Discrete Applied Mathematics 291 (2021) 88–98

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

On the Colijn–Plazzotta numbering scheme for unlabeled binary rooted trees

Noah A. Rosenberg

Department of Biology, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 21 May 2020
Received in revised form 19 November 2020
Accepted 23 November 2020
Available online xxxx

Keywords:
Phylogenetics
Quadratic recursion
Unlabeled trees

ABSTRACT

Colijn and Plazzotta (2018) introduced a scheme for bijectively associating the unlabeled binary rooted trees with the positive integers. First, the rank 1 is associated with the 1-leaf tree. Proceeding recursively, ordered pair (k_1, k_2), $k_1 \geqslant k_2 \geqslant 1$, is then associated with the tree whose left subtree has rank k_1 and whose right subtree has rank k_2. Following dictionary order on ordered pairs, the tree whose left and right subtrees have the ordered pair of ranks (k_1, k_2) is assigned rank $k_1(k_1 - 1)/2 + 1 + k_2$. With this ranking, given a number of leaves n, we determine recursions for a_n, the smallest rank assigned to some tree with n leaves, and b_n, the largest rank assigned to some tree with n leaves. The smallest rank a_n is assigned to the maximally balanced tree, and the largest rank b_n is assigned to the caterpillar. For n equal to a power of 2, the value of a_n is seen to increase exponentially with $2\alpha^n$ for a constant $\alpha \approx 1.24602$; more generally, we show it is bounded $a_n < 1.5^n$. The value of b_n is seen to increase with $2\beta^{(2^n)}$ for a constant $\beta \approx 1.05653$. The great difference in the rates of increase for a_n and b_n indicates that as the index v is incremented, the number of leaves for the tree associated with rank v quickly traverses a wide range of values. We interpret the results in relation to applications in evolutionary biology.

1. Introduction

For a given number of leaves $n \geqslant 2$, the unlabeled binary rooted trees with n leaves can be obtained recursively, beginning from the trivial tree with a single leaf (Table 1). For fixed n, we enumerate all possible pairings of a tree of size k leaves with a tree of size $n - k$ leaves, for each k from 1 to $\lfloor \frac{n}{2} \rfloor$. For each $k < \frac{n}{2}$, each pairing of a tree of size k and a tree of size $n - k$ generates a distinct unlabeled binary rooted tree; for even n and $k = \frac{n}{2}$, we enumerate pairings of distinct trees of size $\frac{n}{2}$ and pairings of identical trees of size $\frac{n}{2}$.

Letting U_n denote the number of unlabeled binary rooted trees with n leaves, we have [10, p. 29]

$$U_n = \begin{cases} 1, & \text{if } n = 1 \\ \sum_{k=1}^{(n-1)/2} U_k U_{n-k}, & \text{if } n \text{ is odd and } n \geqslant 3 \\ [\sum_{k=1}^{(n-2)/2} U_k U_{n-k}] + U_{n/2}(U_{n/2}+1)/2, & \text{if } n \text{ is even.} \end{cases} \quad (1)$$

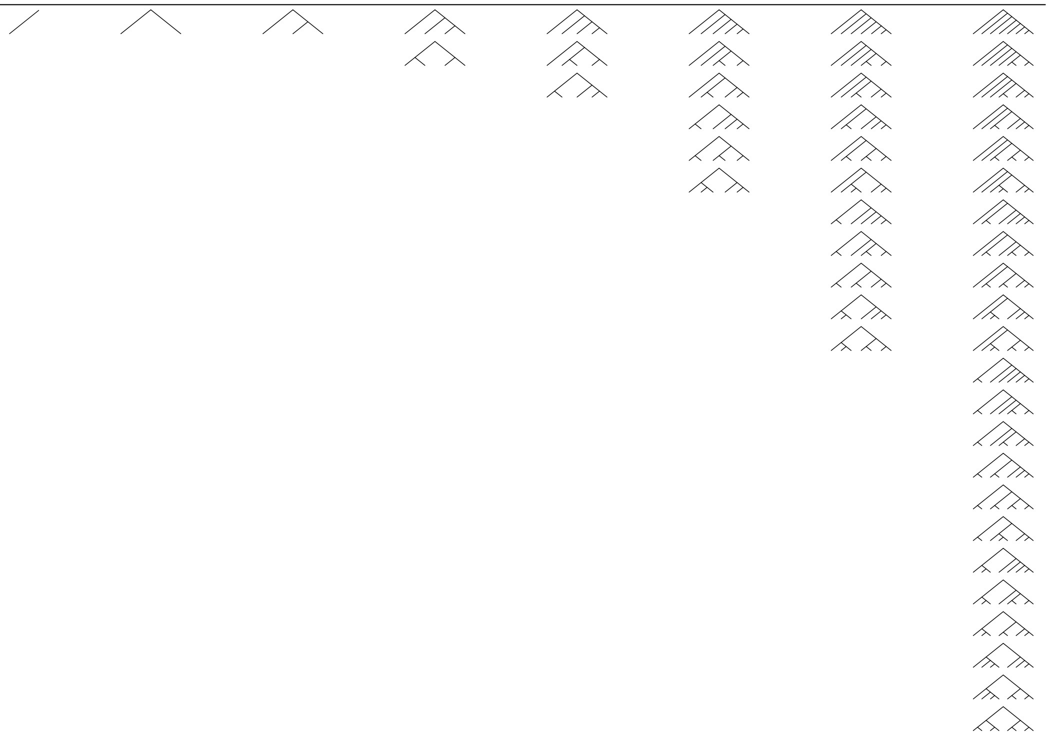
The sequence of values U_n, the Wedderburn–Etherington numbers, begins from $n = 1$ with 1, 1, 1, 2, 3, 6, 11, 23, 46, 98, 207, 451, 983, 2179, 4850 (Table 2, A001194 in OEIS). U_n is straightforward to calculate from $U_1, U_2, \ldots, U_{n-1}$ via the recursion in Eq. (1). However, no closed-form expression is known.

E-mail address: noahr@stanford.edu.

https://doi.org/10.1016/j.dam.2020.11.021

Table 1
Furnas ranks of unlabeled binary rooted trees with $1 \leqslant n \leqslant 8$ leaves.

Furnas rank v	n							
	1	2	3	4	5	6	7	8
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								



For a fixed value of n, the unlabeled binary rooted trees can be enumerated in the sequence in which they appear in the recursion. According to the ranking scheme of Furnas [11] for trees of size n leaves, $k \leqslant \lfloor \frac{n}{2} \rfloor$ is viewed as the size of the left subtree of a tree of size $n \geqslant 2$ and $n-k$ is the size of the right subtree. Trees with n leaves that have a lower value of k are assigned lower rank. Trees with n leaves that have the same value of k are ordered by the rank of their left subtree, and trees with n leaves that have the same value of k and the same left subtree are ordered by the rank of their right subtree. For trees with two distinct subtrees of size $\frac{n}{2}$, the one with lower Furnas rank appears on the left (Table 1).

