

## ON THE VARIETY OF SHAPES ON THE FRINGE OF A RANDOM RECURSIVE TREE

QUNQIANG FENG,\* *University of Science and Technology of China*  
HOSAM M. MAHMOUD,\*\* *The George Washington University*

### Abstract

We consider a variety of subtrees of various shapes lying on the fringe of a recursive tree. We prove that (under suitable normalization) the number of isomorphic images of a given fixed tree shape on the fringe of the recursive tree is asymptotically Gaussian. The parameters of the asymptotic normal distribution involve the shape functional of the given tree. The proof uses the contraction method.

*Keywords:* Random tree; recurrence; contraction method

2000 Mathematics Subject Classification: Primary 05C05; 60C05  
Secondary 60F05

### 1. Introduction

The recognition of patterns in random objects is a fundamental operation in information processing. For example, a linguist may be interested in the frequency of five-letter words in Abraham Lincoln's famous speech:

Four score and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal.

Analysis of pattern matching in text can be found in Mahmoud *et al.* (1997) and Régnier and Szpankowski (1998). The equivalent problem in random trees to the pattern counting problem in texts is to find trees of a certain *size* in a given tree generated randomly. Flajolet *et al.* (1997) dealt with this problem in binary search trees for trees of fixed size on the fringe, and Feng *et al.* (2008) considered it in recursive trees and binary search trees (where there is no restriction on the size of the pattern), so the full spectrum is discussed.

Alternatively, the linguist may be interested in how often the syllable *tion* occurs in the text. The equivalent problem in random trees is to find trees of a certain *shape* in a random tree. This area has already attracted attention in the recent literature on random trees (see Chyzak *et al.* (2008)). In the present paper we investigate this area for one important random tree model. We look at the number of subtrees of a certain shape on the fringe of random recursive trees.

The rest of the paper is organized as follows. In Section 2 we define recursive trees and give examples of the varieties considered on the fringe; we also state the main theorem. In Section 3 we introduce a shape functional (a numerically quantifiable measure) of trees. In Section 4 we discuss the basic stochastic recurrence for the count. In Section 5 we show how in general the moments can be computed, and carry out the program for the mean (Subsection 5.1) and the

---

Received 19 September 2008; revision received 12 November 2009.

\* Postal address: Department of Statistics and Finance, University of Science and Technology of China, Hefei 230026, China. Email address: fengqq@ustc.edu.cn

\*\* Postal address: Department of Statistics, The George Washington University, Washington, DC 20052, USA. Email address: hosam@gwu.edu

variance (Subsection 5.2), and give their exact computation. The task becomes considerably harder for higher moments. Finally, in Section 6 we give a characterization of all the asymptotic moments using a shortcut offered by the contraction method, where a Gaussian result is shown to hold.

### 2. Random recursive trees

The random recursive tree is a naturally growing structure that underlies many stochastic phenomena, such as contagion, and algorithms, such as Burge’s sorting and the union find. It serves as a model for the formation of social networks, and the barest form of preferential attachment schemes, such as the Internet. For numerous applications of recursive trees, we refer the reader to the survey in Smythe and Mahmoud (1995). These are rooted labeled trees in the graph-theoretic sense. The children of the root are unordered without any special plane orientation, unlike, for example, the binary search tree model, which received quite a bit of attention (see Mahmoud (1992, Chapter 2)).

The *random recursive tree* grows out of a single node labeled 1, and nodes are added in stages. At the  $n$ th stage a node in the existing tree is chosen at random as a parent for the  $n$ th entrant (labeled  $n$ ). Here *random* means that all nodes in the tree of size  $n - 1$  are equally likely parents. Per the construction algorithm, the nodes along any root-to-leaf path carry increasing labels; hence, these trees were also dubbed ‘increasing trees’. Figure 1 shows all the recursive trees of size 4.

The model of randomness in the growth of random recursive trees induces a uniform distribution on the trees. All  $(n - 1)!$  recursive trees of size  $n$  are generated with equal probability. Many important properties of recursive trees have been analyzed from the convenient viewpoint of recurrence occurring naturally in the stochastic growth. Nonetheless, the uniform distribution of the trees also gave rise to analyses based on generating functions and analytic methods such as integral transforms (see Bergeron *et al.* (1992)).

