A Class of Random Recursive Tree Algorithms with Deletion

Arnold T. Saunders Jr.¹

Received: 15 January 2020 / Accepted: 17 July 2021 / Published online: 27 July 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
We examine a discrete random recursive tree growth process that, at each time step, either adds or deletes a node from the tree with fixed, complementary probabilities. Node addition follows the usual uniform attachment model. For node removal, we identify a class of deletion rules guaranteeing the current tree conditioned on its size is uniformly distributed over its range. By using generating function theory and singularity analysis, we obtain asymptotic estimates for the expectation and variance of a tree’s size, as well as its expected leaf count and root degree. In all cases, the behavior of such trees falls into three regimes determined by the insertion probability. Interestingly, the results are independent of the specific class member deletion rule used.

Keywords Recursive trees · Random deletions · Generating functions · Singularity analysis

1 Introduction

Tree evolution algorithms supporting both node insertion and deletion are notoriously hard to analyze. Jonassen and Knuth [8] showed deriving the distribution of a mere three-node random binary search tree after a finite series of repeated insertions and deletions required Bessel functions and solving bivariate integral equations. In their words, “the analysis ranks among the more difficult of all exact analyses of algorithms...the problem itself is intrinsically difficult.” Panny [10] later chronicled a near half century of hopeful assumptions and poor intuition about the effect of deletions on binary search tree distribution. In this paper, we study the effect of a class of deletion rules on the evolution of random recursive trees.

¹ Department of Statistics, The George Washington University, Washington, DC 20052, USA
Random recursive trees are stochastic growth processes with diverse applications in modeling searching and sorting algorithms, the spread of rumors, Ponzi schemes and manuscript provenance [13]. The idea behind the model is uncomplicated. Starting from a root node labeled 1, we construct a tree one vertex at a time using sequentially labeled nodes. Each newly introduced node is “randomly” attached to an existing one in the tree.

The insertion or attachment rule we use to construct a random tree determines the distribution over its range. For example, consider an insertion rule where each new node is attached to any of the existing ones with equal probability. The resulting trees are known as uniform recursive trees or uniform attachment trees. In this case, at time \( n \), the random tree is uniformly distributed over the \( n! \) possible recursive trees with \( n + 1 \) nodes. Much research has gone into characterizing the limiting random variables and distributions of functionals on uniform recursive trees such as node degree [3, 7], height [11], leaf count [9], etc.

Motivated by work with random graph models incorporating both insertion and deletion rules [1, 2, 5, 14, e.g.,], we examine the less-studied application of such rules to tree evolution models. Specifically, we start with tree \( T_0 \) containing a single node labeled 1. At each time step \( n \geq 1 \), we either add an incrementally-labeled node to the tree with probability \( p \) or delete an existing node with probability \( q = 1 - p \). After a deletion, we reattach and relabel the remaining nodes so that \( T_{n+1} \) is again a recursive tree. There is one exception to the preceding: we do not allow the tree to vanish. So if \( T_n \) is the single node tree, it remains unchanged with probability \( q \).

We always add nodes using the uniform attachment rule. We will however identify the class of deletion rules guaranteeing \( T_n \), when conditioned on tree size, remains uniformly distributed over its range. We then, using singularity analysis of generating functions, provide a means for deriving the exact and asymptotic expressions of common functionals on \( T_n \) such as tree size, leaf count and root degree.

2 Conditional Equiprobability

A simplifying property of uniform attachment trees is the equiprobability of the range of \( T_n \). Once we introduce deletion, this need not be the case. But if our choice of deletion rule could guarantee—conditioned on tree size—a uniformly distributed \( T_n \), its analysis is greatly simplified. To specify the class of such deletion rules, we must first make concrete the notion of insertion and deletion rules.

Let \( T \) denote the set of all recursive trees. Define the size of a tree in \( T \) to be the number of nodes it possesses. Next define the stratum number of a tree to be one less than its size. Let stratum \( k \), denoted by \( T_k \), be the subset of all trees sharing the common stratum number \( k \). Then \( T_k \) contains \( k! \) trees and we can assign each one a unique integer identifier from 1 to \( k! \) and arrange them in canonical order. We can now capture the probabilities of transitioning from one of the \( k! \) trees in \( T_k \) to one of the \( (k + 1)! \) trees in \( T_{k+1} \) (an insertion) in a single \( k! \times (k + 1)! \) conditional probability matrix \( P_{k,k+1} \). Analogously, we can record the probabilities of transitioning from a tree in \( T_{k+1} \) to a tree in \( T_k \) (a deletion) by \( Q_{k+1,k} \). Insertion and deletion rules then
are simply specifications of the matrices $P_{k,k+1}$ and $Q_{k+1,k}$ for each $k \geq 0$, which we will call insertion and deletion matrices, respectively.

Since we are using uniform attachment as our insertion rule, each insertion matrix $P_{k,k+1}$ has the form

$$P_{k,k+1} = \frac{p}{k+1} D,$$

where $D$ is a 0-1 matrix with row sums $k+1$ and columns sums 1. The exact placement of the 0s and 1s depends on the tree canonicalization used.

