Integrality ratio for Group Steiner Trees and Directed Steiner Trees*

Eran Halperin[†] Guy Kortsarz[‡] Robert Krauthgamer[§] Aravind Srinivasan[¶]
Nan Wang[||]

Abstract

The natural relaxation for the Group Steiner Tree problem, as well as for its generalization, the Directed Steiner Tree problem, is a flow-based linear programming relaxation. We prove new lower bounds on the integrality ratio of this relaxation. For the Group Steiner Tree problem, we show the integrality ratio is $\Omega(\log^2 k)$, where k denotes the number of groups; this holds even for input graphs that are *Hierarchically Well-Separated Trees*, introduced by Bartal [*Symp. Foundations of Computer Science*, pp. 184–193, 1996], in which case this lower bound is tight. This also applies for the Directed Steiner Tree problem. In terms of the number n of vertices, our results for the Directed Steiner problem imply an $\Omega(\frac{\log^2 n}{(\log \log n)^2})$ integrality ratio.

For both problems, these are the first lower bounds on the integrality ratio that are superlogarithmic in the input size. This exhibits, for the first time, a relaxation of a natural optimization problem whose integrality ratio is known to be superlogarithmic but subpolynomial. Our results and techniques have been used by Halperin and Krauthgamer [*Symp. on Theory of Computing*, pp. 585–594, 2003] to show comparable inapproximability results, assuming that NP has no quasi-polynomial Las-Vegas algorithms. We also show algorithmically that the integrality ratio for Group Steiner Tree is much better for certain families of instances, which helps pinpoint the types of instances (parametrized by optimal solutions to their flow-based relaxations) that appear to be most difficult to approximate.

Keywords and phrases. Group Steiner Tree, Directed Steiner Tree, flow-based relaxation, linear programming relaxation, integrality ratio, approximation algorithms.

Proposed running head: Integrality ratio for Group Steiner Trees

1 Introduction

The Group-Steiner-Tree problem is a network design problem that generalizes both Set-Cover and the Steiner-Tree problem. The Directed-Steiner-Tree problem is a further generalization of Group-Steiner-Tree. The natural relaxation for these two problems is a flow-based linear programming

[*] A preliminary version of this work appeared in the *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 275–284, 2003.

[†] International Computer Science Institute, 1947 Center St., Suite 600, Berkeley, CA 94704, USA. Work was done while this author was also affiliated with the Computer Science Division of U.C. Berkeley, supported in part by NSF grants CCR-9820951 and CCR-0121555 and DARPA cooperative agreement F30602-00-2-0601. Email: heran@cs.princeton.edu

[‡] Department of Computer Sciences, Rutgers University, Camden, NJ 08102, USA. Email: guyk@camden.rutgers.edu

[§] IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120, USA. Work was done while this author was with the International Computer Science Institute and with the Computer Science Division of U.C. Berkeley, supported in part by NSF grants CCR-9820951 and CCR-0121555 and DARPA cooperative agreement F30602-00-2-0601. Email: robi@almaden.ibm.com

[¶] Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA. E-mail: srin@cs.umd.edu. Supported in part by NSF Awards CCR-0208005 and CNS-0426683.

[||] Morgan Stanley, Inc., New York, NY, USA. Work done while at the Department of Computer Science, University of Maryland, College Park, MD 20742, USA. Supported in part by NSF Awards CCR-0208005 and CNS-0426683.

relaxation. We show a polylogarithmic (about log squared) lower bound on the integrality ratio of this relaxation. For both problems, these are the first such lower bounds that are superlogarithmic in the input size, and our bounds are, in fact, nearly tight in the important special case of input graphs which are tree networks. Let n be the number of vertices and k the number of groups. Our probabilistic approach and our analysis have been used in [HK03] to show that for every fixed $\epsilon > 0$, the GROUP-STEINER-TREE and DIRECTED-STEINER-TREE problems admit no efficient $\Omega(\log^{2-\epsilon} k)$ and $\Omega(\log^{2-\epsilon} n)$ approximations respectively, unless NP has quasi-polynomial time Las-Vegas algorithms. We also present improved approximation algorithms for certain families of instances of the GROUP-STEINER-TREE problem, shedding light on the type of instances that appear to be most difficult for the flow-based relaxation.

(a). 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) $g_1, g_2, \ldots, g_k$ of 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 each group. 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 a star can be used to model the SET-COVER problem (c.f., [GKR00]).

A natural flow-based relaxation for this problem is the following. Find a *capacity* $x_e \in [0, 1]$ for each edge $e \in E$ so that the capacities can support one unit of flow from r to g_i, separately for each g_i (as opposed to supporting a unit flow simultaneously for all g_i). Subject to this constraint, we want to minimize $\sum_e w_e x_e$. It is easy to check that the feasible solutions which satisfy $x_e \in \{0, 1\}$ for all e, exactly correspond to feasible solutions for the Group Steiner Tree problem; hence, the above flow-based relaxation is indeed a valid linear programming (LP) relaxation for the problem. This is a natural relaxation for this problem (and for some of its generalizations), and is the main subject of investigation in this paper.

We start with a useful definition from [Bar96].

Definition 1.1. *Let $c > 1$. A c-Hierarchically Well-Separated Tree (c-HST) is a rooted weighted tree such that (i) all leaves are at the same distance from the root, (ii) the edges in the same level are equal-weighted, and (iii) the weight of an edge is exactly $1/c$ times the weight of its parent edge.*

(**Remark:** *Item (ii) is slightly stronger than the original definition from [Bar96], but can be assumed without loss of generality due to the analyses of [Bar96, Bar98, KRS01].)*

We simply say "HST" when referring to a c-HST for an arbitrary constant $c > 1$.

The first polylogarithmic approximation algorithm for the GROUP-STEINER-TREE problem was achieved in the elegant work of Garg, Konjevod and Ravi [GKR00]. A brief sketch of their $O(\log n \log \log n \log N \log k)$–approximation algorithm, where $n = |V|$ and $N = \max_i |g_i|$, is as follows. First, the powerful results of [Bar98] are used to reduce the problem to the case where G is a *tree* T, with an $O(\log n \log \log n)$ factor loss in the approximation ratio. T can be furthermore assumed to be a c-HST for any desired constant $c > 1$. Next, solve the flow-based LP relaxation on T and round the fractional solution into an integral solution for T by applying a novel randomized rounding approach that is developed in [GKR00]. It is established in [GKR00] that for any tree T, this randomized rounding leads to an $O(\log N \log k)$–approximation. Thus, for the input graph G, we get an $O(\log n \log \log n \log N \log k)$–approximation. From a technical viewpoint, one of the main difficulties in [GKR00] is that a non-trivial analysis of the randomized process is required. The analysis uses Janson's inequality in an interesting way. The work of [GKR00] has been extended and expanded in several ways: Their algorithm was derandomized in [CCGG98, Sri01]; an alternative (combinatorial) algorithm is devised in [CEK06]; the loss incurred by the reduction to an HST is improved to $O(\log n)$ in [FRT04].

Since the first appearance of a polylogarithmic approximation for GROUP-STEINER-TREE (in the conference version of [GKR00] in 1998), there has been much interest in whether the approximation ratio can be improved. One concrete notable question in this regard has been the following: Can we achieve an approximation ratio better than $O(\log N \log k)$ for trees? This is interesting for at least two reasons. First, since [GKR00] shows a reduction to the case of trees as seen above, an improved approximation for trees (or even for the case of c-HSTs for some constant $c > 1$) would directly lead to an improved approximation for general graphs. Further, even the case where G is a star (which is a tree) captures the SET-COVER problem, for which $o(\log k)$–approximation is hard [LY94, Fei98, RS97], so there is an intriguing gap even on trees.

Our main technical result is that the integrality ratio of the flow-based relaxation for HSTs is $\Omega(\log^2 k)$. This bound is in fact tight – an $O(\log^2 k)$ bound on the integrality ratio holds for HSTs, as we show in Section 2.7. Both bounds hold for c-HSTs where $c > 1$ is any fixed constant. Recall that the upper bound of [GKR00] for trees in general is $O(\log N \log k)$; our methods show an $\Omega(\log N \log k / \log \log N)$ lower bound on the integrality ratio, even for a class of HSTs. The same lower bounds hold also for trees where all weights are the same (i.e., unit-weight trees). Finally, we show randomized rounding algorithms for the flow-based relaxation that lead to improved approximation algorithms for certain special families of HSTs; this sheds light on the type of instances that are most difficult to approximate.

A log-squared lower bound on the integrality ratio for trees had been conjectured circa 1998 by Uri Feige. The specific (randomly constructed) instance he suggested for this purpose (see Section 2 for more details) has proven to be quite difficult to analyze. Our lower bound is shown via a slightly different random instance, which eliminates one source of correlation in the random choices, and makes the construction more amenable to analysis. Nevertheless, we are required to do intricate estimates that delve into low-order terms, in particular when crafting the precise induction hypothesis lying at the heart of the proof.