Using the Furnas ranking, we can construct a simple bijection between the unlabeled binary rooted trees and the positive integers. For each $n \geqslant 1$, we simply list all unlabeled trees with n leaves in increasing order according to their Furnas rank, and we then proceed to trees of $n+1$ leaves. More precisely, for $n \geqslant 1$, we let $S_n = \sum_{k=1}^{n} U_k$ denote the sum of the Wedderburn–Etherington numbers, with $S_0 = 0$ (A173282 in OEIS). In the bijection, the tree of size n with Furnas rank v, $1 \leqslant v \leqslant U_n$, is associated with the integer $S_{n-1} + v$. The trees of size n are then associated with the integers in $[S_{n-1}+1, S_n]$ (Table 2).

This bijection based on the Furnas ranking is convenient as a scheme for indexing trees, but the unavailability of a closed form for U_n and hence for S_n makes it difficult to quickly discern the tree associated with a given integer and vice versa. An alternative ranking scheme—which also bijectively associates the unlabeled binary rooted trees and the positive integers [4]—addresses this problem.

In this alternative scheme, due to Colijn & Plazzotta [4], the 1-leaf tree is given rank 1. For $n \geqslant 2$ leaves, the ordered pair (k_1, k_2), $k_1 \geqslant k_2 \geqslant 1$, is associated with the tree whose left subtree has rank k_1 and whose right subtree has rank k_2. Following the dictionary order on ordered pairs, the tree associated with ordered pair (k_1, k_2) is assigned rank $k_1(k_1-1)/2+1+k_2$. Thus, the Colijn–Plazzotta rank of a tree is obtained recursively from the ranks of its left and right subtrees, and the tree associated with a rank v is obtained by identifying the largest k_1 such that $k_1(k_1-1)/2+1 < v$ and assigning to rank v the tree whose left subtree has rank k_1 and whose right subtree has rank $v-k_1(k_1-1)/2-1$ (Table 3). Note that the left–right orientation of an unlabeled binary rooted tree generally differs for the Furnas and Colijn–Plazzotta rankings.

Colijn & Plazzotta [4] used their ranking to characterize trees from simulations and biological data sets and to define metrics on tree space; the Colijn–Plazzotta ranking is useful when the unlabeled shape of a tree captures features of biological interest. Thus, further investigation of the scheme can give insight into its use in such applications. Here, we study mathematical properties of the Colijn–Plazzotta ranking for unlabeled binary rooted trees. We first describe the

Table 2
Minimal and maximal Furnas and Colijn–Plazzotta ranks among unlabeled binary ranked trees with $1 \leqslant n \leqslant 16$ leaves.

n	U_n	Furnas		Colijn–Plazzotta	
		$S_{n-1}+1$	S_n	a_n	b_n
1	1	1	1	1	1
2	1	2	2	2	2
3	1	3	3	3	3
4	2	4	5	4	5
5	3	6	8	6	12
6	6	9	14	7	68
7	11	15	25	10	2280
8	23	26	48	11	2598062
9	46	49	94	20	3374961778893
10	98	95	192	22	5.70×10^{24}
11	207	193	399	28	1.62×10^{49}
12	451	400	850	29	1.32×10^{98}
13	983	851	1833	53	8.65×10^{195}
14	2179	1834	4012	56	3.74×10^{391}
15	4850	4013	8862	66	6.99×10^{782}
16	10905	8863	19767	67	2.44×10^{1565}

The Wedderburn–Etherington number U_n follows Eq. (1). The minimal rank $S_{n-1}+1$ and maximal rank S_n according to the Furnas ranking are taken from the sums S_n of the Wedderburn–Etherington numbers. The minimal rank a_n and maximal rank b_n according to the Colijn–Plazzotta ranking are taken from Theorems 6 and 9, respectively. For $n \geqslant 10$, b_n is approximated.

ranking and its bijective association with the positive integers (Section 2). For fixed n, we next obtain recursions for the smallest rank a_n assigned to some tree with n leaves (Section 3) as well as the largest rank b_n (Section 4). We then study asymptotic properties of a_n and b_n (Section 5). We conclude with a discussion (Section 6).

2. The Colijn–Plazzotta ranking

We define the Colijn–Plazzotta ranking more formally. Let T_n be the set of unlabeled binary rooted trees with n leaves, and let $T = \cup_{n=1}^{\infty} T_n$ be the set of all unlabeled binary rooted trees. All trees considered here are unlabeled binary rooted trees, and we refer to them simply as *trees*. For a tree $t \in T$, we let $m(t)$ denote its number of leaves. For $m(t) \geqslant 2$, we let $\ell(t)$ and $r(t)$ denote the left and right subtrees of t.

Definition 1. The *Colijn–Plazzotta ranking* for trees $t \in T$ is a function $f : T \to \mathbb{Z}^+$ that satisfies
(a) $f(t) = 1$ if $m(t) = 1$, and
(b) $f(t) = f\big(\ell(t)\big)[f\big(\ell(t)\big) - 1]/2 + 1 + f\big(r(t)\big)$ if $m(t) \geqslant 2$.

We abbreviate the Colijn–Plazzotta ranking as the *CP ranking*. To determine the CP rank of a tree t, we require t to be written in a canonical form in which $f\big(\ell(t)\big) \geqslant f\big(r(t)\big)$. For trees in this canonical form, the number of leaves in the left subtree, $m\big(\ell(t)\big)$, can be greater than, less than, or equal to $m\big(r(t)\big)$ (Table 3). The 1-leaf tree has CP rank 1; hence, if it is a subtree of the root of t and $m(t) \geqslant 3$, then because it has the smallest CP rank among all trees, it is necessarily the right subtree (for $m(t) = 2$, both subtrees have 1 leaf). The 2-leaf tree has CP rank 2, and similarly, if it is a subtree of the root of t and $m(t) \geqslant 5$, then it is the right subtree.

Note that it is possible to eliminate the explicit left–right orientation by referring to the maximal subtrees of the root of t as "first" and "second" subtrees, with the first subtree having CP rank greater than or equal to that of the second. At each internal node of t, subtrees of the node could then be depicted in either order, with the understanding that if one subtree has greater CP rank than the other, then it is "first." For convenience in assigning a canonical form, we continue to use the left–right orientation.

Thus, the dictionary order used in the CP ranking has the implication that for two trees t_1, t_2 in canonical form with $f\big(\ell(t_1)\big) < f\big(\ell(t_2)\big)$, $f(t_1) < f(t_2)$. For two trees t_1, t_2 in canonical form with $f\big(\ell(t_1)\big) = f\big(\ell(t_2)\big)$ and $f\big(r(t_1)\big) < f\big(r(t_2)\big)$, $f(t_1) < f(t_2)$. The CP ranking f gives a bijective map between trees and positive integers [4].

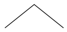
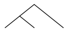
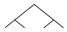
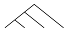
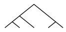
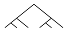
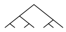
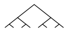
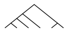
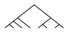
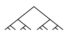
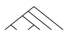
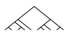
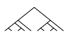
Proposition 2. *The function $f : T \to \mathbb{Z}^+$ is a bijection.*

Proof. For the tree t with $m(t) = 1$, $f(t) = 1$. Consider t with $m(t) \geqslant 2$. By definition of f, $f(t) \geqslant 2$.

