysis of algorithms and problem complexity—*General*

General Terms

Theory

Keywords

Steiner tree, approximation algorithms. hardness of approximation, integrality ratio, polylogarithmic approximation

*Work supported in part by NSF grants CCR-9820951 and CCR-0121555 and DARPA cooperative agreement F30602-00-2-0601.

STOC'03, June 9–11, 2003, San Diego, California, USA.

1. INTRODUCTION

For most NP-hard optimization problems, the known *approximation ratio*[1] falls into one of a few broad classes: (i) arbitrarily good fixed approximation ratio, (ii) specific constant factor approximation, (iii) logarithmic approximation ratio, or (iv) polynomial approximation ratio.[2] Similarly, most of the known *hardness of approximation results* (namely, results excluding a certain approximation ratio under a complexity assumption such as P $\neq$ NP) can be divided into the same categories (with class (i) replaced by strong and weak NP-hardness). In fact, for each of these classes there are (natural) problems whose "optimal" approximation ratio must lie in that class, since we know of both an approximation algorithm and a hardness of approximation result in that class.

Polylogarithmic approximation. Much less understood is the class of polylogarithmic approximation ratios. There are several (natural) problems for which the best approximation ratio currently known is polylogarithmic. However, none of these problems is known to have a hardness of approximation result that excludes the possibility of a logarithmic approximation ratio. The list of problems with such a fundamental gap in our understanding of their approximability includes GROUP-STEINER-TREE, JOB-SHOP-SCHEDULING, MIN-BISECTION, BANDWIDTH, CUTWIDTH, PATHWIDTH, and several problems in bounded-degree graphs (e.g., PLANAR-DRAWING-SIZE and CHORDAL-GRAPH-COMPLETION). We remark that for many of these problems (e.g., BANDWIDTH) no logarithmic hardness of approximation is known, and for some other problems (e.g., MIN-BISECTION, CUTWIDTH and PATHWIDTH) there is even no result excluding the possibility of an arbitrarily good fixed approximation ratio.

So far, progress in closing these gaps came only from the direction of improving the ratio achieved by approximation

[1]The approximation ratio of a polynomial-time algorithm is the worst-case ratio between the value of the solution provided by the algorithm and that of the optimum solution. As usual, the approximation ratio is measured as a function of the input size; however, in graph problems it is more convenient to characterize the input size by the number of vertices in the input graph.

[2]Examples for class (i) are KNAPSACK, BIN-PACKING and EUCLIDEAN-TSP; for class (ii) are METRIC-TSP, MAX-SAT, MAX-CUT, VERTEX-COVER and STEINER-TREE; for class (iii) are SET-COVER and DOMATIC-NUMBER; for class (iv) are MAX-CLIQUE and CHROMATIC-NUMBER.

algorithms. Seymour [22] (essentially) improved the approximation ratio for feedback sets in directed graphs from $O(\log^2 n)$ to $O(\log n \log\log n)$. His technique was extended in [9] and led to a framework that encompasses several other problems in [8]. The approximation ratio was further improved in [18] to $O(\log n)$ for three of these problems (including Total-Linear-Arrangement and Interval-Graph-Completion). These improved approximation algorithms raise the intriguing and basic question regarding the existence of an intermediate hardness (see also [16]). Put differently: *Does every problem with polylogarithmic approximation ratio actually have a logarithmic approximation ratio?* We provide a (negative) answer by showing the first polylogarithmic hardness of approximation result.

The* Group-Steiner-Tree *Problem. The (undirected) Group-Steiner-Tree problem is the following. Given an undirected graph $G = (V, E)$, a collection of subsets (called *groups*) $g_1, g_2, \ldots, g_k \subseteq V$, and a weight $w_e \geq 0$ for each edge $e \in E$, the problem is to construct a minimum-weight tree in G that spans at least one vertex from every group g_i. We can assume without loss of generality that there is a distinguished vertex $r \in V$ (called the root) that must be included in the output tree. The case where $|g_i| = 1$ for all i is just the classical Steiner Tree problem; the case where G is a tree (or even a star) can be used to model the Set-Cover problem, which is known to have a logarithmic (in k) hardness of approximation, see [17, 11, 20].

The first polylogarithmic approximation algorithm for Group-Steiner-Tree was achieved in the elegant work of [13]. First, they show that for an input graph G which is a *tree*, solving a flow-based linear programming relaxation and rounding its fractional solution by a novel randomized rounding approach that they develop, yields an $O(\log k \log N)$–approximation, where $N = \max_i |g_i|$. For a general graph G, they show how to apply the powerful results of [3] to appropriately reduce the problem to the case where G is a tree, with an $O(\log n \log\log n)$ factor loss in the approximation ratio, where $n = |V|$. Thus, [13] achieve an $O(\log n \log\log n \cdot \log k \log N)$–approximation for Group-Steiner-Tree in general graphs.

It is worth noting that by applying the results of Bartal [3] on a graph G one actually obtains a *Hierarchically Well-Separated Tree (HST)*, i.e., a tree in which (i) all leaves are at the same distance from the root, and (ii) the weight of every edge is exactly $1/\tau$ times the weight of its parent edge, where $\tau > 1$ is any desired constant.

The work of [13] has been extended and expanded in several ways: Their algorithm was derandomized in [6, 23]; an alternative (combinatorial) algorithm is devised in [7]; the loss incurred by the reduction to an HST is slightly improved in [4] (at the cost of increasing in the groups' sizes), and an optimal such reduction (namely, with only $O(\log n)$ loss) was recently announced in [10]; for HSTs, [14] show that the fractional solution to the relaxation of [13] can be rounded so as to achieve an $O(\log^2 k)$–approximation, regardless of N and n. In contrast, it is shown in [15] that, even in HSTs, the integrality ratio of this relaxation is $\Omega(\log^2 k) = \Omega(\log k \log N / \log\log N)$. A hardness result in [21] excludes approximation ratio that is better than $O((\log\log n)^{1/6})$ for Group-Steiner-Tree in the Euclidean plane.

The* Directed-Steiner-Tree *Problem. This is the directed version of the (undirected) Steiner-Tree problem. Given an edge-weighted directed graph that specifies a *root* vertex r and k *terminal* nodes $v_1, v_2, \ldots, v_k$, the goal is to construct in G a minimum-weight out-branching tree rooted at r, which spans all the terminals v_i. This problem is easily seen to generalize the undirected Group-Steiner-Tree problem, as well as to be equivalent to the directed Group-Steiner-Tree problem. The polynomial-time approximation ratio currently known for this problem is k^ϵ, for any constant $\epsilon > 0$, due to [5]; their algorithm extends to a polylogarithmic approximation ratio, namely $O(\log^3 n)$, in quasi-polynomial running time. It is shown in [24] that the flow-based linear programming relaxation for this problem has an integrality ratio of $\Omega(\sqrt{k})$, for $k = \Theta(\frac{\log^2 n}{(\log\log n)^2})$. In [15], the integrality ratio of this relaxation is shown to be $\Omega(\frac{\log^2 n}{(\log\log n)^2})$. Directed-Steiner-Tree is known to have an $\Omega(\log n)$ hardness of approximation result, since this problem generalizes the Set-Cover problem.

Our results. We provide the first polylogarithmic hardness of approximation result for a natural problem. Specifically, our main contribution is such a hardness result for Group-Steiner-Tree, as stated in the next theorem. Let $\mathsf{ZTIME}(t)$ denote the class of languages that have a probabilistic algorithm that runs in expected time t (with zero error probability).

Theorem 1.1. *For every fixed $\epsilon > 0$,* Group-Steiner-Tree *cannot be approximated within ratio $\log^{2-\epsilon} k$, unless $\mathsf{NP} \subseteq \mathsf{ZTIME}(n^{\mathrm{polylog(n)}})$; this holds even for HSTs (and in particular, trees).*

We mention several extensions of our results, whose details are omitted from this version of the paper. First, stronger complexity assumptions yield stronger lower bounds on the hardness of approximating Group-Steiner-Tree. For example, our techniques imply an $\Omega(\frac{\log^2 k}{(\log\log k)^2})$ lower bound for approximating Group-Steiner-Tree, assuming that $\mathsf{NP} \not\subseteq \mathsf{ZTIME}(2^{n^\delta})$ for some fixed $\delta > 0$. Second, we can show hardness of approximation for the maximum-coverage variant of Group-Steiner-Tree, where the input contains also a weight (i.e., cost) bound C, and we wish to find a tree of total weight at most C that covers the maximum number of groups. Theorem 1.1 implies, by a simple argument, that for any fixed $\epsilon > 0$, the maximum-coverage version cannot be approximated within ratio $\log^{1-\epsilon} k$, unless $\mathsf{NP} \subseteq \mathsf{ZTIME}(n^{\mathrm{polylog(n)}})$. In our reduction $\log N = \log^{1-\Theta(\epsilon)} k$, so our lower bound above translates to $\log^{1-\epsilon} N$. We note that an $O(\log N)$–approximation for this problem, where $N = \max_i |g_i|$, is essentially devised in [13]. Furthermore, using the same techniques, our reduction shows (directly) that under the weaker assumption that $\mathrm{P} \neq \mathsf{NP}$, for every constant $c > 0$, there is no c–approximation for this problem.