In the next theorem, we identify a necessary and sufficient condition on deletion matrices $Q_{k+1,k}$ for conditional equiprobability and then establish the class of growth algorithms with that property.

**Theorem 1** Conditioned on stratum number, each tree is equiprobable at time $n \geq 1$ if and only if

$$\mathbf{1}_{k+1} Q_{k+1,k} \propto \mathbf{1}_k (0 \leq k \leq n-1),$$

where $\mathbf{1}_k$ is the row vector $\{1, \ldots, 1\}$ of length $k!$ and $u \propto v$ denotes direct proportionality, i.e., there exists a non-zero scalar $c$ such that $u = cv$. Note the above is equivalent to requiring all column sums of $Q_{k+1,k}$ to be identical. In fact, the constant column sum is $(k+1)q$.

**Proof** ($\Rightarrow$) Assume that, conditioned on stratum number, each tree within a stratum is equiprobable at time $n \geq 1$. When $k = 0$, the assertion is trivially true so let us assume $k \geq 1$. Let $T_n$ denote the recursive tree at time $n$ and $S_n$ its stratum number. Additionally, let $t$ be a tree in stratum $S_n$. Then by hypothesis, we have

$$\mathbb{P}(T_n = t \mid S_n = k) = \frac{1}{k!},$$

or equivalently

$$\mathbb{P}(T_n = t) = \frac{1}{k!} \mathbb{P}(S_n = k).$$

If we denote the distribution of $T_n$ within stratum $k$ by $\pi_n^{(k)}$, we can summarize this result succinctly with

$$\pi_n^{(k)} = \frac{1}{k!} \mathbb{P}(S_n = k) \mathbf{1}_k.$$

Consequently, by conditioning on the action (i.e., insertion or deletion) at time $n$, we can express the distribution of the $k$th stratum ($1 \leq k \leq n-1$) at time $n+1$ by the following equality
\[
\frac{1}{k!} \mathbb{P}(S_{n+1} = k) 1_k = \pi_n^{(k)} = \pi_n^{(k-1)} P_{k-1,k} + \pi_n^{(k+1)} Q_{k+1,k}
\]
\[
= \frac{1}{(k-1)!} \mathbb{P}(S_n = k-1) 1_{k-1} P_{k-1,k} + \frac{1}{(k+1)!} \mathbb{P}(S_n = k+1) 1_{k+1} Q_{k+1,k}.
\]

Next, by observing
\[
1_{k-1} P_{k-1,k} = \frac{p}{k} 1_k
\]
and noting the inequality
\[
\mathbb{P}(S_n = k) \geq \mathbb{P} \text{ (staying in place } n-k \text{ iterations, followed by } k \text{ insertions) } = q^{n-k} p^k > 0
\]
holds whenever \(0 \leq k \leq n\), the probability \(\mathbb{P}(S_n = k+1)\) in (3), subject to the given constraint \(1 \leq k \leq n-1\), is positive, we can rearrange the terms on the left and right-hand sides of (3) to obtain
\[
k+1 \mathbb{P}(S_n = k+1) = \mathbb{P}(S_n = k) - p \mathbb{P}(S_n = k-1) \mathbb{P}(S_{n+1} = k) - p 1_k
\]
Finally, when \(1 \leq k \leq n-1\), we have
\[
\mathbb{P}(S_{n+1} = k) = p \mathbb{P}(S_n = k-1) + q \mathbb{P}(S_n = k+1),
\]
and substituting this result into (4) gives us \(1_{k+1} Q_{k+1,k} = (k+1) q 1_k \propto 1_k\) as claimed.

\((\Leftarrow\Rightarrow)\) Assume \(1_{k+1} Q_{k+1,k} \propto 1_k\) for arbitrary \(0 \leq k \leq n-1\). By using mathematical induction on \(n\), we show (2) holds.

Since strata 0 and 1 contain only one tree each, the result is trivially true for \(n = 1\). Next assume it also holds for some arbitrary \(n \geq 1\) and consider the case \(n+1\). Since (2) always holds for \(k = 0\) and \(k = n\) at time \(n\), we can restrict our attention to \(1 \leq k \leq n\) at time \(n+1\). Now, by conditioning on the action at time \(n\), we have
\[
\pi_n^{(k)} = \left\{ \begin{array}{ll}
\pi_n^{(k-1)} P_{k-1,k} + \pi_n^{(k+1)} Q_{k+1,k}, & 1 \leq k \leq n-1 \\
\pi_n^{(k-1)} P_{k-1,k}, & k = n.
\end{array} \right.
\]
Thus when \(k = n\) we have
\[
\pi_n^{(k)} = \pi_n^{(n)} = \pi_n^{(n-1)} P_{n-1,n} \propto 1_{n-1} P_{n-1,n} = \frac{p}{n} 1_n \propto 1_k.
\]
On the other hand, when \(k < n\), we have
\[
\pi_n^{(k)} = \pi_n^{(k-1)} P_{k-1,k} + \pi_n^{(k+1)} Q_{k+1,k} \propto 1_{k-1} P_{k-1,k} + 1_{k+1} Q_{k+1,k} \propto \frac{p}{k} 1_k + 1_k \propto 1_k.
\]
In both cases we have \(\pi_n^{(k)} = \beta 1_k\) for some \(\beta > 0\). Recalling
where $T_k$ is the set of stratum $k$ trees, we conclude $\beta = \frac{1}{k!} \mathbb{P}(S_{n+1} = k)$ as desired. $\square$

A consequence of Theorem 1 is that if our choice of deletion algorithm obeys (1) for arbitrary $n \geq 1$, then for all $n \geq 1$, all trees in the same stratum are equiprobable. We summarize this result with the following corollary.