For injectivity, two distinct trees t_1, t_2 with $m(t_1), m(t_2) \geqslant 2$ differ in their pair of subtrees, $\big(\ell(t_1), r(t_1)\big) \neq \big(\ell(t_2), r(t_2)\big)$, giving rise to distinct values of f, $f(t_1) \neq f(t_2)$.

For surjectivity, each positive integer $v \geqslant 2$ has a unique representation in the form $k_1(k_1 - 1)/2 + 1 + k_2$, with k_1, k_2 positive integers and $k_1 \geqslant k_2$. Hence, given $v \geqslant 2$, the tree whose subtrees have CP ranks k_1, k_2 in this representation is assigned to CP rank v. ■

Table 3
Colijn–Plazzotta ranks of unlabeled binary ranked trees with CP rank $1 \leqslant v \leqslant 37$.

CP rank v	$\big(f(\ell(t)), f(r(t))\big)$	t	$m(t)$	a_n	b_n
1	–		1	$a_1 = 1$	$b_1 = 1$
2	(1,1)		2	$a_2 = 2$	$b_2 = 2$
3	(2,1)		3	$a_3 = 3$	$b_3 = 3$
4	(2,2)		4	$a_4 = 4$	
5	(3,1)		4		$b_4 = 5$
6	(3,2)		5	$a_5 = 6$	
7	(3,3)		6	$a_6 = 7$	
8	(4,1)		5		
9	(4,2)		6		
10	(4,3)		7	$a_7 = 10$	
11	(4,4)		8	$a_8 = 11$	
12	(5,1)		5		$b_5 = 12$
13	(5,2)		6		
14	(5,3)		7		
15	(5,4)		8		
16	(5,5)		8		
17	(6,1)		6		
18	(6,2)		7		
19	(6,3)		8		
20	(6,4)		9	$a_9 = 20$	
21	(6,5)		9		
22	(6,6)		10	$a_{10} = 22$	
23	(7,1)		7		
24	(7,2)		8		
25	(7,3)		9		
26	(7,4)		10		
27	(7,5)		10		
28	(7,6)		11	$a_{11} = 28$	
29	(7,7)		12	$a_{12} = 29$	
30	(8,1)		6		
31	(8,2)		7		
32	(8,3)		8		
33	(8,4)		9		
34	(8,5)		9		
35	(8,6)		10		
36	(8,7)		11		
37	(8,8)		10		

The tree $t = f^{-1}(v)$ and its left and right subtrees $\ell(t)$, $r(t)$ follow Corollary 3, and the number of leaves $m(t)$ follows Corollary 4. Sequence $\{m\big(f^{-1}(v)\big)\}_{v=1}^{\infty}$ follows A064064 in OEIS. The values of a_n and b_n follow Theorems 6 and 9, respectively. The first v for which the number of leaves declines in proceeding from rank v to rank $v+1$ occurs at $v = 7$, so that $m\big(f^{-1}(8)\big) < m\big(f^{-1}(7)\big)$. Thus, $v = 8 \times 7/2 + 1 + 7 = 36$ is the smallest rank for which $f^{-1}(v)$ has fewer leaves in the left subtree than in the right subtree. The next ranks for which the left subtree has fewer leaves than the right subtree are 74, 76, 77, 78.

Given a positive integer $v \geqslant 2$, the proposition gives a characterization of the tree with CP rank v.

Corollary 3. *The function $f^{-1} : \mathbb{Z}^+ \to T$ that gives the tree with specified CP rank satisfies*

(a) $f^{-1}(1)$ is the tree with one leaf, and

(b) for $v \geqslant 2$, $f^{-1}(v)$ is the tree $t \in T$ whose left subtree has CP rank $k_1(v) = \lceil \frac{1+\sqrt{8v-7}}{2} \rceil - 1$ and whose right subtree has CP rank $k_2(v) = v - k_1(v)[k_1(v) - 1]/2 - 1$.

Proof. By definition of the CP ranking, $f(t) = 1$ if $m(t) = 1$. As f is bijective by Proposition 2, $f^{-1}(1)$ is the tree with one leaf.

For $v \geqslant 2$, following the proof of Proposition 2, $f^{-1}(v)$ is the tree $t \in T$ whose left subtree is the tree with CP rank $k_1(v)$, where $k_1(v)$ is the largest integer satisfying $k_1(k_1-1)/2+1 < v$, and whose right subtree is the tree with CP rank $k_2(v) = v - k_1(v)[k_1(v)-1]/2 - 1$. We solve the inequality for k_1. ■

Using the function f^{-1} that gives the tree associated with CP rank v, we obtain a recursion for the number of leaves possessed by the tree of CP rank v.

Corollary 4. *The function $m : \mathbb{Z}^+ \to \mathbb{Z}^+$ that gives the number of leaves in the tree with specified CP rank satisfies*
(a) $m\big(f^{-1}(1)\big) = 1$, and
(b) for $v \geqslant 2$, $m\big(f^{-1}(v)\big) = m\big(f^{-1}(\lceil \frac{\sqrt{8v-7}-1}{2} \rceil)\big) + m\big(f^{-1}(v - \lceil \frac{\sqrt{8v-7}-1}{2} \rceil \lceil \frac{\sqrt{8v-7}-3}{2} \rceil /2 - 1)\big)$.

Proof. The number of leaves in the tree of CP rank $v \geqslant 2$, or $m\big(f^{-1}(v)\big)$, is the sum of the numbers of leaves in its left and right subtrees, or $m\big(k_1(v)\big) + m\big(k_2(v)\big)$. ■

The CP ranking, unlike the Furnas ranking, assigns trees whose numbers of leaves differ substantially to neighboring ranks (Table 3). Also unlike the Furnas ranking, it enables a straightforward calculation of the rank associated with a given tree and the tree associated with a given rank.

3. Smallest CP rank for a fixed number of leaves

Next, we compute the CP ranks of the trees of size n that have the smallest and largest CP ranks. For $n \geqslant 1$, we define $a_n = \min_{t \in T_n} f(t)$ and $b_n = \max_{t \in T_n} f(t)$. The sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ give the minimal and maximal CP rank considering all trees of size n leaves. Let z_n and Z_n respectively denote the trees of size n that achieve the minimal and maximal CP rank, $f(z_n) = a_n$ and $f(Z_n) = b_n$.

We begin with a_n. To determine a recursion for a_n, we first need a lemma that establishes that a_n increases with n.

Lemma 5. *$\{a_n\}_{n=1}^{\infty}$ is a strictly increasing sequence.*

Proof. First, by the definition of the CP ranking and the fact that tree sizes $n = 1$, 2, and 3 each have only one tree, $a_1 = 1$, $a_2 = 2$, and $a_3 = 3$. We show by induction that for each $n \geqslant 3$, $a_{n+1} \geqslant a_n$.