Our proof of Theorem 1.1 immediately implies the following polylogarithmic hardness of approximation result for Directed-Steiner-Tree. This implication follows by standard arguments and is omitted from this version of the paper.

THEOREM 1.2. *For every fixed $\epsilon > 0$,* DIRECTED-STEINER-TREE *cannot be approximated within ratio $\Omega(\log^{2-\epsilon} n)$, unless $\mathsf{NP} \subseteq \mathsf{ZTIME}(n^{\mathrm{polylog(n)}})$.*

Techniques. The proof of Theorem 1.1 consists of the following five ingredients. First, we use (a variant of) the SET-COVER reduction of Lund and Yannakakis [17] to create GROUP-STEINER-TREE instance on a star (i.e., tree of height 1). Then we compose many copies of this reduction in a certain recursive manner to obtain GROUP-STEINER-TREE instance on a tree of larger height. (Clearly, some technical problems arise in this part). Third, our analysis of the reduction uses the polylogarithmic integrality ratio construction of Halperin et al. [15] for GROUP-STEINER-TREE. This is one of the scarce hardness proofs which exploit in a non trivial way the construction and the analysis of the integrality ratio for the same problem; see also Section 5. Another ingredient of our analysis is the known approximation algorithm of Garg, Konjevod and Ravi [13] for GROUP-STEINER-TREE on trees. This somewhat surprising feature is interesting conceptually, and we are not aware of a hardness result whose proof uses the approximation algorithm for the problem. The last (but not least) ingredient is a set of tools designed to combine all these ingredients. For example, we devise an averaging lemma in order to show that the integrality ratio instance 'hides' inside the recursive composition instance.

One complication arising in this recursive composition approach is that our reduction contains many copies of the SET-COVER reduction (namely, every non-leaf vertex together with its children forms a copy of the SET-COVER instance of [17]). However, these copies are not really instances of SET-COVER any more, since in the context of GROUP-STEINER-TREE different copies "interact" with each other. For example, copies located at the same level of the tree may "collaborate" so that each copy covers only some of the groups. Furthermore, copies located at different levels in the tree do not have equal importance, since they have different weights and different coverage magnitudes. We deal with these issues by a suitable averaging over the copies.

Organization. We first describe in Section 2 the random tree instance from [15] and a certain extension of it that will be used in the reduction. The reduction itself is presented in Section 3, and its analysis, which proves Theorem 1.1, is given in Section 4.

2. A SPECIAL RANDOM TREE INSTANCE

In this section we describe a family of random instances (of GROUP-STEINER-TREE on HSTs) that is later used in our reduction's analysis (Section 4). This family of instances extends the one studied in [15]. In fact, we rely on their main technical lemmas, which are stated below.

Preliminaries. Throughout the paper, a tree is said to be of *arity* d if every non-leaf vertex in the tree has d children. A rooted tree has *height* H if all its leaves are at distance H from the root. As usual, the *level of a vertex* is its distance from the root; the root itself is at level 0, and there are $H+1$ levels. The *level of an edge* is h iff it connects a vertex at level $h-1$ to a vertex at level h.

The random tree of [15]. Consider the following *random tree instance* of GROUP-STEINER-TREE. Let T be a complete d-ary tree of height H, where an edge at level $1 \le h \le H$ has weight $1/2^h$. Let $\mathcal{G} = \{g_1, \ldots, g_k\}$ be a collection of k groups, where each group g_j is a subset of leaves chosen randomly as follows. We shall associate with each leaf ℓ a subset $A(\ell) \subseteq \mathcal{G}$ of the groups; we then let each group be the set of leaves ℓ for which $A(\ell)$ contains the group, i.e., $g_j = \{\ell : \ell \text{ is a leaf and } g_j \in A(\ell)\}$. (Thus, a path from the root to a leaf ℓ covers all groups in $A(\ell)$.) To define $A(\cdot)$ for the leaves, we now recursively and randomly define a set $A(v)$ for every node v in the tree. Start with the root r by letting $A(r) = \mathcal{G}$, i.e. $A(r)$ contains every group g_j with probability 1. In general, if $g_j \in A(u)$ for some non-leaf vertex u, then for each child v of u put g_j in $A(v)$ independently with probability $1/2$. Notice that this random process goes top-down in the tree and constructs the groups g_j independently of each other. The groups in $A(v)$ will be referred to as the *groups passing through the vertex* v. For the purpose of defining $A(v)$ we can identify each non-root vertex v with the edge e_v connecting it to its parent, and then $A(v)$ can be referred to as the set of groups passing through the edge e_v.

Recall that a solution to this instance is a *subtree* of T, i.e., a subgraph of T which is a tree. For a subtree S of T, let $p(S)$ denote the probability that S covers all the groups in $\mathcal{G}$, taken over the randomness in constructing the groups $\mathcal{G}$. (Here S is fixed prior to the random construction of the groups.) Since the groups are constructed independently of each other, letting $p'(S)$ denote the probability (over the randomness in constructing the groups $\mathcal{G}$) that S does not cover a particular group in $\mathcal{G}$, we have that $p(S) = (1 - p'(S))^k \le \exp\{-k \cdot p'(S)\}$.

LEMMA 2.1 (HALPERIN ET AL. [15]). *Let T be the random instance as above with height $H \le \frac{1}{2}\log k$, and let S be any fixed subgraph of T with total weight C. Then for a constant $\gamma > 0$ that is sufficiently large, $p'(S) \ge e^{-\gamma C/H^2}$. Thus, $p(S) \le \exp\{-k \cdot e^{-\gamma C/H^2}\}$.*

A subtree S of the random instance T is *minimal* if it has the same root as T and all its leaves are at level H (i.e., they are also leaves of T). Thus, a minimal subtree is defined by the leaves of T that it reaches, yielding the next lemma.

LEMMA 2.2 (HALPERIN ET AL. [15]). *In the tree T there are at most d^{CH2^H} minimal subtrees with total weight at most C.*

The two lemmas above imply the following lower bound (that holds with high probability) on the integral solution for the random instance T.

COROLLARY 2.3 (HALPERIN ET AL. [15]). *Let T be the random instance as above with height $H \le \frac{1}{2}\log k$ and arity $d \le k$. Then with probability at least $1 - e^{-k^{\Omega(1)}}$, any (optimal integral) solution for the random instance T has total weight $\Omega(\frac{1}{\gamma}H^2 \log k)$.*

PROOF. Since all groups in $\mathcal{G}$ contain only leaves of T, an optimal solution to the instance T is necessarily a minimal subtree. For any fixed minimal subtree S of total weight at most C we have by Lemma 2.1 that S is a solution to the random instance T (i.e., covers all the groups) with probability at most $p(S) \le \exp\{-ke^{-\gamma C/H^2}\}$. The number of such minimal subtrees is, by Lemma 2.2, at most d^{CH2^H}. Taking a union bound, we conclude that the probability that T has an (optimal) solution of weight at most $C = \frac{1}{5\gamma}H^2 \ln k$ is at most $d^{CH2^H} \cdot \exp\{-ke^{-\gamma C/H^2}\} \le \exp\{\tilde{O}(\sqrt{k}) - k^{4/5}\} \le \exp\{-k^{3/4}\}$. □

It is also proven in [15] that (with high probability) the value of the fractional solution to the flow-based LP on the random instance T can be upper bounded as follows.