(b). The Directed Steiner Tree problem. This is the directed version of the (undirected) STEINER-TREE problem. Given an edge-weighted directed graph $G = (V, E)$ that specifies a *root* vertex r and k *terminal* nodes $v_1, v_2, \ldots, v_k$, the goal is to construct a minimum-weight out-branching rooted at r, which spans all the terminals. 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. Aside from intrinsic interest, this problem is also of current interest, e.g., in the context of multicasting in the Internet (where inter-node distances are often not symmetric). The polynomial-time approximation ratio currently known for this problem is k^ϵ, for any constant $\epsilon > 0$ [CCC⁺99]; their algorithm extends to a polylogarithmic approximation ratio in quasi-polynomial running time. The flow-based relaxation here is similar: install for every edge $e \in E$ a capacity $x_e \in [0, 1]$, so that a unit of flow can be shipped from r to v_i, separately for any given i. Intriguingly, it was recently shown in [ZK02] that this relaxation has an integrality ratio of $\Omega(\sqrt{k})$, precluding a polylog(k)–approximation algorithm based on this relaxation. However, the examples constructed in [ZK02] have $k = \Theta(\frac{\log^2 n}{(\log \log n)^2})$; hence, the result of [ZK02] does not exclude an $O(\log n)$ integrality ratio. Our GROUP-STEINER-TREE lower-bound result above proves also an $\Omega(\frac{\log^2 n}{(\log \log n)^2})$ lower bound on the integrality gap for the DIRECTED-STEINER-TREE problem. The only lower bound known previously for DIRECTED-STEINER-TREE was $\Omega(\log n)$, since this problem generalizes SET-COVER.

As mentioned above, our results have paved the way to the improved hardness of approximation results of [HK03], which show that, for any fixed $\epsilon > 0$, GROUP-STEINER-TREE cannot be approximated within ratio $\log^{2-\epsilon} k$ and DIRECTED-STEINER-TREE cannot be approximated within ratio $\log^{2-\epsilon} n$, unless NP has quasi-polynomial time Las-Vegas algorithms. The influence of our work on these hardness results is threefold. First, our lower bounds on the integrality ratio has motivated working on a hardness of approximation result. Second, the insights our analysis provides regarding the (edge-weight) structure of instances that are difficult to approximate inspired specific details of the hardness reduction. Third, our main technical lemma (whose proof is rather non-trivial) is in fact made crucial use of in [HK03].

Organization. Our lower bounds on the integrality ratio of Group-Steiner-Tree and Directed-Steiner-Tree are shown in Section 2. We then prove algorithmically that the integrality ratio of the former problem is much better for certain families of instances in Section 3; this pinpoints the type of instances that appear difficult for this relaxation. Finally, concluding remarks are made in Section 4.

2 Lower bounds on the integrality ratio

In this section we prove a lower bound of $\Omega(\log^2 k)$ on the integrality ratio of the flow-based relaxation of the Group-Steiner-Tree problem even on HSTs. In terms of n, the gap is $\Omega(\frac{\log^2 n}{(\log\log n)^2})$. We start (Section 2.1) by describing the linear programming relaxation and constructing a family of 2-HST instances, accompanied by an overview of the analysis; then, the main technical parts (Sections 2.2 and 2.3) analyze the fractional and the integral solutions of this linear program. We also show how simple modifications to this construction extend the integrality ratio to unit-weight trees (Section 2.4) and to c-HSTs for an arbitrary constant $c > 1$ (Section 2.5). We further point out how this immediately leads to a lower bound of $\Omega(\frac{\log^2 n}{(\log\log n)^2})$ on the integrality ratio for the Directed-Steiner-Tree problem (Section 2.6). Finally, the lower bound of $\Omega(\log^2 k)$ is shown to be tight (in Section 2.7).

2.1 The relaxation and the instance

The cut-based relaxation (that is equivalent to the flow-based relaxation) for Group-Steiner-Tree is as follows. (Here, $\delta(S)$ is the set of edges with exactly one endpoint in $S \subset V$.)

$$\begin{array}{lll} \text{Minimize} & \displaystyle\sum_{e\in E} w_e x_e & \\ & \displaystyle\sum_{e\in\delta(S)} x_e \geq 1 \quad , & \forall S \subseteq V \text{ s.t. } r \in S \text{ and } S \cap g_j = \emptyset \text{ for some group } g_j \\ & 0 \leq x_e \leq 1 \quad , & \forall e \in E \end{array} \tag{1}$$

Let $\mathbb{T}_n$ be a 2-HST tree with n nodes defined by the following random process. Let the collection of groups be $\mathcal{G} = \{g_1, g_2, \ldots, g_k\}$. The groups are defined by a random process. The value of k, as well as those of two other parameters H and d, will be defined shortly (in terms of n). The height (i.e., depth) of $\mathbb{T}_n$ is H, and every non-leaf vertex has d children. The root of $\mathbb{T}_n$ is denoted r. The *level* of a vertex is its depth; r is at level 0, and there are $H+1$ levels. An edge is said to be at level i iff it connects a vertex at level $i-1$ to a vertex at level i. Each edge at level i has weight $1/2^i$; thus, for instance, edges incident at r have weight 1/2. Each group g_j is a subset of the leaves, described as follows. We shall associate a subset $A(\ell) \subseteq \mathcal{G}$ of the groups with each leaf ℓ, and define each group g_j to be the set of leaves ℓ for which $g_j \in A(\ell)$. Thus, a solution that reaches a leaf ℓ by a path from r, covers all groups in $A(\ell)$. To define $A(\ell)$ for each leaf ℓ, we now recursively and randomly define a set $A(v)$ for every node v in the tree (including non-leaf nodes), as follows. Proceed *independently* for each group g_j as follows. We start by letting $g_j \in A(r)$ with probability 1. In general, if $g_j \in A(u)$ for some non-leaf node u, then for each child v of u, we independently put g_j in $A(v)$ with probability 1/2. Thus, this random process goes top-down in the tree, independently for each group. Note that the number of vertices in $\mathbb{T}_n$ is $n \simeq d^H$, where H is the height of the tree. Clearly, the expected size of every group is $d^H/2^H$.

Parameters and notation. We set $d = c_0 \log n$ for some universal constant $c_0 > 0$; this will be used in some Chernoff bound arguments in Section 2.2. It then follows that $H = \frac{\log n}{\log d} = \Theta(\log n/\log\log n)$. We further set $k = 2^{2H}$; thus, $\log k = \Theta(\log n/\log\log n)$. Throughout, *with high probability* means with probability that is at least, say, $1 - 1/n$. All probabilities refer to the randomness in constructing the instance $\mathbb{T}_n$.

Feige's instance. The construction suggested by Feige circa 1998 is the following: Take a complete tree of arity 4 (i.e., every non-leaf vertex has 4 children) and height $\log_2 k$; now generate k groups, each containing k leaves, by an independent randomized branching process that starts from the root and randomly picks two out of four children until the leaves are reached. This random instance differs from $\mathbb{T}_n$ in its degree, which is constant rather than logarithmic, and in that the choices made (when generating a single group) at each child of a vertex are correlated (rather than independent).

Overview of the analysis. We first show (Section 2.2) that with high probability the instance $\mathbb{T}_n$ has a feasible fractional solution of cost $O(H)$. In this solution, all edges e in the same level of the tree are assigned the same value x_e, and this value is chosen so that the total cost of every level in the tree is the same, namely, $O(1)$. The feasibility of this solution is shown by employing a sequence of Chernoff bound arguments, and this is the reason why the degree d of the tree must be (at least) logarithmic. This is in contrast to Feige's instance, where the feasibility (of a similar fractional solution) is guaranteed by construction, i.e., with probability 1, and thus the tree can have a constant degree.

We then show (Section 2.3) that with high probability the cost of any integral solution for $\mathbb{T}_n$ is lower bounded by $\Omega(H^2 \log k)$. At a high level, we imitate the argument known for Set-Cover; we show that for any single low-cost integral solution (i.e., subtree of $\mathbb{T}_n$), with high probability over the randomness in constructing the groups this solution is infeasible (i.e., does not cover all the groups), and then we take a union bound over all possible integral solutions. In fact, any single vertex of $\mathbb{T}_n$ together with its children is essentially a "standard" Set-Cover instance with integrality ratio $\Omega(\log k)$.

The main technical work is to estimate the probability that an arbitrary (but fixed in advance) integral solution is feasible. Unlike in the Set-Cover scenario, where this is a straightforward calculation, in the Group-Steiner-Tree instance $\mathbb{T}_n$ the solution's structure comes into play. For instance, the integral solution might not be a regular subtree of $\mathbb{T}_n$, and its cost need not be split evenly among the different levels (of $\mathbb{T}_n$). We prove some upper bound on the probability that this solution for $\mathbb{T}_n$ is feasible. Our analysis shows that to maximize this upper bound on the probability, it is essentially best to have an even "split" of the cost used under a vertex v (i.e., the edges of the solution that belong to the subtree rooted at v). While we do not claim that this is the worst-case for the *exact* probability of feasibility, our upper bound gives a good enough estimate. However, the analysis does not specify how many children of v should have a non-zero cost under them; in fact, this value is very sensitive to lower-order tradeoffs between different levels in the tree.

