On the centroid of recursive trees

J.W. Moon
Mathematical Sciences Department
University of Alberta
Edmonton, Alberta, Canada T6G 2G1

Abstract

It follows from our results that as $n \to \infty$ the average distance between the root and the (nearer) centroid node of a recursive tree T_n tends to 1; and the average value of the label of the (nearer) centroid node tends to 5/2.

1. Introduction

For any node v of a tree T the *branches* of T joined to v are the maximal subtrees of T not containing v. Let $\kappa(v)$ denote the number of nodes in the largest branch joined to v. A node v of a tree T with n nodes is a *centroid* node if $\kappa(v) \leq n/2$. Jordan [3] showed that either (i) T has a single centroid node v and $\kappa(v) < n/2$ or (ii) T has two (adjacent) centroid nodes v_1 and v_2, in which case n is even and $\kappa(v_1) = \kappa(v_2) = n/2$.

A tree T_n with n labelled nodes, rooted at node 1, is a *recursive* tree if $n = 1$ or if T_n can be constructed by joining node n to one of the $n-1$ nodes of some recursive tree T_{n-1}; this is equivalent to requiring that the labels of the nodes encountered along any path leading away from the root node 1 form an increasing sequence. It is easy to see that there are $(n-1)!$ recessive trees T_n. For additional material on recursive trees, see, e.g., [2, 4, 5, 6, 7, 9, 10].

Our object here is to obtain some results pertaining to the centroid of recursive trees. (When we refer to the centroid node of a tree henceforth it is to be understood that if the tree has two centroid nodes we are referring to the centroid node that is nearer to the root.) In particular, it will follow from our results in Sections 2 and 3 that as $n \to \infty$ the average distance between the root and the (nearer) centroid node of a recursive tree T_n tends to 1; and the average value of the label of the (nearer) centroid node tends to 5/2. These results may be contrasted with the fact that for other familiar families of rooted trees T_n — such as the labelled trees, plane trees, or binary trees, for example — it can be shown that the average distance between the root and the centroid is of the order of $n^{1/2}$ (cf. [8] and [1]).

Australasian Journal of Combinatorics 25(2002), pp.211–219

2. The distance from the root to the centroid node

For any recursive tree T_n let $\delta(T_n)$ denote the distance between the root and the centroid node (that is nearer to the root if there are two centroid nodes). We now derive a formula for $D(n)$, the average value of $\delta(T_n)$ over the $(n-1)!$ recursive trees T_n.

Theorem 2.1. *If* $n \geq 1$, *then*

$$D(T_n) = \begin{cases} (n-1)/(n+1), & n \text{ odd}, \\ (n-2)/(n+2), & n \text{ even}. \end{cases} \tag{2.1}$$

Proof. The result certainly holds when $n = 1$ or 2 so we may suppose that $n \geq 3$. If $1 \leq i \leq n-1$, let $s(i, n-i)$ denote the number of recursive trees T_n with a distinguished edge pq that partitions the nodes of T_n into two subsets A and B such that $|A| = i$, $|B| = n-i$, the root node 1 and node p belong to A, and node q belongs to B. (Note that this implies that $p < q$ where, here and elsewhere, we use the same symbol for a node and for its label.)

Let $A^\star$ denote one of the $\binom{n}{i-1}$ subsets of $[n] := (1, 2, \cdots, n)$ of size $i-1$; and let $A = A^\star \cup \{p\}$ where p denotes the smallest element of $[n]$ not in $A^\star$. (Note that element 1 is necessarily in the set A.) Let $B = [n] \backslash A$ and let q denote the smallest element of B. Now let T_i be one of the $(i-1)!$ recursive trees with node-set A, rooted at node 1; and let T_{n-i} be one of the $(n-i-1)!$ recursive trees with node-set B, rooted at node q. Finally, let T_n be the tree obtained by joining node p of the tree T_i to the node q of the tree T_{n-i}. It is not difficult to see, bearing in mind the definitions of p and q and A and B, that the resulting tree T_n is a recursive tree with node-set $[n]$ in which the distinguished edge pq has the required properties. Moreover, when this construction is carried out in all possible ways, each tree T_n is counted separately for each such distinguished edge pq it contains.

