Binary trees with the largest number of subtrees*

L. A. Székely and Hua Wang
Department of Mathematics,
University of South Carolina, Columbia, SC 29208
{szekely,hwang0}@math.sc.edu

February 11, 2004

Abstract

This paper characterizes binary trees with n leaves, which have the greatest number of subtrees. These binary trees coincide with those which were shown by Fischermann *et al.* [2] and Jelen and Triesch [3] to minimize the Wiener index.

1 Terminology

All graphs in this paper will be finite, simple and undirected. A *tree* $T = (V, E)$ is a connected, acyclic graph. We refer to vertices of degree 1 of T as *leaves*. The unique path connecting two vertices v, u in T will be denoted by $P_T(v, u)$. For a tree T and two vertices v, u of T, the *distance* $d_T(v, u)$ between them is the number of edges on the connecting path $P_T(v, u)$. For a vertex v of T, define the *distance of the vertex* as $g_T(v) = \sum_{u \in V(T)} d_T(v, u)$. Then $\sigma(T) = \frac{1}{2} \sum_{v \in V(T)} g_T(v)$ denotes the *Wiener index* of T.

We call a tree (T, r) *rooted at the vertex* r (or just by T if it is clear what the root is) by specifying a vertex $r \in V(T)$. For any two different vertices u, v in a rooted tree (T, r), we say that v is a *successor* of u, if $P_T(r, u) \subset P_T(r, v)$. Furthermore, if u and v are adjacent to each other and $d_T(r, u) = d_T(r, v) - 1$, we say that u is a *parent* of v and v is a *child* of u. A subtree of a tree will often be described by its vertex set.

If v is any vertex of a rooted tree (T, r), let $T(v)$, *the subtree induced by* v, denote the rooted subtree of T that is induced by v and all its successors in T, and is rooted at v.

The *height* of a vertex v of a rooted tree T with root r is $h_T(v) = d_T(r, v)$, and the *height* of a rooted tree T is $h(T) = \max_{v \in T} h_T(v)$, the maximum height of vertices.

A *binary* tree is a tree T such that every vertex of T has degree 1 or 3. A *rooted binary tree* is a tree T with root r, which has exactly two children, while every other vertex of T has degree 1 or 3. A rooted binary tree T is *complete*, if it has height h and 2^h leaves for some $h \geq 0$. In addition, a single vertex tree is also considered a rooted binary tree of height 0.

For a tree T and a vertex v of T, let $f_T(v)$ denote the number of subtrees of T that contain v, let $F(T)$ denote the number of non-empty subtrees of T.

*This research was supported in part by the NSF contracts DMS 007 2187 and 0302307.

If T is a rooted binary tree with root r, and r_1, r_2 are the children of r, then we will simply write T_1 for $T(r_1)$ and T_2 for $T(r_2)$. We assign the labels r_1 and r_2 according to the following rule: $f_{T_2}(r_2) \geq f_{T_1}(r_1)$. T_i will be rooted at r_i, $i = 1, 2$. We define recursively $T_{i_1 i_2 \ldots i_k 1}$ and $T_{i_1 i_2 \ldots i_k 2}$ to be the two rooted binary trees induced by the children of the root of $T_{i_1 i_2 \ldots i_k}$, when $T_{i_1 i_2 \ldots i_k}$ is not a single vertex, where $i_j \in \{1, 2\}$, $j = 1, 2, \ldots, k$. We assign the labels $r_{i_1 i_2 \ldots i_k 1}$ and $r_{i_1 i_2 \ldots i_k 2}$ according to the following rule:

$$f_{T_{i_1 i_2 \ldots i_k 2}}(r_{i_1 i_2 \ldots i_k 2}) \geq f_{T_{i_1 i_2 \ldots i_k 1}}(r_{i_1 i_2 \ldots i_k 1}) \tag{1}$$

We complete the recursive definition by letting $r_{i_1 i_2 \ldots i_k}$ be the root for $T_{i_1 i_2 \ldots i_k}$.

2 Introduction

To present our main results, we have to give more definitions. Call a rooted binary tree *ordered*, if for every $k \geq 1$, the vertices at height k are put in a linear order, such that if u and v are vertices at height $k + 1$, and they have distinct parents, then the order between u and v at height $k + 1$ is the same as the order of their parents at height k.

A rooted binary tree is *good*, if (i) the heights of any two of its leaf vertices differ by at most 1; (ii) the tree can be ordered such that the parents of the leaves at the greatest height make a final segment in the ordering of vertices at the next-to-greatest height. For brevity, we often refer to such trees as *rgood binary trees*. A single-vertex rooted binary tree is also rgood.

A binary tree is *good*, if it is obtained from two rgood binary trees T_1 and T_2 by joining their roots with an edge, if (i) for any two leaves, their respective heights in T_1 and/or T_2 differ by at most 1; (ii) at least one of T_1 and T_2 is complete.

Note that good and rgood binary trees are unique in the following sense: if we have two good (rgood) binary trees with same number of vertices, then we can label their vertices such that they are isomorphic to each other. The concept of *height* can be naturally extended to vertices of good binary trees, as shown on Fig. 1.

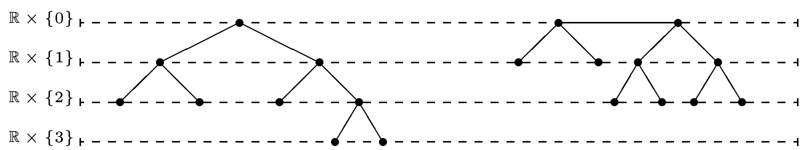

Figure 1: An rgood binary tree (on the left) and a good binary tree (on the right).Vertices at height k of the rgood binary tree and of the two rgood parts of the good binary tree are shown on the line $\mathbb{R} \times k$.

Fischermann *et al.* [2], and independently Jelen and Triesch [3] proved:

Theorem 2.1. *Among binary trees with n leaves, precisely the good binary tree minimizes the Wiener index.*

The goal of this paper is to prove:

Theorem 2.2. *Among binary trees with n leaves, precisely the good binary tree maximizes the number of subtrees.*

In a related paper [5] we discuss an amazing and not yet understood relationship between the Wiener index and the number of subtrees. In [5] we also explain additional motivation for extremal problems about the number of subtrees of trees. Knudsen [4] used this quantity to provide upper bound for the time complexity of his multiple parsimony alignment with affine gap cost using a phylogenetic tree.

