4 An FPTAS for the 1-Hotlink Assignment Problem

In this section we present the first fully polynomial approximation scheme for the 1-Hotlink Assignment Problem. That is, for any instance (T, w) and any real number $\epsilon > 0$, the algorithm finds a 1-assignment for T that costs at most $(1 + \epsilon)$ times the cost of the optimal assignment, and the running time of the algorithm is bounded by a polynomial function of the instance size and $1/\epsilon$. Throughout this section, we assume without loss of generality that $w(u)$ is an integer for all $u \in T$[1].

In the previous section we presented the algorithm PATH that has the following property: for any tree T and integer D, it calculates in $O(n3^D)$ the best assignment for T among the assignments A such that the height of T^A is at most D. In order to achieve a pseudo-polynomial algorithm and then an FPTAS, we have to argue that for each tree T with n nodes and weight function w, there is an optimal assignment A^* such that the height of T^{A^*} is small, more specifically $O(\log w(T) + \log n)$. Therefore, if we set $D = G \cdot (\log w(T) + \log n)$, for a suitable constant G, the algorithm PATH finds an optimal assignment for the 1-Hotlink Assignment Problem in $O(n3^{G(\log w(T) + \log n)}) = poly(n \cdot w(T))$ time. With this pseudo-polynomial algorithm in hand, we scale the weight function $w(T)$ according to ϵ making it such that $w(T) = poly(n/\epsilon)$, achieving polynomial time complexity while increasing the solution by only a $(1 + \epsilon)$ factor.

The crucial step in the development of these results is to guarantee the existence of an optimal tree with small height. Informally, the idea used to establish this property is to pick an optimal enhanced tree T^{A^*} and prove that if we walk c steps down in T^{A^*} we reach trees with (geometrically) reduced weight, otherwise this enhanced tree could be improved and would contradict

[1] By definition, w is a rational function and hence there is an integer η such that $w'(u) = \eta w(u)$ is integer for all $u \in T$. By linearity it follows that $\text{EP}(T, A, w') = \text{EP}(T, A, w)$ for each assignment A. Therefore, an α-approximate solution for (T, w') is also α-approximation for (T, w). In addition, notice that this reduction preserves the time complexity of the FPTAS, which is independent of w. Then we can find an $(1 + \epsilon)$-approximate solution for (T, w) in $poly(n, 1/\epsilon)$ time by executing the FPTAS over the scaled instance (T, w')

its optimality. The argument turns out to be fairly intricate, so we devote Section 4.1 to prove this result, which is stated more formally as follows:

Theorem 2 *Consider an instance (T, w) of the 1-HAP with T rooted at node r and let A^* be an optimal 1-assignment for T. Then, there is a constant $c > 2$ such that for every node u of T with $d(r, u, T + A^*) = c$ we have $w(T_u^{A^*}) \leq \frac{(c-1)w(T)}{c}$.*

Then it can be shown that A^* must contain an optimal assignment for each subtree $T_u^{A^*}$. Consequently, we can use the previous theorem for subtrees of T^{A^*} to argue that every time we walk down c steps in T^{A^*}, the weight of the subtrees are reduced by at least a constant factor of $(c-1)/c$. Because w is an integer function, it follows that in $O(\log w(T))$ steps we reach subtrees of T^{A^*} that have zero weight. As these subtrees with zero weight do not influence the cost of the solution, we can employ the result from [PLS04b] to assume without loss of generality that their heights are at most $O(\log n)$. Consequently, the total height of T^{A^*} is $O(\log w(T) + \log n)$.

Lemma 4 *For any tree T with n nodes and integer valued weight function w, there is an optimal 1-assignment A^* for (T, w) such that T^{A^*} has height at most $O(\log w(T) + \log n)$.*

As an immediate consequence we have a pseudo-polynomial algorithm for 1-HAP:

Theorem 3 *Let (T, w) be an instance for the 1-HAP, where w is an integer valued function. For a suitable constant G, the algorithm PATH set with $D = G(\log w(T) + \log n)$ finds an optimal 1-assignment for T in poly$(n \cdot w(T))$ time.*

Now we show how to reduce the weight of the tree T in order to obtain in polynomial time an arbitrarily close approximation for the 1-Hotlink Assignment Problem. The argument is rather standard and is the same one used to obtain the FPTAS for the knapsack problem.

Theorem 4 *There is a fully polynomial time approximation scheme for the 1-Hotlink Assignment Problem.*

Proof: Assume that T is not a single node, otherwise the result trivially holds. Let W be the weight of the heaviest node of T under w, namely $W = \max_{u \in T}\{w(u)\}$. Define $K = \frac{\epsilon \cdot W}{n^2}$ and the weight function w' such that $w'(u) = \lceil w(u)/K \rceil$ for every node $u \in T$. Analogously, let $W' =$

$\max_{u \in T}\{w'(u)\}$; notice that $W' = \lceil W/K \rceil \leq (n^2)/\epsilon + 1$. Thus, $w'(T)$ is less than $nW' \leq (n^3)/\epsilon + n$. As a consequence, the dynamic programming algorithm runs in polynomial time on n and $1/\epsilon$ over the instance (T, w').

