On the distribution of
distances in recursive trees

BY ROBERT P. DOBROW

National Institute of Standards and Technology

Abstract

Recursive trees have been used to model such things as the spread of epidemics, family trees of ancient manuscripts, and pyramid schemes. A tree $T_n$ with $n$ labeled nodes is a recursive tree if $n = 1$, or $n > 1$ and $T_n$ can be constructed by joining node $n$ to a node of some recursive tree $T_{n-1}$. For arbitrary nodes $i < n$ in a random recursive tree we give the exact distribution of $X_{i,n}$, the distance between nodes $i$ and $n$. We characterize this distribution as the convolution of the law of $X_{i,i+1}$ and $n-i-1$ Bernoulli distributions. We further characterize the law of $X_{i,i+1}$ as a mixture of sums of Bernoullis. For $i = i_n$ growing as a function of $n$, we show that $X_{i_n,n}$ is asymptotically normal in several settings.

1 AMS 1991 subject classifications. Primary 05C05; secondary 60C05

2 Keywords and phrases. recursive trees, Stirling numbers of the first kind.
1 Introduction and summary

A tree on $n$ nodes (vertices) labeled $1, 2, \ldots, n$ is a recursive tree if the node labeled 1 is distinguished as the root, and for each $2 \leq k \leq n$, the labels of the nodes in the unique path from the root to the node labeled $k$ form an increasing sequence. Equivalently, a tree $T_n$ on $n$ nodes is a recursive tree if $n = 1$, or $n > 1$ and $T_n$ is obtained by joining the $n$th node to a node of some recursive tree $T_{n-1}$.

The usual model of randomness on the space of $n$-node recursive trees is to assume that all $(n-1)!$ trees are equally likely. It is easy to see that given a random tree $T_{n-1}$ on $n-1$ nodes, we obtain a random tree on $n$ nodes by choosing a node (a parent) of $T_{n-1}$ uniformly at random and joining a node labeled $n$ (a child) to it.


For nodes $i$ and $j$ in a tree $T$, the distance between $i$ and $j$ is the number of edges on the necessarily unique path between nodes $i$ and $j$. The distance $X_{i,j}$ between nodes $i$ and $j$ in a random recursive tree of order $n$ was studied by Moon (1974) who found the expectation and variance of $X_{i,j}$. Szymański (1990) derived the exact distribution for the height of the node with label $n$, that is, $X_{1,n}$. Devroye (1988) and Mahmoud (1991) have established the asymptotic normality of a normalized version of $X_{1,n}$.

In this paper we study the distribution of the distance between arbitrary nodes in a recursive tree. Note that for $1 \leq i < j \leq n$, the distribution of $X_{i,j}$ does not depend on $n$ and thus without loss of generality we need only consider $X_{i,n}$. In Section 2, the exact distribution of $X_{i,n}$ is derived for arbitrary $i$. This distribution is the convolution of $n-i-1$ Bernoulli distributions and the law of $X_{i,i+1}$. We further exhibit the distribution of $X_{i,i+1}$ as a mixture of sums of Bernoulli distributions. In Section 3, the asymptotic normality of $X_{i,n}$ is established for $i = i_n$ growing as a function of $n$. 

2
2 Exact distribution of distances

In this section we compute \( P(X_{i,n} = d) \), for \( 1 \leq i < n \) and \( 1 \leq d \leq n - 1 \).
This will involve the univariate distributions of \( X_{k,k+1} \), for \( 1 \leq k \leq n - 1 \),
which we give explicitly in Theorem 2 below. The distribution also involves
\( s(n, k) \), the Stirling numbers of the first kind. For nonnegative integers \( n \) and \( k \),
\( s(n, k) \) is the coefficient of \( x^k \) in the product \( x(x - 1) \cdots (x - n + 1) \).

Note that \( P(X_{i,n} = d) = 0 \), for \( d < 1 \) and for \( d > n - 1 \), and \( P(X_{i,i} = 0) = 1 \).

The following lemma, given by Moon (1974), is essential for our development.

Lemma 2.1 If \( 1 \leq i < n \) and \( 1 \leq d \leq n - 1 \) then
\[
P(X_{i,n} = d) = \frac{1}{n - 1} \sum_{k=1}^{n-1} P(X_{i,k} = d - 1).
\]

Proof Condition on the parent of node \( n \). Tree \( T_n \) is obtained from \( T_{n-1} \)
by choosing a node \( k \) uniformly at random from \( 1, 2, \ldots, n - 1 \). The distance
between \( i \) and \( n \) in \( T_n \) is equal to the sum of the distance between \( n \) and \( k \)—which is one—and the distance between \( k \) and \( i \).