The main difficulty in the analysis is to distill the effect of H, the height of the tree, on the feasibility probability. Our proof is by induction on the height of the tree, and uncovers a very delicate tradeoff between the height of a subtree and its cost. This tradeoff eventually translates to the cost of the integral solution for $\mathbb{T}_n$ having, on top of the $\log k$ term which comes from Set-Cover (i.e., a single level), also a linear dependence on the height H. Due to seemingly technical limitations this proof works only for $H \leq \frac{1}{2} \log k$, but we show in Section 2.7 that this is unavoidable. Interestingly, there is an analogy with the approximation algorithm of [GKR00], whose rounding procedure pays, at some intermediate stage, an $O(H \log k)$ factor, and then shows that, in effect, H can be upper bounded by $O(\log N)$.

2.2 The fractional solution

Recall that $d = c_0 \log n$. We start with a couple of propositions which show that if the constant c_0 is sufficiently large, then certain quantities related to our randomly chosen groups stay close to their mean. Henceforth, the phrase "with high probability" will mean "with a probability of $1 - o(1)$".

Proposition 2.1. *Let c_0 be a sufficiently large constant. Then, with high probability, all groups have size at least $(d/2)^H/3$.*

Proof. Fix j. We now show that if c_0 is large enough, then $\Pr\left[|g_j| < (d/2)^H/3\right] \leq 1/n^2$. We may then apply the union bound over all j to conclude the proof.

Let $\delta = 1/4$. Let X_1 be the number of vertices u at level 1 (i.e., children of r) such that $g_j \in A(u)$. Then X_1 has Binomial distribution $X_1 \sim B(d, 1/2)$, so by a Chernoff bound on the lower-tail (see e.g. [MR95]),

$$\Pr\left[X_1 \leq (1-\delta)\frac{d}{2}\right] \leq e^{-\frac{1}{2}\delta^2\cdot\frac{d}{2}}.$$

Let X_2 be the number of vertices u at level 2 such that $g_j \in A(u)$. Then X_2 has binomial distribution $X_2 \sim B(X_1 \cdot d, 1/2)$. Suppose that $X_1 > (1-\delta)\mathbb{E}[X_1] = (1-\delta)\frac{d}{2}$. Then, it is immediate that $X_2 \mid (X_1 > (1-\delta)\frac{d}{2})$ stochastically dominates a random variable $X_2' \sim B((1-\delta)\frac{d}{2}\cdot d, 1/2)$, i.e., $\Pr[X_2 \leq t] \leq \Pr[X_2' \leq t]$ for all t. By applying the Chernoff bound on X_2' we get

$$\Pr\left[X_2' \leq (1-\frac{\delta}{2})(1-\delta)(\frac{d}{2})^2\right] \leq e^{-\frac{1}{2}(\frac{\delta}{2})^2\cdot(1-\delta)(\frac{d}{2})^2}$$

Continue similarly for $i = 3, \ldots, H$, by defining X_i to be the number of vertices u at level i such that $g_j \in A(u)$, and by assuming that $X_{i-1} > (1-\frac{\delta}{2^{i-2}})\cdot\ldots\cdot(1-\frac{\delta}{2})(1-\delta)(\frac{d}{2})^i$. We get by the Chernoff bound that

$$\Pr\left[X_i' \leq (1-\frac{\delta}{2^{i-1}})\cdot\ldots\cdot(1-\frac{\delta}{2})(1-\delta)(\frac{d}{2})^i\right] \leq e^{-\frac{1}{2}(\frac{\delta}{2^{i-1}})^2\cdot(1-\frac{\delta}{2^{i-2}})\cdot\ldots\cdot(1-\frac{\delta}{2})(1-\delta)(\frac{d}{2})^i}. \tag{2}$$

For any $0 < \delta' \leq \frac{1}{2}$ we have $1-\delta' \geq \frac{1}{1+2\delta'} \geq e^{-2\delta'}$. Thus, $(1-\frac{\delta}{2^{i-1}})\cdot\ldots\cdot(1-\frac{\delta}{2})(1-\delta) \geq e^{-\frac{\delta}{2^{i-2}}-\ldots-\delta-2\delta} \geq e^{-4\delta} > \frac{1}{3}$. It follows that the tail-bound obtained in the righthand side of (2) is at most $e^{-\Omega((d/8)^i)}$. Applying the union bound on these H events we get that with high probability none of them happens (if the constant c_0 is sufficiently large), and in particular, $X_H \geq (1-\frac{\delta}{2^{H-1}})\cdot\ldots\cdot(1-\frac{\delta}{2})(1-\delta)(\frac{d}{2})^H \geq \frac{1}{3}(\frac{d}{2})^H$. This concludes the proof of Proposition 2.1. □

The following proposition has a similar proof; the main difference is that we will now employ Chernoff bounds on the upper-tail.

Proposition 2.2. *Suppose that the constant c_0 is large enough. Then with high probability, the following holds for every level i and every group g_j: If a vertex u at level i is such that $g_j \in A(u)$, then the number of leaves ℓ in the subtree rooted at u which satisfy $g_j \in A(\ell)$, is at most $3(d/2)^{H-i}$.*

Proof. Fix a pair (i, j) and a vertex u at level i such that $g_j \in A(u)$. Let $L(u)$ be the set of leaves of the subtree rooted at u. We now show that if c_0 is large enough, then $\Pr\left[|g_j \bigcap L(u)| > 3(d/2)^{H-i}\right] \leq 1/n^3$. We then apply a union bound over all (i, j, u) to conclude the proof.

Let $\delta = 1/4 < \frac{\ln 3}{2}$. Let X_1 be the number of vertices v at level 1 of the subtree rooted at u (i.e., children of u) such that $g_j \in A(v)$. Then X_1 has Binomial distribution $X_1 \sim B(d, 1/2)$, so by a Chernoff bound on the upper-tail (see e.g. [MR95]),

$$\Pr\left[X_1 \geq (1+\delta)\frac{d}{2}\right] \leq e^{-\frac{\delta^2}{3}\cdot\frac{d}{2}}.$$

Let X_2 be the number of vertices v at level 2 of the subtree rooted at u such that $g_j \in A(v)$. Then X_2 has binomial distribution $X_2 \sim B(X_1 \cdot d, 1/2)$. Suppose that $X_1 < (1+\delta)\mathbb{E}[X_1] = (1+\delta)\frac{d}{2}$. Then, it is immediate that $X_2 \mid (X_1 < (1+\delta)\frac{d}{2})$ is stochastically dominated by a random variable $X_2' \sim$

$B((1+\delta)\frac{d}{2} \cdot d, 1/2)$, i.e., $\Pr[X_2 \geq t] \leq \Pr[X_2' \geq t]$ for all t. By applying the Chernoff bound on X_2' we get

$$\Pr\left[X_2' \geq (1+\frac{\delta}{2})(1+\delta)(\frac{d}{2})^2\right] \leq e^{-\frac{1}{3}(\frac{\delta}{2})^2\cdot(1+\delta)(\frac{d}{2})^2}$$

Continue similarly for $l = 3, \dots, H-i$, by defining X_l to be the number of vertices v at level l of the subtree rooted at u such that $g_j \in A(v)$, and by assuming that $X_l < (1+\frac{\delta}{2^{l-1}}) \cdot \ldots \cdot (1+\frac{\delta}{2})(1+\delta)(\frac{d}{2})^l$. We get by the Chernoff bound that

$$\Pr\left[X_l' \geq (1+\frac{\delta}{2^{l-1}}) \cdot \ldots \cdot (1+\frac{\delta}{2})(1+\delta)(\frac{d}{2})^l\right] \leq e^{-\frac{1}{3}(\frac{\delta}{2^{l-1}})^2\cdot(1+\frac{\delta}{2^{l-2}})\cdot\ldots\cdot(1+\frac{\delta}{2})(1+\delta)(\frac{d}{2})^l}. \tag{3}$$

The tail-bound obtained in the righthand side of (3) is clearly at most $e^{-\Omega((d/8)^l)}$. Applying the union bound on these $H-i \leq H$ events we get that with probability at least $1-1/n^3$ none of these events happen, if the constant c_0 is sufficiently large; in particular, $X_{H-i} \leq (1+\frac{\delta}{2^{H-i-1}})\cdot\ldots\cdot(1+\frac{\delta}{2})(1+\delta)(\frac{d}{2})^{H-i} \leq 3(\frac{d}{2})^{H-i}$. (This is because of the following. For any δ' we have $1+\delta' \leq e^{\delta'}$. Thus, $(1+\frac{\delta}{2^{l-1}}) \cdot \ldots \cdot (1+\frac{\delta}{2})(1+\delta) \leq e^{\frac{\delta}{2^{l-1}}+\ldots+\frac{\delta}{2}+\delta} \leq e^{2\delta} < 3$.) This concludes the proof of Proposition 2.2. □

We now upper bound the value of LP (1) for the tree $\mathbb{T}_n$ by exhibiting a feasible solution for it: Let each edge e at each level i have value $\hat{x}_e = 9 \cdot (2/d)^i$.