Corollary 1 (The Class of Conditional Equiprobable Growth Algorithms) If the deletion matrix $Q_{k+1,k}$ for a given growth algorithm satisfies (1) for arbitrary $k \geq 0$, then it supports conditional equiprobability. Moreover since any conditional equiprobable algorithm possesses this property, this criterion describes the class of such algorithms. Finally, this class is nonempty.

Proof To show the class is not empty consider the “last in, first out” (LIFO) deletion rule. When invoked, we delete the last node inserted into the tree. Then each row in an arbitrary deletion matrix $Q_{k+1,k}$ contains exactly one nonzero entry, $q$. Each column of this matrix represents a stratum $k$ tree. By adding a node to this tree, we obtain $k+1$ trees in stratum $k+1$. Hence each of the column sums is $(k+1)q$, satisfying condition (1). $\square$

3 Tree Size Generating Functions and Asymptotics

Having established the class of deletion rules ensuring $T_n$ given $\{S_n = k\}$ is equally likely to be any one of the $k!$ trees in stratum $k$, we next explore the distribution and moments of $S_n$, as well as those of several functions of $S_n$.

Proposition 1 Let $P_{n,k}$ denote the probability of obtaining a stratum $k$ tree on the $n$th iteration. If we mark the iteration with $z$ and the stratum number with $u$, then the bivariate generating function $P(z, u)$ for the double sequence $\{P_{n,k}; n, k \geq 0\}$ is

$$P(z, u) = \sum_{n,k} P_{n,k} u^k z^n = \frac{q(1-u)zP(z,0) - u}{qz - u(1-puz)},$$

where

$$P(z,0) = \frac{2}{1 - 2qz + \sqrt{1 - 4pqz^2}}$$

is the generating function for the sequence $\{P_{n,0}; n \geq 0\}$.

Proof To derive the generating function $P(z, u)$, we begin with the recurrence condition
For all $i \geq 0$, let us define a corresponding formal power series

$$P^{(i)}(z) \equiv \sum_{n \geq 0} P_{n,i} z^n$$

so that multiplying (7) through by $z^n$ and summing over $n \geq 0$ yields

$$\sum_{n \geq 0} P_{n+1,k} z^n = pP^{(k-1)}(z) + qP^{(k+1)}(z).$$

The left side of the preceding can be written

$$\frac{1}{z} \sum_{n \geq 0} P_{n+1,k} z^n = \frac{1}{z} \sum_{n \geq 1} P_{n,k} z^n = \frac{1}{z} \left[ P^{(k)}(z) - P_{0,k} \right] = \frac{1}{z} P^{(k)}(z),$$

since $P_{0,k} = 0$ when $k \geq 1$. Substituting this result into (8), multiplying through by $u^k$ and summing over $k \geq 1$ gives us

$$\frac{1}{z} \sum_{k \geq 1} P^{(k)}(z) u^k = p \sum_{k \geq 1} P^{(k-1)}(z) u^k + q \sum_{k \geq 1} P^{(k+1)}(z) u^k$$

or equivalently

$$\frac{1}{z} \left[ P(z, u) - P(z, 0) \right] = puP(z, u) + \frac{q}{u} \left[ P(z, u) - P(z, 0) - uP^{(1)}(z) \right],$$

since $P(z, u) = \sum_{k \geq 0} P^{(k)}(z) u^k$ and $P(z, 0) = P^{(0)}(z)$. Next, from the recurrence

$$P_{n+1,0} = qP_{n,0} + qP_{n,1} \quad (n \geq 0)$$

we obtain

$$P^{(i)}(z) = \left( \frac{1}{qz} - 1 \right) P(z, 0) - \frac{1}{qz}.$$ 

Substituting this result into (9) leads us to (5).

The generating function for $P(z, 0)$ can be found by applying the kernel method [12]. The denominator of (5) can be factored as $(u - r_-(z))(u - r_+(z))$ where

$$r_-(z) = \frac{1 - \sqrt{1 - 4pqz^2}}{2pz} \quad \text{and} \quad r_+(z) = \frac{1 + \sqrt{1 - 4pqz^2}}{2pz}.$$ 

Since $P(z, u)$ has a power series representation about $(0, 0)$ while $1/(u - r_-(z))$ does not, $u - r_-(z)$ must also be a factor of the numerator of (5), implying

$$q(1 - r_-(z))zP(z, 0) - r_-(z) = 0.$$ 

Solving for $P(z, 0)$ gives us (6).
Before proceeding, let us consider the probability the algorithm returns to its initial configuration at time $n$, that is, $T_n$ is a single node tree. Since this quantity is $P_{n,0}$, we obtain the following estimate.