For a random variable \( X \) let \( \mathcal{L}(X) \) denote the distribution (law) of \( X \).
For independent random variables \( X \) and \( Y \) we write \( X \oplus Y \) for the sum
of \( X \) and \( Y \). Let \( \text{Be}(p) \) denote a Bernoulli random variable with success
probability \( p \).

Theorem 1 For \( 1 \leq i < n \),
\[
\mathcal{L}(X_{i,n}) = \mathcal{L}(Y \oplus X_{i,i+1}) = \mathcal{L} \left( \left( \oplus_{k=i+1}^{n-1} \text{Be}(1/k) \right) \oplus X_{i,i+1} \right),
\]
where
\[
P(Y = y) = \frac{i!}{(n - 1)!} \sum_{k=y}^{n-i-1} |s(n - i - 1, k)| \frac{k}{y^k},
\]
for \( y = 0, \ldots, n - i - 1 \). The distribution of \( X_{i,i+1} \) is given in Theorem 2
below.
Proof With the standard convention that \( s(0, 0) = 1 \), the result is trivial for \( i = n - 1 \). Suppose that \( i < n - 1 \). Then for \( 1 \leq d \leq n - 1 \), Lemma 2.1 gives

\[
P(X_{i,n} = d) = \left( \frac{n-2}{n-1} \right) \frac{1}{n-2} \sum_{k=1}^{n-2} P(X_{k,i} = d - 1) + \frac{1}{n-1} P(X_{i,n-1} = d - 1) = \frac{n-2}{n-1} P(X_{i,n-1} = d) + \frac{1}{n-1} P(X_{i,n-1} = d - 1). \tag{3}
\]

Fixing \( i \), let \( F_k(z) := \sum_{d=1}^{n-1} P(X_{i,k} = d)z^d \) be the probability generating function of \( X_{i,k} \). Then by multiplying (3) by \( z^d \) and summing over \( d = 2, \ldots, n-1 \),

\[
F_n(z) - P(X_{i,n} = 1)z = \frac{n-2}{n-1} \left[ F_{n-1}(z) - P(X_{i,n-1} = 1)z \right] + \frac{z}{n-1} \left[ F_{n-1}(z) - P(X_{i,n-1} = n-1)z^{n-1} \right]. \tag{4}
\]

Clearly \( P(X_{i,n} = 1) = 1/(n-1) \) and \( P(X_{i,n-1} = n-1) = 0 \). Thus

\[
F_n(z) - \frac{z}{n-1} = \frac{n-2}{n-1} \left[ F_{n-1}(z) - \frac{z}{n-2} \right] + \frac{z}{n-1} F_{n-1}(z),
\]

and so

\[
F_n(z) = \left( \prod_{k=1}^{n-i-1} \frac{z + i + k - 1}{i + k} \right) F_{i+1}(z) =: A(z)F_{i+1}(z). \tag{5}
\]

Note that \( A(z) \) is the probability generating function of the convolution of \( n - i - 1 \) independent Bernoulli random variables with success probabilities \( 1/(i + k) \). To obtain (2) we compute the coefficient of \( z^y \) in \( A(z) \).

The signless Stirling number \( |s(n,k)| \) is the coefficient of \( x^k \) in the product \( x(x + 1) \cdots (x + n - 1) \). Thus

\[
A(z) = \frac{i!}{(n-1)!} \sum_{k=0}^{n-i-1} |s(n-i-1,k)|(z+i)^k
\]
\[
\sum_{k=0}^{n-i-1} \frac{i!}{(n-1)!} \binom{k}{y} z^y i^{k-y}
= \frac{i!}{(n-1)!} \sum_{y=0}^{n-i-1} \left[ \sum_{k=y}^{n-i-1} |s(n-i-1, k)| \binom{k}{y} i^{k-y} \right] z^y.
\] (6)

As an immediate corollary we obtain the distribution of \(X_{1,n}\), the depth of the last node inserted.

**Corollary 2.1**

\[
\mathcal{L}(X_{1,n}) = \mathcal{L} \left( \oplus_{k=1}^{n-1} \text{Be}(1/k) \right).
\] (7)

\[
P(X_{1,n} = d) = \frac{1}{(n-1)!} \sum_{k=d-1}^{n-2} |s(n-2,k)| \binom{k}{d-1} = \frac{|s(n-1,d)|}{(n-1)!}.
\] (8)

Note that the second equality in (8) is a well-known identity for Stirling numbers of the first kind (cf., Graham, et al. (1989)).