We discuss here the scope of the main result. It is concerned with rooted subtrees. Any connected subgraph of a recursive tree is a *rooted subtree*, and its root is the vertex with the smallest label. Given a rooted tree  $\Gamma$  with size  $|\Gamma| = \gamma$ , let  $R_n := R(n, \Gamma)$  denote the number of rooted subtrees (simply referred to as subtrees henceforth) on the fringe of a random recursive tree of order  $n$  that are isomorphic to  $\Gamma$ . A subtree of size  $|\Gamma|$  is ‘on the fringe’ if it has no proper subtree that is also of size  $|\Gamma|$ . For instance,  $\Gamma$  may be the rooted tree shown in Figure 2 (where the arrow points to the root), with the recursive tree being that shown in Figure 3. In this recursive tree there are three occurrences on the fringe of the tree with shape isomorphic to  $\Gamma$  (the subtrees with the labels (2, 4, 6, 9), (10, 11, 14, 17), and (12, 13, 15, 16)). Thus, in this example  $R(17, \Gamma) = 3$ ; the three isomorphic images of  $\Gamma$  lying on the fringe of the recursive tree of Figure 3 are pointed to by arrows. We draw the reader’s attention to the fact that the

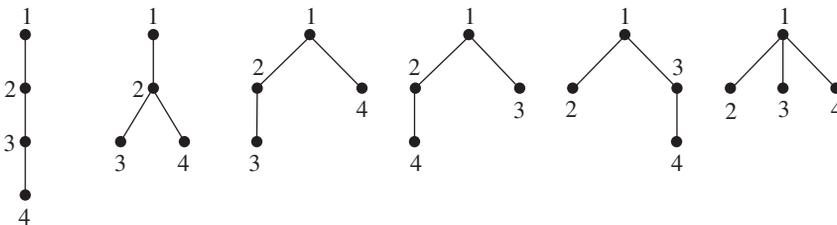


FIGURE 1: Recursive trees of size 4.

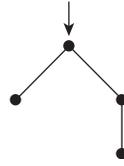


FIGURE 2.

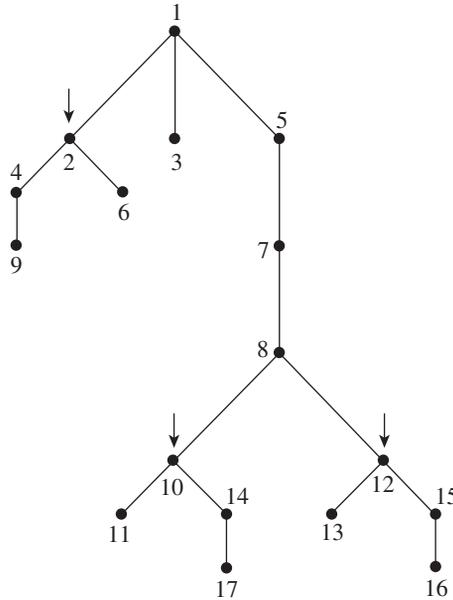


FIGURE 3: A recursive tree with three isomorphic copies (on its fringe) of a given pattern.

rooted tree  $\Gamma$  in Figure 2 is one of two possible drawings (the other one coincides with the third tree from the left in Figure 1). Thus, as a rooted tree shape,  $\Gamma$  is isomorphic to the third, fourth, and fifth recursive trees in Figure 1.

We will demonstrate the asymptotic normality of  $R_n$ . We will work out the exact mean and variance of  $R_n$ , which include a constant called the shape functional. The main result is as follows.

**Theorem 1.** *Let  $R_n$  be the number of occurrences of a rooted fixed-shape tree  $\Gamma$  on the fringe of a random recursive tree of size  $n$ . Then,*

$$\frac{R_n - \mu_\Gamma n}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, \sigma_\Gamma^2),$$

where  $\mu_\Gamma$  and  $\sigma_\Gamma^2$  are constants (given explicitly below in Propositions 1 and 2) that depend only on  $\Gamma$ , and ‘ $\xrightarrow{D}$ ’ denotes convergence in distribution.