3 Lemmas about arbitrary trees

Lemma 3.1. *For any rooted tree T with root r, and any $r' \in V(T)$ $(r' \neq r)$, consider the induced subtree $T' = T(r')$ rooted at r'. Then we have*

$$f_T(r) > f_{T'}(r'). \tag{2}$$

If T'' is obtained from T by deleting some vertices, but not r, then

$$f_T(r) > f_{T''}(r). \tag{3}$$

□

In the rest of this section we prove two lemmas. Consider the tree T in Fig. 2, with leaves x and y, and $P_T(x,y) = xx_1 \ldots x_n z y_n \ldots y_1 y$ $(xx_1 \ldots x_n y_n \ldots y_1 y)$ if $d_T(x,y)$ is even (odd).

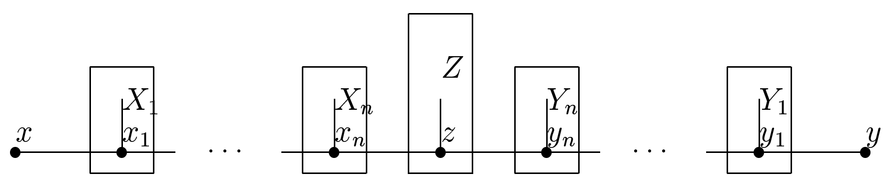

Figure 2: Path $P_T(x,y)$ connecting leaves x and y.

After the deletion of all the edges of $P_T(x,y)$ from T, some connected components will remain. Let X_i denote the component that contains x_i, let Y_j denote the component that contains y_j, for $i,j = 1,2,\ldots,n$, and let Z denote the component that contains z. Set
$a_i = f_{X_i}(x_i)$ for $i = 1,\ldots,n$, $(n \geq 0)$
$b_j = f_{Y_j}(y_j)$ for $j = 1,\ldots,n$,
$c = f_Z(z)$.

Lemma 3.2. *In the situation described above, if $a_i \geq b_i$ for $i = 1,2,\ldots,n$, then $f_T(x) \geq f_T(y)$. Furthermore, $f_T(x) = f_T(y)$ if and only if $n = 0$ or $a_i = b_i$ for all i.* □

Proof. With the above notations, if z and Z occur, we have

$$f_T(x) = 1 + \sum_{k=1}^{n}(\prod_{i=1}^{k} a_i) + c(\prod_{i=1}^{n} a_i) + c(\prod_{i=1}^{n} a_i)(\sum_{k=1}^{n}(\prod_{j=n+1-k}^{n} b_j)) + N;$$

$$f_T(y) = 1 + \sum_{k=1}^{n}(\prod_{j=1}^{k} b_j) + c(\prod_{j=1}^{n} b_j) + c(\prod_{j=1}^{n} b_j)(\sum_{k=1}^{n}(\prod_{i=n+1-k}^{n} a_i)) + N;$$

(Here $N = c\prod_{i=1}^n (a_i b_i)$ is the number of subtrees that contain both x and y.)

Then we have $f_T(x) - f_T(y) =$

$$\sum_{k=1}^{n}(\prod_{i=1}^{k} a_i - \prod_{j=1}^{k} b_j) + c(\prod_{i=1}^{n} a_i - \prod_{j=1}^{n} b_j) + c\sum_{k=1}^{n}(\prod_{i=1}^{n-k} a_i - \prod_{j=1}^{n-k} b_j)\prod_{l=n+1-k}^{n} a_l b_l \geq 0,$$

with strict inequality if $a_i > b_i$ for any $i \in \{1, 2, \ldots, n\}$.

A similar argument works if z and Z do not occur. $\square$

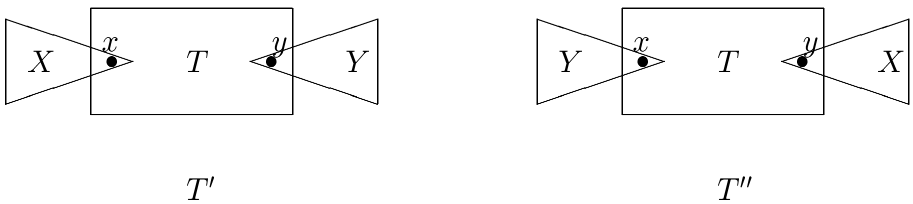

Figure 3: Switching subtrees rooted at x and y.

If we have a tree T with leaves x and y, and two rooted trees X and Y, then we can build two new trees, first T', by identifying the root of X with x and the root of Y with y, second T'', by identifying the root of X with y and the root of Y with x. Under the circumstances below we can tell which composite tree has more subtrees.

Lemma 3.3. *If $f_T(x) > f_T(y)$ and $f_X(x) < f_Y(y)$, then we have $F(T'') > F(T')$.* $\square$

Proof. When T' changes to T'', the number of subtrees which contain both or neither of x and y do not change, so we only need to consider the number of subtrees which contain precisely one of x and y. For T', the number of subtrees which contain x but not y is

$$f_X(x)(f_T(x) - N),$$

the number of the subtrees which contain y but not x is

$$f_Y(y)(f_T(y) - N),$$

where N is the number of subtrees of T that contain both x and y. Similarly, for T'', these two numbers are

$$f_Y(y)(f_T(x) - N) \quad \text{and} \quad f_X(x)(f_T(y) - N).$$

We have

$$F(T'') - F(T') = (f_Y(y) - f_X(x))(f_T(x) - f_T(y)) > 0.$$

$\square$

4 Basic properties of good and rgood binary trees

The following 4 lemmas immediately follow from the definitions and we leave the proofs to the Reader.