Lemma 2.3. *With high probability, $\hat{x}$ is a feasible solution to LP (1). Its value is $9H$.*

Proof. Observe that $\hat{x}$ satisfies the constraints of LP (1) if (see also [GKR00]), for every group g_j, every cut $(S, \bar{S})$ separating r from all the vertices of g_j has capacity at least 1, where the capacity of each edge e is $\hat{x}_e$. By the (single-source) max-flow min-cut theorem (or, say, weak duality) it suffices to show that for every group g_j, a unit of flow can be shipped from the root r to the vertices of g_j while obeying the "capacity" $\hat{x}_e$ of each edge e. To this end, fix a group g_j and define the flow f as follows. For every vertex v in g_j (i.e., for every leaf v such that $g_j \in A(v)$), ship $3 \cdot (2/d)^H$ units of flow along the unique simple path from r to v. By Proposition 2.1, the total flow shipped to g_j is at least $|g_j| \cdot 3 \cdot (2/d)^H \geq 1$ with high probability. Next, consider an edge connecting a node u at some level i to its parent. If $g_j \notin A(u)$, no flow is shipped through this edge; if $g_j \in A(u)$, the total flow shipped through this edge (i.e., through u) is, by Proposition 2.2, at most $3(d/2)^{H-i} \cdot 3(2/d)^H = 9(2/d)^i$ with high probability. In both cases, the flow along the edge obeys the edge's capacity. We conclude that with high probability $\hat{x}$ is a feasible solution to LP (1).

The value of the solution $\hat{x}$ is $\sum_{i=1}^{H} d^i \cdot 1/2^i \cdot 9(2/d)^i = 9H$ since each level i contains d^i edges of weight $1/2^i$. □

2.3 The integral solution

We now show that with probability $1-o(1)$ (over the random choice of the groups), all integral solutions have value $\Omega(H^2 \log k)$. Whenever we say that some T' is a subtree of $\mathbb{T}_n$, we allow T' to be an arbitrary connected subgraph of $\mathbb{T}_n$. Since $\mathbb{T}_n$ is rooted, any subtree T' of $\mathbb{T}_n$ is also thought of as rooted in the obvious way: the node in T' of the smallest depth is the root of T' (and is denoted $\mathrm{root}(T')$). Also, when we say that some T' is a subtree of $\mathbb{T}_n$ with root u, we allow T' to be an arbitrary connected subgraph of $\mathbb{T}_n$ with root u.

Let $M(c)$ be the number of subtrees of $\mathbb{T}_n$ which are rooted at r and have total weight at most c. Fix $g_j \in \mathcal{G}$. For any given subtree T' of $\mathbb{T}_n$, let $p_j(T')$ be the probability that no leaf of T' belongs to the group g_j, conditioned on the event that $g_j \in A(\mathrm{root}(T'))$. Since by symmetry $p_j(T') = p_i(T')$ for all i, j, we will simply denote it by $p(T')$. We now define a key value $f(H,i,c)$ as follows. Choose an arbitrary vertex u at level i. Then $f(H,i,c)$ is the minimum value of $p(T')$, taken over all possible subtrees T' that are rooted at u and have total weight at most c. (If there is no such T', then $f(H,i,c)=1$. Also, it is easy to see by symmetry that $f(H,i,c)$ does not depend upon the choice of j or u.) Let P_c be the probability that there exists an integral solution of weight c. We wish to show that $P_c = o(1)$ for c that is smaller than a certain threshold of the order $H^2 \log k$. Using the independence between the different groups and applying a union bound over all possible subtrees rooted at r that have total weight c, we obtain

$$P_c \leq M(c)(1-f(H,0,c))^k. \tag{4}$$

We now have to lower bound f and upper bound M. We employ the following crude bound on $M(c)$. Note that it suffices to count only subtrees of $\mathbb{T}_n$ that span distinct sets of leaves (since the groups g_i contain only leaves). Observing that $\mathbb{T}_n$ has d^H leaves, and a subtree of total weight at most c spans at most $c2^H$ leaves (since each spanned leaf requires a distinct edge at level H), we get that

$$M(c) \leq \binom{d^H}{c2^H} \leq d^{cH2^H}. \tag{5}$$

Our goal in the next subsection is to prove the following key lemma.

Lemma 2.4. *For $H \leq \frac{1}{2}\log k$ and a constant $\gamma > 0$ that is sufficiently large, we have $f(H,0,c) \geq e^{-\gamma c/H^2}$. Thus $P_c \leq M(c)(1-f(H,0,c))^k \leq M(c)\cdot \exp\{-k\cdot e^{-\gamma c/H^2}\}$.*

2.3.1 Bounding $f(H,0,c)$.

We start with some preliminaries. The main technical result is Lemma 2.7 below. It gives a more general bound for f than stated in Lemma 2.4, and hence the proof of the latter would follow quite easily.

Proposition 2.5. *Let $l \geq 2$ and $\beta > 0$. Then the minimum of $\sum_{S\subseteq\{1,\dots,l\}}\prod_{i\in S} e^{-\beta x_i}$ over all $(x_1,\dots,x_l)$ with a given $\sum_{i=1}^{l} x_i$ is attained when all x_i are equal.*

Proof. The minimum is clearly attained at some point $(x_1,\dots,x_l)$, so assume to the contrary that at this point not all x_i are equal, say without loss of generality that $x_1 > \sum_i x_i/l > x_2$. We will show that changing both x_1 and x_2 to $\frac{x_1+x_2}{2}$ decreases the above sum while maintaining $\sum_i x_i$, which contradicts the assumption that $(x_1,\dots,x_l)$ is a minimum point. Actually, it suffices to prove that

$$\sum_{S'\subseteq\{1,2\}}\prod_{i\in S'} e^{-\beta x_i} > \sum_{S'\subseteq\{1,2\}}\prod_{i\in S'} e^{-\beta\cdot\frac{x_1+x_2}{2}}\ , \tag{6}$$

since multiplying (6) by $\prod_{i\in S''} e^{-\beta x_i}$ and summing over all $S'' \subseteq \{3,\dots,l\}$ shows that changing x_1, x_2 indeed decreases the above-mentioned sum. To prove (6), observe that it simplifies to

$$e^{-\beta x_1} + e^{-\beta x_2} > 2e^{-\beta(x_1+x_2)/2}$$

which follows from the arithmetic mean-geometric mean inequality since $x_1 \neq x_2$. This completes the proof of Proposition 2.5. □

It is easy to check (by considering higher derivatives) that for all $B \geq 0$,

$$e^{-B} \geq 1 - B + \frac{B^2}{2} - \frac{B^3}{6}. \tag{7}$$

Proposition 2.6. *For all $B_0 > 0$ there exists $\delta > 0$ such that for all $B \geq B_0$ we have $\frac{1+e^{-B}}{2} \geq e^{-\frac{B}{2+\delta}}$.*

Proof. Fix $B_0 > 0$. We first make sure that the inequality holds at B_0. By the arithmetic mean-geometric mean inequality $\frac{1+e^{-B_0}}{2} > e^{-B_0/2}$ (since $B_0 > 0$) so a sufficiently small $\delta > 0$ satisfies $\frac{1+e^{-B_0}}{2} > e^{-B_0/(2+\delta)}$. It now suffices to make sure that for all $B \geq B_0$ the derivative of the lefthand side is at least that of the righthand side, i.e., that $-\frac{1}{2}e^{-B} \geq -\frac{1}{2+\delta}e^{-B/(2+\delta)}$. This holds for any $0 < \delta < B_0$ since $\frac{2+\delta}{2} = 1 + \delta/2 \leq 1 + B_0/2 \leq e^{B_0/2} \leq e^{B/2} \leq e^{B-B/(2+\delta)}$, completing the proof of Proposition 2.6. □

Lemma 2.7. *Let γ be a sufficiently large constant. Then $f(H,h,c) \geq \exp(-\frac{\gamma c 2^h}{(H-h)^2})$ for all $c > 0$ and all $0 \leq h \leq H-1$.*

Proof. Fix B_0 to be an arbitrary positive constant, and let $\delta > 0$ be the corresponding constant from Proposition 2.6.