Consequently,

$$s(i, n-i) = \binom{n}{i-1} (i-1)!(n-i-1)! = n!\big((n-1)(n-i+1)\big)^{-1}. \tag{2.2}$$

(We remark that relation (2.2) is equivalent to Lemma 1 in [8]; but the derivation given here is more direct.)

Consider one of the $s(i, n-i)$ recursive trees T_n with a distinguished edge pq that partitions the nodes of T into subsets A and B such that $|A| = i$, $|B| = n-i$, nodes 1 and p belong to A, and node q belongs to B. If u and v are any nodes of A and B, respectively, then $\kappa(u) \geq n-i$ and $\kappa(v) \geq i$. So if $i > n/2$, then no node of B can be a centroid node and, hence, the centroid node(s) must belong to A. Similarly, if $i < n/2$, the centroid node(s) must belong to B. And if $i = n/2$, then p and q are each centroid nodes. It follows, therefore, that the distinguished edge pq is on the path from the root of T_n to the (nearer) centroid node of T_n if and only if $1 \leq i < n/2$. Consequently, if we sum $s(i,\ n-i)$ over i, for $1 \leq i < n/2$, each tree T_n is counted $\delta(T_n)$ times. This implies that

$$\sum_{i=1}^{M} s(i,\ n-i) = D(n) \cdot (n-1)! \tag{2.3}$$

where $M = [(n-1)/2]$. (We remark that the basic idea underlying identity (2.3) is a slight refinement of an observation given by Wiener [11; p. 17, para. 4]; he pointed out that the sum over all edges of a tree of the product of the number of pairs of nodes separated by the edge is equal to the sum of the distances between all pairs of nodes of the tree.)

When we combine relations (2.2) and (2.3) and simplify, we find that

$$\begin{aligned} D(n) &= n \sum_{i=1}^{M} \big((n-i)(n-i+1)\big)^{-1} \\ &= n\{(n-M)^{-1} - n^{-1}\} = M/(n-M), \end{aligned}$$

and this implies conclusion (2.1). □

Let $D_2(n)$ denote the second factorial moment of $\delta(T_n)$ over the $(n-1)!$ recursive trees T_n. The argument used in Theorem 2.1 can be extended to show that if $n \geq 3$, then

$$D_2(n) = (2n/N) \cdot \sum_{h=N+2}^{n} h^{-1} - 2n/(N+1) + 2$$

where $N = [n/2]+1$. Consequently, $D_2(n) \to 4\log 2 - 2$ as $n \to \infty$. Furthermore, it can be shown (by an extension of the argument that will be used to prove Lemma 3.1 in the next section) that if $0 \leq k \leq (n-1)/2$, then

$$Pr\{\delta(T_n) \geq k\} = \sum{}' (h_1 \cdots h_k)^{-1},$$

when the sum is over all k-tuples of integers $h_1, \cdots, h_k$ such that

$$n/2 < h_1 < \cdots < h_k \leq n-1$$

and where an empty sum equals one. Consequently,

$$Pr\{\delta(T_n) \geq k\} \to (\log 2)^k / k!$$

for each fixed non-negative integer k as $n \to \infty$.

3. The label of the centroid node

For any recursive tree T_n let $\alpha(T_n)$ denote the label of the centroid node (that is nearer to the root if there are two centroid nodes). Our main object in this section is to derive a formula for $A(n)$, the average value of $\alpha(T_n)$ over the $(n-1)!$ recursive trees. First, however, we introduce some more definitions and prove two lemmas.

Suppose $\alpha(T_n) = a > 1$. Then the *root branch* of T_n is the branch joined to the centroid node a that contains the root of T_n. Let $\beta(T_n)$ denote the number of nodes in the root branch of T_n. If $a = 1$ then the root of T_n is the centroid node and we say that T_n has an empty root branch and we let $\beta(T_n) = 0$. It follows readily from these definitions and Jordan's theorem that $\beta(T_n) < n/2$.