LEMMA 2.4 (HALPERIN ET AL. [15]). *With probability at least $1 - d^H k/e^{\Omega(d)}$, the random instance T has a fractional solution with value $O(H)$.*

We remark that an integrality ratio of $\Omega(\log^2 k)$ is shown in [15] by combining Corollary 2.3 with Lemma 2.4. Notice also that the randomized rounding procedure of [13] upper bounds the integrality ratio in T by $O(\log(d^H)\log k) = O(H \log d \log k)$. (Their proof actually yields the better bound $O(H \log k)$ for any tree instance with height H and k groups, but this improved bound is not necessary for our purposes.) By combining this integrality ratio upper bound with Lemma 2.4 we obtain that with high probability T has an integral solution of weight $O(H^2 \log d \log k)$. We shall use this upper bound in the proof of Lemma 2.6.

An extended random tree. We now consider an extended version of the random tree T, in which some pairs of edges may be "identical" or "complementary", and then the random choices in constructing the groups in this pair of edges are not independent, as follows. Let every vertex of T have one of three labels: (i) use a fresh coin; (ii) use the same coin as some sibling vertex that uses a fresh coin; or (iii) use the opposite coin of some sibling vertex that uses a fresh coin. We stress that the same labels are used for all the groups in $\mathcal{G}$. Two edges are called *identical* if they have a common parent and they use the same coin tosses (e.g., they both use the opposite of a coin tossed by a common sibling); thus, for identical edges, the set of groups passing through them are identical. Two edges are *complementary* if they have a common parent and they use opposite coins tosses (e.g., one of the edges tosses the coins, and the other edge uses the opposite result); thus, for a complementary pair, the set of groups passing through the two edges forms a partition of the set of groups passing through their common parent.

We call this tree an *extended random tree*. We assume further that each non-leaf vertex has at least $\hat{d} = d^{\Omega(1)}$ children that use fresh coins (and are thus independent of each other). The following two lemmas provide a step-by-step extension of Corollary 2.3 to this tree.

LEMMA 2.5. *Let T be an extended random tree as above with height $H \le \frac{1}{2}\log k$ and arity $d \le k$. Then with probability at least $1 - e^{-k^{\Omega(1)}}$ any (integral) solution for T that contains no complementary pairs has total weight $\Omega(\frac{1}{\gamma}H^2 \log k)$.*

PROOF. Let S be a fixed subgraph of T with total weight C and let $p(S)$ be the probability that S covers all the groups in $\mathcal{G}$, taken over the randomness in constructing the groups $\mathcal{G}$. We first claim that if S has no complementary edges then $p(S) \le \exp\{-ke^{-\gamma C/H^2}\}$. For two edges e_1 and e_2 with a common parent e_0, and with subtrees T_1, T_2 rooted at e_1 and e_2 respectively, the contraction of the pair (e_1, e_2) is done by deleting e_2 and adding T_2 as subtree rooted at e_1. To prove the claim, we construct a tree $\hat{T}$ from T by contracting every pair of identical edges, and let $\hat{S}$ be the subtree of $\hat{T}$ that is induced by S. It is easy to see that $p(S) = p(\hat{S})$ and that the total weight of $\hat{S}$ is at most C. The tree $\hat{T}$ might end up being not regular, but we can view $\hat{S}$ as a subtree of a bigger tree, which is regular. By applying Lemma 2.1 (whose bound does not depend on the degree of the tree) we obtain $p(S) = p(\hat{S}) \le \exp\{-ke^{-\gamma C/H^2}\}$.

Finally, observe that the upper bound of Lemma 2.2 applies also for an extended random tree T. Thus, the proof follows by the same union bound as in the proof of Corollary 2.3. □

LEMMA 2.6. *Let T be an extended random tree as above with height $H \le \frac{1}{2}\log k$ and arity $d \le k$. Assume that $\hat{d} \ge \max\{\log^2 k, d^{\Omega(1)}\}$, and let $\beta > 0$ be a sufficiently small constant. Then with probability at least $1 - e^{-d^{\Omega(1)}}$ any (integral) solution for T in which the number m_h of complementary pairs at level h satisfies $\sum_{h=1}^{H} \frac{m_h}{2^h} \le \frac{\beta}{\gamma \log d}$ has total weight $\Omega(\frac{1}{\gamma}H^2 \log k)$.*

PROOF. We first claim that with probability at least $1 - e^{-d^{\Omega(1)}}$, for every vertex v at level h the subtree of T rooted at v has an integral solution that contains no pair of complementary edges and has total weight $O(H^2/2^h \cdot \log d \log k)$. To prove the claim, fix a vertex v at some level h. The subtree of T rooted at v contains a (non-extended) random tree (similar to T but) with height $H - h$, arity $\hat{d} = d^{\Omega(1)}$, at most k groups, and its edge weights are scaled down by a factor of 2^h. It follows from Lemma 2.4 that with probability at least $1 - \hat{d}^H k/e^{\Omega(\hat{d})}$ this scaled instance has a fractional solution of value at most $O(H/2^h)$. By the randomized rounding procedure of [13], such a fractional solution implies an integral solution (i.e., a subtree covering all the groups passing through v) with total weight $O(H/2^h \cdot H \log d \log k)$. Taking a union bound over less than d^H non-leaf vertices v, the overall probability of failure is at most $d^H \cdot \hat{d}^H k/e^{\Omega(\hat{d})} \le e^{O(\log k \log d) - \Omega(\hat{d})} \le e^{-d^{\Omega(1)}}$, which proves the claim.

For the rest of the proof assume that the event in the claim above happens. We then show that for any solution S (for T) with total weight C and $\sum_{h=1}^{H} \frac{m_h(S)}{2^h} \le \frac{\beta}{\gamma \log d}$ (recall that $m_h(S)$ is the number of complementary pairs at level h that S contains) there exists a solution S' that has total weight at most $C + O(\frac{\beta}{\gamma}H^2 \log k)$ and contains no complementary pairs. To prove this, construct S' from S by replacing every complementary pair with a subtree as in the claim above (with no complementary pairs) under their common parent. It follows that every complementary pair at level h is replaced with a subtree of total weight $O(H^2/2^h \cdot \log d \log k)$ that covers all the groups covered by the common parent. S' clearly covers every group that S covers and is thus a solution for T. Observe that the weight of S' is at most $C + \sum_{h=1}^{H} m_h(S) \cdot (H^2/2^h \cdot \log d \log k) \le C + O(\frac{\beta}{\gamma}H^2 \log k)$.

Finally, assume that the event described in Lemma 2.5 also happens. Notice that by a union bound, these two events (the one from Lemma 2.5 and the one above) both happen with probability at least $1 - e^{-d^{\Omega(1)}}$ (recall that $k \ge$

d). In this case, we have from Lemma 2.5 that the total weight of S' is $\Omega(\frac{1}{\gamma}H^2 \log k)$; combining this lower bound with the upper bound $C + O(\frac{\beta}{\gamma}H^2 \log k)$ from above, we conclude that for a sufficiently small constant $\beta > 0$, the total weight of S is $C \geq \Omega(\frac{1}{\gamma}H^2 \log k)$. □

3. THE REDUCTION

In this section we introduce a reduction from an arbitrary NP-hard problem (say SAT) to Group-Steiner-Tree. We first recall (a variant of) the reduction of Lund and Yannakakis [17] that yields a logarithmic gap for Set-Cover (see also [1, 11]); however, we view it as a reduction to Group-Steiner-Tree in a tree of height 1 by using the equivalence between these two problems (Set-Cover and Group-Steiner-Tree in tree of height 1). We then show how to recursively compose many copies of this reduction, thus obtaining a reduction to Group-Steiner-Tree in a tree of large height.

3.1 The parallel repetition starting point

The starting point for our reduction is the following PCP system (which is motivated by one-round two-prover interactive proof systems.) A *witness* to the PCP is an assignment to the $2M$ variables $X_1^1, \ldots, X_M^1$ and $X_1^2, \ldots, X_M^2$, all taking values from the same set A (called *answers*). A *verifier* is any polynomial time procedure that decides whether to accept or reject a given witness as follows: the verifier uses its private random coins to pick $q_1, q_2 \in \{1, \ldots, M\}$ (called *queries*), reads from the witness the values of the two corresponding variables $X_{q_1}^1, X_{q_2}^2$, and then decides whether to accept or reject the witness. The next theorem follows essentially by applying the Parallel Repetition Theorem of Raz [19] on the so-called PCP Theorem of [2] (see also [1]).