The distribution defined in (7) and (8) arises in several settings. It is the distribution of: the number of cycles in a random permutation of \(n-1\) objects; the number of records in an exchangeable sequence of \(n-1\) unequal random variables; the number of sides in the greatest convex minorant of an \(n-1\) step random walk. See Goldie (1989) for details and related results.

In our case we can specifically identify the Bernoulli random variables in the distribution (7). That is,

\[
X_{1,n} = \sum_{k=1}^{n-1} \mathbf{1}(A_k),
\]

where \(A_k\) is the event that node \(k\) is on the path from the root to node \(n\), and where \(\mathbf{1}(A)\) denotes the indicator of event \(A\). It is not hard to see that the random variables \(\mathbf{1}(A_1), \ldots, \mathbf{1}(A_{n-1})\) are independent and \(P(A_k) = 1/k\) for \(k = 1, \ldots, n-1\).

In the more general setting of Theorem 1, we give an interpretation of the Bernoulli random variables in (1). Consider the following dynamic construction of a random recursive tree: Given tree \(T_{n-1}\) on \(n-1\) nodes, pick a node uniformly at random. If node \(n-1\) is picked, then tree \(T_n\) is formed.
by making node \( n \) a child of \( n - 1 \). If node \( k \neq n - 1 \) is picked, make node \( n \) a child of \( k \) and then swap the labels on nodes \( n - 1 \) and \( n \). The resulting tree \( T_n \), it is easily checked, is a random recursive tree. By this construction, conditional on the tree \( T_{n-1} \),

\[
P(X_{i,n} = X_{i,n-1} + 1) = \frac{1}{n-1} = 1 - P(X_{i,n} = X_{i,n-1}),
\]

for \( n > i + 1 \). For \( k = i + 1, \ldots, n - 1 \), define \( A_k \) to be the event, conditional on \( T_k \), that \( X_{i,k+1} = X_{i,k} + 1 \). Then

\[
\mathcal{L}(X_{i,n}) = \mathcal{L}
\left(\begin{bmatrix} X_{i-1,n-1}^{n-1} \end{bmatrix} \odot \left(A_k + X_{i,i+1} \right) \right).
\]

It now remains to compute the distribution of \( X_{i,i+1} \), which we show to be fundamentally a mixture of independent Bernoulli distributions. Let \( \delta_k \) denote point mass at \( k \). Let \( i \wedge j := \min\{i,j\} \).

**Theorem 2** (a) For \( i \geq 1 \),

\[
\mathcal{L}(X_{i,i+1}) = \frac{1}{i} \sum_{k=1}^{i \wedge 2} \delta_k + \sum_{j=0}^{i-3} \frac{2}{(i-j)(i-j-1)} \mathcal{L}
\left(3 + \sum_{k=0}^{j-1} \text{Be}
\left(\frac{2}{i-k}\right) \right).
\]

(b) For fixed \( i \) and \( 1 \leq d \leq i \),

\[
P(X_{i,i+1} = d) =
\begin{cases}
  1/i, & d = 1, 2 \\
  (2(i-2))/(i(i-1)), & d = 3 \\
  (2^{d-2})/(i!) \sum_{j=d-4}^{i-4} (i-j-3)! \sum_{k=d-3}^{j+1} s(j+1,k) \binom{k}{d-3} (i-2)^{k-d+3}, & d > 3.
\end{cases}
\]

**Proof** The theorem is obviously true for \( i = 1, 2 \), so assume \( i \geq 3 \). The case \( d = 1 \) is clear. The case \( d = 2 \) follows easily from Lemma 2.1. Suppose that \( 3 \leq d \leq i \). Then the event \( \{X_{i,i+1} = d\} \) is equal to the event that for some \( 1 \leq j, k \leq i - 1 \), \( i \) is a child of \( j \), \( i + 1 \) is a child of \( k \), and \( X_{j,k} = d - 2 \). By independence,
\[ P(X_{i,i+1} = d) = \sum_{1 \leq j,k \leq i-1} P(i+1 \text{ is a child of } k)P(i \text{ is a child of } j)P(X_{j,k} = d - 2) \]

\[ = \frac{2}{i(i-1)} \sum_{1 \leq j < k \leq i-1} P(X_{j,k} = d - 2) \]

\[ =: \frac{2}{i(i-1)} f(i,d). \]

An application of Lemma 2.1 gives

\[ f(i,d) = \sum_{l \leq j < k \leq i-1} \frac{1}{k-1} \sum_{l \leq j \leq k-1} P(X_{l,j} = d - 3) \]

\[ + \sum_{1 \leq j < k \leq i-1} \frac{1}{k-1} P(X_{j,j} = d - 3) \]  

\[ = \left( \sum_{k=2}^{i-1} \frac{2}{k-1} \sum_{l \leq j < l \leq k-1} P(X_{j,l} = d - 3) \right) + (i-2)1(d = 3) \]

\[ = \left( \sum_{k=2}^{i-1} \frac{2}{k-1} f(k,d-1) \right) + (i-2)1(d = 3). \]  