The proof is by backward induction on h, i.e., we assume that the claim holds for $h+1$ and prove it for h, where $h \leq H-2$. We will consider the base case, which is the case that $h \geq H - 1 - \frac{6}{\delta}$, later on. In order to bound f, we derive a recurrence relation for $f(H,h,c)$. Recall the definition of $f(H,h,c)$: fix an arbitrary vertex u at level h, and take the minimum value of $p(T')$, over all possible subtrees T' rooted at u such that the total weight of T' is at most c. We bound $f(H,h,c)$ by considering all possibilities of u having $l = 1,2,\ldots,d$ children and all possible partitions $\vec{x}^{(l)} = (x_1, x_2, \ldots, x_l)$ of the weight c to (the subtrees under) these l children; since the edge from u to each of its children has weight $\frac{1}{2^{h+1}}$, we get that $\sum_{i=1}^{l} x_i = c - \frac{l}{2^{h+1}}$. We then get that

$$f(H,h,c) \geq \min_{1\leq l\leq d}\{ \min_{\vec{x}^{(l)}: x_i\geq 0, \sum_i x_i = c - \frac{l}{2^{h+1}}} \{\frac{1}{2^l} \sum_{S\subseteq\{1,\ldots,l\}} \prod_{i\in S} f(H,h+1,x_i)\}\};$$

since once the l children of u are chosen, we only need to consider the subset S of all children with g_j in their $A(\cdot)$ set. (Each such set S occurs with probability $1/2^l$.) Plugging the induction hypothesis in, we get that

$$f(H,h,c) \geq \min_{1\leq l\leq d}\{ \min_{\vec{x}^{(l)}: x_i\geq 0, \sum x_i = c - \frac{l}{2^{h+1}}} \{\frac{1}{2^l} \sum_{S\subseteq\{1,\ldots,l\}} \prod_{i\in S} \exp(-\frac{\gamma x_i 2^{h+1}}{(H-h-1)^2})\}\}. \tag{8}$$

For any l, we have by Proposition 2.5 that the righthand side of (8) is minimized when all x_i are equal to $\frac{c}{l} - \frac{1}{2^{h+1}}$. We thus get that

$$\begin{aligned} f(H,h,c) &\geq \min_{1\leq l\leq d} \frac{1}{2^l} \sum_{S\subseteq\{1,\ldots,l\}} \left(\exp\left(-\frac{\gamma(\frac{c}{l} - \frac{1}{2^{h+1}})2^{h+1}}{(H-h-1)^2}\right)\right)^{|S|} \\ &= \min_{1\leq l\leq d} \frac{1}{2^l} \sum_{i=0}^{l} \binom{l}{i} \left(\exp\left(-\frac{\gamma(\frac{c}{l} - \frac{1}{2^{h+1}})2^{h+1}}{(H-h-1)^2}\right)\right)^{i} \\ &= \min_{1\leq l\leq d} \left(\frac{1 + \exp\left(-\frac{\gamma(\frac{c}{l} - \frac{1}{2^{h+1}})2^{h+1}}{(H-h-1)^2}\right)}{2}\right)^{l}. \end{aligned}$$

Fix l arbitrarily such that $1 \leq l \leq d$. Let $B = \dfrac{\gamma(\frac{c}{l} - \frac{1}{2^{h+1}})2^{h+1}}{(H-h-1)^2}$ and $C = \dfrac{\gamma\frac{c}{l}2^h}{(H-h)^2}$. To complete the induction, we want to prove that $\left(\frac{1+e^{-B}}{2}\right)^l \geq e^{-Cl}$, i.e., that

$$\frac{1+e^{-B}}{2} \geq e^{-C}. \tag{9}$$

We have four cases.

Case 1: In this case we assume that $C \geq \frac{B}{2}$. By the arithmetic mean-geometric mean inequality we have that $\frac{1+e^{-B}}{2} \geq e^{-B/2} \geq e^{-C}$, which proves (9).

Case 2: In this case we assume that $B \geq B_0$ and $\frac{B}{2+\delta} \leq C \leq \frac{B}{2}$; recall that δ is the constant from Proposition 2.6. Then we have from Proposition 2.6 that $\frac{1+e^{-B}}{2} \geq e^{-\frac{B}{2+\delta}} \geq e^{-C}$, which proves (9).

Case 3: In this case we assume that $C \leq \frac{B}{2}$ and $B < B_0$. Then by (7) we have $\frac{1+e^{-B}}{2} \geq 1 - \frac{B}{2} + \frac{B^2}{4} - \frac{B^3}{12}$. Since $C \geq 0$, we have (by Taylor's Theorem) that $e^{-C} \leq 1 - C + \frac{C^2}{2}$. Thus, it suffices to prove that

$$1 - \frac{B}{2} + \frac{B^2}{4} - \frac{B^3}{12} \geq 1 - C + \frac{C^2}{2}.$$

Since $B < B_0 \leq \frac{1}{2}$ we have that $\frac{B^3}{12} \leq \frac{B^2}{24}$, and then since $2C \leq B$, we have that $\frac{B^2}{4} - \frac{B^3}{12} \geq \frac{5B^2}{24} \geq \frac{5C^2}{6}$. It therefore suffices to prove that

$$C + \frac{C^2}{3} \geq \frac{B}{2}.$$

Note that

$$\frac{B}{2} - C \leq \frac{\gamma 2^{h+1}\frac{c}{l}}{(H-h-1)^2(H-h)} - \frac{\gamma}{2(H-h-1)^2}.$$

Plugging in the values of B and C and simplifying we get that it suffices to prove that

$$\frac{\gamma 2^{h+1}\frac{c}{l}}{(H-h-1)^2(H-h)} \leq \frac{\gamma}{2(H-h-1)^2} + \frac{\gamma^2 \frac{c^2}{l^2} 2^{2h}}{3(H-h)^4}.$$

If $\frac{2^{h+2}\frac{c}{l}}{H-h} \leq 1$, then the desired inequality indeed holds since $\frac{\gamma 2^{h+1}\frac{c}{l}}{(H-h-1)^2(H-h)} \leq \frac{\gamma}{2(H-h-1)^2}$. Otherwise, the inequality holds for any $\gamma \geq 96$, since then,

$$\frac{\gamma^2 \frac{c^2}{l^2} 2^{2h}}{3(H-h)^4} = \frac{\gamma}{6} \cdot \frac{2^{h+2}\frac{c}{l}}{H-h} \cdot \frac{\gamma \frac{c}{l} 2^{h-1}}{(H-h)^3} \geq 16 \frac{\gamma \frac{c}{l} 2^{h-1}}{(H-h)^3} \geq \frac{\gamma 2^{h+1}\frac{c}{l}}{(H-h-1)^2(H-h)}.$$

Case 4: In this case we assume that $C < \frac{B}{2+\delta}$. Note that for $h \leq H-2$,

$$2 + \delta \leq \frac{B}{C} = 2\frac{\frac{c}{l} - \frac{1}{2^{h+1}}}{\frac{c}{l}} \cdot \frac{(H-h)^2}{(H-h-1)^2} \leq 2\frac{(H-h)^2}{(H-h-1)^2} \leq 2 + \frac{6}{H-h-1}.$$

Thus, $h \geq H - 1 - \frac{6}{\delta}$. Since $\delta > 0$ is a constant, this is really the base case of the induction, which we shall prove directly. Consider a subtree T' of weight at most c that is rooted at a vertex u at level h. Since u has at most $c2^{h+1}$ children in T', each *not* having the group g_j in its $A(\cdot)$ set independently with probability 1/2, with probability at least $2^{-c2^{h+1}}$ the subtree T' does not cover g_j. Thus, $f(H, h, c) \geq e^{-c2^{h+1}}$. Choosing a constant $\gamma \geq 2(1+\frac{6}{\delta})^2$, we get that $\gamma \geq 2(H-h)^2$, and thus

$$f(H, h, c) \geq e^{-c2^{h+1}} \geq \exp(-\frac{c2^h \cdot \gamma}{(H-h)^2}).$$

This concludes the proof of Lemma 2.7. □

Proof of Lemma 2.4. Lemma 2.7 implies that $f(H,0,c) \geq e^{-\gamma c/H^2}$. Plugging this into (4), we get that

$$P_c \leq M(c)(1 - f(H,0,c))^k \leq M(c) \cdot \exp\{-k \cdot e^{-\gamma c/H^2}\}.$$

□

2.3.2 Bounding the weight of an integral solution.

By Lemma 2.4 in conjunction with (5), we have

$$P_c \leq \exp\{cH2^H \log d - ke^{-\gamma \frac{c}{H^2}}\}.$$

Now, suppose that $c \leq \frac{1}{4\gamma}H^2 \ln k$. Then $cH2^H = O(2^H H^3 \log k)$. Recalling that $H = \frac{1}{2}\log k$, we have

$$P_c \leq \exp\{\tilde{O}(\sqrt{k}) - \Omega(k^{3/4})\} = o(1).$$

We conclude that with high probability no subtree of weight at most $\frac{1}{4\gamma}H^2 \log k$ covers all the groups, and thus an optimal integral solution has value at least $\Omega(H^2 \log k)$. Since LP (1) has a fractional feasible solution of value $9H$, we get:

Theorem 2.8. *The integrality ratio of the relaxation (1) for* Group-Steiner-Tree *is* $\Omega(\log^2 k)$. *In terms of* N, k, *the integrality gap is* $\Omega(\log k \log N / \log\log N)$ *and in terms of* n *it is* $\Omega\left(\frac{\log^2 n}{(\log\log n)^2}\right)$.