### 3. The shape functional

To study the random variable  $R_n$ , we need to ‘quantify’ the shape of a tree, that is, describe it by one numerical value that captures the relevant essence of the structure. For a certain rooted

tree  $\Gamma$ , let  $V(\Gamma)$  denote the set of vertices and let  $|\Gamma| := |V(\Gamma)|$  denote the total number of vertices in it. For simplicity, we will use  $\gamma = |\Gamma|$  throughout. For any vertex  $v \in V(\Gamma)$ , let  $S(v)$  be the set of immediate children of  $v$  and let  $s(v)$  be the size of  $S(v)$ , that is,  $s(v)$  is the outdegree of  $v$ . In particular, if  $s(v) = 0$  then  $v$  is a leaf. Let  $\tau(v)$  be the subtree with root  $v \in V(\Gamma)$ . If  $v_1, \dots, v_{s(v)}$  are the immediate children of  $v$ , some of the graphs  $\tau(v_1), \dots, \tau(v_{s(v)})$  may be isomorphic images of  $\Gamma$ . We define  $d(v)$  to be the number of different subtrees rooted at  $v$ . After renumbering the vertices  $v_1, \dots, v_{s(v)}$  in an appropriate way, there are numbers  $b_1(v), \dots, b_{d(v)}(v)$  such that

$$\begin{aligned} \tau(v_1) &\stackrel{\text{ISO}}{=} \dots \stackrel{\text{ISO}}{=} \tau(v_{b_1(v)}), \\ \tau(v_{b_1(v)+1}) &\stackrel{\text{ISO}}{=} \dots \stackrel{\text{ISO}}{=} \tau(v_{b_1(v)+b_2(v)}), \\ &\vdots \\ \tau(v_{1+\sum_{i=1}^{d(v)-1} b_i(v)}) &\stackrel{\text{ISO}}{=} \dots \stackrel{\text{ISO}}{=} \tau(v_{s(v)}), \end{aligned}$$

where  $\stackrel{\text{ISO}}{=}$  denotes isomorphism. Of course, the definition of  $b_i(v)$  is not unique, since the different groups of subtree can be rearranged. However, the function

$$\psi : V(\Gamma) \rightarrow [0, 1]; \quad \psi(v) = \begin{cases} \left( \prod_{i=1}^{d(v)} b_i(v)! \prod_{\tilde{v} \in S(v)} |\tau(\tilde{v})| \right)^{-1} & \text{if } S(v) \neq \phi, \\ 1 & \text{otherwise,} \end{cases}$$

is well defined.

Let

$$C(\Gamma) := \prod_{v \in V(\Gamma)} \psi(v).$$

This shape functional is similar to those that appear in positional trees, such as binary and  $m$ -ary (see Dobrow and Fill (1996), Fill (1996), and Fill and Kapur (2005)), and in digital trees (see Knuth (1998, Chapter 6)). The essential difference is the factorial terms, which appear here because recursive trees basically allow a new entrant to appear in any number (of possibly isomorphic) subtrees, whereas only a single position is prescribed for a key in binary and  $m$ -ary search trees, depending on the key's order statistic.

**Lemma 1.** *Let  $\Gamma$  be a fixed rooted tree. A randomly grown recursive tree of size  $|\Gamma|$  is isomorphic to  $\Gamma$  with probability  $C(\Gamma)$ .*

*Proof.* We use induction on  $\gamma = |\Gamma|$ . For  $\gamma = 1$ , we have  $C(\Gamma) = 1$ . Assume now that the lemma holds for  $\gamma' \leq \gamma - 1$ , where  $\gamma \geq 2$  is an integer. Note that, for a tree  $\Gamma'$  of size  $\gamma'$ , the number  $(\gamma' - 1)! P(\Gamma')$  counts the recursive trees that are isomorphic to  $\Gamma'$ . Now suppose that a fixed tree  $\Gamma$  of size  $\gamma \geq 1$  is given. It suffices to compute  $T(\Gamma)$ , the number of recursive trees of size  $\gamma$  that are isomorphic to  $\Gamma$ . To have a recursive tree of size  $\gamma$  being isomorphic to  $\Gamma$ , its root must have the same number of children as  $\Gamma$ 's root. Suppose that vertex  $r$  is the root of  $\Gamma$  and that  $r_1, \dots, r_{s(r)}$  are the immediate children of  $r$ . We call a subtree rooted at any of the children of the root of a tree a *branch* of that tree. For any vertex  $v \in S(r)$ ,  $|\tau(v)| < \gamma$ , and the induction hypothesis applies to each branch.