Lemma 4.1. *For any rgood binary tree T, all the induced rooted subtrees $T_1, T_2, T_{11}, T_{12}, T_{21}, T_{22}, \ldots$ are rgood as well.* $\square$

Lemma 4.2. *For any two rgood binary trees T and T' with roots r and r' respectively, we have*

$$h(T) > h(T') \quad \Rightarrow \quad |V(T)| > |V(T')|; \tag{4}$$

$$|V(T)| \geq |V(T')| \quad \Rightarrow \quad h(T) \geq h(T'); \tag{5}$$

$$f_T(r) > f_{T'}(r') \Leftrightarrow |V(T)| > |V(T')| \quad \text{and} \quad f_T(r) = f_{T'}(r') \Leftrightarrow |V(T)| = |V(T')|. \tag{6}$$

Thus, when trying to compare the number of subtrees containing the roots of some rgood trees, it suffices to compare their sizes. $\square$

Lemma 4.3. *Assume that in a rooted binary tree T, the induced subtrees at the children of the root, T_1 and T_2, are rgood. Now T is rgood if and only if one of the following conditions hold:*
i) $h(T_1) = h(T_2)$, and T_2 is complete;
ii) $h(T_1) = h(T_2) - 1$, and T_1 is complete. $\square$

Lemma 4.4. *Let us be given two rgood binary trees, T' and T'', such that $h(T') \leq h(T'')$. Join with an edge the roots of T' and T'' to obtain the binary tree T. Now T is good if and only if one of the following conditions hold:*
i) $h(T') = h(T'')$, and one or both of T' and T'' is complete;
ii) $h(T') = h(T'') - 1$, and T' is complete. $\square$

Lemma 4.5. *If T is an rgood binary tree, then (T_1, r_1) is isomorphic to a subtree of (T_2, r_2), and consequently $(T_{1i_1\ldots i_k}, r_{1i_1\ldots i_k})$ is isomorphic to a subtree of $(T_{2i_1\ldots i_k}, r_{2i_1\ldots i_k})$ for every $i_j \in \{1, 2\}$ such that $r_{1i_1\ldots i_k}$ exists.*

Proof. An immediate consequence of Lemma 4.3 $\square$

Lemma 4.6. *For any rgood binary tree T and any $k \geq 0$, we have*

$$f_{T_1}(v_1) \geq f_{T_{\underbrace{2\ldots21}_{k\ 2's}}}(v_{\underbrace{2\ldots21}_{k\ 2's}}). \tag{7}$$

Proof. For $k = 0$, (7) holds with identity. For $k \geq 1$, we consider two cases:
If $h(T_1) = h(T_2)$, then $h(T_1) > h(T_{21}) \geq h(T_{\underbrace{2\ldots21}_{k\ 2's}})$, and (7) holds by (4) and (6).

If $h(T_1) = h(T_2) - 1$, then by Lemma 4.3, T_1 is complete. Notice that $h(T_1) = h(T_2) - 1 \geq h(T_{\underbrace{2\ldots21}_{k\ 2's}})$ for $k \geq 1$, hence (3) applies to the rooted trees T_1 and $T_{\underbrace{2\ldots21}_{k\ 2's}}$. Hence, (7) holds. $\square$

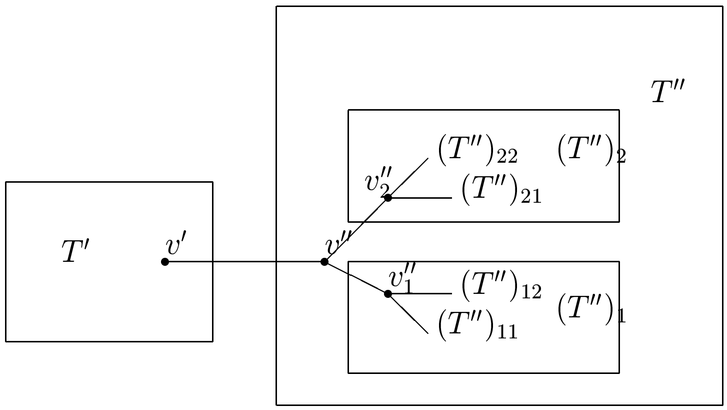

Figure 4: Dividing a binary tree T into two rooted binary trees.

5 The structure of optimal binary trees

For brevity, we will call a binary tree maximizing the number of subtrees among binary trees with the same number of leaves *optimal*. We will show several lemmas describing parts of optimal binary trees. For any binary tree T, the deletion of an edge $v'v''$ divides T into two rooted binary trees T' and T'' with roots v' and v'' respectively.

Lemma 5.1. *Assume T is an optimal binary tree. Assume that T is divided into two rooted subtrees T', T'' by the removal of the edge $v'v''$ as shown in Fig. 4. Then, if for all $k \geq 1$ the inequalities*

$$f_{T'}(v') > f_{(T'')\underbrace{2\ldots21}_{k\ 2's}}(v''_{\underbrace{2\ldots21}_{k\ 2's}}), \tag{8}$$

hold as far as vertex $v''_{\underbrace{2\ldots21}_{k\ 2's}}$ exists, then T'' is rgood.

Note: We understand that (8) holds if $(T'')_{21}$ does not exist. Then $(T'')_2$ is a single vertex, and by (1) $(T'')_1$ is also a single vertex. Therefore T'' is rgood as Lemma 5.1 requires.

Proof. The proof goes by induction on $|V(T'')|$. The base case: if $|V(T'')| = 1$, then by definition, T'' is rgood. Now, suppose that Lemma 5.1 holds for any induced subtree in place of T'' with fewer vertices. We are going to show the following:

Claim 5.1. $(T'')_1$ *and* $(T'')_2$ *are rgood.*

Proof. Consider $(T'')_1$ and $(T'')_2$ with roots v_1'' and v_2''. For $(T'')_1$, consider T as being divided into $T''' = ((T'')_1, v_1'')$ and $T^* = (T' \cup (T'')_2 \cup \{v''\}, v'')$. Notice that for any $k \geq 1$,

$$\begin{aligned} f_{T^*}(v'') &>^{(2)} f_{(T'')_2}(v_2'') \geq^{(1)} f_{(T'')_1}(v_1'') \\ &>^{(2)} f_{(T'')\underbrace{12\ldots21}_{k\ 2's}}(v''_{\underbrace{12\ldots21}_{k\ 2's}}) = f_{(T''')\underbrace{2\ldots21}_{k\ 2's}}(v''_{\underbrace{12\ldots21}_{k\ 2's}}), \end{aligned}$$

thus (8) holds for T^* and T'''. By hypothesis, it follows that $(T'')_1$ is rgood. (We fall into the habit of superscripting some inequalities for a reference to their proofs.)