2.4 Integrality ratio for unit-weight trees

The above analysis gives a lower bound on the integrality gap for Group-Steiner-Tree in HSTs. A consequent interesting question is whether the LP is tighter for unit-weight trees. We show that a slight modification of the trees described above gives the same integrality ratio lower bounds for unit-weight trees. The idea is very simple – recall that in our random construction $\mathbb{T}_n$, edges at level i had weight $1/2^i$; replacing each such edge by a path of 2^{H-i} unit-weight edges does not really change our integrality ratio argument, because the resulting instance $\mathbb{T}'_n$ is essentially equivalent to the instance $\mathbb{T}_n$ with edge weights scaled up by a factor of 2^H. Formally, it is easy to verify that our fractional solution for $\mathbb{T}_n$ naturally yields a fractional solution for $\mathbb{T}'_n$ with value $9H2^H$, and that an optimal integral solution for $\mathbb{T}'_n$ with value OPT' corresponds to an integral solution with value $\mathrm{OPT}'/2^H$ for $\mathbb{T}_n$. Since we know the latter value is at least $\Omega(H^2 \log k)$, we conclude that $\mathrm{OPT}' = \Omega(H^2 2^H \log k)$, and the integrality ratio of $\mathbb{T}'_n$ is $\Omega(H \log k) = \Omega(\log^2 k) = \Omega(\log k \log N / \log\log N)$. Furthermore, the total number of vertices in $\mathbb{T}'_n$ is at most $2^H n = O(n^2)$, and hence the integrality ratio is also $\Omega(\frac{\log^2 n}{(\log\log n)^2})$ in terms of the number of vertices in $\mathbb{T}'_n$.

2.5 Integrality ratio for c-HSTs

Straightforward modifications of our integrality ratio proof for Group-Steiner-Tree in 2-HSTs lead to the same lower bounds for c-HSTs, for arbitrary constant $c > 1$. Here, we take an alternative approach; instead of going through the whole proof, we show that our lower bounds for 2-HSTs imply (in a black-box manner) similar bounds for c-HSTs, where $c > 1$ is an arbitrary constant. We first consider the case $c > 2$ and then use it to handle the case $1 < c < 2$.

An arbitrary constant $c > 2$. In this case our 2-HST instance $\mathbb{T}_n$ can be modified into a c-HST instance $\mathbb{T}'_n$ as follows. The set of vertices of $\mathbb{T}'_n$ is a subset of the vertices of $\mathbb{T}_n$. For every $j = 0, 1, 2, \ldots$ iteratively (up to about $\log_c 2^H$), let $i = i(j)$ be the (unique) integer such that $2^i \le c^j < 2^{i+1}$, and include all the level i vertices of $\mathbb{T}_n$ as the level j vertices in $\mathbb{T}'_n$. For example, at $j = 0$, we include the root of $\mathbb{T}_n$ as the root of $\mathbb{T}'_n$ because $1 \le c^0 < 2$. We may assume that the height of $\mathbb{T}_n$ is chosen so that at some iteration j_0, we include in $\mathbb{T}'_n$ the leaves of $\mathbb{T}_n$ (i.e., $i(j_0) = H$), at which point the iterations are stopped. With this assumption (and since all the groups in $\mathbb{T}_n$ contain only leaves), we also get that $\mathbb{T}_n$ and $\mathbb{T}'_n$ have exactly the same k groups. Finally, two vertices at two consecutive levels $j-1, j$ in $\mathbb{T}'_n$ are connected by an edge of weight $1/c^j$ whenever, in $\mathbb{T}_n$, one of the two vertices is an ancestor of the other one. For example, the edges incident at the root of $\mathbb{T}'_n$ have weight $1/c$. Notice that $\mathbb{T}'_n$ is a c-HST with height $j_0 \simeq \log_c 2^H = H/\log_2 c$.

A fractional solution LP for $\mathbb{T}_n$, of value LP, say, naturally induces a fractional solution to $\mathbb{T}'_n$ with value at most LP. Indeed, we let the fractional value of an edge connecting a vertex u to its parent v' in $\mathbb{T}'_n$ be the equal to the fractional value of the edge connecting (the same vertex) u to its parent v in $\mathbb{T}_n$. It is easy to see that whenever this fractional solution for $\mathbb{T}'_n$ pays (fractionally) $1/c^j$ for an edge, the solution LP pays $1/2^i$ for the corresponding edge in $\mathbb{T}_n$, with $1/c^j \le 1/2^i$. Since the corresponding edges in $\mathbb{T}_n$ are distinct, the value of the constructed solution for $\mathbb{T}'_n$ is at most LP.

An optimal integral solution OPT$'$ for $\mathbb{T}'_n$, of value OPT', say, naturally induces an integral solution INT for $\mathbb{T}_n$ with value at most $O(c) \cdot$ OPT'. Indeed, simply take in $\mathbb{T}_n$ the (minimal) subtree that spans exactly the same leaves (that are spanned by the solution for $\mathbb{T}'_n$). It is easy to see that whenever OPT$'$ pays $1/c^j$ for an edge connecting a vertex u to its parent v' in $\mathbb{T}'_n$, the solution INT has to pay for the path between u and its ancestor v' in $\mathbb{T}_n$. The total weight of this path is at most $1/2^{i'} + 1/2^{i'-1} + \cdots + 1/2^i = O(1/2^{i'}) = O(c) \cdot 1/2^i = O(c) \cdot 1/c^j$. We thus get an integral solution for $\mathbb{T}_n$ with value $O(c) \cdot$ OPT'.

Combining these arguments with our bounds on the fractional and integral solutions for $\mathbb{T}_n$ yields a lower bound of $\Omega(\frac{1}{c} H \log k)$ on the integrality ratio in c-HSTs. Notice also that the number of vertices in $\mathbb{T}'_n$ is similar to that in $\mathbb{T}_n$ because they have the same leaves. For fixed $c > 2$, we thus get the same integrality ratio lower bounds in c-HSTs as in 2-HSTs, namely, $\Omega(\log^2 k) = \Omega(\log k \log N/\log\log N)$ and also $\Omega(\frac{\log^2 n}{(\log\log n)^2})$ in terms of the number of vertices in $\mathbb{T}'_n$.

An arbitrary constant $1 < c < 2$. Let t be the smallest integer such that $c^t > 2$, and define $q = 1 + c + c^2 + \ldots + c^{t-1}$. The above construction then yields a c^t-HST instance $\mathbb{T}'_n$. Now replace every edge of weight $1/(c^t)^j$ in $\mathbb{T}'_n$ with a path of t edges having weights $1/(qc^{tj-(t-1)}), 1/(qc^{tj-(t-2)}), \ldots, 1/(qc^{tj})$. Clearly, the resulting instance $\mathbb{T}''_n$ is a c-HST. Notice further that the total weight of the above t-path in $\mathbb{T}''_n$ is $(c^{t-1} + c^{t-2} + \ldots + 1)/(qc^{tj}) = 1/c^{tj}$, i.e., equal to the edge in $\mathbb{T}'_n$ it replaced. Hence, any fractional solution for $\mathbb{T}'_n$ yields a fractional solution for $\mathbb{T}''_n$ with the same value, and also an optimal integral solution in $\mathbb{T}'_n$ yields an integral solution for $\mathbb{T}''_n$ with the same value. Since the number of vertices in $\mathbb{T}''_n$ is larger than that in $\mathbb{T}'_n$ by only a constant factor of $t \le \log_c 4 = O(\frac{1}{c-1})$, we get the same integrality ratio lower bounds in c-HSTs as in 2-HSTs, namely, $\Omega(\log^2 k) = \Omega(\log k \log N/\log\log N)$ and also $\Omega(\frac{\log^2 n}{(\log\log n)^2})$ in terms of the number of vertices in $\mathbb{T}'_n$.

2.6 Integrality ratio for Directed Steiner Tree

The above results immediately lead to a lower bound of $\Omega(\frac{\log^2 n}{(\log\log n)^2})$ on the integrality ratio for the Directed-Steiner-Tree problem. Let $\mathbb{T}_n$ be an instance as described above with $\Omega(\frac{\log^2 n}{(\log\log n)^2})$ integrality ratio for Group-Steiner-Tree, and construct a Directed-Steiner-Tree instance as follows. Orient all the edges of $\mathbb{T}_n$ away from the root r. Then, introduce new nodes $v_1, v_2, \ldots, v_k$, and for each j and each $u \in g_j$, introduce a zero-weight arc from u to v_j. This defines a Directed-Steiner-Tree instance I which is essentially the same as I: fractional solutions for the two problems map bijectively, with identical total weights, and the same thing holds also for integral solutions. Observe that the number of vertices in the resulting graph is $n + k \leq 2n$, and thus the lower bound of $\Omega(\frac{\log^2 n}{(\log\log n)^2})$ on the integrality ratio for $\mathbb{T}_n$ holds also for I.

2.7 An $O(\log^2 k)$-approximation for Group Steiner Tree in HSTs

We now show a tight $O(\log^2 k)$-approximation algorithm for the Group-Steiner-Tree problem on HSTs. This was obtained jointly with Anupam Gupta and R. Ravi, and we thank them for allowing us to include this algorithm here.