Now we argue that the solution for (T, w') is an $(1+\epsilon)$-approximation for (T, w). Let A^* be an optimal solution for (T, w) and A be the solution returned by the algorithm. Clearly for each node u we have $K \cdot w'(u) \geq w(u)$. Therefore:

$$\begin{aligned} K \cdot \mathrm{EP}(T, A, w') &= \sum_{u \in T} d(r, u, T+A) K \cdot w'(u) \\ &\geq \sum_{u \in T} d(r, u, T+A) w(u) = \mathrm{EP}(T, A, w) \end{aligned} \tag{1}$$

Analogously, $K \cdot w'(u) \leq w(u) + K$ and hence $K \cdot \mathrm{EP}(T, A^*, w') \leq \mathrm{EP}(T, A^*, w) + \sum_u d(r, u, T + A^*) \cdot K$. Clearly the distance between any pairs of nodes in $T + A^*$ is at most n, thus $K \cdot \mathrm{EP}(T, A^*, w') \leq \mathrm{EP}(T, A^*, w) + n^2 \cdot K \leq \mathrm{EP}(T, A^*, w) + \epsilon \cdot W$. From the optimality of A^* it follows that $K \cdot \mathrm{EP}(T, A, w') \leq \mathrm{EP}(T, A^*, w) + \epsilon \cdot W$. Recalling that only leaves of T have nonzero weight, it follows that the cost of the optimal solution for (T, w) is at least W, and consequently:

$$K \cdot \mathrm{EP}(T, A, w') \leq \mathrm{EP}(T, A^*, w) + \epsilon \cdot \mathrm{EP}(T, A^*, w) = (1+\epsilon)\mathrm{OPT}(T, w) \tag{2}$$

By chaining inequalities (1) and (2) we complete the proof:

$$\mathrm{EP}(T, A, w) \leq K \cdot \mathrm{EP}(T, A, w') \leq (1+\epsilon)\mathrm{OPT}(T, w) \tag{3}$$

In sum, by executing the pseudo-polynomial algorithm stated in Theorem 3 for the scaled instance (T, w') we have a $(1+\epsilon)$-approximation for the original instance (T, w) in $poly(n/\epsilon)$ time. ■

In order to establish these results, we need to prove the claims made in Theorem 2 and Lemma 4. The former is done in the next section. Although the arguments used to prove Lemma 4 were already sketched in previous paragraphs, the actual proof is very technical and we defer it to the appendix.

4.1 Proof of Theorem 2

Before starting the proof itself, we need to introduce some notation. We define $\mathbf{T}_u(A)$ as the subtree of T_u left after some parts of it have been 'adopted' by proper ancestors of u due to an assignment A (Figure 4.1). More formally, we have the following definitions whose equivalence follows from the greedy user model:

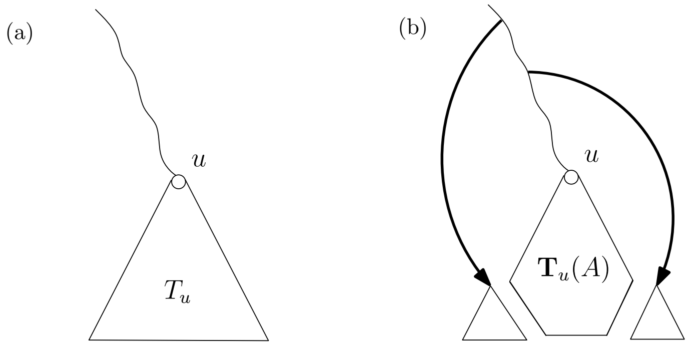

Figure 4.1: Illustration of tree $\mathbf{T}_u(A)$

Definition 1 *Consider a tree T and a non-crossing assignment A for it. Let u be a node in T. Then we have the following equivalent definitions:*

(i) Let $U = \{v \in T :$ user path from $r(T)$ to v in $T + A$ contains the node $u\}$. Then $\mathbf{T}_u(A)$ is the subgraph of T induced by U.

(ii) Let $U = \{v \in T : (w, v) \in A$, w is a proper ancestor of u and v is a proper descendant of $u\}$. Then $\mathbf{T}_u(A) = T_u - \left(\bigcup_{v \in U} T_v\right)$.

Although the trees T_u^A and $\mathbf{T}_u(A)$ are different (for instance, T_u^A has hotlinks of A as arcs), it follows from part (i) of the above definition that the set of nodes of T_u^A is equal to the set of nodes of $\mathbf{T}_u(A)$. Due to this correspondence, henceforth we use the notation $\mathbf{T}_u(A)$ instead of T_u^A, as the former allows us to visualize all subsequent constructions directly on the original tree T.