For $(T'')_2$, consider T as being divided into $T''' = ((T'')_2, v''_2)$ and $T^* = (T' \cup (T'')_1 \cup \{v''\}, v'')$. We have for any $k \geq 1$

$$f_{T^*}(v'') >^{(2)} f_{T'}(v') >^{(8)} f_{(T'')_{\underbrace{2\ldots21}_{k+1\ 2's}}}(v''_{\underbrace{2\ldots21}_{k+1\ 2's}}) = f_{(T''')_{\underbrace{2\ldots21}_{k\ 2's}}}(v''_{\underbrace{2\ldots21}_{k+1\ 2's}}),$$

thus (8) holds for T^* and T'''. By hypothesis, it follows that $(T'')_2$ must be rgood. $\square$

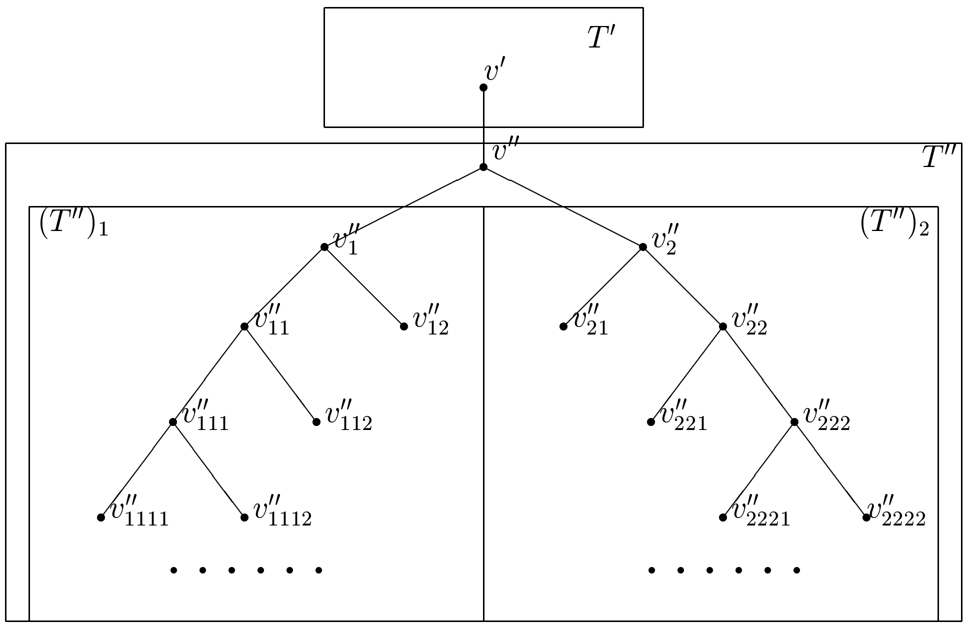

Figure 5: Considering subtrees of T''.

Knowing that $(T'')_1$ and $(T'')_2$ are rgood, we return to the inductive step in the proof of Lemma 5.1. We consider the following cases: (i) $h((T'')_1) < h((T'')_2)$ and (ii) $h((T'')_1) = h((T'')_2)$. (Note that the third inequality $h((T'')_1) > h((T'')_2)$ is impossible by the rgoodness of $(T'')_1$ and $(T'')_2$, (1) and Lemma 4.2).

Case (i): $h((T'')_1) < h((T'')_2)$.

By (6), (4) and Claim 5.1, we have $|V((T'')_2)| > |V((T'')_1)|$ and $f_{(T'')_2}(v''_2) > f_{(T'')_1}(v''_1)$.

Claim 5.2. *For any $k \geq 0$ such that $(T'')_{\underbrace{1\ldots1}_{k}}$ is not empty, we have*

$$|V((T'')_{\underbrace{1\ldots1}_{k}})| \geq |V((T'')_{\underbrace{22\ldots2}_{k+1}})|. \tag{9}$$

Proof. The proof goes by induction on k. The base case $k = 0$ is trivial. For the inductive step, suppose that (9) holds for $k = 0, 1, 2, \ldots, l$. We are going to prove that (9) also holds for $k = l + 1$, if $(T'')_{\underbrace{1\ldots1}_{l+1}}$ is not empty. We need that for $k = 0, 1, 2, ..., l$

$$|V((T'')_{\underbrace{1\ldots12}_{k\ 1's}})| \geq |V((T'')_{\underbrace{22\ldots21}_{k+1\ 2's}})|. \tag{10}$$

Indeed, $|V((T'')_{\underbrace{1\ldots 1}_{k\ 1's}2})| \geq \frac{1}{2}(|V((T'')_{\underbrace{1\ldots 1}_{k}})|-1)$, since by Claim 5.1 and Lemma 4.1 all rooted subtrees of $(T'')_1$ and $(T'')_2$ are rgood, and therefore convention (1) and formula (6) apply. A similar argument shows $\frac{1}{2}(|V((T'')_{\underbrace{22\ldots 2}_{k+1}})|-1) \geq |V((T'')_{\underbrace{22\ldots 2}_{k+1\ 2's}1})|$. Combining these with the hypothesis (9) for $k=l$, we obtain (10).