(9)

Since \( f(i,d) = 0 \) for \( d < 3 \) and \( d > i \), (10) holds for all positive \( d \). Let \( F_i(z) := \sum_{d=1}^{\infty} f(i,d) z^d \) be the generating function of \( f(i,d) \). We have

\[ F_i(z) - f(i,1)z = 2z \sum_{d=2}^{\infty} \sum_{k=2}^{i-1} \frac{1}{k-1} f(k,d-1) z^{d-1} + (i-2)z^3 \]

\[ = 2z \sum_{k=2}^{i-1} \frac{1}{k-1} F_k(z) + (i-2)z^3. \]

Thus,

\[ F_i(z) = 2z \sum_{k=2}^{i-1} \frac{1}{k-1} F_k(z) + (i-2)z^3 \]

\[ = 2z \sum_{k=2}^{i-2} \frac{1}{k-1} F_k(z) + (i-3)z^3 + \frac{2z}{i-2} F_{i-1}(z) + z^3 \]

\[ = \frac{2z + i - 2}{i - 2} F_{i-1}(z) + z^3. \]
Since $F_2(z) = 0$, this gives for $i \geq 3$,

$$F_i(z) = z^3 + z^3 \sum_{j=0}^{i-4} \prod_{k=0}^{j} \frac{2z + i - k - 2}{i - k - 2}$$

(11)

$$= z^3 + z^3 \sum_{j=0}^{i-4} \frac{(i-j-3)!}{(i-2)!} \prod_{k=0}^{j} (2z + i - 2 - k)$$

$$= z^3 + z^3 \sum_{j=0}^{i-4} \frac{(i-j-3)!}{(i-2)!} \sum_{k=0}^{j+1} s(j+1,k) \sum_{l=0}^{k} \left( \frac{k}{l} \right) (2z)^l (i-2)^{k-l}$$

$$= z^3 + z^3 \sum_{l=0}^{i-3} \left[ 2^l \sum_{j=0}^{i-4} \frac{(i-j-3)!}{(i-2)!} \sum_{k=l}^{j+1} s(j+1,k) \left( \frac{k}{l} \right) (i-2)^{k-l} \right] z^l.$$  

Part (b) follows after computing the coefficient of $z^d$ in the above expression for $d \geq 3$.

The righthand side of (11), suitably normalized, is the generating function of a mixture of sums of Bernoulli distributions and point mass at 3. It is now straightforward (we omit details) to show part (a) of the theorem.

3 Asymptotic distribution of $X_{i,n}$

In this section we let $i = i_n$ grow as a function of $n$ and consider the asymptotic distribution of $X_{i,n}$.

Moon (1974) gives the following formulas for the expectation and variance of $X_{i,n}$:

$$E[X_{i,n}] = H_i + H_{n-1} - 2 + \frac{1}{i}$$

(12)

$$\text{Var}[X_{i,n}] = H_i + H_{n-1} - 3H_i^{(2)} - H_{n-1}^{(2)} + 4 - \frac{4H_i}{i} + \frac{3}{i} - \frac{1}{i^2}.$$  

(13)

where $H_k := \sum_{j=1}^{k} j^{-1}$ is the $k$th harmonic number and $H_k^{(2)} := \sum_{j=1}^{k} j^{-2}$.

Mahmoud (1991) shows that $X_n^* := (X_{1,n} - \ln n) / \sqrt{\ln n}$ converges in distribution to a standard normal random variable. By the triangle inequality,

$$|X_{i,n} - X_{1,n}| \leq X_{1,i}$$
and it follows easily that for fixed \(i\), \((X_{i,n} - \ln n)/\sqrt{\ln n}\) converges in distribution to a standard normal random variable.

A similar argument as in Mahmoud (1991) shows that \(X_{n-1,n}\) is asymptotically normal. By (12) and (13), \(E[X_{n-1,n}] = 2\ln n + O(1)\) and \(\text{Var}[X_{n-1,n}] = 2\ln n + O(1)\). (Thus for “nearly all” recursive trees the distance between nodes \(n-1\) and \(n\) is about twice the distance between the root and node \(n\). Roughly, this means that the nodes which are common ancestors of \(n-1\) and \(n\) are “high up” on the tree, implying that “nearly all” recursive trees are “short” and “wide.”)