Our algorithm uses the rounding procedure of [GKR00] as a subroutine and takes advantage of the geometrically-decreasing-weights property of HSTs. Let T be the HST instance of the Group-Steiner-Tree problem. We assume this tree is a 2-HST, i.e., the weight of each edge in the $(i+1)$-st level equals half the weight of its parent edge in the i-th level, and each edge in the first level has weight exactly *one*; the algorithm extends in a simple way to c-HSTs. We also assume that the height of the tree is $H > \log k$; otherwise, the approximation ratio $O(H \log k)$ in [GKR00] already implies the $O(\log^2 k)$ upper bound. It is not difficult to arrange that all members of each group are at the leaves of the HST (with only a constant factor increase in the optimum value). Our algorithm is as follows:

1. Create a new tree T' consisting of only the first $H' = \log k$ levels of T. For each leaf ℓ' in T' find its corresponding node ℓ in level $\log k$ of T, and assign to ℓ' all groups that appear in the subtree of rooted at ℓ in T.
2. Run the approximation algorithm of [GKR00] on the Group-Steiner-Tree instance T'. Let SOL' be the value of the solution obtained and let OPT' be the value of the optimal solution in T'. The analysis of [GKR00] shows that $\mathrm{SOL}' = O(H' \log k \cdot \mathrm{OPT}') = O(\log^2 k \cdot \mathrm{OPT}')$.
3. From the solution SOL' (which is a subtree of T'), we construct a solution SOL for T as follows:
   (a) Find the subtree S in T that corresponds to SOL', and include S in SOL.
   (b) For each group $g \in \mathcal{G}$, repeat the following steps. Find a leaf ℓ' in SOL' that belongs to g (there must be such a leaf since SOL' covers g), and let ℓ be the level H' vertex in T corresponding to ℓ'. Now find in the subtree under ℓ in T a leaf u that belongs to g (there must be such leaf because of the way we assigned groups in T'), and add to SOL the path that connects ℓ to its descendant u.

It is easy to verify that the above procedure produces a valid solution SOL to the 2-HST instance T. We claim that $\mathrm{SOL} = O(\log^2 k \cdot \mathrm{OPT})$, where OPT is the optimum solution value in T. Indeed, SOL consists of SOL' and at most k paths (one path per group) added in step 3(b). Because T is an HST, each of these paths has total weight $O(1/k)$ (since its edges have geometrically decreasing weights), and thus, $\mathrm{SOL} = \mathrm{SOL}' + O(1)$. Since the optimal solution must contain at least one edge in the first level (and thus having weight *one*), we get that $\mathrm{OPT} \geq 1$. It is also obvious that $\mathrm{OPT}' \leq \mathrm{OPT}$, and therefore, $\mathrm{SOL} = \mathrm{SOL}' + O(1) = O(\log^2 k \cdot \mathrm{OPT}') + O(1) = O(\log^2 k \cdot \mathrm{OPT})$.

3 Improved approximations for certain families of trees

To better understand the approximability of the Group-Steiner-Tree problem, one may consider the following question: What are the instances (in particular, trees) that are difficult to approximate better than within ratio $O(\log k \log N)$? We partially answer this question by presenting a significantly better approximation ratio for a certain family of trees, which differs from the trees constructed in Section 2 in a crucial way.

This improved approximation also sheds light on the instances $\mathbb{T}_n$ constructed in Section 2. For example, it may be tempting to believe, at a first glance, that the edge weights pose an unnecessary complication to $\mathbb{T}_n$. Notice that the *uniform weight* version of $\mathbb{T}_n$ (i.e., a tree similar to $\mathbb{T}_n$, except that all its edges have the same weight), has the same fractional solutions as $\mathbb{T}_n$. Furthermore, it can be verified that for both $\mathbb{T}_n$ and its uniform weight version, the fractional solution presented in Section 2.2 is, with high probability, near-optimal, and that applying to it the rounding procedure of [GKR00] yields an integral solution with value larger by a $\Theta(\log k \log N)$ factor than the relaxation value. However, in the uniform weight version of $\mathbb{T}_n$, the contribution of level-i edges to the relaxation value increases significantly with i; and as we shall soon see, this implies that an approximation ratio better than $O(\log k \log N)$ is possible. This explains why the weights in $\mathbb{T}_n$ are necessary – they make every level have the same contribution to the relaxation value. This also elucidates the disparity between the performance of a rounding procedure for a relaxation and the integrality ratio of the relaxation – the uniform weight version of $\mathbb{T}_n$ exhibits a large ratio according to the former measure, but a significantly smaller ratio according to the latter.

Technically, fix a Group-Steiner-Tree instance T on a tree of height H, and an optimal solution x_e to its flow-based relaxation LP (1). Define z_i^* to be the total contribution of the edges at level i (of T) to the objective function of the relaxation. We show that the relationship between the different z_i^* plays a crucial role in the strength/weakness of the LP: if for some constant $\alpha > 1$ we have $z_{i+1}^* \geq \alpha z_i^*$ for all i, then we can achieve an $O(\log k \cdot \log\log(kN)/\log\alpha) = O(\log k \cdot \log\log(kN))$–approximation. This may suggest that instances with $z_i^* \simeq z_{i+1}^*$ for all/most i are among the worst cases for the relaxation.

The following lemma proves the improved approximation ratio for the case where $\alpha = 2$. The argument easily extends to any constant factor $\alpha > 1$. We sometimes refer to a valid integral solution simply as a cover.

Lemma 3.1. *If $z_{i+1}^* \geq 2z_i^*$ for all i, then we can find an integral solution of value $O(z^* \cdot \log k \cdot \log\log(kN))$, where $z^* = \sum_{i=1}^H z_i^*$ denotes the optimal LP value.*

Before getting into the formal proof, let us outline the main ideas. We separate the tree T into a lower part, that contains the lowest $\Theta(\log\log(kN))$ levels of the tree, and an upper part, that contains the rest of the tree. Let $z^*(U), z^*(L)$ be the contributions of the upper and lower parts to the fractional solution value z^*. Notice that $z^*(U) \leq z^*/\text{polylog}(kN)$ since $z_{i+1}^* \geq 2z_i^*$ for all i. We can thus take care of the upper part as follows. We use the same randomized rounding as in [GKR00] for the upper part, only that now we repeat the process about $O(\log N)$ more times (multiplicatively) – this results in a solution that is considerably more expensive with respect to $z^*(U)$, but it is still not too much with respect to the total fractional solution z^*. Since we repeat the rounding procedure more times, we cover each group more "times" (in a way that is formalized in the proof). Now every leaf of the upper part that we managed to cover, can be regarded, in the lower part, as the root of a subtree. This allows us to apply the algorithm of [GKR00] to some of the subtrees in the lower part, namely, to those subtrees whose root was covered by our upper part solution. By the analysis of [GKR00] we only need to pay proportionally to the height of the lower part (times $O(\log k)$), i.e. the lower part solution has value $O(z^*(L) \log k \log\log(kN))$. In the case where $z_{i+1}^* \geq \alpha z_i^*$ for all i, we define the lower part to be the lowest $\Theta(\log\log(kN)/\log\alpha)$ levels of the tree,

Proof. We may assume that all groups contain only leaves of T, by adding zero weight edges. Let L_i be the set of edges at level i. Let $h = 2\log\log(kN)$. Let $U = \{e : e \in L_i \text{ for } i \leq H - h\}$ and $L = \{e : e \in L_i \text{ for } i > H - h\}$. For every $e \in U$, let y_e be x_e rounded upwards to the nearest power of 2, increasing the LP value by a factor of at most 2.

We first find a cover of U, as follows. Let $c_1 > 0$ be a suitably large constant. For every $e \in U$, assign $\hat{x}_e = \min\{1, x_e \cdot c_1 \log k \log^2(kN)\}$, and use one iteration of the rounding scheme presented in

[GKR00] (w.r.t. $\hat{x}_e$) to solve the problem in U. The expected total weight of this solution is at most $O(z^*(U) \log k \log^2(kN)) \leq O(z^* \log k)$, where $z^*(U) = \sum_{i=1}^{H-h} z_i^*$ is the total contribution to z^* of the edges in U.

Using arguments similar to those in [GKR00], we now wish to show that from the perspective of U, every group g is covered "sufficiently many times", with high probability. Let $e_1, \ldots, e_m$ be the leaves of (the subtree induced on) U that "lead" to g (i.e., g contains at least one of their descendants in T'). A unit amount of flow can be shipped in T from the root to g, under the LP values x_e (as capacities), because x_e is a feasible solution. Let $f_1, \ldots, f_m$ be the corresponding flows on the edges $e_1, \ldots, e_m$. Clearly, $\sum_j f_j = 1$. Partition the m flows, letting $A_i = \{j : \frac{1}{2^i} < f_j \leq \frac{1}{2^{i-1}}\}$. Let $B(g) = \{i : i \leq 2 \log N \text{ and } \sum_{j \in A_i} f_j > 1/4 \log N\}$. It is easy to see that the flow in the remaining sets A_i is at most $(\frac{2}{N^2} + \frac{2}{2N^2} + \cdots) + (2 \log N) \frac{1}{4 \log N} < 3/4$, and thus $\sum_{i \in B(g)} \sum_{j \in A_i} f_j \geq 1/4$. We can therefore ignore the remaining sets and focus on the flows in $B(g)$. Fix $i \in B(g)$ and let V_i be the set of leaves (of U) e_j, for $j \in A_i$, that are chosen by the [GKR00] procedure in U according to $\hat{x}_e$.