Now we are ready to start the proof of Theorem 2. Consider an instance (T, w) of the 1-HAP and an optimal 1-assignment A^* for it. (In order to simplify the notation, $\mathbf{T}_u$ is used as a shorthand for $\mathbf{T}_u(A^*)$.) We assume that $w(T) > 0$, otherwise the theorem trivially holds. Moreover, we assume without loss of generality[2] that: (i) A^* is a non-crossing assignment; (ii) it only contains proper hotlinks, that is hotlinks of the form (u, v) where $v \neq u$ and v is not a child of u in T; (iii) there is at most one hotlink in A^* pointing to each node of T.

The proof goes by contradiction. Assume that the theorem does not hold for A^*, that is, A^* satisfies the following hypothesis:

Hypothesis 1 *For the constant $c > 2$ given by Theorem 2, there is a node $h \in T$ such that $d(r, h, T + A^*) = c$ and $w(\mathbf{T}_h) > (c-1)w(T)/c$.*

[2] It is easy to see that for each assignment A^* that do not satisfy some of these hypothesis, there is an assignment $A^{*\prime}$ that satisfies them and is such that the user path from r to every node $u \in T$ is the same in $T + A^*$ and $T + A^{*\prime}$. Therefore, if the enhanced tree $T^{A^{*\prime}}$ complies with the statement of the theorem, then so does the enhanced tree T^{A^*}.

(Notice that due to the correspondence between nodes of the trees $\mathbf{T}_h$ and $T_h^{A^*}$, in the hypothesis we used $w(\mathbf{T}_h)$ instead of $w(T_h^{A^*})$.)

In the sequel we find a 1-assignment A which costs strictly less than the optimal assignment A^*, thus reaching the contradiction.

So let h be a node of T satisfying Hypothesis 1 and let $Q = (q_1 \rightarrow \ldots \rightarrow q_{|Q|} = h)$ be the user path from r to h in $T + A^*$. We note that Q cannot have two consecutive hotlinks in A^*, otherwise one could obtain an assignment better than A^* through a fairly simple transformation (proof in the appendix). This property of Q will be useful at the end of our analysis.

Proposition 3 *There cannot be two consecutive hotlinks in Q.*

We construct the assignment A by removing and adding some hotlinks to A^*. The idea behind this construction is to bring subtrees of the 'heavy' tree T_h closer to the root of T by adding new hotlinks to nodes in Q. In this process, paths that reach nodes outside T_h may be lengthier in $T + A$ than in $T + A^*$, but because T_h has most of the weight of T (Hypothesis 1) the expected path length in $T + A$ is less than in $T + A^*$.

The general idea is the following. Consider a hotlink from a proper ancestor u of h to a proper descendant v of h. Due to Hypothesis 1, most users have to traverse node h and only a few users will make use of this hotlink. So we can construct an assignment that replaces (u, v) by (u, v'), where v' is a proper ancestor of v. This way, more users will use the new hotlink instead of traversing the long path Q and consequently the overall expected path should be reduced. Finally, in order to ensure that the path from r to v does not increase much in the new assignment, we add the hotlink (v', v) to it. However, this simple construction has two major problems. First, the expected path in the new enhanced tree may not be shortened due to crossing hotlinks. In addition, this new assignment may contain two hotlinks in node v', and hence may not be a valid 1-assignment. Great part of the work is directed to employ this construction strategy while overcoming these drawbacks.

In order to simplify the analysis, we construct the assignment A as the union of two auxiliary assignments A^1 and A^2. Informally, the assignment A^1 is responsible for removing from A^* hotlinks of nodes in Q, and the assignment A^2 adds new hotlinks to these nodes pointing to 'balanced' subtrees of T_h. In addition, because the assignments A^1 and A^2 are built on two different 'parts' of T, we can analyze them separately in order to bound the cost of A.

Construction and analysis of A^1***.*** For each node $q_i \in Q - \{h\}$ and for each child j of q_i, with $j \neq q_{i+1}$, we define $T_i^j = T_j - T_{q_{i+1}}$ (Figure 4.2). We start with A^1 as the set of hotlinks of A^* that have both endpoints in Q. Then, for each T_i^j, we add to A^1 an optimal non-crossing assignment A_i^j for T_i^j (Figure 4.3). Note that some nodes in Q do not have hotlinks in A^1.

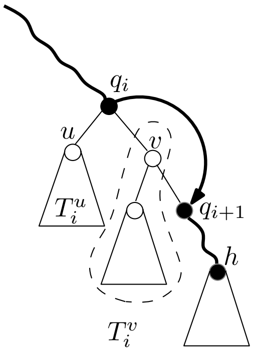

Figure 4.2: Construction of trees T_i^j. The path Q is depictured in bold.

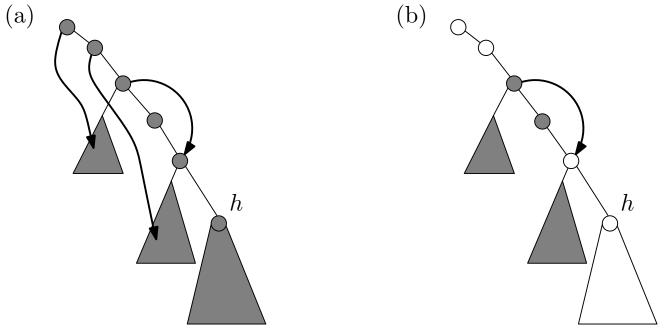

Figure 4.3: (a) Illustration of enhanced tree $T+A^*$. (b) Illustration of enhanced tree $T + A^1$. The shaded nodes and subtrees may contain hotlinks, but not the blank ones.