Consider a tree t of size $n+1$. We must show $f(t) > a_n$, as it would then follow that $a_{n+1} = \min_{t \in T_{n+1}} f(t) > a_n$. We consider two cases. (i) Suppose the two subtrees of the root of t have sizes n and 1. Then, requiring t to be in canonical form, the left subtree of t has size $\ell(t) = n$ and the right subtree has size 1, and

$$\begin{aligned} f(t) &= f\big(\ell(t)\big)[f\big(\ell(t)\big) - 1]/2 + 2 \\ &\geqslant \frac{a_n(a_n-1)}{2} + 2 \\ &> a_n. \end{aligned}$$

Here, the first inequality uses $f\big(\ell(t)\big) \geqslant a_n$ by the definition of a_n, and the second follows from the quadratic inequality $x(x-1)/2 + 2 > x$. Thus, each tree t of size $n+1$ with subtrees of size n and 1 has $f(t) > a_n$.

(ii) Suppose t instead has subtrees of size m, $\lceil \frac{n+1}{2} \rceil \leqslant m \leqslant n-1$, and $n+1-m \leqslant m$. The subtrees of t are $\ell(t)$ and $r(t)$, one of which has size m and the other of which has size $n+1-m$ (possibly $m = n+1-m$ for odd n). As it is not yet specified which subtree is $\ell(t)$ and which is $r(t)$, we consider both left–right arrangements, in each exhibiting a tree t' of size n with $a_n \leqslant f(t') < f(t)$.

Suppose that $\ell(t)$ has size m. Then $f\big(\ell(t)\big) \geqslant f(z_{n+1-m})$ by the inductive assumption: if $\ell(t)$ has size m, then $f\big(\ell(t)\big) \geqslant a_m \geqslant a_{n+1-m}$. Consider a tree t' of size n whose two subtrees are $\ell(t)$ and z_{n-m}. Note that $f\big(\ell(t)\big) \geqslant a_m > a_{n-m} = f\big(z_{n-m}\big)$ by the inductive assumption, so that the canonical form for t' has $\ell(t') = \ell(t)$ and $r(t') = z_{n-m}$. We then have

$$\begin{aligned} f(t) &= f\big(\ell(t)\big)[f\big(\ell(t)\big) - 1]/2 + 1 + f\big(r(t)\big) \\ &\geqslant f\big(\ell(t)\big)[f\big(\ell(t)\big) - 1]/2 + 1 + a_{n+1-m} \\ &> f\big(\ell(t)\big)[f\big(\ell(t)\big) - 1]/2 + 1 + a_{n-m} \\ &= f\big(\ell(t')\big)[f\big(\ell(t')\big) - 1]/2 + 1 + f\big(r(t')\big) \\ &= f(t'). \end{aligned}$$

The first inequality follows from the definition of a_n, and the second follows from the inductive assumption. Thus, $f(t) > f(t') \geqslant a_n$.

Now suppose instead that $\ell(t)$ has size $n+1-m$. Let the two subtrees of t' be $r(t)$ and z_{n-m}. Then $f\big(r(t)\big) \geqslant a_m > a_{n-m} = f\big(z_{n-m}\big)$, so that in canonical form, t' has $\ell(t') = r(t)$ and $r(t') = z_{n-m}$. We have

$$f(t) = f\big(\ell(t)\big)[f\big(\ell(t)\big)-1]/2+1+f\big(r(t)\big)$$
$$f(t') = f\big(r(t)\big)[f\big(r(t)\big)-1]/2+1+a_{n-m}.$$

Rearranging terms, it follows that $f(t) > f(t')$ is equivalent to $[f\big(\ell(t)\big)-f\big(r(t)\big)][f\big(\ell(t)\big)+f\big(r(t)\big)-3] > 2[a_{n-m}-f\big(\ell(t)\big)]$. This latter inequality holds, as $f\big(\ell(t)\big)-f\big(r(t)\big) \geqslant 0$ for any t, $f\big(\ell(t)\big)+f\big(r(t)\big)-3 \geqslant 0$ for any t with $m(t) \geqslant 3$, and $f\big(\ell(t)\big) \geqslant a_{n+1-m} > a_{n-m}$ by the inductive hypothesis. Thus, $f(t) > f(t') \geqslant a_n$.

We conclude that for each tree t of size $n+1$ with subtrees of size m and $n+1-m$, $\lceil \frac{n+1}{2} \rceil \leqslant m \leqslant n-1$, we can find a tree t' of size n for which $f(t) > f(t')$. As $f(t') \geqslant a_n$, it follows that $f(t) > a_n$. ■

The computation of a_n encodes a result that the tree with minimal CP rank is obtained by appending two subtrees of minimal CP rank for their size to a shared root. These subtrees are identical for even n, and they differ in size by one leaf for odd n.

Theorem 6. *The sequence $\{a_n\}_{n=1}^{\infty}$ of values of the minimal CP rank across trees of fixed size n satisfies*
(a) $a_1 = 1$,
(b) $a_{2n} = a_n(a_n-1)/2+1+a_n$ for $2n \geqslant 2$, and
(c) $a_{2n-1} = a_n(a_n-1)/2+1+a_{n-1}$ for $2n-1 \geqslant 3$.

Proof. The base case of $a_1 = 1$ is trivial, as are the cases of $a_2 = 2$ and $a_3 = 3$. Consider a tree t with an even number of leaves $2n \geqslant 4$.

We claim that if $m\big(\ell(t)\big) < n$, then $t \neq z_{2n}$. Suppose the left subtree of t has $n^* < n$ leaves. The right subtree then has at least $n+1$ leaves, so that $f\big(\ell(t)\big) > f\big(r(t)\big) \geqslant a_{n+1}$. Then $\ell(t)$ cannot equal z_{n^*}, as $f(z_{n^*}) = a_{n^*} < a_{n+1}$ by Lemma 5. We could then construct a tree of $2n$ leaves whose left subtree is $r(t)$ and whose right subtree is z_{n^*}. This tree would have a lower CP rank than t, as the inequality

$$\frac{f\big(\ell(t)\big)[f\big(\ell(t)\big)-1]}{2}+1+f\big(r(t)\big) > \frac{f\big(r(t)\big)[f\big(r(t)\big)-1]}{2}+1+a_{n^*}$$

is algebraically equivalent to $[f\big(\ell(t)\big)-f\big(r(t)\big)][f\big(\ell(t)\big)+f\big(r(t)\big)-3] > 2[a_{n^*}-f\big(\ell(t)\big)]$; this latter inequality holds as its left side is nonnegative and its right size is negative. Thus, $m\big(\ell(z_{2n})\big) \geqslant n$.

Having established that the canonical form of z_{2n} has $m\big(\ell(z_{2n})\big) \geqslant n$, we have $\ell(z_{2n}) \in T_n \cup T_{n+1} \cup \cdots \cup T_{2n-1}$. We now argue that z_{2n} is the tree t^* whose left subtree is z_n and whose right subtree is also z_n.