For the sake of upper bounding the lower-tail of $|V_i|$, we may assume that the capacity x_e on every edge e equals the total flow shipped along the edge e (since a larger capacity x_e just increases $|V_i|$). Thus, the expectation of $|V_i|$ is $\mu_i = \sum_{j \in A_i} \min\{1, f_j \cdot c_1 \log k \log^2(kN)\}$. If $\frac{c_1}{2^i} \cdot \log k \log^2(kN) \geq 1$, then $|V_i| = |A_i| = \mu_i$ with probability 1. Otherwise, $\mu_i \geq \sum_{j \in A_i} \frac{c_1 f_j}{2} \cdot \log k \log^2(kN) \geq \frac{c_1 \log k \log(kN)}{8}$, and by Janson's inequality [Jan90],

$$\Pr[|V_i| \leq \mu_i/2] \leq e^{-\Omega(\frac{\mu_i}{2+\Delta_i/\mu_i})},$$

where $\Delta_i = \sum_{e \sim e'} \Pr[e \text{ and } e' \text{ are chosen}]$; here, the sum is over pairs of distinct edges $e \in A_i$ and $e' \in A_i$ whose events of being chosen are not independent. By the proofs in [GKR00, KRS02], it is easy to see that $\Delta_i \leq O(\mu_i \log k)$, where the constant in the "$O(\cdot)$" is an absolute constant that is independent of c_1. Thus,

$$\Pr[|V_i| \leq \mu_i/2] \leq e^{-\Omega(\mu_i/\log k)} \leq e^{-\Omega(c_1 \log(kN))},$$

where the constants in the "$\Omega(\cdot)$" are absolute constants that are independent of c_1. There are only k groups, and for each one $|B(g)| \leq O(\log N)$. Thus, by choosing c_1 sufficiently large, we get by the union bound (over a polynomial in kN number of V_i's) that with high probability, for every group g and every $i \in B(g)$, $|V_i| \geq \mu_i/2$. Recall that at least 1/4 of the total flow in $f_1, \ldots, f_m$ is shipped through sets A_i with $i \in B(g)$, and that the flows among each such set A_i are all equal, up to a constant factor. Hence, at least $\Omega(1)$ of the unit amount of flow into g must be shipped using the leaves of U chosen by the [GKR00] procedure.

We next apply the rounding algorithm of [GKR00] to L with the values x_e, starting from every chosen vertex of U. Since we know from above that one can ship to any group g, an $\Omega(1)$ amount of flow from the level $H - h$ vertices chosen in the solution for U, we get that x_e satisfies the LP constraints, up to a constant factor. It is proven in [GKR00], that after $O(h \log k)$ iterations of the rounding scheme, with high probability all the groups are covered. We now claim that the expected cost of each such iteration is at most z^*. Indeed, the probability to choose and edge e is proportional to its fractional value x_e, and the claim follows by the linearity of expectation.

Therefore, the expected cost of this solution is $O(z^* h \log k + z^* \log k) = z^* \cdot O(\log k \cdot \log\log(kN))$. □

4 Discussion

Our results improve the current understanding of the integrality ratio of the flow-based relaxation for the Group Steiner Tree problem, but some very intriguing gaps still remain. Although for HSTs our $\Omega(\log^2 k)$ lower bound is tight, for general trees there is a slight slack between our $\Omega(\log k \log N / \log\log N)$

lower bound and the $O(\log k \log N)$ upper bound of [GKR00]. Interestingly, an $O(\log^2(kN)/\log\log(kN))$–approximation by a quasi-polynomial time algorithm is devised in [CEK06]; their algorithm is combinatorial (i.e., not LP-based). Does their algorithm hint that the known upper bound on the integrality ratio in trees is not tight? Or perhaps there is a separation between polynomial and quasi-polynomial (approximation) algorithms? A possible step towards closing this small gap (in the integrality ratio on trees) is to analyze the random instance suggested by Uri Feige (see our description in Section 2).

For general graphs, there is an even bigger slack, as the known upper bound is $O(\log n \log k \log N)$ [GKR00, FRT04] and the lower bound is just the lower bound for trees described above. It is worth noting that a significantly better upper bound can be achieved in (general) graphs of small diameter. In particular, an $O(\log k)$ upper bound for expander graphs is shown in [BM04]; this bound is tight since expanders contain a large star metric. We therefore set forth the following question, which was formulated together with Yair Bartal: What is the integrality ratio of the Group-Steiner-Tree problem on the (say two-dimensional) grid graph?

The shortest-path metric of a grid contains, up to constant distortion, an HST which is a complete regular tree (see e.g. [BBM01]). This tree is similar to our tree $\mathbb{T}_n$ (and to Feige's tree described above), but differs in parameters like arity and edge-weight; thus, one may suspect that the integrality ratio in grids is at least as large as in HSTs. In comparison, the best upper bound that we are aware of for two-dimensional grids is the one known for general graphs.

A broader message of our paper is that in some cases the study of inapproximability lower bounds is well-served by proving a preliminary integrality gap result. If for no other reason, the proof of an integrality gap might be somewhat easier. Recall that in our case, the hardness of approximation result of [HK03] uses our integrality gap as a "gadget". Recently, in yet another breakthrough along these lines (an integrality gap result that was strengthened into an inapproximability result) an $\Omega(\log^* n)$ hardness was obtained for the Asymmetric k-Center problem [CGH$^+$05] . It would be nice to see how far this paradigm can be taken.

On the other hand, if a proof of integrality gap does not seem to be possible, one should perhaps try to improve the approximation algorithm. For example, for the Covering-Steiner problem (another generalization of Group-Steiner-Tree), a logarithmic term appearing in the approximation ratio of [KRS02] was not known to have a counterpart in the integrality ratio. Recently it turned out that this term is spurious–the approximation algorithm is improved in [GS06] by designing a better rounding procedure.

Acknowledgments. Our work has been influenced by discussions with several people. We thank Moses Charikar, Chandra Chekuri, Uri Feige, Anupam Gupta, Sanjeev Khanna, Adam Meyerson, Seffi Naor, R. Ravi, and Leonid Zosin for very helpful conversations. We also thank the conference and journal reviewers for their very helpful comments.

References

[Bar96] Y. Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *37th Annual Symposium on Foundations of Computer Science*, pages 184–193. IEEE, 1996.

[Bar98] Y. Bartal. On approximating arbitrary metrics by tree metrics. In *30th Annual ACM Symposium on Theory of Computing*, pages 161–168. ACM, 1998.

[BBM01] Y. Bartal, B. Bollobás, and M. Mendel. A Ramsey-type theorem for metric spaces and its applications for metrical task systems and related problems. In *42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 396–405, October 2001.

[BM04] Y. Bartal and M. Mendel. Multiembedding of metric spaces. *SIAM J. Comput.*, 34(1):248–259, 2004.

[CCC$^+$99] 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.

[CCGG98] 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.

[CEK06] C. Chekuri, G. Even, and G. Kortsarz. A greedy approximation algorithm for the Group Steiner problem. *Discrete Applied Mathematics*, 154:15–34, 2006.

[CGH$^+$05] J. Chuzhoy, S. Guha, E. Halperin, S. Khanna, G. Kortsarz, R. Krauthgamer, and J. Naor. Asymmetric k-center is $\log^* n$-hard to approximate. *J. ACM*, 52(4):538–551, 2005.

[Fei98] U. Feige. A threshold of $\ln n$ for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[FRT04] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *J. Comput. Syst. Sci.*, 69(3):485–497, 2004.

[GKR00] N. Garg, G. Konjevod, and R. Ravi. A polylogarithmic approximation algorithm for the Group Steiner tree problem. *J. Algorithms*, 37(1):66–84, 2000.

[GS06] A. Gupta and A. Srinivasan. An Improved Approximation Ratio for the Covering Steiner Problem. On the Covering Steiner problem. *Theory of Computing*, 2:53–64, 2006.

[HK03] E. Halperin and R. Krauthgamer. Polylogarithmic inapproximability. In *Proceedings of the 35th ACM Symposium on Theory of Computing*, pages 585–594, 2003.

[Jan90] S. Janson. Poisson approximations for large deviations. *Random Structures & Algorithms*, 1:221–230, 1990.

[KRS01] G. Konjevod, R. Ravi, and F. S. Salman. On approximating planar metrics by tree metrics. *Inform. Process. Lett.*, 80(4):213–219, 2001.

[KRS02] G. Konjevod, R. Ravi, and A. Srinivasan. Approximation algorithms for the covering Steiner problem. *Random Structures & Algorithms*, 20(3):465–482, 2002.

[LY94] C. Lund and M. Yannakakis. On the hardness of approximating minimization problems. *J. ACM*, 41(5):960–981, 1994.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[RS97] 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.

[Sri01] A. Srinivasan. New approaches to covering and packing problems. In *12th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 567–576, 2001.

[ZK02] L. Zosin and S. Khuller. On directed Steiner trees. In *13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 59–63, 2002.