Because we have assumed A^* to be non-crossing, our construction implies that A^1 is non-crossing as well. In addition, we have the following important property, which is mostly due to the fact A^1 contains all hotlinks of A^* with both endpoints in Q (proof in the appendix):

Lemma 5 *For any node $q \in Q$, the user path from r to q is the same in $T+A^1$ and in $T + A^*$.*

Now we show that the cost of reaching nodes of $T - T_h$ in the enhanced tree $T + A^1$ is at most $w(T)$ units greater than the cost of reaching these nodes in $T+A^*$. As the trees $\{T_i^j\}$ define a partition of the leaves in $T-T_h$, it suffices to analyze the cost of reaching nodes in these trees.

Fix a tree T_i^j. We claim that the path in $T + A^1$ from r to a node $u \in T_i^j$ has the form $(r \rightsquigarrow q_i \rightarrow j \rightsquigarrow u)$. By the definition of the assignment, no proper ancestor of q_i can have a hotlink in A^1 pointing to a node in T_i^j. Therefore, using Definition 1-(ii) it follows that $T_i^j \subseteq \mathbf{T}_{q_i}(A^1)$, or alternatively, that q_i belongs to the user path in $T + A^1$ from r to u. Using an analogous argument, it is easy to see that j also belongs to the user path from r to u in $T + A^1$. Hence, the user path in $T + A^1$ from r to u can be written as $(r \rightsquigarrow q_i \rightarrow j \rightsquigarrow u)$ and the claim follows.

Again considering the fixed tree T_i^j, the definition of the trees $\{T_{i'}^{j'}\}$ implies that the node q_i belongs to $Q - h$. It then follows by the choice of h that $d(r, q_i, T + A^*) < d(r, h, T + A^*) = c$. Combining with this fact with Lemma 5, we have can bound the distance from r to q_i in $T + A^1$ as $d(r, q_i, T + A^1) \leq c - 1$. Weighting the paths $(r \rightsquigarrow q_i \rightarrow j \rightsquigarrow u)$ for all $u \in T_i^j$ we have:

$$\begin{aligned}
\mathrm{EP}(T, A^1)_{T_i^j} &= \sum_{u \in T_i^j} \big(d(r, q_i, T + A^1) + 1 + d(j, u, T + A^1)\big)\, w(u) \\
&= \sum_{u \in T_i^j} \big(d(r, q_i, T + A^1) + 1 + d(j, u, T_i^j + A_i^j)\big)\, w(u) \\
&\leq c \cdot w(T_i^j) + \mathrm{OPT}(T_i^j)
\end{aligned}$$

where the second equality follows from Proposition 1.

On the other hand, Lemma 2 assures that the contribution of T_i^j to $\mathrm{EP}(T, A^*)$ is not smaller than $\mathrm{OPT}(T_i^j)$, hence $\mathrm{EP}(T, A^1)_{T_i^j} \leq c \cdot w(T_i^j) + \mathrm{EP}(T, A^*)_{T_i^j}$. Therefore, the total increase in the cost of reaching nodes in $\{T_i^j\}$ (and consequently the nodes of $T - T_h$) is:

$$\begin{aligned}
\mathrm{EP}(T, A^1)_{T - T_h} - \mathrm{EP}(T, A^*)_{T - T_h} &= \sum_{i,j} \mathrm{EP}(T, A^1)_{T_i^j} - \sum_{i,j} \mathrm{EP}(T, A^*)_{T_i^j} \\
&\leq c \sum_{i,j} w(T_i^j) = c \cdot w(T - T_h) \\
&\leq w(T)
\end{aligned} \tag{4}$$

where the last inequality follows from Hypothesis 1.

Construction of A^2***.*** Now we define the first step of the construction of the assignment A^2, which is illustrated in Figure 4.4. This step is responsible for bringing nodes of $\mathbf{T}_h$ closer to the root of T. Let D be the nodes of $Q - \{h\}$ which do not have hotlinks in A^* pointing to nodes of Q. Moreover, define d_i as the ith node of D, that is, the ith closest to the root of T. Let $\{H_1, \ldots, H_k\}$ be the partition of T_h given by Lemma 3 with respect to w when $U = T_h$ and $\alpha = 4w(T)/|D|$. We can assume that the trees of our partition are labeled such that for all $i < j$, $r(H_i)$ is not an ancestor of $r(H_j)$. Then the first step of the construction of A^2 consists of the following: initially define A^2 as the set of hotlinks of A^* with both endpoints in Q; then add the hotlinks $\bigcup_{i=1}^{k}(d_i, r(H_i))$ to A^2 (Figure 4.4).