While  $\Gamma$  is a fixed unlabeled tree, numerous recursive trees of the same size can be isomorphic to it, by permuting the labels in the branches. We thus need to choose  $|\tau(r_1)|$  labels for the

branch rooted at  $r_1$ ,  $|\tau(r_2)|$  labels for the branch rooted at  $r_2$ , ...,  $|\tau(r_{s(r)})|$  labels for the subtree rooted at  $r_{s(r)}$ , and this can be done in

$$\binom{\gamma - 1}{|\tau(r_1)|, \dots, |\tau(r_{s(r)})|}$$

ways. Subtrees in the  $i$ th group can further be permuted in  $b_i(r)!$  ways, and still the recursive tree obtained is isomorphic to  $\Gamma$ . Finally, we impose the isomorphism on the branches (i.e. inductively count the number of ways that each branch can be configured to be isomorphic to a corresponding branch of  $\Gamma$ ). We proceed with

$$\begin{aligned} T(\Gamma) &= \frac{(\gamma - 1)!}{\prod_{i=1}^{s(r)} |\tau(r_i)|!} \frac{1}{b_1(r)! \cdots b_{d(r)}(r)!} \prod_{j=1}^{s(r)} (|\tau(r_j)| - 1)! P(\tau(r_j)) \\ &= \frac{(\gamma - 1)!}{\prod_{i=1}^{d(r)} b_i(r)! \prod_{i=1}^{s(r)} |\tau(r_i)|} \prod_{j=1}^{s(r)} P(\tau(r_j)) \\ &= (\gamma - 1)! \psi(r) \prod_{v \in V(\Gamma) \setminus \{r\}} \psi(v) \quad (\text{by the induction hypothesis}) \\ &= (\gamma - 1)! \prod_{v \in V(\Gamma)} \psi(v) \\ &= (\gamma - 1)! C(\Gamma). \end{aligned}$$

#### 4. Counting the variety of shapes

To formulate a recurrence for  $R_n$ , for  $\gamma < n$ , we employ a decomposition. It is shown in van der Hofstad *et al.* (2002) that the distribution of the size of the branch rooted at 2 is uniform on the set  $\{1, \dots, n - 1\}$ . Let  $U_n$  be such a discrete random variable. If the branch rooted at 2 is removed from the tree, what is left is isomorphic to a random recursive tree on its respective size  $n - U_n$ .

Consider the branch rooted at 2 to be ‘special’, and consider the rest of the tree to be a nonspecial subtree. A node rooting a subtree of size  $\gamma$  on the fringe may fall in the special branch or outside. The totality  $R_n$  is synthesized from the number of size  $\gamma$  subtree roots in the special branch and of shape  $\Gamma$  on the fringe (that is,  $R_{U_n}$ ) plus the number of size  $\gamma$  subtree roots on the fringe of the nonspecial subtree and of shape  $\Gamma$ , and, thus, has a number of subtrees of shape  $\Gamma$  distributed like  $R_{n-U_n}$ , and conditionally independent of their number in the special branch. This decomposition has one exception, which is the case  $U_n = n - \gamma$ . When we assemble in this case the special and nonspecial trees into one whole recursive tree (which has size  $n > \gamma$ ), the root 1 is no longer a root of a subtree of size  $\gamma$ , and we must subtract 1 from the combined count, if the nonspecial subtree is isomorphic to  $\Gamma$ .

Let  $\text{Ber}(p)$  be a Bernoulli random variable, with success rate  $p$ , and let  $\mathbf{1}_{\mathcal{E}}$  be an indicator of the event  $\mathcal{E}$  that assumes the value 1 if  $\mathcal{E}$  occurs, and assumes the value 0 otherwise. We have the distributional equality (valid for  $n > \gamma$ )

$$R_n \stackrel{D}{=} R_{U_n} + \tilde{R}_{n-U_n} - \mathbf{1}_{\{n-U_n=\gamma\}} \text{Ber}(C(\Gamma)), \tag{1}$$

where  $\tilde{R}_{n-U_n} \stackrel{D}{=} R_{n-U_n}$ , and  $R_{U_n}$  and  $\tilde{R}_{n-U_n}$  are conditionally independent. Conditional independence in this context means that while  $R_{U_n}$  and  $\tilde{R}_{n-U_n}$  are dependent (via their joint

dependence on  $U_n$ ), they are conditionally independent given the value of  $U_n$ , that is,  $R_i$  and  $\tilde{R}_j$  are independent for all  $i, j \geq 0$ . The two variables  $\mathbf{1}_{\{n-U_n=\gamma\}}$  and  $\text{Ber}(C(\Gamma))$  are independent.

### 5. Exact moments

The relation (1) is useful for the direct computation of moments, although the task becomes considerably more demanding for each higher moment. We can use it for the first few moments. We illustrate the procedure on the mean and variance. For third and higher moments, the procedure becomes too cumbersome, and it is prudent to seek transparent methods that shortcut these calculations.

#### 5.1. The mean

We have the following result for the mean.