Theorem 3.1. *Let L be any NP-complete language. Then there exist constants a, c_0, c_1 and an algorithm that, given an instance $\mathcal{I}$ for L and an integer $l \geq 1$, produces in time $|\mathcal{I}|^{O(l)}$ a verifier V for the above PCP system, such that: $|M| = m^l$ for $m \leq |\mathcal{I}|^{c_1}$ and $|A| = a^l$; if $\mathcal{I} \in L$ then there exists a witness that V accepts with probability 1; if $\mathcal{I} \notin L$ then any witness is accepted by V with probability at most $2^{-c_0 l}$.*

We remark that for Set-Cover, it suffices that the number of parallel repetitions is $l = O(\log \log m)$. In our reduction we shall take $l = \log m$.

Additional properties. For our purposes, it is important that the PCP system of Theorem 3.1 has some additional properties. We say that the verifier is *D-regular* if the queries pair (q_1, q_2) is chosen uniformly at random from a collection of $D \cdot M$ pairs, called the *plausible pairs*, such that each possible value for q_1 appears in exactly D pairs, and each possible value for q_2 appears in exactly D pairs. Note that a D-regular verifier chooses each of q_1 and q_2 from a uniform distribution, but these two distributions are (generally) not independent. It is shown in [11] (and also in [12]) that Theorem 3.1 can be proven with the additional property that the verifier V is d^l-regular for some constant $d > 0$.

We say that the verifier is a *projection-equality tester* if it decides whether to accept or reject a witness as follows. The verifier has for each variable X_t^p (where $p \in \{1,2\}$ and $t \in \{1, \ldots, M\}$) a function $\pi_t^p : A \to \tilde{A}$ called a *projection.* The verifier accepts the witness if the projections of the two variables it reads from the witness are equal, namely, if $\pi_{q_1}^1(X_{q_1}^1) = \pi_{q_2}^2(X_{q_2}^2)$. The known proofs of Theorem 3.1 give a verifier that is a projection-equality tester with $|\tilde{A}| = \tilde{a}^l$ for some constant $\tilde{a}$. (In fact, $\tilde{a} = 2$ in [11] and $\tilde{a} = 3$ in [12].) We assume without loss of generality that $\tilde{A} = \{1, \ldots, \tilde{a}^l\}$.

3.2 Reduction to Group-Steiner-Tree in a tree of height one.

We next describe the randomized reduction from the PCP system from Section 3.1 to the Group-Steiner-Tree problem in a tree of height one (i.e., a star graph). This reduction is essentially (a variant of) the Lund and Yannakakis [17] reduction to Set-Cover (see also [1, 11]). The reduction produces an instance $\mathcal{T}$ of the Group-Steiner-Tree problem on a rooted tree of height 1 (i.e. a star), as follows. Let $\mathcal{T}$ have one edge (and thus also one leaf vertex) for every possible assignment of a value in A for every variable X_t^p. Thus, the total number of edges in $\mathcal{T}$ (i.e., the degree of the root in this tree of height 1) is $2 \cdot m^l \cdot a^l$. For every plausible pair of queries the instance $\mathcal{T}$ has a *ground set of u groups*, where u will be determined later. To simplify notation, let us enumerate the plausible pairs of queries by $i = 1, \ldots, (md)^l$ and denote the corresponding ground sets by U_i. Let $\hat{U} := \cup_i U_i$ be the set of all groups in the instance $\mathcal{T}$. Then $|U_i| = u$ for all i and $|\hat{U}| = (md)^l \cdot u$.

As a means to define the members of each group in this instance, we give every leaf vertex a set of *labels*. Each label is of the form $D_{i,j}$ or $\bar{D}_{i,j}$, where $i \in \{1, \ldots, (md)^l\}$ and $j \in \tilde{A}$. A leaf that corresponds to assigning a value $r \in A$ to a variable X_t^1 is labeled by $D_{i,\pi_t^1(r)}$ for every plausible pair of queries i that contains the variable X_t^1. Similarly, a leaf that corresponds to assigning a value $r \in A$ to a variable X_t^2 is labeled by $\bar{D}_{i,\pi_t^2(r)}$ for every plausible pair of queries i that contains the variable X_t^2. Since each variable X_t^p appears in exactly d^l pairs of queries, we have that every leaf has exactly d^l labels. (These labels are distinct since they have different values for i.) Note that the labeling is deterministic.

Finally, we randomly construct the groups using the labels. For every group $g \in U_i$ and every $j \in \tilde{A}$, choose uniformly at random (i.e., with equal probabilities) and independently of all other events, one of the two labels $D_{i,j}$ and $\bar{D}_{i,j}$. Then let each group g contain all the leaves that have the label chosen by g.

3.3 Reduction to Group-Steiner-Tree by recursive composition.

We now construct an instance $\mathcal{T}'$ of the group Steiner tree from (many copies of) the instance $\mathcal{T}$ (of Section 3.2). $\mathcal{T}'$ is a complete (rooted) tree of arity $2 \cdot m^l \cdot a^l$ and height H (that we will determine later to be polylog(m)). That is, every non-leaf vertex has $2 \cdot m^l \cdot a^l$ children, and all the leaves of the tree are at distance H from the root. As usual, the *level of a vertex* is its distance from the root; the root itself is at level 0, and there are $H + 1$ levels. The *level of an edge* is h iff it connects a vertex at level $h - 1$ to a vertex at level h. We then let every edge at level h have weight w^h, where $w := 1/(2m^l)$.

The set $\hat{U}$ of all groups in the instance $\mathcal{T}'$ has the following structure. $\hat{U}$ is partitioned into $(md)^l$ what we shall call *level*

1 ground sets, each of size u_1 (that we will choose later). Each level 1 ground set is further partitioned into m^l *level* 2 ground sets of size $u_2 = u_1/m^l$. More generally, each level h ground set, for $1 \le h \le H-1$, is further partitioned into m^l *level* $h+1$ ground sets of size $u_{h+1} = u_h/m^l$. Thus, for every $1 \le h \le H$ the number of level h ground sets is $(md)^l \cdot m^{l(h-1)} = d^l m^{lh}$.

We define the group members so that each non-leaf vertex will essentially form, together with its children, a copy of $\mathcal{T}$. (However, the groups will actually contain only leaves of $\mathcal{T}'$.) First, let the root vertex and its children form a copy of $\mathcal{T}$ using the level 1 ground sets. This is possible since the degree of the root is $2 \cdot m^l \cdot a^l$, just like in the tree $\mathcal{T}$, and since the number of level 1 ground sets is $(md)^l$, just like the number of ground sets in $\mathcal{T}$. This copy of $\mathcal{T}$, with ground sets' size $u = u_1$, is said to be at level 1. We stress that at this point we do not determine which groups are chosen to each edge, but rather determine only the ground sets and labels that are associated with each edge; at a later stage we will use the labels to decide which groups of these ground sets actually pass through each edge. Consider next a vertex v at level 1. Since v is a leaf in the copy of $\mathcal{T}$ at level 1, exactly d^l level 1 ground sets appears in it; each such ground set is partitioned into m^l level 2 ground sets, yielding a total of $(md)^l$ level 2 ground sets "relevant" to v. We then let v, together with its $2 \cdot m^l \cdot a^l$ children, form a copy of $\mathcal{T}$ using these $(md)^l$ level 2 ground sets. This copy of $\mathcal{T}$, with ground sets' size $u = u_2$, is said to be at level 2. More generally, each vertex v at level $1 \le h \le H-1$ is a leaf in one copy of $\mathcal{T}$ at level h, and thus exactly d^l level h ground sets appears in it. Each such ground set is partitioned into m^l level $h+1$ ground sets, yielding a total of $(md)^l$ level $h+1$ ground sets "relevant" to v. We then let v, together with its $2 \cdot m^l \cdot a^l$ children, form a copy of $\mathcal{T}$ using these $(md)^l$ level $h+1$ ground sets. This copy of $\mathcal{T}$, with ground sets' size $u = u_{h+1}$, is said to be at level $h+1$.