**Proposition 2** An asymptotic estimate for $P_{n,0}$ as a function of $p$ as $n \to \infty$ is

$$P_{n,0} = \begin{cases} \frac{q-p}{q} + O((2\sqrt{pq} + \epsilon)^n), & p < \frac{1}{2} \\ \sqrt{\frac{2}{\pi n}} + \sqrt{\frac{2}{\pi n^3}} + O(2^{-n/3/2}), & p = \frac{1}{2} \\ \sqrt{\frac{2}{\pi}} p \left( \frac{1}{(\sqrt{p} - \sqrt{q})^2} + \frac{(-1)^n}{(\sqrt{p} + \sqrt{q})^2} \right) (2\sqrt{pq})^{n-3/2} + O((2\sqrt{pq})^{n-5/2}), & p > \frac{1}{2}. \end{cases}$$

**Proof** For an asymptotic estimate of $P_{n,0}$, we begin by noting the singularities of $P(z, 0)$ occur at branch points $z = \pm 1/(2\sqrt{pq})$ and, if $p < 1/2$, also at a simple pole $z = 1$. The branch points are on the unit circle when $p = 1/2$. Otherwise $pq < 1/4$ holds, implying the branch points lie beyond it.

Consider the case $p < 1/2$. Since the branch points fall outside the unit circle, $P(z, 0)$ is meromorphic within a disk of radius $R$, where $1 < R < 1/(2\sqrt{pq})$. Hence we can expand $P(z, 0)$ about the simple pole $z = 1$ to obtain the Laurent series representation

$$P(z, 0) = \frac{q-p}{q} \left( \frac{1}{1-z} \right) + g(z),$$

where $g$ is some function analytic at $z = 1$ and therefore has radius of convergence $1/(2\sqrt{pq})$. From the expansion of meromorphic functions [4, Theorem IV.10, p.258] we can derive

$$P_{n,0} = \frac{q-p}{q} + O((2\sqrt{pq} + \epsilon)^n),$$

where $2\sqrt{pq} < 1$ and $\epsilon > 0$ is an arbitrarily small positive number.

Next consider the case $p > 1/2$. Here the radius of convergence is determined by branch points on opposite sides of the imaginary axis. The function $P(z, 0)$ is star-continuable [4, Theorem VI.5, p.398] and, in the vicinity of its singularities, we have

$$P(z, 0) = \begin{cases} \frac{2\sqrt{p}}{\sqrt{p} - \sqrt{q}} - \frac{2\sqrt{p}}{(\sqrt{p} - \sqrt{q})^2}(1 - 2\sqrt{pq})^{1/2} + O(1 - 2\sqrt{pq}), & z \to \frac{1}{2\sqrt{pq}} \\ \frac{2\sqrt{p}}{\sqrt{p} + \sqrt{q}} - \frac{2\sqrt{p}}{(\sqrt{p} + \sqrt{q})^2}(1 + 2\sqrt{pq})^{1/2} + O(1 + 2\sqrt{pq}), & z \to -\frac{1}{2\sqrt{pq}} \end{cases}$$

from which we obtain, by Big-Oh transfer [4, Theorem VI.3, p.390], the desired estimate. Interestingly, the main term of this estimate is $1/\sqrt{2pq}$ times larger for even $n$ than odd $n$, indicating the reflecting boundary at the trivial tree is not traversed frequently enough to break periodicity.
Finally for the case \( p = 1/2 \), we observe that \( P(z, 0) = (1/2)M(z/2) \) where \( M(z) \) is the generating function for the number of meanders of length \( n \), a scaling of the \( n \)th Catalan number (see OEIS A001405). Hence we deduce

\[
P_{n,0} = [z^n] \frac{1}{2} M(z/2) = \frac{1}{2^{n+1}} \left( \frac{n}{\binom{n+1}{2}} \right) = \sqrt{\frac{1}{2\pi n}} + \sqrt{\frac{2}{\pi n^3}} + O(2^{-n}n^{-3/2}).
\]

Returning to (5), if we let \( S_n \) denote the stratum number of a tree at time \( n \), then \( P_n(u) \equiv [z^n] P(z, u) \) is the probability generating function of \( S_n \). Thus we immediately have \( \mathbb{E}[S_n] = P'_n(1) \) and \( \text{Var}(S_n) = P''_n(1) + P'_n(1) - [P'_n(1)]^2 \). This idea leads to the generating functions for the first and second factorial moments of \( S_n \) and asymptotic estimates for \( \mathbb{E}[S_n] \) and \( \text{Var}(S_n) \).