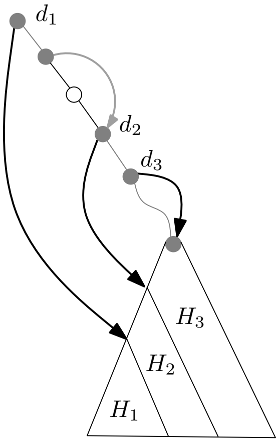

Figure 4.4: Enhanced tree $T + A^2$ after the first part of the construction, with path Q in gray.

The second and last step of the construction of A^2 is designed to control the expansion of paths that reach nodes of $T_h - \mathbf{T}_h$ (Figure 4.5). Before presenting this step, we need to define a few subtrees of T_h which are illustrated in Figure 4.5.a. Let S be the set of nodes in T_h that are endpoints of hotlinks of A^* departing from nodes in D, namely $S = \{s \in T_h : (d, s) \in A^* \text{ for some } d \in D\}$. For each meaningful pair i, j, we use H_i^j to denote the maximal subtree of H_i rooted at the child j of $r(H_i)$. A key property for our construction, which is imposed by the choice of α in the decomposition of T_h, states that for every $s \in S$ the tree $\mathbf{T}_s$ is a subtree of some H_i^j (proof in the appendix).

Lemma 6 *Consider a node $s \in S$. Then $\mathbf{T}_s$ is a subtree of a tree H_i^j.*

Lastly, define $\overline{H}_i^j = H_i^j - (\bigcup_{s \in S \cap H_i^j} \mathbf{T}_s)$. We remark that each $\overline{H}_i^j \neq \emptyset$ is a tree rooted at node j (see Lemma 20 in the appendix).

The second step of the construction of A^2 consists of: (i) adding to A^2 a non-crossing optimal assignment A_s for each tree $\mathbf{T}_s$ with $s \in S$; (ii) adding a non-crossing optimal assignment $\overline{A}_i^j$ for each of the trees $\{\overline{H}_i^j\}$; (iii) adding, for each meaningful pair i, j, the hotlinks $\bigcup_{s \in S \cap H_i^j}(r(H_i^j), s)$ (Figure 4.5.b).

Due to the assumed order of the trees $\{H_i\}$, it is easy to see that A^2 does not contain any crossing pair of hotlinks. Notice however that the roots of the trees $\{H_i^j\}$ may have more than one hotlink in A^2, but we will remove them later.

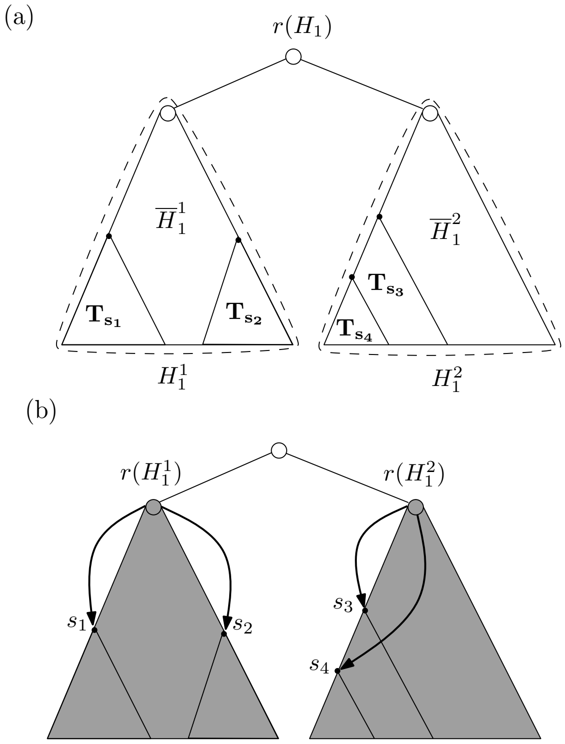

Figure 4.5: Illustration of the second part of the construction of A^2. (a) Tree H_1 and its subtrees $\{H_i\}$, $\{H_i^j\}$, $\{\overline{H}_i^j\}$ and $\{\mathbf{T}_s\}_{s \in S}$. (b) Addition of hotlinks $(r(H_i^j), s)$ for $s \in S \cap H_i^j$ and optimal assignments for the subtrees of $H_i^{j_1}$ and $H_i^{j_2}$.

Analysis of A^2***.*** The following property is analogous to Lemma 5 presented during the construction of A^1 and can be proved with the same argument:

Lemma 7 *For any node $q \in Q$, the user path from r to q is the same in $T + A^2$ and in $T + A^*$.*

Now we compare the cost of reaching nodes of T_h in the enhanced trees $T + A^*$ and $T + A^2$. A first observation which will be used to bound the length of paths in $T + A^2$ is that users will necessarily traverse a node of D in order to reach any node in T_h. Informally, users trying to reach a node u in H_i traverse the path Q until finding a node in d_j which takes them closer to u and then follow the path $(d_j \to r(H_j) \rightsquigarrow u)$. However, due to our ordering on the trees $\{H_{i'}\}$, all nodes in D which are proper ancestors of d_i have hotlinks pointing to nodes $\{r(H_{i'})\}$ which are non-ancestors of $r(H_i)$ and consequently, because H_i is a tree, non-ancestors of u. Therefore, d_i is the first node of D that has a hotlink which takes such users closer to u and hence d_i belongs to the user path from r to u in $T + A^2$. This result is formalized in the following lemma which is proved in the appendix:

Lemma 8 *Consider a tree H_i and a node $u \in H_i$. Then the user path in $T+A^2$ from r to u contains the node d_i.*

We remark that the trees $\{\mathbf{T}_s\}_{s\in S}$ and $\mathbf{T}_h$ form a partition of T_h (see Lemma 19 in the appendix). Therefore, it suffices to analyze separately the cost of reaching nodes of these trees.