For t, t' in canonical form, $f\big(\ell(t)\big) < f\big(\ell(t')\big)$ implies $f(t) < f(t')$ (Section 2); for t, t' in canonical form with $f\big(\ell(t)\big) = f\big(\ell(t')\big)$ and $f\big(r(t)\big) < f\big(r(t')\big)$, $f(t) < f(t')$. By Lemma 5, $a_n \leqslant a_{n+1} \leqslant \cdots \leqslant a_{2n-1}$, so that $z_n = \arg\min_{t \in T_n \cup T_{n+1} \cup \cdots \cup T_{2n-1}} f(t)$. Combining these results, each tree $t \neq t^*$ with $t \in T_{2n}$ and $\ell(t) \in T_n \cup T_{n+1} \cup \cdots \cup T_{2n-1}$, written in canonical form, has $f(t) > f(t^*)$: if $\ell(t) \neq z_n$, then $f(t) > f(t^*)$; if $\ell(t) = z_n$ and $r(t) \neq z_n$, then $f(t) > f(t^*)$. We conclude $\ell(z_{2n}) = r(z_{2n}) = z_n$ and $a_{2n} = a_n(a_n-1)/2+1+a_n$.

For trees of size $2n-1 \geqslant 5$, the same argument applies: we show $m\big(\ell(t)\big) \geqslant n$, then we argue that z_{2n-1} is the tree with left subtree z_n and right subtree z_{n-1}, producing $a_{2n-1} = a_n(a_n-1)/2+1+a_{n-1}$. ■

The first terms of $\{a_n\}_{n=1}^{\infty}$ are 1, 2, 3, 4, 6, 7, 10, 11, 20, 22, 28, 29, 53, 56, 66, 67 (Table 2). The recursion for a_n constructs the trees z_n. For odd n, the two subtrees immediately descended from the root of the tree $z_n = f^{-1}(a_n)$ have numbers of leaves that differ by 1 (Table 3). For even n, $z_n = f^{-1}(a_n)$ has two identical subtrees descended from the root.

An internal node is termed *balanced* if the numbers of leaves in its two immediate subtrees differ by 0 or 1 [6,13]; hence, the root of z_n is a balanced node. The unique tree each of whose internal nodes is balanced is termed *maximally balanced* [6,13]; recursively following the proof of Theorem 6, in both the odd and even cases, for each internal node, the two subtrees immediately descended from the node differ by at most 1 in their numbers of leaves. We thus have the following corollary.

Corollary 7. *Among trees of size n, the tree z_n that achieves the minimal CP rank is the maximally balanced tree.*

Proof. The construction of z_n in the proof of Theorem 6 demonstrates that for $n \geqslant 2$, z_n is balanced at each internal node. ■

Note that in the case that n is a power of 2, $n = 2^k$ for $k \geqslant 1$, the tree that has minimal CP rank is the fully symmetric tree.

4. Largest CP rank for a fixed number of leaves

We now turn to $\{b_n\}_{n=1}^{\infty}$, the sequence of values of the maximal CP rank among trees with n leaves. In a similar manner to Section 3, before giving the recursion for b_n, we introduce a lemma that demonstrates that b_n increases with n.

Lemma 8. *$\{b_n\}_{n=1}^{\infty}$ is a strictly increasing sequence.*

Proof. We show $b_{n+1} > b_n$ for $n \geqslant 1$.

For $n \geqslant 1$, we append Z_n and Z_1 to a shared root to obtain a tree t. Then $f(t) = b_n(b_n - 1)/2 + 2$. The inequality $b_n(b_n - 1)/2 + 2 > b_n$ always holds, as $b_n^2 - 3b_n + 4$ is an upward-facing parabola with vertex at a positive value, $(\frac{3}{2}, \frac{7}{4})$. Thus, we have constructed a tree of $n + 1$ leaves with CP rank greater than that of the tree of n leaves with largest CP rank. ■

Next, to obtain b_n, we show that the tree of size n with maximal CP rank is obtained by appending the tree of maximal CP rank with size $n - 1$ and a single leaf to a shared root.

Theorem 9. *The sequence $\{b_n\}_{n=1}^{\infty}$ of values of the maximal CP rank across trees of fixed size n satisfies*
(a) $b_1 = 1$,
(b) $b_n = b_{n-1}(b_{n-1} - 1)/2 + 2$ for $n \geqslant 2$.

Proof. The base case $b_1 = 1$ is trivial, as are the cases of $b_2 = 2$ and $b_3 = 3$.

Let $n \geqslant 4$ and consider a tree t with $m(t) = n$. We claim that if $m\big(\ell(t)\big) < \lceil \frac{n}{2} \rceil$, then $t \neq Z_n$. Suppose the left subtree of t has $n^* < \lceil \frac{n}{2} \rceil$ leaves. The right subtree then has at least $\lceil \frac{n}{2} \rceil$ leaves. Then $\ell(t)$ cannot be Z_{n^*}, as $f(Z_{n^*}) = b_{n^*} < b_{\lceil \frac{n}{2} \rceil}$ by Lemma 8. We could then construct a tree t' of size n whose left subtree is Z_{n-n^*} and whose right subtree is Z_{n^*}. This tree would have a greater CP rank than t, as

$$\frac{f\big(\ell(t)\big)[f\big(\ell(t)\big) - 1]}{2} + 1 + f\big(r(t)\big) < \frac{b_{n^*}(b_{n^*} - 1)}{2} + 1 + b_{n-n^*} < \frac{b_{n-n^*}(b_{n-n^*} - 1)}{2} + 1 + b_{n^*} = f(t'),$$

where we use $b_{n^*} < b_{n-n^*}$ by Lemma 8. Thus, $m\big(\ell(Z_n)\big) \geqslant \lceil \frac{n}{2} \rceil$.

Having established that the canonical form of Z_n has $m\big(\ell(Z_n)\big) \geqslant \lceil \frac{n}{2} \rceil$, we have $\ell(Z_n) \in T_{\lceil \frac{n}{2} \rceil} \cup T_{\lceil \frac{n}{2} \rceil + 1} \cup \cdots \cup T_{n-1}$. We now argue that Z_n is the tree t^* whose left subtree is Z_{n-1} and whose right subtree is Z_1.

For t, t' in canonical form, $f\big(\ell(t)\big) < f\big(\ell(t')\big)$ implies $f(t) < f(t')$ (Section 2). By Lemma 8, $b_{\lceil \frac{n}{2} \rceil} \leqslant b_{\lceil \frac{n}{2} \rceil + 1} \leqslant \cdots \leqslant b_{n-1}$, so that $Z_{n-1} = \arg\max_{t \in T_{\lceil \frac{n}{2} \rceil} \cup T_{\lceil \frac{n}{2} \rceil + 1} \cup \cdots \cup T_{n-1}} f(t)$. Combining these results, each tree $t \neq t^*$ with $t \in T_n$ and $\ell(t) \in T_{\lceil \frac{n}{2} \rceil} \cup T_{\lceil \frac{n}{2} \rceil + 1} \cup \cdots \cup T_{n-1}$, written in canonical form, has $f(t) < f(t^*)$. We conclude $\ell(Z_n) = Z_{n-1}$ and $r(Z_n)$ necessarily is Z_1. Hence $b_n = b_{n-1}(b_{n-1} - 1)/2 + 2$. ■

The first values of $\{b_n\}_{n=1}^{\infty}$ are 1, 2, 3, 5, 12, 68, 2280, 2598062 (Table 2, A108225 in OEIS). Because the tree Z_n with maximal CP rank is obtained by successively appending the tree Z_{n-1} with maximal CP rank and a single leaf to a shared root, the tree of n leaves that achieves maximal CP rank is the *caterpillar* tree—the tree in which there exists an internal node that descends from all other internal nodes (Table 3).