**Proposition 3** Let \( S_n \) denote the stratum number of the tree generated by the \( n \)th iteration of the algorithm. The generating function \( \mu(z) \) for the sequence \( \{\mathbb{E}[S_n] : n \geq 0\} \) is

\[
\mu(z) = \frac{g z P(z,0)}{1-z} + \frac{(p-q)z}{(1-z)^2}. \tag{10}
\]

An asymptotic estimate of \( \mu_n \equiv [z^n] \mu(z) = \mathbb{E}[S_n] \) as a function of \( p \) as \( n \to \infty \) is given by

\[
\mu_n = \begin{cases} 
\frac{p}{q-p} + O\left((2\sqrt{pq} + \varepsilon)^n\right), & p < \frac{1}{2} \\
\sqrt{\frac{2n}{\pi}} - \frac{1}{2} + \frac{1}{2\sqrt{2\pi n}} + O(n^{-3/2}), & p = \frac{1}{2} \\
(p-q)n + \frac{q}{p-q} + O\left((2\sqrt{pq} + \varepsilon)^n\right), & p > \frac{1}{2}.
\end{cases}
\]

**Proof** Since \( \mu(z) \equiv \partial_u P(z, u)|_{u=1} \), the expression (10) can be obtained from (5) in a straightforward manner.

For the asymptotic analysis, we have from (10) the expression

\[
\mu_n = q[z^{n-1}] \frac{P(z,0)}{1-z} + (p-q)n,
\]

so it remains only to consider the term \( P(z,0)/(1-z) \). From Proposition 2, we deduce for cases \( p < 1/2 \) and \( p > 1/2 \) that the dominant singularity is an order 2 pole and a simple pole, respectively, at \( z = 1 \). Thus \( P(z,0)/(1-z) \) is meromorphic within a disk \( 1 < |z| < 1/(2\sqrt{pq}) \) and so we can compute asymptotic estimates accordingly.

For the case \( p = 1/2 \), the singularities are branch points at \( z = \pm 1 \). The function \( \mu(z) \) simplifies to
\[ \mu(z) = \frac{1}{2} \left[ \sqrt{\frac{1 + z}{(1 - z)^3}} - \frac{1}{1 - z} \right], \]

implying

\[ \mu_n = \frac{1}{2} [z^n] \left\{ \frac{(1 + z)^{1/2}}{(1 - z)^{3/2}} \right\} - \frac{1}{2}. \]  

(11)

In the neighborhood of the singularities, we have

\[ \frac{(1 + z)^{1/2}}{(1 - z)^{3/2}} = \begin{cases} \sqrt{2}(1 - z)^{-3/2} - \frac{1}{2\sqrt{2}}(1 - z)^{-1/2} + O(\sqrt{1 - z}), & z \to 1 \\ O(\sqrt{1 + z}), & z \to -1, \end{cases} \]

and therefore by Big-Oh transfer

\[ [z^n] \left\{ \frac{(1 + z)^{1/2}}{(1 - z)^{3/2}} \right\} = 2\sqrt{\frac{2n}{\pi}} + \frac{1}{2\sqrt{2\pi n}} + O(n^{-3/2}). \]

Substituting this result into (11) gives us

\[ \mu_n = \sqrt{\frac{2n}{\pi}} - \frac{1}{2} + \frac{1}{2\sqrt{2\pi n}} + O(n^{-3/2}). \]

Proposition 4 Let \( S_n \) denote the stratum number of the tree generated by the \( n \)th iteration of the algorithm. The generating function \( \mu^{(2)}(z) \) for the second factorial moment of \( S_n \), namely \( \mathbb{E}[S_n(S_n - 1)] \), is

\[ \mu^{(2)}(z) = \frac{2q(2pz - 1)zP(z, 0)}{(1 - z)^2} + \frac{2(4p - 3)pz^2}{(1 - z)^3} + \frac{2qz}{(1 - z)^3}. \]  

(12)

An asymptotic estimate of \( \mu^{(2)}(z) = \mathbb{E}[S_n(S_n - 1)] \) as a function of \( p \) as \( n \to \infty \) is given by

\[ \mu_n^{(2)} = \begin{cases} 2\left( \frac{p}{q-p} \right)^2 + O((2\sqrt{pq} + \epsilon)^n), & p < \frac{1}{2} \\ n - 2\sqrt{\frac{2n}{\pi}} + 1 + \frac{1}{\sqrt{2\pi n}} + O(n^{-3/2}), & p = \frac{1}{2} \\ (p-q)^2n^2 - (4p^2 - 3)n + \frac{2q(1-3p)}{(p-q)^2} + O((2\sqrt{pq} + \epsilon)^n), & p > \frac{1}{2}. \end{cases} \]

Proof Equation (12) follows directly from the relation \( \mu^{(2)}(z) = \partial_u^2 P(z, u) \big|_{u=1}. \)

The asymptotic analysis mirrors that of Proposition 3, that is, from (12) we deduce
\[ \mu_n^{(2)} = 4pq[z^{n-2}] \frac{P(z,0)}{(1-z)^2} - 2q[z^{n-1}] \frac{P(z,0)}{(1-z)^2} + (p-q)^2n^2 + (1+2p-4p^2)n, \]

and focus on the term \( P(z,0)(1-z)^{-2} \).