First we argue that the path in $T + A^2$ for reaching nodes of the trees $\{\mathbf{T}_s\}_{s\in S}$ is not much longer than the path for reaching these nodes in the optimal enhanced trees $T + A^*$. For some node $s \in S$, consider the tree $\mathbf{T}_s$. From Lemma 6, $\mathbf{T}_s$ is contained in a tree, say, H_i^j. Based on Lemma 8 and on the fact that $r(H_i)$ does not have any hotlink in A^2, it is not difficult to see that the user path in $T + A^2$ from r to a node $u \in \mathbf{T}_s$ has the form $(r \rightsquigarrow d_i \rightarrow r(H_i) \rightarrow r(H_i^j) \rightarrow s \rightsquigarrow u)$. Therefore, weighting the paths $(r \rightsquigarrow u)$ over all nodes u of $\mathbf{T}_s$, we have:

$$\begin{aligned}
\mathrm{EP}(T, A^2)_{\mathbf{T}_s} &= \sum_{u\in\mathbf{T}_s} \big(d(r, d_i, T + A^2) + 3 + d(s, u, T + A^2)\big)\, w(u) \\
&= \sum_{u\in\mathbf{T}_s} \big(d(r, d_i, T + A^2) + 3 + d(s, u, \mathbf{T}_s + A_s)\big)\, w(u) \\
&= d(r, d_i, T + A^*)w(\mathbf{T}_s) + 3w(\mathbf{T}_s) + \mathrm{OPT}(\mathbf{T}_s)
\end{aligned}$$

where the second equality follows from Proposition 1 and the third equality follows from Lemma 7.

On the other hand, Lemma 2 asserts that the cost of reaching nodes of $\mathbf{T}_s$ in the optimal enhanced tree $T + A^*$ is at least $\mathrm{OPT}(\mathbf{T}_s)$. Using the fact that the trees $\{\mathbf{T}_s\}_{s\in S}$ define a partition of the nodes of $T_h - \mathbf{T}_h$, we have that the total difference in the cost of reaching nodes of $T_h - \mathbf{T}_h$ in $T + A^2$ and $T + A^*$ is:

$$\begin{aligned}
\mathrm{EP}(T, A^2)_{T_h-\mathbf{T}_h} - \mathrm{EP}(T, A^*)_{T_h-\mathbf{T}_h} &= \sum_{s\in S} \mathrm{EP}(T, A^2)_{\mathbf{T}_s} - \sum_{s\in S} \mathrm{EP}(T, A^*)_{\mathbf{T}_s} \\
&\leq \sum_{i=1}^{k} \sum_{s\in S\cup H_i} d(r, d_i, T + A^*)w(\mathbf{T}_s) + 3w(T_h - \mathbf{T}_h) \quad (5)
\end{aligned}$$

Now we need to bound the cost of reaching nodes of $\mathbf{T}_h$ in the enhanced tree $T + A^2$. We first argue that the trees $\{\overline{H}_i^j\}$ and $\{r(H_i)\}$ form a partition of the nodes of $\mathbf{T}_h$. Because the trees $\mathbf{T}_h$ and $\{\mathbf{T}_s\}_{s\in S}$ form a partition of T_h, it follows that the nodes of $\mathbf{T}_h$ are exactly the nodes of $T_h - \bigcup_{s\in S} \mathbf{T}_s$. Noticing that the trees $\{H_i^j\}$ and $\{r(H_i)\}$ form a partition of the nodes of T_h, it follows from Lemma 6 that the nodes of $\mathbf{T}_h$ are exactly the nodes of $\{\overline{H}_i^j\} \cup \{r(H_i)\}$.

Therefore, in order to bound the cost of reaching nodes of $\mathbf{T}_h$, it suffices to analyze separately the nodes of the trees $\{\overline{H}_i^j\}$ and $\{r(H_i)\}$.