Consider the subtree of T_n rooted at the centroid node a, i.e., the subtree determined by node a and all nodes u such that the path from the root node to u contains node a. This subtree has $n - \beta(T_n)$ nodes and all these nodes — apart from node a — have labels larger than a. Hence, $n - \beta(T_n) - 1 \le n - a$ or $\alpha(T_n) \le \beta(T_n) + 1$.

Let $F(n; a, b)$ denote the number of recursive trees T_n such that $\alpha(T_n) = a$ and $\beta(T_n) = b$. We now derive a preliminary result that we shall use in obtaining a formula for $F(n; a, b)$.

Lemma 3.1. *Let m and h be integers such that $m/2 \le h \le m-1$, and let $N(m, h)$ denote the number of recursive trees T_m with a (unique) branch of size h joined to the root. Then*

$$N(m, h) = (m-1)! \cdot h^{-1}$$

Proof. If $m/2 \le h \le m-1$, then the branch of size h joined to the root is clearly unique. The nodes in the branch of size h can be selected in $\binom{m-1}{h}$ ways, since the node labelled 1 cannot be one of the selected nodes. There are $(h-1)!$ ways of forming a recursive tree T_h on these h nodes and $(m-h-1)!$ ways of forming a recursive tree T_{m-h} on the remaining $m-h$ nodes. When we join the root-node of the tree T_h to the root-node of the tree T_{m-h}, we obtain a recursive tree T_n in which the root is joined to a branch of size h. It follows, therefore, that

$$N(m, h) = \binom{m-1}{n} (h-1)!(m-h-1)! = (m-1)! \cdot h^{-1},$$

as required. □

We now derive a formula for $F(n; a, b)$.

Lemma 3.2. *If $a = 1$ and $b = 0$, then*

$$F(n; 1, 0) = (n-1)!\Big\{1 - \sum_{n/2 < h \le n-1} h^{-1}\Big\}. \tag{3.1}$$

If $2 \le a \le b+1$ and $1 \le b < n/2$, then

$$F(n; a, b) = (a-1)\binom{n-a}{n-b-1}(b-1)!(n-b-1)! \cdot \Big\{1 - \sum_{n/2 < h \le n-b-1} h^{-1}\Big\}. \tag{3.2}$$

Proof. If $\alpha(T_n) = 1$ and $\beta(T_n) = 0$, then the root-node of T_n is a centroid node and, hence, is not incident with any branches of size h, for $n/2 < h \leq n-1$. Therefore, by Lemma 3.1,

$$F(n;1,0) = (n-1)! - \sum{}' N(n,h) = (n-1)!\left\{1 - \sum{}' h^{-1}\right\},$$

where, here and elsewhere, the sums $\sum'$ are over h such that $n/2 < h \leq n-1$. This proves (3.1).

If $\alpha(T_n) = a$ and $\beta(T_n) = b$ where $2 \leq a \leq b+1$ and $1 \leq b < n/2$, then the root-node of T_n is not a centroid node. The $n-b-1$ nodes of the subtree of T_n rooted at the centroid node a — other than the node a itself — can be selected in $\binom{n-a}{n-b-1}$ ways; this follows from the earlier observation that the labels of these nodes must exceed a. The number of ways of forming a recursive tree T_{n-b} on these selected nodes plus node a in which the root-node a is *not* joined to any branches of size h, where $n/2 < h \leq n-b-1$, is equal to

$$(n-b-1)!\left\{1 - \sum_{n/2<h\leq n-b-1} h^{-1}\right\},$$

in view of Lemma 3.1. The number of ways of forming a recursive tree T_b on the remaining b nodes is $(b-1)!$. Finally, there are $a-1$ ways of joining the root-node a of T_{n-b} to a node with a smaller label in the tree T_b to form a recursive tree T_n such that $\alpha(T_n) = a$ and $\beta(T_n) = b$. Combining these observations, we find that

$$F(n;a,b) = (a-1)\binom{n-a}{n-b-1}(b-1)!(n-b-1)! \cdot \left\{1 - \sum_{n/2<h\leq n-b-1} h^{-1}\right\},$$

as required. □

We now derive a formula for $A(n)$, the average value of $\alpha(T_n)$ over the $(n-1)!$ recursive trees T_n.