Corollary 10. *Among trees of size n, the tree Z_n that achieves the maximal CP rank is the caterpillar tree.*

Proof. In the proof of Theorem 9, it is observed that for $n \geqslant 2$, Z_n is obtained by appending Z_{n-1} and Z_1 to a shared root. As caterpillar trees C_n are constructed by appending the previous caterpillar tree C_{n-1} and the single-leaf tree C_1 to a shared root, $Z_n = C_n$. ■

A relationship exists between entries of $\{a_n\}_{n=1}^{\infty}$ and entries of $\{b_n\}_{n=1}^{\infty}$. We write $d_n = a_{2^n}$ for $n \geqslant 0$.

Proposition 11. *For $n \geqslant 0$, $d_n + 1 = b_{n+2}$.*

Proof. We demonstrate the result by induction. We have $d_0 + 1 = a_1 + 1 = 2$ and $b_2 = 2$. For the inductive step, we assume $d_n + 1 = b_{n+2}$ and show $d_{n+1} + 1 = b_{n+3}$.

By Theorem 6, for $n \geqslant 1$, $d_{n+1} = a_{2^n}(a_{2^n} - 1)/2 + 1 + a_{2^n} = d_n(d_n - 1)/2 + 1 + d_n = d_n(d_n + 1)/2 + 1$. At the same time, $b_{n+3} = b_{n+2}(b_{n+2} - 1)/2 + 2$ by Theorem 9. By the inductive hypothesis, we then have $b_{n+3} = (d_n + 1)d_n/2 + 2 = d_{n+1} + 1$. ■

The sequence $\{d_n\}_{n=0}^{\infty} = \{a_{2^n}\}_{n=0}^{\infty}$ begins 1, 2, 4, 11, 67, 2279, 2598061 (A006894 in OEIS). As a result of Theorem 9 and Proposition 11, as we traverse ranks in the interval $[b_n, b_{n+1})$, flanked by the largest ranks for trees with n and $n+1$ leaves, we encounter ranks for trees representing numbers of leaves as high as 2^{n-1}. The CP ranking can place trees with quite different numbers of leaves in adjacent ranks. We characterize this difference in the following proposition.

Proposition 12. *For $n \geqslant 1$, all trees with CP rank in $[b_n, b_{n+1})$ have sizes in $[n, 2^{n-1}]$. The smallest size for a tree with CP rank in $[b_n, b_{n+1})$ is n, and the largest size for a tree with CP rank in $[b_n, b_{n+1})$ is 2^{n-1}.*

Proof. The interval $[b_n, b_{n+1})$, ranging from the largest CP rank of a tree with n leaves to one less than the largest CP rank of a tree with $n + 1$ leaves, contains the smallest CP rank of a tree with 2^{n-1} leaves ($a_{2^{n-1}} = d_{n-1} = b_{n+1} - 1$ by

Proposition 11). Because $\{b_n\}_{n=1}^{\infty}$ is increasing by Lemma 8, $b_{n-1} < b_n$, so that no trees of size $n-1$ leaves or fewer have CP rank in $[b_n, b_{n+1})$. Because $\{a_n\}_{n=1}^{\infty}$ is increasing by Lemma 5, $a_{2^{n-1}} < a_{2^{n-1}+1}$, and no trees of size $2^{n-1}+1$ or greater have CP rank in $[b_n, b_{n+1})$. ■

5. Asymptotics

We now evaluate the asymptotic behavior of the sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$. We use the method of Aho and Sloane [1]. This method converts a quadratic recursion into a linear recursion for a logarithm. The linear recursion is then solved with the help of a constant that is evaluated numerically.

Theorem 13. *$d_n \sim 2\alpha^{(2^n)}$ for a constant $\alpha \approx 1.24602$.*

Proof. In Proposition 11, $d_n = \frac{1}{2}d_{n-1}^2 + \frac{1}{2}d_{n-1} + 1$ for $n \geqslant 1$, with $d_0 = 1$. Substituting $d_n = 2x_n - \frac{1}{2}$, we obtain $x_n = x_{n-1}^2 + \frac{11}{16}$, with $x_0 = \frac{3}{4}$.

We take $y_n = \log x_n$ in this quadratic recursion for x_n. We then have, for $n \geqslant 1$, $y_n = 2y_{n-1} + \alpha_{n-1}$, where $\alpha_{n-1} = \log[1 + 11/(16x_{n-1}^2)]$. Applying the method of Aho and Sloane [1],

$$\begin{aligned} y_n &= 2^n y_0 + \sum_{i=0}^{n-1} 2^{n-i-1}\alpha_i \\ &= 2^n\left(y_0 + \sum_{i=0}^{\infty} 2^{-i-1}\alpha_i\right) - \sum_{i=n}^{\infty} 2^{n-i-1}\alpha_i. \end{aligned}$$

Exponentiating both sides, we obtain

$$\begin{aligned} x_n &= \left[x_0 \exp\left(\sum_{i=0}^{\infty} 2^{-i-1}\alpha_i\right)\right]^{(2^n)} \exp\left(-\sum_{i=n}^{\infty} 2^{n-i-1}\alpha_i\right) \\ &= \alpha^{(2^n)} \exp\left(-\sum_{i=n}^{\infty} 2^{n-i-1}\alpha_i\right), \end{aligned}$$

where α is the constant $\alpha = x_0 \exp(\sum_{i=0}^{\infty} 2^{-i-1}\alpha_i)$. Inserting the first terms of the recursive sequence $\{x_n\}_{n=0}^{\infty}$, we have $(x_0, x_1, x_2, x_3, \ldots) = (\frac{3}{4}, \frac{5}{4}, \frac{9}{4}, \frac{23}{4}, \ldots)$. From these values, we have $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \ldots) = (\log \frac{20}{9}, \log \frac{36}{25}, \log \frac{92}{81}, \log \frac{540}{529}, \ldots)$. Numerically evaluating the constant α by iterating the recursion to obtain the first 10 terms, we obtain $\alpha \approx \frac{3}{4}\exp(\frac{1}{2}\log\frac{20}{9} + \frac{1}{4}\log\frac{36}{25} + \frac{1}{8}\log\frac{92}{81} + \frac{1}{16}\log\frac{540}{529} + \cdots)$, or $\alpha \approx 1.24602083298366$.

Then

$$\frac{x_n}{\alpha^{(2^n)}} = \exp\left(-\sum_{i=n}^{\infty} 2^{n-i-1}\alpha_i\right).$$

As $n \to \infty$, the sum $\sum_{i=n}^{\infty} 2^{n-i-1}\alpha_i$ can be bounded $0 \leqslant \sum_{i=n}^{\infty} 2^{n-i-1}\alpha_i \leqslant \alpha_n \sum_{i=n}^{\infty} 2^{n-i-1} = \alpha_n$. Because $x_n \to \infty$ as $n \to \infty$, $\alpha_n \to 0$ as $n \to \infty$. Hence $\lim_{n\to\infty}[x_n/\alpha^{(2^n)}] = 1$.