Consider a nonempty tree $\overline{H}_i^j$. By the same arguments of previous analysis, it is not difficult to see that the path in $T + A^2$ from r to a node $u \in \overline{H}_i^j$ has the form $(r \rightsquigarrow d_i \rightarrow r(H_i) \rightarrow j \rightsquigarrow u)$. Weighting these paths over all nodes u of $\overline{H}_i^j$ gives the cost of reaching the nodes of $\overline{H}_i^j$ in $T + A^2$:

$$\begin{aligned}
\mathrm{EP}(T, A^2)_{\overline{H}_i^j} &= \sum_{u \in \overline{H}_i^j} (d(r, d_i, T + A^2) + 2 + d(j, u, T + A^2))w(u) \\
&= \sum_{u \in \overline{H}_i^j} (d(r, d_i, T + A^2) + 2 + d(j, u, \overline{H}_i^j + \overline{A}_i^j))w(u) \\
&\leq d(r, d_i, T + A^2)w(\overline{H}_i^j) + 2w(\overline{H}_i^j) + \mathrm{OPT}(\overline{H}_i^j)
\end{aligned}$$

where the second equality follows from Proposition 1 (recall that $\overline{H}_i^j \neq \emptyset$ implies that j is the root of $\overline{H}_i^j$).

Due to the uniqueness of user paths on enhanced trees, the above argument implies that the user path from r to $r(H_i)$ in $T + A^2$ is $(r \rightsquigarrow d_i \rightarrow r(H_i))$, and consequently $\mathrm{EP}(T, A^2)_{r(H_i)} = d(r, d_i, T + A^2)w(r(H_i)) + w(r(H_i))$.

Adding the costs of reaching nodes of the trees $\{\overline{H}_i^j\}$ and $\{r(H_i)\}$ we conclude that the cost of reaching all nodes of $\mathbf{T}_h$ in $T + A^2$ is at most:

$$\begin{aligned}
\mathrm{EP}(T, A^2)_{\mathbf{T}_h} &\leq \sum_{i=1}^{k} d(r, d_i, T + A^2)(\sum_j w(\overline{H}_i^j) + w(r(H_i))) + 2w(\mathbf{T}_h) + \sum_{i,j} \mathrm{OPT}(\overline{H}_i^j) \\
&\leq \sum_{i=1}^{k} d(r, d_i, T + A^*)(\sum_j w(\overline{H}_i^j) + w(r(H_i))) + 2w(\mathbf{T}_h) + \mathrm{OPT}(\mathbf{T}_h)
\end{aligned}$$

where the first and the last terms of the last inequality follow from Lemma 7 and Corollary 1, respectively.

On the other hand, consider the path in $T + A^*$ from r to a node u in $\mathbf{T}_h$. Recalling that $\mathbf{T}_h = \mathbf{T}_h(A^*)$, it follows that this path has the form $(r \rightsquigarrow h \rightsquigarrow u)$, which, by Hypothesis 1, has length $c + d(h, u, T + A^*)$. Weighting this path for all nodes $u \in \mathbf{T}_h$, it follows that the cost of reaching nodes of $\mathbf{T}_h$ in $T + A^*$ is at least $c \cdot w(\mathbf{T}_h) + \mathrm{OPT}(\mathbf{T}_h)$. Hence, subtracting the bounds for $\mathrm{EP}(T, A^2)_{\mathbf{T}_h}$ and $\mathrm{EP}(T, A^*)_{\mathbf{T}_h}$ we have:

$$\begin{aligned}
&\mathrm{EP}(T, A^2)_{\mathbf{T}_h} - \mathrm{EP}(T, A^*)_{\mathbf{T}_h} \leq \\
&\sum_{i=1}^{k} d(r, d_i, T + A^*)(\sum_j w(\overline{H}_i^j) + w(r(H_i))) + 2w(\mathbf{T}_h) - c \cdot w(\mathbf{T}_h) \qquad (6)
\end{aligned}$$

Then adding inequalities (5) and (6) and employing Lemma 6, we have the difference between the cost of reaching nodes of T_h in $T+A^*$ and in $T+A^2$:

$$\begin{aligned}
\mathrm{EP}(T,A^2)_{T_h} - \mathrm{EP}(T,A^*)_{T_h} &\leq \sum_{i=i}^{k} d(r,d_i,T+A^*)w(H_i) + 3w(T_h) - c \cdot w(\mathbf{T}_h) \\
&\leq \sum_{i=i}^{k} d(r,d_i,T+A^*)w(H_i) + 3w(T_h) \\
&\quad -(c-1)\cdot w(T) \qquad (7)
\end{aligned}$$

where the last inequality follows from Hypothesis 1.

Now we need to bound the first term of the right-hand side of inequality (7). Because there cannot be two consecutive hotlinks in Q, it is not difficult to see that the distance $d(r,d_i,T+A^*)$ can be upper bounded by $2i$. In addition, because at least $k-1$ of the trees $\{H_i\}$ have weight greater than $4w(T)/|D|$, it follows that k can be at most $(|D|/4)+1$ (otherwise $\sum_i w(H_i) > w(T)$). Therefore, for all $i \leq k$ we have $d(r,d_i,T+A^*) \leq 2i \leq 2k \leq (|D|/2)+2$. Then the first term of the right-hand side of inequality (7) can be bounded by:

$$\sum_{i=i}^{k} d(r,d_i,T+A^*)w(H_i) \leq \left(\frac{|D|}{2}+2\right)\sum_{i=1}^{k} w(H_i) \leq \left(\frac{|D|}{2}+2\right) w(T) \leq \left(\frac{c}{2}+2\right) w(T)$$

Finally, employing this bound to inequality (7) we have:

$$\mathrm{EP}(T,A^*)_{T_h} - \mathrm{EP}(T,A^2)_{T_h} \;\geq\; \left(\frac{c}{2}-6\right) w(T) \qquad (8)$$

Analysis of $A^1 \cup A^2$**.** Let A' be the set union of the assignments A^1 and A^2, that is, hotlinks that appear in both sets are only included once in A'.