Theorem 3.1. *If* $n \geq 1$, *then*

$$A(n) = \begin{cases} (5n+3)/(2n+6), & n \text{ odd}, \\ (5n^2+10n+8)/2(n+2)(n+4), & n \text{ even}. \end{cases} \tag{3.3}$$

Proof. The result certainly holds when $n = 1$ or 2 so we may assume that $n \geq 3$. It follows from the definitions of $F(n;a,b)$ and $A(n)$ that

$$A(n) = \left\{F(n;1,0) + \sum_{b=1}^{M}\sum_{a=2}^{b+1} aF(n;a,b)\right\} \div (n-1)!, \tag{3.4}$$

where $M := [(n-1)/2]$.

We observe that

$$\sum_{a=2}^{b+1} a(a-1)\binom{n-a}{n-b-1} \tag{3.5}$$
$$= \frac{(n+1)!}{(b-1)!(n-b-1)!}\cdot\left\{\frac{1}{n-b}-\frac{2}{n-b+1}+\frac{1}{n-b+2}\right\}.$$

This follows, after simplification, upon writing

$$a(a-1) = n(n+1) - 2(n+1)(n-a+1) + (n-a+2)(n-a+1)$$

and then using the identity

$$\sum_{a=2}^{b+1}\binom{Q-a}{Q-1-b} = \binom{Q-1}{Q-b}$$

with $Q = n,\ n+1$, and $n+2$. Relations (3.2) and (3.5) imply that

$$\sum_{a=2}^{b+1} aF(n;a,b) \tag{3.6}$$
$$= (n+1)!\left\{\frac{1}{n-b}-\frac{2}{n-b+1}+\frac{1}{n-b+2}\right\}\cdot\left\{1 - \sum_{n/2<h\le n-b-1} h^{-1}\right\}$$

for $1 \le b \le M$.

We now sum relation (3.6) over the relevant values of b. This yields the relation

$$\sum_{b=1}^{M}\sum_{a=2}^{b+1} aF(n;a,b) = (n+1)!\left\{S_1 - S_2 + 2S_3 - S_4\right\} \tag{3.7}$$

where $S_1,\ S_2,\ S_3$, and S_4 are as follows:

$$S_1 := \sum_{b=1}^{M} \left\{ \frac{1}{n-b} - \frac{2}{n-b+1} + \frac{1}{n-b+2} \right\}$$
$$= \frac{1}{n-M} - \frac{1}{n-M+1} - \frac{1}{n} + \frac{1}{n+1};$$

$$S_2 := \sum_{1\le b<n/2} \frac{1}{n-b} \cdot \sum_{n/2<h\le n-b-1} \frac{1}{h}$$
$$= \sum_{n/2<h_1<h_2\le n-1} (h_1 h_2)^{-1};$$

$$S_3 := \sum_{1\le b<n/2} \frac{1}{n-b+1} \cdot \sum_{n/2<h\le n-b-1} \frac{1}{h}$$
$$= S_2 + n^{-1} {\sum}' \ h^{-1} - {\sum}' \ \big(h(h+1)\big)^{-1};$$

and, finally,

$$S_4 = \sum_{1\le b<n/2} \frac{1}{n-b+2} \cdot \sum_{n/2<h\le n-b-1} \frac{1}{h}$$
$$= S_2 + (n+1)^{-1} {\sum}' h^{-1} + n^{-1} {\sum}' h^{-1} - {\sum}' \big(h(h+1)\big)^{-1} - {\sum}' \big(h(h+2)\big)^{-1}.$$