**Proposition 1.** *Let  $R_n$  be the number of subtrees on the fringe of a random recursive tree of size  $n$  that are isomorphic to a given rooted tree  $\Gamma$  (of size  $\gamma$ ). For  $n > \gamma$ ,*

$$E[R_n] = \frac{C(\Gamma)}{\gamma(\gamma + 1)}n.$$

*Proof.* Let us use the symmetry  $U_n \stackrel{D}{=} n - U_n$ . Condition the distributional relation (1) on  $U_n = j$ , and write

$$\begin{aligned} E[R_n] &= E[R_{U_n}] + E[\tilde{R}_{n-U_n}] - E[\mathbf{1}_{\{n-U_n=\gamma\}}] E[\text{Ber}(C(\Gamma))] \\ &= \frac{2}{n-1} \sum_{j=1}^{n-1} E[R_j] - P(U_n = n - \gamma)C(\Gamma) \\ &= \frac{2}{n-1} \sum_{j=1}^{n-1} E[R_j] - \frac{C(\Gamma)}{n-1}. \end{aligned}$$

Take the difference between a version of the recurrence for  $(n - 2) E[R_{n-1}]$  from that for  $(n - 1) E[R_n]$ . Observe that, because recurrence (1) is valid for  $n > \gamma$ , we must also take  $n - 1 > \gamma$  to obtain

$$E[R_n] = \frac{n}{n-1} E[R_{n-1}],$$

valid for  $n > \gamma + 1$ . The latter recurrence then unwinds as

$$E[R_n] = \frac{n}{n-1} \frac{n-1}{n-2} E[R_{n-2}] = \dots = \frac{n}{\gamma+1} E[R_{\gamma+1}].$$

Unless all the nodes labeled higher than 1 in a recursive tree of size  $\gamma + 1$  appear in the special branch, and that branch takes the form of  $\Gamma$ , the recursive tree cannot have any subtrees isomorphic to  $\Gamma$  on its fringe. That is,  $R_{\gamma+1}$  is a Bernoulli random variable that is the product  $\mathbf{1}_{\{U_n=\gamma\}} \text{Ber}(C(\Gamma))$ . The average value of this product (of independent random variables) is  $C(\Gamma)/\gamma$ , in view of the uniform decomposition.

**5.2. The variance**

For the second moment calculation, square both sides of (1) and take the expectation. Employing several symmetries we obtain

$$\begin{aligned}
 E[R_n^2] &= 2 E[R_{U_n}^2] + P(U_n = n - \gamma)C(\Gamma) + 2 E[R_{U_n} \tilde{R}_{n-U_n}] \\
 &\quad - 2 E[R_{U_n} \mathbf{1}_{\{n-U_n=\gamma\}} \text{Ber}(C(\Gamma))] \\
 &\quad - 2 E[\tilde{R}_{n-U_n} \mathbf{1}_{\{n-U_n=\gamma\}} \text{Ber}(C(\Gamma))] \\
 &= \frac{2}{n-1} \sum_{j=1}^{n-1} E[R_j^2] + \frac{C(\Gamma)}{n-1} + \frac{2}{n-1} \sum_{j=1}^{n-1} E[R_j \tilde{R}_{n-j}] \\
 &\quad - \frac{2}{n-1} \sum_{j=1}^{n-1} E[R_j \mathbf{1}_{\{n-j=\gamma\}} \text{Ber}(C(\Gamma))] \\
 &\quad - \frac{2}{n-1} \sum_{j=1}^{n-1} E[\tilde{R}_{n-j} \mathbf{1}_{\{n-j=\gamma\}} \text{Ber}(C(\Gamma))] \\
 &= \frac{2}{n-1} \sum_{j=1}^{n-1} E[R_j^2] + \frac{C(\Gamma)}{n-1} + \frac{2}{n-1} \sum_{j=1}^{n-1} E[R_j] E[R_{n-j}] \\
 &\quad - \frac{2C(\Gamma)}{n-1} E[R_{n-\gamma}] - \frac{2C(\Gamma)}{n-1} E[R_\gamma],
 \end{aligned}$$

where the rearrangement in the last sum follows from the conditional independence of  $R_{U_n}$  and  $\tilde{R}_{n-U_n}$ .