**Proposition 5** Let \( S_n \) denote the stratum number of the tree generated by the \( n \)th iteration of the algorithm. An asymptotic estimate of the variance of \( S_n \) as a function of \( p \) as \( n \to \infty \) is given by

\[
\mathbb{Var}(S_n) = \begin{cases} 
\frac{pq}{(p-q)^2} + O((2\sqrt{pq} + \epsilon)^n), & p < \frac{1}{2} \\
\left(1 - \frac{2}{\pi}\right)n - \frac{1}{4} + \sqrt{\frac{2}{\pi n}} + O(n^{-1}), & p = \frac{1}{2} \\
4pqn - \frac{3pq}{(p-q)^2} + O((2\sqrt{pq} + \epsilon)^n), & p > \frac{1}{2}.
\end{cases}
\]

**Proof** Noting \( \mathbb{Var}(S_n) = \mu_n^{(2)} + \mu_n - (\mu_n)^2 \), the result is an immediate consequence of Propositions 3 and 4.

Harmonic numbers commonly arise in the moments of functionals of recursive trees. For example, in Sect. 4.3, while studying the expected degree of the root node at time \( n \), we require the expectation of the \((S_n + 1)\)th harmonic number. To that end, we derive an asymptotic expression for that quantity. But first, we prove two useful lemmas.

**Lemma 1** The generating function \( H(z) \) for the expected value of \( H_{S_n} \), where \( H_0 \equiv 0 \) and \( H_k \) denotes the \( k \)th harmonic number \((k \geq 1)\), is

\[ H(z) = \frac{1}{1-z} \log \left( \frac{1 + \sqrt{1 - 4pqz^2}}{1 - 2pz + \sqrt{1 - 4pqz^2}} \right). \]

An asymptotic estimate of \([z^n] H(z) = \mathbb{E}[H_{S_n}] \) as a function of \( p \) as \( n \to \infty \) is given by

\[
\mathbb{E}[H_{S_n}] \sim \begin{cases} 
\log \left(\frac{q}{q-p}\right), & p < \frac{1}{2} \\
\log \sqrt{n}, & p = \frac{1}{2} \\
\log n, & p > \frac{1}{2}.
\end{cases}
\]

**Proof** We first note the expectation \( \mathbb{E}[H_{S_n}] \) can be written as

\[
\sum_{j \geq 1} \frac{1}{k} \mathbb{P}(S_n = j) = \sum_{k \geq 1} \frac{1}{k} \sum_{j \geq k} \mathbb{P}(S_n = j) = \sum_{k \geq 1} \frac{1}{k} \mathbb{P}(S_n \geq k).
\]

Thus if we set

\[
F(z, u) \equiv \sum_{n \geq 0} \sum_{k \geq 0} \mathbb{P}(S_n \geq k) u^{k} z^{n} = \frac{1}{(1-u)(1-z)} - \frac{uP(z, u)}{1-u},
\]

\( Springer \)
the desired generating function is

\[ H(z) = \int_0^1 \frac{1}{u} \left[ F(z, u) - \frac{1}{1 - z} \right] du = \int_0^1 \left[ \frac{1}{(1 - u)(1 - z)} - \frac{P(z, u)}{1 - u} \right] du. \]

To facilitate the evaluation of this integral, we write

\[ P(z, u) = \frac{A - (A + 1)u}{pz(u - B)(u - C)}, \]

where

\[ A = qzP(z, 0), \quad B = \frac{1 + \sqrt{1 - 4pqz^2}}{2pz} \quad \text{and} \quad C = \frac{1 - \sqrt{1 - 4pqz^2}}{2pz}. \]

The integral (with respect to \( u \)) now follows immediately after a partial fractions expansion of \( P(z, u)(1 - u)^{-1} \) and some simplification of the result.

To obtain an asymptotic estimate for \([z^n]H(z)\), we consider three cases. First, when \( p < 1/2 \), the generating function \( H(z) \) is meromorphic within a disk \( 1 < |z| < 1/(2\sqrt{pq}) \), where \( z = 1 \) is a simple pole. Within this disk, \( H(z) \) has the Laurent expansion

\[ H(z) = \log \left( \frac{q}{q - p} \right) \frac{1}{1 - z} + g(z), \]

with \( g(z) \) analytic in the disk \( |z| < 1/(2\sqrt{pq}) \). The meromorphic function expansion gives us the desired result.

When \( p > 1/2 \), the branch point at \( z = 1 \) dominates. Within a neighborhood of that singularity, we have

\[ H(z) \sim \frac{1}{1 - z} \log \left( \frac{1}{1 - z} \right) \quad \text{as} \quad z \to 1 \]

and so by sim-transfer [4, Corollary VI.I, p.392], we have \([z^n]H(z) \sim \log n\).