For contradiction, assume that (9) does not hold for $k=l+1$, i.e.

$$|V((T'')_{\underbrace{1\ldots 11}_{l+1}})| < |V((T'')_{\underbrace{22\ldots 22}_{l+2}})|. \tag{11}$$

Through Claim 5.1, Lemma 4.1, and (6), formula (11) implies

$$f_{(T'')_{\underbrace{1\ldots 11}_{l+1}}}(v''_{\underbrace{1\ldots 11}_{l+1}}) < f_{(T'')_{\underbrace{22\ldots 22}_{l+2}}}(v''_{\underbrace{22\ldots 22}_{l+2}}). \tag{12}$$

Observe that

$$|V((T'')_{\underbrace{1\ldots 1}_{l\ 1's}2})| + |V((T'')_{\underbrace{1\ldots 11}_{l+1}})| = |V((T'')_{\underbrace{1\ldots 11}_{l}})| - 1$$

$$\geq^{(9,k=l)} |V((T'')_{\underbrace{22\ldots 2}_{l+1}})| - 1 = |V((T'')_{\underbrace{22\ldots 2}_{l+1\ 2's}1})| + |V((T'')_{\underbrace{22\ldots 22}_{l+2}})|,$$

and therefore (11) implies that strict inequality holds in (10) when $k=l$, i.e.

$$|V((T'')_{\underbrace{1\ldots 1}_{l\ 1's}2})| > |V((T'')_{\underbrace{22\ldots 2}_{l+1\ 2's}1})|. \tag{13}$$

Now we are in the position to apply Lemma 3.2 in the following setting:

$$x \leftarrow v''_{\underbrace{1\ldots 11}_{l+1}};\, x_i \leftarrow v''_{\underbrace{1\ldots 1}_{l+1-i}};\, x_{l+1} \leftarrow v'';\, y_{l+1} \leftarrow v''_2;\, y_i \leftarrow v''_{\underbrace{22\ldots 2}_{l+2-i}};\, y \leftarrow v''_{\underbrace{22\ldots 22}_{l+2}}$$

for $i=1,2,\ldots,l$. For the subtrees, the substitution is

$$X \leftarrow ((T'')_{\underbrace{1\ldots 11}_{l+1}}, v''_{\underbrace{1\ldots 11}_{l+1}});\quad X_i \leftarrow ((T'')_{\underbrace{1\ldots 1}_{l+1-i\ 1's}2} \cup \{v''_{\underbrace{1\ldots 1}_{l+1-i}}\}, v''_{\underbrace{1\ldots 1}_{l+1-i}});$$

$$X_{l+1} \leftarrow (T' \cup \{v''\}, v'');\quad Y_{l+1} \leftarrow ((T'')_{21} \cup \{v''_2\}, v''_2);$$

$$Y_i \leftarrow ((T'')_{\underbrace{22\ldots 2}_{l+2-i\ 2's}1} \cup \{v''_{\underbrace{22\ldots 2}_{l+2-i}}\}, v''_{\underbrace{22\ldots 2}_{l+2-i}});\quad Y \leftarrow ((T'')_{\underbrace{22\ldots 22}_{l+2}}, v''_{\underbrace{22\ldots 22}_{l+2}}),$$

$$S \leftarrow (T\backslash(X \cup Y)) \cup \{x, y\},$$

for where $i=1,2,\ldots,l$. Using the notation in Lemma 3.2, we have

$$a_i = f_{(T'')_{\underbrace{1\ldots 1}_{l+1-i\ 1's}2}}(v''_{\underbrace{1\ldots 1}_{l+1-i\ 1's}2}) + 1 \geq f_{(T'')_{\underbrace{22\ldots 2}_{l+2-i\ 2's}1}}(v''_{\underbrace{22\ldots 2}_{l+2-i\ 2's}1}) + 1 = b_i \tag{14}$$

for $i=1,2,\ldots,l$, by (10) and (6). In fact, strict inequality holds in (14) for $i=1$ by (13). We also have

$$a_{l+1} = f_{T'}(v') + 1 > f_{(T'')_{21}}(v''_{21}) + 1 = b_{l+1}$$

by (8). From here, we obtain the conclusion of Lemma 3.2, which is exactly the first condition of Lemma 3.3 as well:

$$f_S(x) > f_S(y).$$

We also have the other condition of Lemma 3.3

$$f_X(x) = f_{(T'')\underbrace{1\dots 11}_{l+1}}(v''_{\underbrace{1\dots 11}_{l+1}}) < f_{(T'')\underbrace{22\dots 22}_{l+2}}(v''_{\underbrace{22\dots 22}_{l+2}}) = f_Y(y)$$

from (12). Thus, by Lemma 3.3, interchanging X and Y increases $F(T)$, contradicting the optimality of T. Hence (9) holds for $k = l+1$, and we completed the induction proof. □

Since $(T'')_{\underbrace{1\dots 1}_{k}}$ and $(T'')_{\underbrace{22\dots 2}_{k+1}}$ are rgood trees, (9) implies through (5) that

$$h((T'')_{\underbrace{1\dots 1}_{k}}) \geq h((T'')_{\underbrace{22\dots 2}_{k+1}}) \tag{15}$$

for any $k \geq 1$ such that $(T'')_{\underbrace{1\dots 1}_{k}}$ is not empty. On the other hand, since we are in the case $h((T'')_1) < h((T'')_2)$, we have

$$h((T'')_1) \leq h((T'')_2) - 1 = h((T'')_{22}),$$

$$h((T'')_{11}) \leq h((T'')_1) - 1 \leq h((T'')_{22}) - 1 = h((T'')_{222}),$$

$$\dots,$$

$$h((T'')_{\underbrace{1\dots 1}_{k}}) \leq h((T'')_{\underbrace{22\dots 2}_{k+1}}) \tag{16}$$

for any $k \geq 1$ such that $(T'')_{\underbrace{1\dots 1}_{k}}$ is not empty. Comparing (15) and (16), we conclude that equality holds all the way in (15) and (16) until both $(T'')_{11\dots 1}$ and $(T'')_{222\dots 2}$ turns into a single vertex. In this case $(T'')_1$ is complete and of height $h((T'')_2) - 1$. By Lemma 4.3, T'' is rgood. End of Case (i).

Case (ii): $h((T'')_1) = h((T'')_2)$.