Finally, the groups are constructed randomly by using independent coins tosses at the different copies of $\mathcal{T}$ in $\mathcal{T}'$ and letting only the leaves of $\mathcal{T}'$ be group members. More specifically, at every copy of $\mathcal{T}$ at every level $1 \le h \le H$, we have (just like in $\mathcal{T}$) for every "relevant" level h ground set U_i, every group $g \in U_i$, and every $j \in \tilde{A}$, one fair coin. This coin chooses for the group g (in this copy of $\mathcal{T}$) uniformly at random (i.e., with equal probabilities) and independently of all other events, one of the two labels $D_{i,j}$ and $\bar{D}_{i,j}$. We stress that all these coins are mutually independent. Consider a group $g \in \hat{U}$ and a leaf v in $\mathcal{T}'$, and denote by v_h the ancestor of v at level h. Then g passes through v_h if it passes through its parent vertex and v_h has the label chosen by g in the corresponding copy of $\mathcal{T}$. That is, the leaf v is a member of the group g if at every level $1 \le h \le H$ the vertex v_h has the label chosen by g in the corresponding level h copy of $\mathcal{T}$.

4. ANALYSIS OF THE GAP

In this section we show that the reduction creates a polylogarithmic gap between the case where $\mathcal{I} \in L$ and $\mathcal{I} \notin L$. Specifically, Section 4.1 shows that if $\mathcal{I} \in L$ (called a YES instance) then the instance $\mathcal{T}'$ has a solution of weight H, and Section 4.2 shows that if $\mathcal{I} \notin L$ (called a NO instance) then with high probability (over the randomness in the reduction) any solution of the instance $\mathcal{T}'$ has weight $\Omega(H^2 \log k)$. (Recall that H is a parameter of the reduction.) Thus, the reduction produces a gap of $\Omega(H \log k)$ on the approximation ratio of GROUP-STEINER-TREE on trees. Section 4.3 concludes by setting the parameter H so that $H = (\log k)^{1-\epsilon}$ for an arbitrarily small constant $\epsilon > 0$. The crux is that approximation ratio $o(H \log k)$ for GROUP-STEINER-TREE on trees yields a randomized randomized algorithm for deciding whether $\mathcal{I} \in L$; the guarantees of this algorithm are like in co-RP, but its running is quasi-polynomial (with our intended choice of l and H). By standard arguments we then get for $L \in$ NP also a ZPP-like decision algorithm, i.e., that NP $\subseteq$ ZTIME$(n^{\mathrm{polylog(n)}})$. completing the proof of Theorem 1.1.

4.1 YES instance

Consider the case $\mathcal{I} \in L$. Then the PCP system described in Section 3.1 has a witness for which the verifier V always accepts. Thus, there exists an assignment to the variables $X^1_1, \dots, X^1_M$ and $X^2_1, \dots, X^2_M$, such that $\pi^1_{q_1}(X^1_{q_1}) = \pi^2_{q_2}(X^2_{q_2})$ for all plausible pairs of queries (q_1, q_2). This assignment defines for the star instance $\mathcal{T}$ a solution with $2m^l$ edges, as follows. Let S be the set of $2m^l$ edges (in $\mathcal{T}$) that correspond to the values of the variables $X^1_1, \dots, X^1_M$ and $X^2_1, \dots, X^2_M$ in the assignment. To see that S is a solution for $\mathcal{T}$, consider a ground set U_i. This ground set corresponds to a plausible pair of queries (q_1, q_2). The assignment mentioned above gives values to $X^1_{q_1}$ and $X^2_{q_2}$, such that $\pi^1_{q_1}(X^1_{q_1}) = \pi^2_{q_2}(X^2_{q_2})$. Let us denote the latter value by $j \in \tilde{A}$; then the assignments to $X^1_{q_1}$ and $X^2_{q_2}$ correspond, to an edge of S whose non-root endpoint vertex is labeled by $D_{i,j}$ and to an edge of S whose non-root endpoint vertex is labeled by $\bar{D}_{i,j}$, respectively. Thus, every group $g \in U_i$ is covered by one of these two edges of S (regardless of the random choice that g makes among these two labels). It follows that S always covers every group g in the instance $\mathcal{T}$.

The assignment mentioned above also defines a solution of total weight H for $\mathcal{T}'$, as follows. First, in the level 1 of $\mathcal{T}'$ (i.e., in the level 1 copy of $\mathcal{T}$) let S' contain the $2m^l$ edges that correspond to the values of the variables $X^1_1, \dots, X^1_M$ and $X^2_1, \dots, X^2_M$ in the assignment. These edges reach $2m^l$ level 2 copies of $\mathcal{T}$ in $\mathcal{T}'$. At each of these copies, let S' contain the $2m^l$ edges that correspond to the values of the variables $X^1_1, \dots, X^1_M$ and $X^2_1, \dots, X^2_M$ in the assignment. Continue similarly at every level $h \le H$. Then the set S' contains $(2m^l)^h$ edges at each level h. Since each edge at level h has weight w^h, the total weight of S' is $\sum_{h=1}^{H} (2m^l)^h \cdot w^h = H$. We now argue that S' is a solution for $\mathcal{T}'$. Consider a level H ground set U, and denote by U_h the level h ground set that contains U. By the arguments above for the star instance $\mathcal{T}$ we have that S' contains a complete binary tree $\mathcal{T}'_U$ of height H (i.e. from the root to the leaves of $\mathcal{T}'$), in which every non-leaf vertex at level h has, for $i = U_h$ and for some $j \in \tilde{A}$, one of its children labeled by $D_{i,j}$ and the other one labeled by $\bar{D}_{i,j}$. Thus, every group $g \in U$ is covered by one of the 2^H leaves in this complete binary tree $\mathcal{T}'_U$ (regardless of the random choice that g makes among these two labels). It follows that S' always covers every group g in the instance $\mathcal{T}'$.

4.2 NO instance

Consider the case $\mathcal{I} \notin L$, and assume for contradiction that $\mathcal{T}'$ has a solution (i.e., a subtree) S of total weight $C \le \delta H^2 \log k$ for a sufficiently small constant $\delta > 0$ (that will be determined later). We shall choose the size of the level H ground sets to be $u_H := d^l m^{lH}$, so that $k = d^l m^{lH} \cdot u_H = (u_H)^2$ and thus $C \leq 2\delta H^2 \log u_H$. Below we use the solution S to construct randomly a witness π_R (for the PCP system from Section 3.1) which the verifier accepts with probability higher than $2^{-c_0 l}$. (This probability is taken over both the randomness in constructing the witness π_R and the randomness of the verifier.) It follows that there exists a witness π which the verifier accepts with probability higher than $2^{-c_0 l}$ (this probability is taken only over the randomness of the verifier), which contradicts Theorem 3.1. Technically, this whole argument only goes through when the (randomized) reduction is "successful", and we shall show that this event happens with high probability (over the randomness in the reduction).

An averaging lemma. Consider a level H ground set U. Then $U = U_H \subset U_{H-1} \subset \cdots \subset U_1$, where U_h is a level h ground set. Let $\mathcal{T}'_U$ be the subtree of $\mathcal{T}'$ that is induced on the vertices in which any of $U_1, \ldots, U_H$ appears (i.e., all vertices at level h with U_h appearing in them, for all h). Let $s_h(U)$ denote the number of edges at level h in $\mathcal{T}'_U$ that belong to (i.e., are used by) the solution S. The next lemma will allow us to restrict our attention to a single level H ground set and analyze it separately.

LEMMA 4.1. *Let U be a level H ground set chosen uniformly at random. Then $\mathbb{E}_U\left[\sum_{h=1}^{H} \frac{s_h(U)}{2^h}\right] = C$.*

PROOF. For any edge e, let X_e be 1 if $e \in S$ and 0 otherwise. Let L_h denote the set of edges at level h of $\mathcal{T}'$. Recall that the number of level h ground sets that appear in a vertex at level h is d^l, and that each level h ground set is partitioned into $m^{l(H-h)}$ level H ground sets. Letting U range over all level H ground sets we thus have:

$$\sum_U \sum_{h=1}^{H} \frac{s_h(U)}{2^h} = \sum_{h=1}^{H} \frac{1}{2^h} \sum_U s_h(U) = \sum_{h=1}^{H} \frac{1}{2^h} \sum_{e \in L_h} X_e \cdot d^l m^{l(H-h)}.$$

Dividing by the number of level H ground sets, which is $d^l m^{lH}$, we obtain

$$\mathbb{E}_U\left[\sum_{h=1}^{H} \frac{s_h(U)}{2^h}\right] = \sum_{h=1}^{H} \sum_{e \in L_h} \frac{X_e}{2^h m^{lh}} = \sum_{h=1}^{H} \sum_{e \in L_h} X_e \cdot w^h = C.$$

□

The restricted instance $\mathcal{T}'_U$. Lemma 4.1 suggests that S provides a "low-weight" solution to the problem of covering the groups of a "typical" ground set U of level H. More precisely, for each level H ground set U the instance $\mathcal{T}'$ defines the following GROUP-STEINER-TREE instance on the subtree $\mathcal{T}'_U$: This is the problem of covering the groups of U in the subtree $\mathcal{T}'_U$, where edges at level $1 \leq h \leq H$ have weight $1/2^h$. We shall call this the *restricted instance $\mathcal{T}'_U$*. Since S is a solution for $\mathcal{T}'$, the edges of S in $\mathcal{T}'_U$ form a solution for the restricted instance $\mathcal{T}'_U$. Lemma 4.1 shows that if U is a randomly chosen level H ground set, then the solution for $\mathcal{T}'_U$ provided by S has expected weight C.