When we substitute these expressions for $S_1,\ S_2,\ S_3$, and S_4 into relation (3.7), combine the telescoping sums, and simplify, we find that

$$\text{(3.8)} \qquad \sum_{b=1}^{M} \sum_{a=2}^{b+1} aF(n;a,b) = (n-1)! \left\{ {\sum}' \ h^{-1} - \frac{1}{2} + \frac{n(n+1)}{2(n-M)(n+1-M)} \right\}.$$

It now follows from (3.4), (3.1), and (3.8) that

$$A(n) = \frac{1}{2} + \frac{n(n+1)}{2(n-M)(n+1-M)}$$

where $M = [(n-1)/2]$; and it is easy to see, considering odd and even values of n separately, that this implies conclusion (3.3). □

We remark that when n is even there are $4(n+2)^{-1} \cdot (n-1)!$ recursive trees T_n with two centroid nodes. If we restrict our attention to these trees, then it can be shown that the average value of the label of the centroid node closer to the root is

$2(n+1)/(n+4)$ and the expected value of the label of the further centroid node is $4(n+1)/(n+4)$.

We conclude by stating without proof some other results that can be deduced from Lemma 3.2. It follows from (3.1) that

$$Pr\{\alpha(T_n)=1\} = 1 - \sum_{n/2<h\le n-1} h^{-1} \to 1-\log 2,$$

where the probability is over all the $(n-1)!$ recursive trees T_n. More generally, it can be shown that if a is any fixed positive integer, then

$$\begin{aligned} Pr\{\alpha(T_n)=a\} &\to (1/2)^{a-1} + \sum_{i=1}^{a-1} \frac{1}{i}(1/2)^i - \log 2 \\ &= (1/2)^{a-1} - \sum_{i=a}^{\infty} \frac{1}{i}(1/2)^i \end{aligned}$$

as $n \to \infty$.

Moreover, it can be shown that if $b > 0$, then

$$Pr\{\beta(T_n)=b\} = \frac{n}{(n-b)(n-b+1)} \cdot \Big\{1 - \sum_{n/2<h\le n-b-1} h^{-1}\Big\}.$$

From this it follows that if $B(n)$ denotes the average value of $\beta(T_n)$ over the $(n-1)!$ recursive trees T_n, then

$$\begin{aligned} B(n)/n &= \sum_{n/2<h_1<h_2\le n} (h_1h_2)^{-1} \\ &\to \frac{1}{2}\log^2 2 = .2402\ldots \end{aligned}$$

as $n \to \infty$.

References

[1] D. Aldous, *Recursive self-similarity for random trees, random triangulations and Brownian excursions*, Ann. Probab. **22** (1994), 527-545.

[2] M. Dondajewski, P. A. Kirschenhofer, and J. Szymanski, *Vertex-degrees in strata of a random recursive tree*, North-Holland Math. Studies **118** (1985), 99-105.

[3] C. Jordan, *Sur les assemblages des lignes*, J. Reine Angew. Math. **70** (1869), 185-190.

[4] H. M. Mahmoud, *Limiting distributions for path lengths in recursive trees*, Probab. Engrg. Inform. Sci. **5** (1991), 53-59.

[5] H. M. Mahmoud, R. T. Smythe, and J. Szymanski, *On the structure of random plane-oriented recursive trees and their branches*, Random Structures Algorithms **4** (1993), 151-176.

[6] A. Meir and J. W. Moon, *Cutting down recursive trees*, Math. Biosciences **21** (1974), 173-181.

[7] A. Meir and J. W. Moon, *On the altitudes of nodes in random trees*, Canad. J. Math. **30** (1978), 97-1015.

[8] J. W. Moon, *On the expected distance from the centroid of a tree*, Ars Combin. **20A** (1985), 263-276.

[9] R. T. Smythe and H. M. Mahmoud, *A survey of recursive trees*, Th. Probab. Math. Statist. **51** (1995), 1-27.

[10] J. Szymanski, *Branches in recursive trees*, Fasc. Math. **29** (1999), 139-147.

[11] H. Wiener, *Structural determination of paraffin boiling points*, J. Amer. Chem. Soc. **69** (1947), 17-20.

(Received 10 Jan 2001)