Claim 5.3. *For any $k \geq 0$ such that $(T'')_{\underbrace{21\dots 1}_{k\ 1's}}$ is not empty, we have*

$$|V((T'')_{\underbrace{21\dots 1}_{k\ 1's}})| \geq |V((T'')_{\underbrace{12\dots 2}_{k\ 2's}})| \tag{17}$$

Proof. The proof goes by induction on k. The base case $k = 0$ follows from Lemma 4.2 and Claim 5.1. For the inductive step, suppose that (17) holds for $k = 0, 1, 2, \dots, l$. We are going to prove that (17) also holds for $k = l+1$, if $(T'')_{\underbrace{21\dots 1}_{k\ 1's}}$ is not empty. Hypothesis $|V((T'')_{\underbrace{21\dots 1}_{k\ 1's}})| \geq |V((T'')_{\underbrace{12\dots 2}_{k\ 2's}})|$ implies that

$$|V((T'')_{\underbrace{21\dots 12}_{k\ 1's}})| \geq |V((T'')_{\underbrace{12\dots 21}_{k\ 2's}})| \tag{18}$$

through the facts that these trees are rgood by Claim 5.1, labelled according to the convention (1), and formula (6). For contradiction, assume that (17) does not hold for $k = l+1$, i.e.

$$|V((T'')\underbrace{21\ldots 11}_{l+1\ 1's})| < |V((T'')\underbrace{12\ldots 22}_{l+1\ 2's})|. \tag{19}$$

Notice that

$$|V((T'')\underbrace{21\ldots 12}_{l\ 1's})| + |V((T'')\underbrace{21\ldots 11}_{l+1\ 1's})| = |V((T'')\underbrace{21\ldots 1}_{l\ 1's})| - 1$$

$$\geq^{(17,k=l)} |V((T'')\underbrace{12\ldots 2}_{l\ 2's})| - 1 = |V((T'')\underbrace{12\ldots 21}_{l\ 2's})| + |V((T'')\underbrace{12\ldots 22}_{l+1\ 2's})|.$$

Therefore (19) implies that strict inequality holds in (18) for $k = l$, i.e.

$$|V((T'')\underbrace{21\ldots 12}_{l\ 1's})| > |V((T'')\underbrace{12\ldots 21}_{l\ 2's})|. \tag{20}$$

Now we are in the position to apply Lemma 3.2 in the following setting:

$$x \leftarrow v''_{\underbrace{21\ldots 11}_{l+1\ 1's}};\ x_i \leftarrow v''_{\underbrace{21\ldots 1}_{l+1-i\ 1's}};\ z \leftarrow v''; y_i \leftarrow v''_{\underbrace{12\ldots 2}_{l+1-i\ 2's}};\ y \leftarrow v''_{\underbrace{12\ldots 22}_{l+1\ 2's}};$$

$$X \leftarrow ((T'')\underbrace{21\ldots 11}_{l+1\ 1's}, v''_{\underbrace{21\ldots 11}_{l+1\ 1's}}),\ X_i \leftarrow ((T'')\underbrace{21\ldots 12}_{l+1-i\ 1's} \cup \{v''_{\underbrace{21\ldots 1}_{l+1-i\ 1's}}\}, v''_{\underbrace{21\ldots 1}_{l+1-i\ 1's}});$$

$$Z \leftarrow (T' \cup \{v''\}, v'');$$

$$Y_i \leftarrow ((T'')\underbrace{12\ldots 21}_{l+1-i\ 2's} \cup \{v''_{\underbrace{12\ldots 2}_{l+1-i\ 2's}}\}, v''_{\underbrace{12\ldots 2}_{l+1-i\ 2's}});\ Y \leftarrow ((T'')\underbrace{12\ldots 22}_{l+1\ 2's}, v''_{\underbrace{12\ldots 22}_{l+1\ 2's}});$$

$$S \leftarrow (T\backslash(X \cup Y)) \cup \{x, y\},$$

for $i = 1, 2, \ldots, l+1$. Using the notation in Lemma 3.2, we have

$$a_i = f_{(T'')\underbrace{21\ldots 12}_{l+1-i\ 1's}}(v''_{\underbrace{21\ldots 12}_{l+1-i\ 1's}}) + 1 \geq f_{(T'')\underbrace{12\ldots 21}_{l+1-i\ 2's}}(v''_{\underbrace{12\ldots 21}_{l+1-i\ 2's}}) + 1 = b_i \tag{21}$$

for $i = 1, 2, \ldots, l+1$, by (18) and (6). In fact, strict inequality holds in (21) for $i = 1$ by (20), and therefore $a_1 > b_1$. From here, we obtain the conclusion of Lemma 3.2, which is exactly the first condition of Lemma 3.3 as well:

$$f_S(x) > f_S(y).$$

By (19) (also using Claim 5.1, Lemma 4.1, and (6)) we also have the second condition of Lemma 3.3:

$$f_X(x) = f_{(T'')\underbrace{21\ldots 11}_{l+1\ 1's}}(v''_{\underbrace{21\ldots 11}_{l+1\ 1's}}) < f_{(T'')\underbrace{12\ldots 22}_{l+1\ 2's}}(v''_{\underbrace{12\ldots 22}_{l+1\ 2's}}) = f_Y(y).$$

Thus, Lemma 3.3 applies, interchanging X and Y increases $F(T)$, contradicting the optimality of T. Hence (17) holds for $k = l+1$. Using induction, we proved Claim 5.3. $\square$

Notice that the trees mentioned in (17) are rgood by Claim 5.1 and Lemma 4.1, and therefore (17) implies through (5) that

$$h((T'')_{\underbrace{21\dots1}_{k\ 1's}}) \geq h((T'')_{\underbrace{12\dots2}_{k\ 2's}}) \tag{22}$$

for any $k \geq 1$ such that $(T'')_{\underbrace{21\dots1}_{k\ 2's}}$ is not empty. On the other hand, since we are in the case $h((T'')_1) = h((T'')_2)$, we must have

$$h((T'')_{21}) \leq h((T'')_2) - 1 = h((T'')_1) - 1 = h((T'')_{12}),$$

$$h((T'')_{211}) \leq h((T'')_{21}) - 1 \leq^{(22)} h((T'')_{12}) - 1 = h((T'')_{122}),$$

$$\dots,$$

$$h((T'')_{\underbrace{21\dots1}_{k\ 1's}}) \leq^{(22)} h((T'')_{\underbrace{12\dots2}_{k\ 2's}}) \tag{23}$$

for any $k \geq 1$ such that $(T'')_{\underbrace{21\dots1}_{k\ 1's}}$ is not empty.