Finally, when \( p = 1/2 \), the generating function simplifies to

\[ H(z) = \frac{1}{1 - z} \log \left( \frac{1 + \sqrt{1 - z^2}}{1 - z - \sqrt{1 - z^2}} \right), \]

a star-continuable function with dominant singularities at \( z = \pm 1 \). Near these singularities we have

\[ H(z) \sim \begin{cases} \frac{1}{1-z} \log \left( \frac{1}{\sqrt{1-z}} \right), & z \to 1 \\ constant, & z \to -1, \end{cases} \]

and so by sim-transfer, \([z^n]H(z) \sim \log \sqrt{n}\). \( \square \)
Lemma 2  Let $Z_n$ denote the size of the tree generated by the $n$th iteration of the algorithm, i.e., $Z_n = S_n + 1$. The generating function $h(z)$ for the mean of the reciprocal of $Z_n$, that is, $\mathbb{E}[Z_n^{-1}]$ is

$$h(z) = \frac{1 + \sqrt{1 - 4pqz^2}}{pz[1 - 2qz + \sqrt{1 - 4pqz^2}]} \log \left( \frac{1 + \sqrt{1 - 4pqz^2}}{1 - 2pqz + \sqrt{1 - 4pqz^2}} \right).$$

An asymptotic estimate of $[z^n] h(z) = \mathbb{E}[Z_n^{-1}]$ as a function of $p$ as $n \to \infty$ is given by

$$\mathbb{E}[Z_n^{-1}] \sim \begin{cases} \frac{q-p}{p} \log \left( \frac{q}{q-p} \right), & p < \frac{1}{2} \\ \frac{1}{\sqrt{2\pi n}}, & p = \frac{1}{2} \\ \frac{1}{(p-q)n}, & p > \frac{1}{2}. \end{cases}$$

**Proof** To derive the generating function, define

$$h(z, u) \equiv \sum_{n \geq 0} \sum_{k \geq 0} \frac{1}{k+1} \mathbb{P}(S_n = k) u^k z^n,$$

so that $h(z, 1) = h(z)$. Then

$$\partial_u uh(z, u) = P(z, u),$$

giving us

$$h(z, u) = \frac{1}{u} \int_0^u P(z, s)ds,$$

implying $h(z) = \int_0^1 P(z, s)ds$. This integral can be evaluated using the same partial fraction expansion given in Lemma 1 for $P(z, u)$.

The asymptotic estimate derivation also parallels that of Lemma 1. Specifically, when $p < 1/2$, the function $h(z)$ is meromorphic in a disk $1 < |z| < 1/(2\sqrt{pq})$, where $z = 1$ is a simple pole. Within this disk, $h(z)$ has the Laurent expansion

$$h(z) = \frac{q-p}{p} \log \left( \frac{q}{q-p} \right) \frac{1}{1-z} + g(z).$$

When $p > 1/2$, the branch point at $z = 1$ again dominates. Near this singularity, we have

$$h(z) \sim \frac{1}{p-q} \log \left( \frac{1}{1-z} \right) \text{ as } z \to 1$$

and so by sim-transfer, we have

$$h(z) \sim \left( \frac{1}{p-q} \right) \frac{1}{n}. $$
Lastly, when \( p = 1/2 \), the generating function simplifies to

\[
h(z) = \frac{2}{z} \cdot \frac{1 + \sqrt{1 - z^2}}{1 - z + \sqrt{1 - z^2}} \log \left( \frac{1 + \sqrt{1 - z^2}}{1 - z + \sqrt{1 - z^2}} \right)
\]

a star-continuable function with dominant singularities at \( z = \pm 1 \). This yields

\[
h(z) \sim \begin{cases} 
\sqrt{\frac{2}{1-z}} \log \left( \frac{1}{\sqrt{2-2z}} \right), & z \to 1 \\
\text{constant}, & z \to -1,
\end{cases}
\]

with

\[
[z^n] h(z) \sim \frac{\log n}{\sqrt{2\pi n}}
\]

by sim-transfer.

\[\square\]

**Proposition 6** Let \( H_{Z_n} \) denote the \( Z_n \)th harmonic number, where \( Z_n = S_n + 1 \) is the tree size after the \( n \)th iteration of the algorithm. The generating function of \( \mathbb{E}[H_{Z_n}] \) is \( H(z) + h(z) \) as given in Lemmas 1 and 2. An asymptotic estimate of \( \mathbb{E}[H_{Z_n}] \) as a function of \( p \) as \( n \to \infty \) is

\[
\mathbb{E}[H_{Z_n}] \sim \begin{cases} 
quotient{q}{p} \log \left( \frac{q}{q-p} \right), & p < \frac{1}{2} \\
\log \sqrt{n} + \frac{\log n}{\sqrt{2\pi n}}, & p = \frac{1}{2} \\
\log n + \frac{1}{(p-q)n}, & p > \frac{1}{2}.
\end{cases}
\]

**Proof** The results follow immediately from the identity \( H_{Z_n} = H_{S_n} + \frac{1}{Z_n} \). \[\square\]

**4 Application of Results to Tree Functionals**

**4.1 Tree Size**

Given we use uniform attachment for the addition rule and any member of the class defined in Corollary 1 for deletion, we immediately have from Propositions 3 and 5 the asymptotics of the expected tree size and corresponding variance of tree \( T_n \), namely \( \mathbb{E}[Z_n] = \mathbb{E}[S_n] + 1 \) and \( \text{Var}(Z_n) = \text{Var}(S_n) \). We note there are three behavioral regimes determined by whether insertion probability \( p \) is less than, equal to, or exceeds 1/2.
4.2 Leaf Count