**Theorem 3** Let

\[
X_n^* := \frac{X_{n-1,n} - 2\ln n}{\sqrt{2\ln n}}.
\]

Then \(X_n^*\) converges in distribution to a standard normal random variable.

**Proof** Let \(M_n(t) := E[e^{X_n^* t}]\) be the moment generating function of \(X_n^*\). Then

\[
M_n(t) = \sum_{d=1}^{n-1} \exp \left( \frac{d - 2\ln n}{\sqrt{2\ln n}} t \right) P(X_{n-1,n} = d)
\]

\[
= \frac{1}{n!} \sum_{d=4}^{n-1} \exp \left( \frac{d - 2\ln n}{\sqrt{2\ln n}} t \right) 2^{d-2} \times \sum_{k=d-4}^{d-3} \sum_{l=d-3}^{d+1} (n-k-3)! (n-k-3)\left(\frac{l}{d-3}\right) (n-2)^{l-d+3} + o(1)
\]

\[
= \frac{1}{n!} e^{-t\sqrt{2\ln n}} \sum_{k=0}^{n-4} \sum_{l=1}^{k+1} (n-k-3)! s(k+1, l) \left( \frac{l}{d-3} \right) (n-2)^{l-d+3} + o(1)
\]

\[
= \frac{2}{n!} e^{-t\sqrt{2\ln n}} \sum_{k=0}^{n-4} (n-k-3)! \sum_{l=1}^{k+1} s(k+1, l) \left( n-2+2e^{t/\sqrt{2\ln n}} \right)^{l} + o(1)
\]

\[
= 2e^{-t\sqrt{2\ln n}} \Gamma(n+1+\left(2e^{t/\sqrt{2\ln n}} - 2\right)) \Gamma(n+1) \times \sum_{k=0}^{n-4} \frac{\Gamma(n-k-2)}{\Gamma(n+k+2e^{t/\sqrt{2\ln n}} - 2)} + o(1)
\]

\[
= 2e^{-t\sqrt{2\ln n}} \Gamma(n+1+\left(2e^{t/\sqrt{2\ln n}} - 2\right)) \Gamma(n+1)
\]
By Stirling’s approximation and a Taylor expansion, the first ratio of gamma functions in the last expression is asymptotic to

\[
n^2e^t/\sqrt{2\ln n} \sim \exp(t\sqrt{2\ln n} + (t^2/2)) \left(1 + O((\ln n)^{-1/2})\right).
\]

Let \( S_n \) denote the sum in the last expression. Then

\[
\frac{\Gamma(4)}{\Gamma(2 + 2e^t/\sqrt{2\ln n})} \sum_{k=0}^{n-4} \frac{1}{(n-k)(n-k-2)(n-k-1)} \leq S_n \leq \sum_{k=0}^{n-4} \frac{1}{(n-k)(n-k-2)(n-k-1)} = \frac{1}{2} - \frac{1}{n-1}.
\]

Taking limits as \( n \to \infty \) shows \( S_n \to 1/2 \). This gives \( M_n(t) \to e^{t^2/2} \) as \( n \to \infty \). The limit is the moment generating function of the standard normal distribution.

Theorems 1 and 3 afford an easy proof of the asymptotic normality of \( X_{i_n,n} \) when \( i_n \) grows linearly in \( n \).

**Theorem 4** For \( 0 < \lambda < 1 \) and \( i_n := \lfloor \lambda n \rfloor \), let

\[
X_{i_n,n}^* := \frac{X_{i_n,n} - 2\ln n}{\sqrt{2\ln n}}.
\]

Then \( X_{i_n,n}^* \) converges in distribution to a standard normal random variable as \( n \to \infty \).

**Proof** By Theorem 1,

\[
X_{i_n,n}^* \stackrel{d}{=} \frac{Y_n}{\sqrt{2\ln n}} + \frac{X_{i_n,i_n+1} - 2\ln n}{\sqrt{2\ln n}},
\]

where \( Y_n \) is distributed as in (2). By Markov’s inequality, for \( t > 0 \),

\[
P(Y_n > t\sqrt{2\ln n}) \leq \frac{H_{n-1} - H_{i_n}}{t\sqrt{2\ln n}} = O((\ln n)^{-1/2}).
\]
Thus $Y_n/\sqrt{2\ln n} \to 0$ in probability. It follows easily from Theorem 3 that $(X_{i_n,i_n+1} - 2\ln n)/\sqrt{2\ln n}$ converges in distribution to a standard normal random variable.

4 Acknowledgements

I thank Hosam Mahmoud for introducing me to recursive trees and several helpful discussions. I thank Jim Fill for all his advice and suggestions.

5 References