Because $d_n = 2x_n - \frac{1}{2}$, we conclude $d_n \sim 2\alpha^{(2^n)}$. ■

The connection between $d_n = a_{2^n}$ and b_n (Proposition 11) quickly gives the following result.

Corollary 14. *$b_n \sim 2\beta^{(2^n)}$ for a constant $\beta \approx 1.05653$.*

Proof. By Proposition 11, $b_n \sim d_{n-2}$, and by Theorem 13, $d_{n-2} \sim 2\alpha^{(2^{n-2})}$. Hence, $b_n \sim 2\alpha^{(2^{n-2})}$. Writing $\beta = \alpha^{1/4} \approx 1.05652876566960$, the result follows. ■

We have obtained an asymptotic equivalence for $\{d_n\}_{n=0}^{\infty}$ in Theorem 13, giving the increase of $\{a_n\}_{n=1}^{\infty}$ for the subsequence $n = 1, 2, 4, 8, 16, \ldots$. We now place a bound on the increase in $\{a_n\}_{n=1}^{\infty}$ more generally.

Proposition 15. *$a_n < (\frac{3}{2})^n$ for $n \geqslant 1$.*

Proof. We use induction. The result holds for $n = 1$ ($a_1 = 1 < \frac{3}{2}$), $n = 2$ ($a_n = 2 < \frac{9}{4}$), $n = 3$ ($a_3 = 3 < \frac{27}{8}$), and $n = 4$ ($a_4 = 4 < \frac{81}{16}$). We assume that the inequality holds for each n from 1 to $2k-2$. The inductive step is separated into even and odd cases.

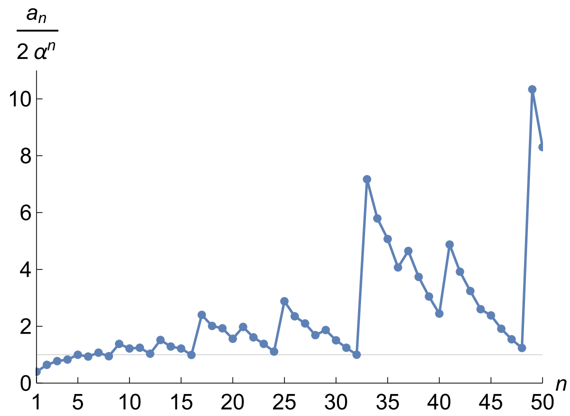


Fig. 1. The ratio $a_n/(2\alpha^n)$, $\alpha \approx 1.24602$. This ratio has limit 1 for subsequence $\{a_{2^k}\}_{k=0}^{\infty}$ (Theorem 13).

For even $2k \geqslant 4$, applying Theorem 6 and the inductive hypothesis,

$$\begin{aligned} a_{2k} &= \frac{a_k(a_k-1)}{2} + 1 + a_k \\ &< \frac{(\frac{3}{2})^k[(\frac{3}{2})^k - 1]}{2} + 1 + \left(\frac{3}{2}\right)^k \\ &= \frac{1}{2}\left(\frac{9}{4}\right)^k + \frac{1}{2}\left(\frac{3}{2}\right)^k + 1. \end{aligned}$$

To demonstrate $a_{2k} < (\frac{3}{2})^{2k}$, we must show $\frac{1}{2}(\frac{9}{4})^k + \frac{1}{2}(\frac{3}{2})^k + 1 < (\frac{3}{2})^{2k}$, or equivalently, $(\frac{3}{2})^k + 2 < (\frac{9}{4})^k$. This latter inequality holds: $g(k) = (\frac{9}{4})^k - (\frac{3}{2})^k - 2$ is an increasing function for $k > 0$, with $g(2) = \frac{13}{16} > 0$, and $g(k)$ therefore remains positive for $k \geqslant 2$.

For odd $2k-1 \geqslant 5$, applying Theorem 6 and the inductive hypothesis,

$$\begin{aligned} a_{2k-1} &= \frac{a_k(a_k-1)}{2} + 1 + a_{k-1} \\ &< \frac{(\frac{3}{2})^k[(\frac{3}{2})^k - 1]}{2} + 1 + \left(\frac{3}{2}\right)^{k-1} \\ &= \frac{1}{2}\left(\frac{9}{4}\right)^k - \frac{1}{2}\left(\frac{3}{2}\right)^k + 1 + \left(\frac{3}{2}\right)^{k-1}. \end{aligned}$$

To demonstrate $a_{2k-1} < (\frac{3}{2})^{2k-1}$, we must show $\frac{1}{2}(\frac{9}{4})^k - \frac{1}{2}(\frac{3}{2})^k + 1 + (\frac{3}{2})^{k-1} < (\frac{3}{2})^{2k-1}$, or equivalently, $(\frac{3}{2})^k + 6 < (\frac{9}{4})^k$. Again, the function $g(k) = (\frac{9}{4})^k - (\frac{3}{2})^k - 6$ is increasing for $k > 0$, with $g(3) = \frac{129}{64} > 0$. Hence, $g(k)$ remains positive for $k \geqslant 3$. ■

In Figs. 1 and 2, we examine the ratios $a_n/(2\alpha^n)$ and $a_n/(\frac{3}{2})^n$ for small values of n. In Fig. 1, the ratio $a_n/(2\alpha^n)$, which has limit 1 for the subsequence $n = 1, 2, 4, 8, 16, \ldots$ (Theorem 13), generally exceeds 1. It has a non-monotonic pattern, returning to near 1 when n is equal to a power of 2. In Fig. 2, the ratio $a_n/(\frac{3}{2})^n$ lies substantially below 1, indicating that $(\frac{3}{2})^n$ is a relatively loose upper bound for a_n.

6. Discussion

The Colijn–Plazzotta ranking provides a convenient method for obtaining the rank associated with a given tree and the tree associated with a given rank. We have obtained recursions for the minimal and maximal CP rank across trees with n leaves (Theorems 6 and 9), analyzing their asymptotic behavior (Section 5). This analysis demonstrates that as the CP rank increases, the numbers of leaves in the associated trees traverse a wide range of values. In fact, for $n \geqslant 1$, the interval bounded by the largest rank across trees with n leaves and the largest rank across trees with $n+1$ leaves contains ranks for trees with as many as 2^{n-1} leaves (Proposition 12). Unlike for the Furnas ranking, the CP ranking has the property that the trees associated with sequential ranks do not necessarily differ in size by either 1 or 0 leaves; the difference in size between trees with sequential ranks is $2^n - n - 2$ in the transition from rank a_{2^n} to rank $a_{2^n} + 1 = b_{n+2}$. Asymptotically, the largest rank across trees with n leaves increases with $2\beta^{(2^n)}$ for a constant $\beta \approx 1.05653$ (Corollary 14),