A first observation is that, due to their construction, both A^1 and A^2 have the same set of hotlinks with both endpoints in Q, and consequently so does A'. As A^2 do not have hotlinks with endpoints in $\{T_i^j\}$, it follows that A^1 and A' have the same set of hotlinks with both endpoints in $Q \cup \{T_i^j\}$, or alternatively in $T - T_h$. Thus, employing Proposition 1 with $T' = T - T_h$ we have that $\mathrm{EP}(T,A')_{T-T_h} = \mathrm{EP}(T,A^1)_{T-T_h}$.

Now let P be the path in $T+A^2$ from r to a node $u \in T_h$. According to previous discussions, the path P has the form $(r \rightsquigarrow d_i \rightarrow r(H_i) \rightsquigarrow u)$ for some $d_i \in D$. Moreover, because d_i belongs to Q, it is easy to see that the path in $T+A^*$ from r to d_i is the subpath $(r = q_1 \rightarrow q_2 \rightarrow \ldots \rightarrow q_j = d_i)$ of Q. Then Lemma 7 implies that P has the form $(q_1 \rightarrow q_2 \rightarrow \ldots \rightarrow q_j \rightarrow r(H_i) \rightsquigarrow u)$. Because A^1 do not have hotlinks departing from nodes in $T_h \supseteq H_i$, is easy to see that nodes in P have the same hotlinks in A^2 and in A'. It then follows from Proposition 2 that $\mathrm{EP}(T, A')_{T_h} = \mathrm{EP}(T, A^2)_{T_h}$.

Therefore, using these observations and employing the bounds from inequalities (4) and (8), we have:

$$\begin{aligned} \mathrm{EP}(T, A^*) - \mathrm{EP}(T, A') &= (\mathrm{EP}(T, A^*)_{T_h} - \mathrm{EP}(T, A')_{T_h}) \\ &\quad +(\mathrm{EP}(T, A^*)_{T-T_h} - \mathrm{EP}(T, A')_{T-T_h}) \\ &\geq \left(\frac{c}{2} - 7\right) w(T) \qquad (9) \end{aligned}$$

Notice we can choose the value of c as a sufficiently large constant such that the right-hand side of the previous inequality is strictly positive, which implies that $\mathrm{EP}(T, A^*) > \mathrm{EP}(T, A')$. However, this does not contradict the optimality of A^* as A' may have more than one hotlink at the roots of $\{H_i^j\}$ and therefore may not be a valid 1-assignment.

Removing multiple hotlinks. First notice that, due to its construction, only the roots of the trees H_i^j can have more than one hotlink in A'. More specifically, a node $r(H_i^j)$ has exactly $|S \cap H_i^j|$ hotlinks in A'. Let $A_i'^j$ be the set of hotlinks of A' with both endpoints in H_i^j. Lemma 1 asserts there is a 1-assignment A_i^j for H_i^j such that $\mathrm{EP}(H_i^j, A_i^j) \leq \mathrm{EP}(H_i^j, A_i'^j) + |S \cap H_i^j| w(H_i^j)$. In order to obtain a 1-assignment for T, we can replace each $\{A_i'^j\}$ in A' by A_i^j, that is, we define $A = A' - (\bigcup_{i,j} A_i'^j) \cup (\bigcup_{i,j} A_i^j)$. Fortunately, as proved in the appendix by means Lemma 21 and Corollary 2, the cost of A has a very natural relation to the cost of A': $\mathrm{EP}(T, A) = \mathrm{EP}(T, A') + \sum_{i,j} |S \cap H_i^j| w(H_i^j)$. Due to the construction of the trees $\{H_i\}$, $w(H_i^j) \leq 4w(T)/|D|$ for all meaningful i, j. Therefore:

$$\begin{aligned} \mathrm{EP}(T, A) - \mathrm{EP}(T, A') &\leq \sum_{i,j} |S \cap H_i^j| w(H_i^j) \leq \frac{4w(T)}{|D|} \sum_{i,j} |S \cap H_i^j| \\ &= \frac{4w(T)|S|}{|D|} \leq 4w(T) \end{aligned}$$

where the last inequality follows from the definition of S.

Finally, combining the previous inequality with (9) we can compare the costs of A and A^*:

$$\mathrm{EP}(T, A^*) - \mathrm{EP}(T, A) \geq \left(\frac{c}{2} - 11\right) w(T)$$

By choosing $c = 23$ the right-hand side becomes positive, implying that $\mathrm{EP}(T, A^*) > \mathrm{EP}(T, A)$. Because A is a valid 1-assignment for T, this contradicting the optimality of A^* and concludes the proof of the theorem.