Complementary edges. Two edges in $\mathcal{T}$ form a *complementary pair in $\mathcal{T}$ with respect to a ground set U_i* if for some j, one of these edges is labeled by $D_{i,j}$ and the other one is labeled by $\bar{D}_{i,j}$. Observe that two such complementary edges in $\mathcal{T}$ define values for the two variables $X^1_{q_1}, X^2_{q_2}$ that correspond to U_i, such that if the verifier's pair of queries is (q_1, q_2) then these values of $X^1_{q_1}, X^2_{q_2}$ cause the verifier to accept. Two edges in $\mathcal{T}'$ form a *complementary pair in $\mathcal{T}'$ at level h* if they belong to the same copy of $\mathcal{T}$ at level h of $\mathcal{T}'$ and they are complementary in this copy of $\mathcal{T}$. Two edges in $\mathcal{T}'_U$ form a *complementary pair in $\mathcal{T}'_U$* if they belong to the same copy of $\mathcal{T}$ and they are complementary in this copy of $\mathcal{T}$ with respect to a ground set that contains U.

An "easy" case - no complementary edges. For ease of exposition (and to demonstrate the role of Lemma 4.1), we first prove the special case where S contains no two edges which are complementary in $\mathcal{T}'$. We shall use the following lemma, whose proof essentially follows from that of the integrality ratio of [15] (see Section 2).

LEMMA 4.2. *Let $l \ll H \leq \frac{1}{2}\log u_H$. Then with high probability (namely, $1 - e^{-u_H^{0.1}}$) (over the randomness in the reduction), the following holds for all level H ground sets U: In the restricted instance $\mathcal{T}'_U$, every solution that contains no complementary pairs has weight $\Omega(\frac{1}{\gamma}H^2 \log u_H)$.*

PROOF. For any fixed level H ground set U we can apply Lemma 2.5 on $\mathcal{T}'_U$, since $\mathcal{T}'_U$ is an extended random tree with height H, arity $2a^l$ and u_H groups. Thus, with probability at least $1 - e^{-u_H^{\Omega(1)}}$ any (integral) solution for $\mathcal{T}'_U$ that contains no complementary pairs has total weight $\Omega(\frac{1}{\gamma}H^2 \log k)$. The proof then follows by taking a union bound over all the $d^l m^{lH} \leq u_H$ level H ground sets U. □

Suppose that no two edges in S are complementary in $\mathcal{T}'$. By Lemma 4.1 there exists a restricted instance $\mathcal{T}'_U$ for which S provides a solution of cost $\sum_{h=1}^{H} \frac{s_h(U)}{2^h} \leq C$. But since S contains no two complementary edges in $\mathcal{T}'_U$, we have from Lemma 4.2 that with high probability, the cost of this solution is $\sum_{h=1}^{H} \frac{s_h(U)}{2^h} \geq \Omega(\frac{1}{\gamma}H^2 \log u_H)$. It follows that $\Omega(H^2 \log u_H) \leq \sum_{h=1}^{H} \frac{s_h(U)}{2^h} \leq C \leq 2\delta H^2 \log u_H$. For a sufficiently small constant $\delta > 0$, this yields a contradiction. Note that the only restriction on H is that $H \leq \frac{1}{2}\log u_H$; thus, in this "easy" case we can choose $H = \Omega(\log k)$, yielding a gap of $\Omega(\log^2 k)$.

A successful reduction. We say that the reduction is *successful* if the events described in Lemmas 4.3 and 4.4 below both happen. These two lemmas will be used in the proof of the general case. The first one extends Lemma 4.2 to solutions (for a restricted instance $\mathcal{T}'_U$) that contain relatively few complementary edges. The second one shows that any solution for $\mathcal{T}'$ must have edges in at least $(1/w)^{h-1}$ copies of $\mathcal{T}$ at any level h (i.e., similar to the situation in the YES instance).

LEMMA 4.3. *Let $H \leq \frac{1}{2}\log u_H$ and $\lambda' \log\log u_H \leq l \ll H$ for a sufficiently large constant $\lambda' > 0$, and let $\beta > 0$ be a sufficiently small constant. Then with probability at least $1 - e^{-u_H}$ (over the randomness in the reduction), the following holds for all level H ground sets U: In the restricted instance $\mathcal{T}'_U$, every solution that contains at most $\frac{\beta}{2\gamma \log a \cdot lH} 2^h = O(\frac{2^h}{H^2})$ complementary pairs at every level h has weight $\Omega(\frac{1}{\gamma}H^2 \log u_H)$.*

Proof. For any level H ground set U the restricted instance $\mathcal{T}'_U$ is an extended random tree with height H, arity $2a^l$ and u_H groups. Note also that the number of "independent" children of any non-leaf vertex (called there $\hat{d}$) is indeed at least $\tilde{a}^l \geq (2a^l)^{\Omega(1)}$. To apply Lemma 2.6 on $\mathcal{T}'_U$, notice that the requirement $(2a^l)^{\Omega(1)} \geq \log^2 u_H$ holds if λ' is a sufficiently large constant, and that our assumption that $m_h \leq \frac{\beta}{2\gamma \log a \cdot lH} 2^h$ for all h implies that $\sum_{h=1}^{H} \frac{m_h}{2^h} \leq H \cdot \frac{\beta}{2\gamma \log a \cdot lH} \leq \frac{\beta}{\gamma \log(2a^l)}$. Then by Lemma 2.6, with probability at least $1 - e^{-(2a^l)^{\Omega(1)}}$ any (integral) solution for $\mathcal{T}'_U$ in which the number m_h of complementary pairs at level h satisfies $\sum_{h=1}^{H} \frac{m_h}{2^h} \leq \frac{\beta}{\gamma \log(2a^l)}$, has total weight $\Omega(\frac{1}{\gamma} H^2 \log k)$. The proof then follows by taking a union bound over all $d^l m^{lH} \leq u_H$ level H ground sets U. If λ' is sufficiently large the total failure probability is at most $u_H \cdot e^{-(2a^l)^{\Omega(1)}} \leq u_H \cdot e^{-\log^2 u_H} \leq e^{-u_H}$. □

Lemma 4.4. *With probability at least* $1 - e^{-u_H^{\Omega(1)}}$ *(over the randomness in the reduction), any solution for the instance* $\mathcal{T}'$ *contains edges in* $\Omega((1/w)^{h-1})$ *copies of* $\mathcal{T}$ *at every level* h.

We need the following claim for the proof of Lemma 4.4.

Claim 4.5. *With probability at least* $1 - e^{-u_H^{\Omega(1)}}$, *every vertex* v *at every level* h *contains* $\Theta(u_H/2^h)$ *groups of each of the* $d^l m^{l(H-h)}$ *level* H *ground sets it contains.*

Proof of Claim 4.5. Consider a level H ground set U that appears in v. The path from the root to v has length h, and thus each group $g \in U$ passes through this path independently with probability $1/2^h$. Therefore, the number of groups of U that v contains is a binomially distributed random variable $X \sim B(u_H, 1/2^h)$. By Chernoff bounds, $X = \Theta(u_H/2^h)$ with probability at least $1 - e^{-\Theta(u_H/2^h)}$. Taking a union bound over the $\Theta((2a^l m^l)^H) \leq m^{2lH} \leq u_H^2$ vertices v and $d^l m^{lH} = u_H$ level H ground sets U, the total probability of failure is at most $u_H \cdot u_H^2 \cdot e^{-\Theta(u_H/2^h)} \leq e^{-u_H^{\Omega(1)}}$, which proves the claim. □

We are now ready to prove Lemma 4.4:

Proof of Lemma 4.4. Let us assume that the event described in Claim 4.5 indeed happens. Notice that edges in a copy of $\mathcal{T}$ at level h can cover only groups that appear in the level $h-1$ vertex v that is the root of this copy of $\mathcal{T}$; by our assumption any number of edges from the same copy of $\mathcal{T}$ covers together $O(u_H/2^{h-1})$ groups in each of the $d^l m^{l(H-h+1)}$ level H ground sets appearing in v. Thus, a solution for $\mathcal{T}'$ that contains edges from at most z distinct copies of $\mathcal{T}$ at level h can cover at most $z \cdot O(u_H/2^{h-1}) \cdot d^l m^{l(H-h+1)}$ level H groups. Since there are $k = d^l m^{lH} \cdot u_H$ level H groups in $\mathcal{T}'$ and a solution covers all of them, we get that $z \geq \Omega(2^{h-1} \cdot m^{l(h-1)}) = \Omega((1/w)^{h-1})$. □

The general case. We describe below how the solution S yields a randomized witness π_R (for the PCP system from Section 3.1). We will then show that if the reduction is successful (i.e., that the events described in Lemmas 4.3 and 4.4 both happen) then the verifier accepts π_R with probability higher than $2^{-c_0 l}$.