Comparing (22) and (23), we conclude that equality holds all the way in (22) and (23) until both $(T'')_{21\dots1}$ and $(T'')_{12\dots2}$ turns into a single vertex. In this case $(T'')_2$ is complete and $h((T'')_2) = h((T'')_1)$. By Lemma 4.3, T'' is rgood. End of Proof to Lemma 5.1. □

Now consider an optimal binary tree T which maximizes $F(T)$ among n-leaf binary trees. Divide T into two rooted binary trees (T', v') and (T'', v'') by deleting an edge $v'v''$. We obtain the following two lemmas.

Lemma 5.2. *If $|h(T'') - h(T')| \leq 1$, then T' and T'' both must be rgood.*

Note that if we choose a longest path P and choose (v', v'') as the closest to middle edge on P, we obtain such a T' and T''.

Proof. Without loss of generality, we can assume $f_{T''}(v'') \geq f_{T'}(v')$ (see Lemma 4.2). First, it is easy to see that for any $k \geq 1$

$$f_{T''}(v'') \geq f_{T'}(v') >^{(2)} f_{(T')_{\underbrace{2\dots21}_{k\ 2's}}}(v'_{\underbrace{2\dots21}_{k\ 2's}}).$$

Thus condition (8) holds, and by Lemma 5.1, T' is rgood.

On the one hand, since T' is rgood, T' must contain a complete rooted binary tree T^*, with the same root, of height at least $h(T') - 1 \geq h(T'') - 2$. On the other hand, $(T'')_{\underbrace{2\dots21}_{k\ 2's}}$ is of height at most $h(T'') - 2$ and is isomorphic to a subtree of T' (sharing the same root). Therefore

$$f_{T'}(v') \geq^{(4,6,3)} f_{(T'')_{\underbrace{2\dots21}_{k\ 2's}}} \tag{24}$$

for $k \geq 1$. In fact, (24) is always a strict inequality, since T' has some other vertices than those in the complete rooted binary tree with height $h(T') - 1$. So condition (8) holds, T'' is also rgood. □

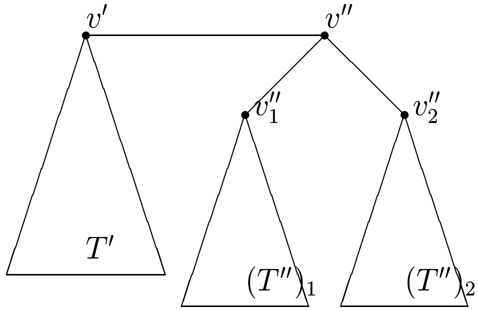

Figure 6: The optimal binary tree T, which maximizes $F(T)$.

Let T be divided into T' and T'' by deleting the closest to middle edge as described after Lemma 5.2. By Lemma 5.2, T' and T'' are both rgood. Without loss of generality we may assume that $f_{T''}(v'') \geq f_{T'}(v')$ (and also $h(T'') \geq h(T')$ by (4) and (6)).

Lemma 5.3. *T' is complete or $T^* = (T' \cup (T'')_1 \cup \{v''\}, v'')$ is rgood.*

Proof. Assume that T' is not complete, and therefore $f_{(T')_1}(v'_1) < \frac{1}{2}[f_{T'}(v') - 1]$. We have that $f_{(T'')_2}(v''_2) \geq^{(1)} \frac{1}{2}[f_{T''}(v'') - 1]$ and $1 \leq f_{T'}(v') \leq f_{T''}(v'')$; and therefore

$$f_{(T')_1}(v'_1) < f_{(T'')_2}(v''_2). \tag{25}$$

Consider T as being divided into T^* and $(T'')_2$. Since T' is rgood by Lemma 5.2, Lemma 4.6 yields for any $k \geq 0$

$$f_{(T')\underbrace{2\ldots21}_{k\ 2's}}(v'_{\underbrace{2\ldots21}_{k\ 2's}}) \leq^{(7)} f_{(T')_1}(v'_1). \tag{26}$$

Combining (25) with (26) yields for any $k \geq 0$

$$f_{(T')\underbrace{2\ldots21}_{k\ 2's}}(v'_{\underbrace{2\ldots21}_{k\ 2's}}) < f_{(T'')_2}(v''_2). \tag{27}$$

Similarly, notice that $(T'')_1$ is rgood, and then for $k \geq 0$,

$$f_{(T'')_2}(v''_2) \geq^{(1)} f_{(T'')_1}(v''_1) >^{(2)} f_{(T'')_{11}}(v''_{11}) \geq^{(7)} f_{(T'')\underbrace{12\ldots21}_{k\ 2's}}(v''_{\underbrace{12\ldots21}_{k\ 2's}}). \tag{28}$$

Combining (27) and (28), we obtain that for any $k \geq 0$,

$$f_{(T'')_2}(v''_2) > \max\left(f_{(T')\underbrace{2\ldots21}_{k\ 2's}}(v'_{\underbrace{2\ldots21}_{k\ 2's}}), f_{(T'')\underbrace{12\ldots21}_{k\ 2's}}(v''_{\underbrace{12\ldots21}_{k\ 2's}}) \right). \tag{29}$$

Since $(T^*)_2 = T'$ or $(T'')_1$, we have from (29) that

$$f_{(T'')_2}(v''_2) > f_{(T^*)\underbrace{2\ldots21}_{k+1\ 2's}}(r^*) \text{ for } k \geq 0,$$

where r^* is the root of $(T^*)_{\underbrace{2\ldots21}_{k+1\ 2's}}$. So (8) holds, T^* is rgood by Lemma 5.1. $\square$