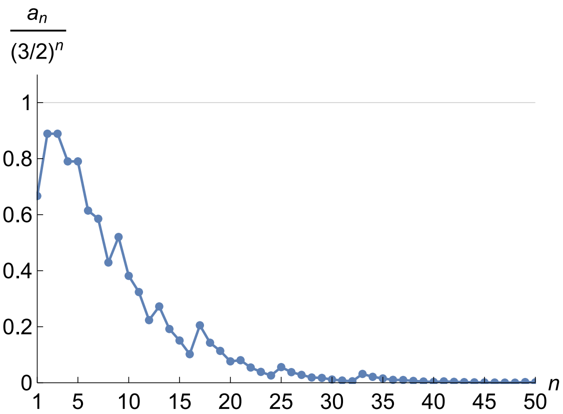


Fig. 2. The ratio $a_n/(\frac{3}{2})^n$. This ratio lies below 1 for all n (Proposition 15).

and the smallest rank across trees with n leaves is bounded above by the substantially smaller $(\frac{3}{2})^n$ (Proposition 15), with asymptotic equivalence to $2\alpha^n$, $\alpha \approx 1.24602$, for the subsequence $\{a_{2^n}\}_{n=0}^{\infty}$ (Theorem 13).

The computations of a_n and b_n construct the trees z_n and Z_n that respectively have the smallest and largest CP ranks among n-leaf trees. The smallest rank belongs to the maximally balanced tree (Corollary 7), and the largest rank belongs to the caterpillar (Corollary 10). Measures of tree balance—the extent to which an unlabeled shape resembles balanced shapes [2,10,12,15]—often have this property that their most extreme values among trees of size n are represented by a balanced tree and an unbalanced caterpillar tree. CP rank itself can potentially be used as a measure of increasing imbalance for a tree t, or it can be normalized to produce an imbalance index I_{CP} for which the maximally balanced tree with a specified number of leaves has value 0 and the caterpillar has value 1:

$$I_{CP}(t) = \frac{f(t) - a_{m(t)}}{b_{m(t)} - a_{m(t)}}. \tag{2}$$

Taking into account the growth of b_n with $2\beta^{(2^n)}$ (Corollary 14) and the slower growth of $a_n < (\frac{3}{2})^n$ (Proposition 15), another sensible imbalance index could instead rely on logarithms:

$$J_{CP}(t) = \frac{\log f(t) - \log a_{m(t)}}{\log b_{m(t)} - \log a_{m(t)}} \approx \frac{\log f(t) - \log a_{m(t)}}{2^{m(t)} \log \beta + \log 2 - \log a_{m(t)}}. \tag{3}$$

It will be useful to compare CP rank and the normalizations in Eqs. (2) and (3) to existing measures of tree balance. Because the tree of minimal CP rank is maximally balanced with absolute difference 0 or 1 between the sizes of the two subtrees for each internal node, it is of particular interest to compare to the Colless imbalance index [3,5,6,14]—which for each node sums the absolute difference in the numbers of descendants of the two subtrees of the node and which has larger values for unbalanced trees. Notably, whereas the Colless index admits tied values for distinct equal-sized trees [6], a feature of imbalance indices based on CP rank is that no two equal-sized trees are tied with equal values.

The study augments recent results examining unlabeled binary rooted trees that possess maximal or minimal features in scenarios arising from consideration of evolutionary problems [6–8,13]. Curiously, Theorem 13 has a close connection with an analysis of "non-equivalent ancestral configurations," structures that are used in characterizing relationships of pairs of trees [9,16]. For non-equivalent ancestral configurations associated with the completely balanced trees—the same trees that produce the smallest CP rank in the case that n is a power of 2—Section 4.2 of Disanto & Rosenberg [9] gives a recursion for a quantity γ_n, with $\gamma_0 = 0$, which can be transformed by $\gamma_n = 2x_n - \frac{3}{2}$ to produce the recursion $x_n = x_{n-1}^2 + \frac{11}{16}$ with $x_0 = \frac{3}{4}$ seen in the proof of Theorem 13. Thus, Disanto & Rosenberg [9] obtain the same asymptotic result $2\alpha^{(2^n)}$ we observed for d_n, but for the growth of a different quantity γ_n, the number of non-equivalent ancestral configurations with increasing numbers of leaves 2^n in completely balanced trees.

The CP ranking encodes an innovative scheme that facilitates computations with unlabeled binary rooted trees. It has potential for use in summarizing trees that arise in probabilistic models and data sets, as shown by Colijn & Plazzotta [4] in their construction of metrics for unlabeled binary rooted trees and their use of these metrics to study evolutionary trees of strains of infectious agents. Further analysis of the mathematical properties of the CP ranking can potentially inform its applications.

Acknowledgments

I thank Egor Alimpiev and three reviewers for comments. Support was provided by NIH, USA grant R01 GM131404.

References

[1] A.V. Aho, N.J.A. Sloane, Some doubly exponential sequences, Fibonacci Q. 11 (1973) 429–437.
[2] M.G.B. Blum, O. François, On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited, Math. Biosci. 195 (2005) 141–153.
[3] G. Cardona, A. Mir, F. Rosselló, Exact formulas for the variance of several balance indices under the Yule model, J. Math. Biol. 67 (2013) 1833–1846.
[4] C. Colijn, G. Plazzotta, A metric on phylogenetic tree shapes, Syst. Biol. 67 (2018) 113–126.
[5] D.H. Colless. Phylogenetics, The theory and practice of phylogenetic systematics, Syst. Zool. 31 (1982) 100–104.
[6] T.M. Coronado, M. Fischer, L. Herbst, F. Rosselló, K. Wicke, On the minimium value of the Colless index and the bifurcating trees that achieve it, J. Math. Biol. 80 (2020) 1993–2054.
[7] F. Disanto, N.A. Rosenberg, Enumeration of ancestral configurations for matching gene trees and species trees, J. Comput. Biol. 24 (2017) 831–850.
[8] F. Disanto, N.A. Rosenberg, Enumeration of compact coalescent histories for matching gene trees and species trees, J. Math. Biol. 78 (2019) 155–188.
[9] F. Disanto, N.A. Rosenberg, On the number of non-equivalent ancestral configurations for matching gene trees and species trees, Bull. Math. Biol. 81 (2019) 384–407.
[10] J. Felsenstein, Inferring Phylogenies, Sinauer, Sunderland, MA, 2004.
[11] G.W. Furnas, The generation of random, binary unordered trees, J. Classification 1 (1984) 187–233.
[12] S.B. Heard, Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees, Evolution 46 (1992) 1818–1826.
[13] A. Mir, F. Rosselló, L. Rotger, A new balance index for phylogenetic trees, Math. Biosci. 241 (2013) 125–136.
[14] A. Mir, L. Rotger, F. Rosselló, Sound Colless-like balance indices for multifurcating trees, PLoS One 13 (2018) e0203401.
[15] M. Steel, Phylogeny: Discrete and Random Processes in Evolution, Society for Industrial and Applied Mathematics, Philadelphia, 2016.
[16] Y. Wu, Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood, Evolution 66 (2012) 763–775.