The solution S defines a randomized witness π_R for the PCP system, as follows. First choose at random a level $h_R \in \{1, \ldots, H\}$. Then choose at random a copy of $\mathcal{T}$ uniformly at random from all the copies at level h_R of $\mathcal{T}'$ that contain at least one edge of the solution S. Denote this copy of $\mathcal{T}$ by $\mathcal{T}_R$. Recall that each edge in $\mathcal{T}_R$ corresponds to a value to one of the $2M$ variables $X_1^1, \ldots, X_M^1, X_1^2, \ldots, X_M^2$. Thus, the edges in $\mathcal{T}_R$ that belong to the solution S define for each of these $2M$ variables a set of values from A. We call this a *multi-assignment* $\tilde{\pi}_R$ to the $2M$ variables. This multi-assignment is *sparsified* into an assignment π_R by choosing for each variable X_t^p one value uniformly at random from the set of values assigned to X_t^p by the multi-assignment $\tilde{\pi}_R$. If this set of values is empty then the variable is assigned an arbitrary value.

We wish to lower bound the probability that the verifier accepts the above randomized witness π_R. (This probability is taken over the randomness in constructing the witness and the randomness of the verifier.) Denote by (q_{1R}, q_{2R}) the plausible pair of queries chosen randomly by the verifier. Each plausible pair of queries is identified with a ground set in $\mathcal{T}$, and thus also with a level h_R ground set U_R in the copy $\mathcal{T}_R$ (used in the witness construction). Thus, the randomized witness and the verifier define together a random tuple $< \mathcal{T}_R, U_R >$ at level h_R. This tuple motivates our following definition. A *permissible tuple* at level h is a tuple $< \mathcal{T}_z, U >$ where U is a level h ground set that appears in $\mathcal{T}_z$, and $\mathcal{T}_z$ is a copy of $\mathcal{T}$ at level h of $\mathcal{T}'$ that contains at least one edge of S (note that it is possible that S does not contain any edge which corresponds to U in T_z). Observe that a tuple is permissible if and only if it has a positive probability to be chosen as the random tuple $< \mathcal{T}_R, U_R >$. Below, we first lower bound the probability of choosing any particular permissible tuple, and then identify a relatively large set of permissible tuples that are likely to cause the verifier to accept.

Lemma 4.6. *Every permissible tuple* $< \mathcal{T}_z, U >$ *at any level* h *is chosen to be the random tuple* $< \mathcal{T}_R, U_R >$ *with probability at least* $\frac{w^{h-1}}{HC(md)^l}$.

Proof. The probability that h_R is chosen to be h is $1/H$. The number of copies of $\mathcal{T}$ at level h that contain at least one edge of S is at most $C(1/w)^{h-1}$, because each of them requires that S contains a distinct edge at level $h-1$. Thus, each such copy $\mathcal{T}_z$ is chosen with probability at least $(1/C) \cdot w^{h-1}$. Recall that U_R is defined by choosing one of $(md)^l$ plausible pairs of queries uniformly at random. Thus, the probability to choose a particular U is $1/(md)^l$. The proof follows. □

An *S-complementary tuple* is a permissible tuple $< \mathcal{T}_z, U >$ in which $\mathcal{T}_z$ contains two edges of S that form a complementary pair with respect to the ground set U. We next show that there are many S-complementary tuples.

Lemma 4.7. *If the reduction is successful, then there exists a level* h_0 *such that* $\mathcal{T}'$ *contains at least* $\Omega(\frac{d^l m^{lh_0} 2^{h_0}}{H^3})$ *distinct* S*-complementary tuples at level* h_0.

Proof. Assume that the reduction is successful. From Lemma 4.1 we have, using Markov's inequality, that for at least half of all level H ground sets U, we have $\sum_{h=1}^{H} \frac{s_h(U)}{2^h} \leq 2C$, i.e., S provides a solution with weight at most $2C =$

$4\delta H^2 \log u_H$ for the restricted instance $\mathcal{T}'_U$. Consider one of these ground sets U. For a sufficiently small constant δ, this weight $4\delta H^2 \log u_H$ is smaller than the term $\Omega_\gamma(H^2 \log u_H)$ in Lemma 4.3, and so (by this lemma) the solution S contains, for some level h_U, at least $\Omega(\frac{2^{h_U}}{H^2})$ complementary pairs in $\mathcal{T}'_U$. The levels h_U (for different ground sets U) can be spread over at most H levels; thus, there exists a level h_0 such that for at least $\frac{1}{2H}$-fraction of all level H ground sets U (namely, $\frac{d^l m^{lH}}{2H}$ ground sets), the solution S contains at least $\Omega(\frac{2^{h_0}}{H^2})$ complementary pairs at level h_0 in $\mathcal{T}'_U$. Observe that each of these complementary pairs defines an S-complementary tuple $< \mathcal{T}_z, W >$, where $\mathcal{T}_z$ is the copy of $\mathcal{T}$ at level h_0 in $\mathcal{T}'_U$ that contains the complementary pair of S and W is the level h_0 ground set that contains U. We thus obtain at least $\frac{d^l m^{lH}}{2H} \cdot \Omega(\frac{2^{h_0}}{H^2})$ S-complementary tuples at level h_0; however, these tuples need not be distinct because different level H ground sets U may correspond to the same level h_0 ground set W. Since each level h_0 ground set W is partitioned into $m^{l(H-h_0)}$ level H ground sets U, we have counted every $< \mathcal{T}_z, W >$ at most $m^{l(H-h_0)}$ times. We conclude that there are at least $\Omega\left(\frac{d^l m^{lH} 2^{h_0}}{H^3} \cdot \frac{1}{m^{l(H-h_0)}}\right) = \Omega\left(\frac{d^l m^{lh_0} 2^{h_0}}{H^3}\right)$ S-complementary tuples $< \mathcal{T}_z, W >$ at level h_0. □

COROLLARY 4.8. *With probability at least* $\Omega\left(\frac{1}{H^4 C}\right)$ *the random tuple* $< \mathcal{T}_R, U_R >$ *is an* S*-complementary tuple.*

PROOF. By Lemmas 4.6 and 4.7 we have that the random tuple $< \mathcal{T}_h, U_h >$ is an S-complementary tuple at level h_0 with probability at least

$$\Omega\left(\frac{d^l m^{lh_0} 2^{h_0}}{H^3} \cdot \frac{w^{h_0 - 1}}{HC(md)^l}\right) = \Omega\left(\frac{1}{H^4 C}\right),$$

where the equality follows from $w = 1/(2m^l)$. □

Recall that $\mathcal{T}_R$ corresponds to the multi-assignment $\tilde{\pi}_R$ and U_R corresponds to the plausible pair of queries (q_{1R}, q_{2R}). Thus, Corollary 4.8 says that with probability at least $\Omega\left(\frac{1}{H^4 C}\right)$, the multi-assignment $\tilde{\pi}_R$ gives the queried variables $X^1_{q_{1R}}, X^2_{q_{2R}}$ sets of values which contain a pair of values for which the verifier accepts. We thus only need to lower bound the probability that the sparsification procedure chooses this specific pair of values to be the values given by the assignment π_R. The next lemma shows that the number of values given by the multi-assignment $\tilde{\pi}_R$ is likely to be relatively small.