6 The proof of Theorem 2.2

Proof. Let T be an optimal binary tree on n leaves. For contradiction, suppose that T is not good. Divide T into T' and T'' by deleting the closest to middle edge as described before Lemma 5.3. By Lemma 5.2, both T' and T'' are rgood. We assume that $f_{T''}(v'') \geq f_{T'}(v')$, and also $h(T'') \geq h(T')$ by (4), (5) and (6). (Figs. 4, 5, and 6 explain how the vertices are labelled.) Since T'' is rgood,

$$h(T'') - 2 \leq h((T'')_1) \leq h(T'') - 1 = h((T'')_2). \tag{30}$$

By definition, $h(T'') - 1 \leq h(T') \leq h(T'')$. According to Lemma 4.4, T' is not complete, and if $h(T') = h(T'')$, then T'' is not complete either. Define $T^* = (T' \cup (T'')_1 \cup \{v''\}, v'')$ (as in Lemma 5.3). Since T' is not complete, T^* must be rgood (Lemma 5.3)) and so, by Lemma 4.3,

$$(T'')_1 \text{ must be complete.} \tag{31}$$

If $h(T') = h(T'')$, then since T'' is not complete and (31), we must have $h((T'')_1) = h((T'')_2) - 1 = h(T') - 2$. But this contradicts the rgoodness of T^*, (it would have leaves at heights differing by 2), therefore we must have

$$h(T') = h(T'') - 1. \tag{32}$$

Assume at this point for a second $h((T'')_1) = h((T'')_2)$. Applying Lemma 4.3 to T'' yields that $(T'')_2$ must be complete, and consequently, by (31), T'' must be complete. Now, let $T''' = (T' \cup (T'')_2 \cup \{v''\}, v'')$. Then $h(T''') = h(T') + 1 = h((T'')_2) + 1 = h((T'')_1) + 1$, the completeness of $(T'')_2$ and T' indicates that T''' is complete. $(T'')_1$ is complete by (31), and observe $h(T''') = h(T') + 1 = h((T'')_2) + 1 = h((T'')_1) + 1$. Apply Lemma 4.4 (ii) for joining T''' and $(T'')_1$ to obtain T, and observe that T is good, a contradiction. Therefore we have $h((T'')_1) = h((T'')_2) - 1$. Assume now for a second that $(T'')_2$ is complete. Now draw T by placing the edge $v''v''_2$ to the line $\mathbb{R} \times 0$ and observe that T is good, a contradiction. Therefore we may assume for the rest of the proof that

$$(T'')_2 \text{ is not complete, and } h((T'')_1) = h((T'')_2) - 1 = h(T'') - 2 = h(T') - 1. \tag{33}$$

Set $T''' = (T' \cup (T'')_2 \cup \{v''\}, v'')$. Consider now T as being divided into T''' and $(T'')_1$, and note that T''' is not rgood as neither T' nor $(T'')_2$ are complete. First we will show that for all $k \geq 0$, we have

$$f_{(T'')_1}(v''_1) > f_{(T')\underbrace{2\ldots21}_{k\ 2's}}(v'\underbrace{2\ldots21}_{k\ 2's}). \tag{34}$$

Now

$$h((T')\underbrace{22\ldots21}_{k\ 2's}) \leq h(T') - (k+1) =^{(33)} h((T'')_1) - k, \tag{35}$$

so (34) holds for $k \geq 1$ by (4) and (6).

Also if $h((T')_1) = h((T')_2) - 1 <^{(33)} h((T'')_1)$, then (34) holds for $k = 0$ by (4) and (6). Therefore we only need to show that (34) holds for $k = 0$ when $h((T')_1) = h((T')_2) = h(T') - 1 =^{(33)} h(T''_1)$. But since T' is not complete, if $h((T')_1) = h((T')_2)$ then $(T')_1$ must not be complete, and since $(T'')_1$ is complete, we get from (6) that $f_{(T'')_1}(v''_1) > f_{(T')_1}(v'_1)$, and therefore (34) is true.

Similarly to the above, we will also show that for every $k \geq 1$ we have

$$f_{(T'')_1}(v''_1) > f_{(T'')\underbrace{22\ldots21}_{k\ 2's}}(v''\underbrace{22\ldots21}_{k\ 2's}). \tag{36}$$

As before, $h((T'')\underbrace{22\ldots21}_{k\ 2's}) \leq h(T'') - (k+1) =^{(33)} h((T'')_1) - (k-1)$ so (36) holds for $k \geq 2$ by (4). Also if $h((T'')_{21}) = h((T'')_{22}) - 1 = h(T'') - 3 <^{(33)} h((T'')_1)$, then (36) holds for $k = 1$ by (4).

So, since $(T'')_2$ is rgood, all we need to show is that (36) holds for $k = 1$ when $h((T'')_{21}) = h((T'')_{22}) =^{(33)} h((T'')_1)$. But since $(T'')_2$ is not complete, from (6) we have in this case that $f_{(T'')_1}(v''_1) > f_{(T'')_{21}}(v''_{21})$ as required.

Combining (34) with (36), we obtain that for any $k \geq 0$,

$$f_{(T'')_1}(v''_1) > \max\left(f_{(T')\underbrace{2\ldots21}_{k\ 2's}}(v'\underbrace{2\ldots21}_{k\ 2's}), f_{(T'')\underbrace{22\ldots21}_{k+1\ 2's}}(v''\underbrace{22\ldots21}_{k+1\ 2's})\right). \tag{37}$$

Since $(T''')_2 = T'$ or $(T'')_2$, we have from (37) that

$$f_{(T'')_1}(v''_1) > f_{(T''')\underbrace{2\ldots21}_{k+1\ 2's}}(r) \text{ for } k \geq 0,$$

where r is the root of $(T''')\underbrace{2\ldots21}_{k+1\ 2's}$. So (8) holds, but T''' is not rgood as neither of T' or $(T'')_2$ is complete, contradiction to Lemma 5.1.

Thus, we must have that T is good.

□

Acknowledgment. The authors thank Éva Czabarka for her careful reading of an earlier version of the manuscript and useful remarks.

References

[1] Dobrynin, A.A., Entringer, R., Gutman, I., Wiener index of trees: Theory and applications, *Acta Appl. Math.* **66** (3) (2001), 211–249.

[2] Fischermann, M., Hoffmann, A., Rautenbach, D., Székely, L.A., Volkmann, L., Wiener index versus maximum degree in trees, *Discrete Appl. Math.* **122** (1–3) (2002), 127–137.

[3] Jelen, F., Triesch, E., Superdominance order and distance of trees with bounded maximum degree, *Discrete Appl. Math.* **125** (2–3) (2003), 225–233.

[4] Knudsen, B., Optimal multiple parsimony alignment with affine gap cost using a phylogenetic tree, Lecture Notes in Bioinformatics 2812, Springer Verlag, 2003, 433–446.

[5] Székely, L.A., Wang, H., On subtrees of trees, submitted.