By conditioning on stratum number, we can obtain similar results for other tree functionals. To see this, suppose the distribution of $T_n$, conditioned on $\{S_n = k\}$, is equiprobable. Then we have the equality

$$\mathbb{P}_q(T_n = t \mid S_n = k) = \frac{[t \in T_k]}{k!} = \mathbb{P}_0(T_k = t),$$

where $\mathbb{P}_x$ is the probability of an event given a deletion probability of $x$ and $[\cdot]$ is Iverson bracket notation for an indicator function. The conditional expectation of $f(T_n)$ is thus given by

$$\mathbb{E}_q[f(T_n) \mid S_n = k] = \sum_{t \in T} f(t) \mathbb{P}_q(T_n = t \mid S_n = k)$$

$$= \sum_{t \in T_k} f(t) \mathbb{P}_q(T_n = t \mid S_n = k) = \sum_{t \in T_k} f(t) \mathbb{P}_0(T_k = t) = \mathbb{E}_0[f(T_k)],$$

where $\mathbb{E}_x$ is the expectation of an event given a deletion probability of $x$.

Since the expectation $\mathbb{E}_0[f(T_k)]$ depends only on $k$, we can “ignore” the effect of deletion probability $q$ on the probabilistic behavior of the tree functional and, by iterating expectation, exploit the useful result

$$\mathbb{E}_q[f(T_n)] = \mathbb{E}_q[\mathbb{E}_0[f(T_{S_n})]]. \quad (13)$$

Interestingly, this results holds regardless of the specific deletion rule from the class chosen. Whether it is LIFO or something more intricate, the moments are the same.

If we let $L_n$ denote the number of leaves in the tree at time $n$, we have the well-known result [6, pp. 326–327]

$$\mathbb{E}_0[L_n] = \frac{n+1}{2} \llbracket n>0 \rrbracket + \llbracket n=0 \rrbracket.$$  

By using (13) we can deduce

$$\mathbb{E}_q[L_n] = \frac{1 + \mathbb{E}_q[S_n] + \mathbb{P}_q(S_n = 0)}{2}.$$  

Applying Propositions 2 and 3 yields the generating function and asymptotic estimate.

4.3 Root Degree

Let $D_n$ denote the degree of the root node at time $n$. Since

$$\mathbb{E}_0[D_n] = H_{n+1},$$

where $H_n$ is the $n$th harmonic number [6, pp 323-324], we find
\[ \mathbb{E}_q [D_n] = \mathbb{E}_q [H_{S_n+1}] = \mathbb{E}_q [H_{Z_n}], \]
and obtain the corresponding generating function and asymptotics from Proposition 6.

5 Summary

By allowing for the possibility of node removal during the course of their evolution, we can extend the utility of random recursive trees models. The analysis of such trees; however, is complicated by the fact that the tree size at time \( n \) is no longer deterministic. Nevertheless, for the class of deletion rules identified by Corollary 1, we showed that a random tree \( T_n \), conditioned on its size, is uniformly distributed over its range. This reduces the problem of studying \( T_n \) to that of studying its stratum number. By using generating function theory, we obtain several results for the expected tree size, leaf count and root degree of tree \( T_n \).

Declarations

Conflicts of interest  The author declares that he has no conflict of interest.

References

1. Ben-Naim, E., Krapivsky P.L.: Addition-deletion networks. The Journal of Physics A: Mathematical and Theoretical 40, 8607–8619 (2007)
2. Deo, N., Cami, A.: Preferential deletion in dynamic models of web-like networks. Information Processing Letters 102, 156–162 (2007)
3. Dondajewski, M., Szymanski, J.: On the distribution of vertex-degrees in a strata of a random recursive tree. Bulletin de l’Académie Polonaise des Sciences: Série des Sciences Mathématiques 30, 205–209 (1982)
4. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press, Cambridge (2009)
5. Ghosal, G., Chi, L., Barabási, A.L.: Uncovering the role of elementary processes in networking evolution. Scientific Reports 3, 2920 (2013)
6. Hofri, M., Mahmoud, H.: Algorithms of Nonuniformity: Tools and Paradigms. CRC Press, Boca Raton (2018)
7. Janson, S.: Asymptotic degree distribution in random recursive trees. Random Structures and Algorithms 26, 69–83 (2005)
8. Jonassen, A.R., Knuth, D.E.: A trivial algorithm whose analysis isn’t. Journal of Computer and System Sciences 16, 301–322 (1978)
9. Najock, D., Heyde, C.C.: On the number of terminal vertices in certain random trees with an application to stemma construction in philology. Journal of Applied Probability 19, 675–680 (1982)
10. Panny, W.: Deletions in random binary search trees: a story of errors. Journal of Statistical Planning and Inference 140, 2335–2345 (2010)
11. Pittel, B.: Note on the height of random recursive trees and m-ary search trees. Random Structures and Algorithms 5, 337–347 (1994)
12. Prodinger, H.: The kernel method: a collection of examples. Séminarie Lotharingien de Combinatoire 50, Article B50f (2003)
13. Smythe R.T., Mahmoud, H.M.: A survey of recursive trees. Theory of probability and mathematical
statistics 51, 1–27 (1995)
14. Zhang, X., He, Z., Rayman-Bacchus, L.: Random birth-and-death networks. Journal of Statistical
Physics 162, 842–854 (2016)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published
maps and institutional affiliations.