LEMMA 4.9. *With probability at least* $1 - O\left(\frac{1}{H^5 C}\right)$*, the number of values given by* $\tilde{\pi}_R$ *to the (actually queried) variables* $X^1_{q_{1R}}, X^2_{q_{2R}}$ *is at most* $H^5 C^2$*.*

PROOF. Suppose that h_R has been chosen. By Lemma 4.4 we have that the solution S has edges in at least $\Omega((1/w)^{h_R - 1})$ copies of $\mathcal{T}$ at level h_R. Thus, $\mathcal{T}_R$ is chosen (randomly) from at least $\Omega((1/w)^{h_R-1})$ copies of $\mathcal{T}$. It follows that the expected weight of S that is contained in the copy $\mathcal{T}_R$ is at most $O(Cw^{h_R - 1})$. The weight of an edge at level h_R is w^{h_R}, and so the expected number of edges of S in $\mathcal{T}_R$ is at most $O(Cw^{h_R-1})/w^h_R = O(C/w)$.

Let Y_p, for $p = 1, 2$ be the number of edges of S in $\mathcal{T}_R$ that correspond to values for the variables $X^p_1, \ldots, X^p_M$. By the above, $\mathbb{E}[Y_1 + Y_2] \le O(C/w)$. Let Z_p denote the number of edges of S in $\mathcal{T}_R$ that correspond to values for the variable $X^p_{q_{pR}}$ actually queried. Since each q_{pR} has a uniform distribution over m^l possible queries, we have that $\mathbb{E}[Z_p] = \mathbb{E}[\mathbb{E}[Z_p | Y_p]] = \mathbb{E}[Y_p/m^l]$ and thus $\mathbb{E}[Z_1 + Z_2] = \mathbb{E}[Y_1 + Y_2]/m^l \le O(C/w)/m^l = O(C)$. By applying Markov's inequality we get $\Pr[Z_1 + Z_2 \ge H^5 C^2] \le O(\frac{1}{H^5 C})$. Observing that $Z_1 + Z_2$ is just the number of values given by $\tilde{\pi}_R$ to the (actually queried) variables $X^1_{q_{1R}}, X^2_{q_{2R}}$, the lemma follows. □

By Corollary 4.8 and Lemma 4.9, with probability at least $\Omega(\frac{1}{H^4 C})$ the multi-assignment $\tilde{\pi}_R$ gives the queried variables $X^1_{q_{1R}}, X^2_{q_{2R}}$ sets of values such that (i) these sets are of size at most $H^5 C^2$. and (ii) these sets contain a pair of values for which the verifier accepts. We thus conclude that the multi-assignment $\tilde{\pi}_R$ yields a sparsified assignment π_R which the verifier accepts, with probability at least $\Omega(\frac{1}{H^4 C} \cdot \frac{1}{(H^5 C^2)^2}) = \frac{\Omega(1)}{H^{14} C^5}$.

4.3 Parameter setting

Finally, we need to set the parameters so that $\frac{\Omega(1)}{H^{14} C^5} > 2^{-c_0 l}$, i.e., that $c_0 l > 14 \log H + 5 \log C$. To this end, let us choose $l = \log m$ and $H = (\log m)^\alpha$ where α is an arbitrarily large constant (recall that $m = |\mathcal{I}|^{c_1}$). Recall that $k = u_H^2$ and $u_H = d^l m^{lH}$; thus $\log k = 2 \log u_H = \Theta(lH \log m) = \Theta(\log^{\alpha+2} m) = \Theta(H^{1+2/\alpha})$. It is easy to verify that for our choice of $C \le \delta H^2 \log k$ we have $\log H = \Theta(\log \log m) \ll l$ and $\log C = \Theta(\log H + \log \log k) = \Theta(\log \log m) \ll l$. We thus conclude that the hardness factor shown is $\Omega(H \log k) = \Omega((\log k)^{2-\epsilon})$ for any fixed $\epsilon > 0$. This completes the proof of Theorem 1.1.

5. CONCLUDING REMARKS

This paper gives the first polylogarithmic inapproximability result for a natural NP-hard problem. It is reasonable to ask whether there are other polylogarithmic thresholds, especially en route to understanding the structure of this phenomenon (e.g., in the sense of [16]). One candidate for that could be GROUP-STEINER-TREE in general graphs, for which the best upper bound currently known is $O(\log^2 k \log n)$–approximation. In fact, it is plausible (now more than before) to suspect that we are not aware of various other "intermediate" thresholds that arise naturally.

The construction of our reduction relies on the intuition given by the integrality ratio for GROUP-STEINER-TREE due to [15], and our analysis actually uses that result in a nontrivial manner. In the past, very few hardness results were devised or proven by building on the integrality ratio results. Nevertheless, this line of attack still appears to be of the most natural ones for many problems.

6. REFERENCES

[1] S. Arora and C. Lund. Hardness of approximations. In D. Hochbaum, editor, *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Company, 1996.

[2] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximation problems. *J. ACM*, 45(3):501–555, 1998.

[3] Y. Bartal. On approximating arbitrary metrics by tree metrics. In *30th Annual ACM Symposium on Theory of Computing*, pages 161–168. ACM, 1998.

[4] Y. Bartal and M. Mendel. Multi-embeddings and path approximation of metric spaces. In *14th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 424–433, 2003.

[5] M. Charikar, C. Chekuri, T. Cheung, Z. Dai, A. Goel, S. Guha, and M. Li. Approximation algorithms for directed Steiner problems. *J. Algorithms*, 33(1):73–91, 1999.

[6] M. Charikar, C. Chekuri, A. Goel, and S. Guha. Rounding via trees: deterministic approximation algorithms for group Steiner trees and k-median. In *30th Annual ACM Symposium on Theory of Computing*, pages 114–123. ACM, New York, 1998.

[7] C. Chekuri, G. Even, and G. Kortsarz. An approximation algorithm for the group Steiner problem. Manuscript. (Preliminary version appeared in SODA 2002), 2002.

[8] G. Even, J. Naor, S. Rao, and B. Schieber. Divide-and-conquer approximation algorithms via spreading metrics. *J. ACM*, 47(4):585–616, 2000.

[9] G. Even, J. Naor, B. Schieber, and M. Sudan. Approximating minimum feedback sets and multicuts in directed graphs. *Algorithmica*, 20(2):151–174, 1998.

[10] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. To appear in *35th Annual ACM Symposium on Theory of Computing*, 2003.

[11] U. Feige. A threshold of $\ln n$ for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[12] U. Feige, M. M. Halldórsson, G. Kortsarz, and A. Srinivasan. Approximating the domatic number. *SIAM J. Comput.*, 2002. To appear.

[13] N. Garg, G. Konjevod, and R. Ravi. A polylogarithmic approximation algorithm for the group Steiner tree problem. *J. Algorithms*, 37(1):66–84, 2000.

[14] A. Gupta, E. Halperin, G. Kortsarz, R. Krauthgamer, R. Ravi, A. Srinivasan, and N. Wang. Private communication. 2002.

[15] E. Halperin, G. Kortsarz, R. Krauthgamer, A. Srinivasan, and N. Wang. Integrality ratio for group Steiner trees and directed Steiner trees. In *14th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 275–284, Jan. 2003.

[16] S. Khanna, M. Sudan, L. Trevisan, and D. P. Williamson. The approximability of constraint satisfaction problems. *SIAM J. Comput.*, 30(6):1863–1920, 2001.

[17] C. Lund and M. Yannakakis. On the hardness of approximating minimization problems. *J. ACM*, 41(5):960–981, 1994.

[18] S. Rao and A. W. Richa. New approximation techniques for some ordering problems. In *9h Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 211–218, New York, 1998. ACM.

[19] R. Raz. A parallel repetition theorem. *SIAM J. Comput.*, 27(3):763–803, 1998.

[20] R. Raz and S. Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *29th Annual ACM Symposium on Theory of Computing*, pages 475–484. ACM, 1997.

[21] S. Safra and O. Schwartz. On the complexity of approximating TSP with neighborhoods and related problems. Manuscript, June 2002.

[22] P. D. Seymour. Packing directed circuits fractionally. *Combinatorica*, 15(2):281–288, 1995.

[23] A. Srinivasan. New approaches to covering and packing problems. In *12th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 567–576, 2001.

[24] L. Zosin and S. Khuller. On directed Steiner trees. In *13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 59–63, 2002.