We can use the boundary conditions  $R_j \equiv 0$  for  $j < \gamma$ ,  $R_\gamma = \text{Ber}(C(\Gamma))$ , and Proposition 1 for  $E[R_{n-\gamma}]$ . To apply this proposition, we must take  $n - \gamma > \gamma$ , i.e.  $n > 2\gamma$ , a condition that we assume holds in the sequel. The recurrence now has the form

$$\begin{aligned}
 E[R_n^2] &= \frac{2}{n-1} \sum_{j=\gamma}^{n-1} E[R_j^2] + \frac{C(\Gamma)}{n-1} - \frac{2C(\Gamma)}{n-1} (E[R_{n-\gamma}] + E[R_\gamma]) \\
 &\quad + \left( \frac{4}{n-1} E[R_\gamma] E[R_{n-\gamma}] + \frac{2C^2(\Gamma)}{n-1} \sum_{j=\gamma+1}^{n-\gamma-1} \frac{j}{\gamma(\gamma+1)} \frac{n-j}{\gamma(\gamma+1)} \right) \\
 &= \frac{2}{n-1} \sum_{j=\gamma}^{n-1} E[R_j^2] + \frac{C(\Gamma)}{n-1} - \frac{2C^2(\Gamma)(3\gamma^2 + 4\gamma - 1)}{3\gamma(\gamma+1)(n-1)} \\
 &\quad + \frac{C^2(\Gamma)(n^3 - n)}{3\gamma^2(\gamma+1)^2(n-1)} \\
 &:= \frac{2}{n-1} \sum_{j=\gamma}^{n-1} E[R_j^2] + \frac{h(n)}{n-1}.
 \end{aligned}$$

Taking the difference between a version of the recurrence for  $(n - 2) E[R_{n-1}^2]$  from that for  $(n - 1) E[R_n^2]$ , we simplify the recurrence to

$$E[R_n^2] = \frac{n}{n-1} E[R_{n-1}^2] + \frac{h(n) - h(n-1)}{n-1}.$$

Recurrence equations of the general form

$$v_n = \frac{n}{n-1}v_{n-1} + w_n$$

with a given  $v_{2\gamma+1}$  have a transparent solution:

$$v_n = \frac{v_{2\gamma+1}}{2\gamma+1}n + n \sum_{j=2\gamma+2}^n \frac{w_j}{j}.$$

Subsequently,

$$E[R_n^2] = \frac{C^2(\Gamma)n^2}{\gamma^2(\gamma+1)^2} + \sigma_\Gamma^2 n,$$

where

$$\sigma_\Gamma^2 := \frac{E[R_{2\gamma+1}^2]}{2\gamma+1} - \frac{(2\gamma+1)C^2(\Gamma)}{\gamma^2(\gamma+1)^2}.$$

**Lemma 2.** *The random variable  $R_{2\gamma+1}$  has the distribution*

$$R_{2\gamma+1} = \begin{cases} 0 & \text{with probability } 1 + \frac{(\gamma+1)C^2(\Gamma) - 2\gamma(2\gamma+1)C(\Gamma)}{2\gamma^2(\gamma+1)}, \\ 1 & \text{with probability } \frac{\gamma(2\gamma+1)C(\Gamma) - (\gamma+1)C^2(\Gamma)}{\gamma^2(\gamma+1)}, \\ 2 & \text{with probability } \frac{C^2(\Gamma)}{2\gamma^2}. \end{cases}$$

*Proof.* The random variable  $R_{2\gamma+1}$  cannot exceed 2, for if it were 3 or more, there would be at least three subtrees of size  $\gamma$  each, and none of them rooted at 1 (which roots a tree of size  $2\gamma+1$ ), and the number of nodes in the recursive tree would be at least  $3\gamma+1 > 2\gamma+1$ . The range of  $R_{2\gamma+1}$  is thus  $\{0, 1, 2\}$ .

The easiest of the three probabilities is  $P(R_{2\gamma+1} = 2)$ . To have the event  $\{R_{2\gamma+1} = 2\}$ , we must have an even split. The root node 1 with two branches, each with the shape of  $\Gamma$  (hence, each with size  $\gamma$ ). The special branch will have size  $\gamma$ , occurring with probability  $1/(2\gamma)$  (according to the decomposition property), and when that size is realized, it will have the shape of  $\Gamma$  with probability  $C(\Gamma)$ .

The nonspecial subtree, of size  $\gamma+1$ , should consist of the root 1, and a single child of it which fathers all the other descendants in the nonspecial tree, which should then assume the shape of  $\Gamma$ . Again, by the decomposition property, the subtree of the root of the nonspecial subtree must be isomorphic to a random recursive tree of size  $\gamma$ , and, for that child to father a tree of shape  $\Gamma$ , the event has probability  $C(\Gamma)/\gamma$ . The probability

$$P(R_{2\gamma+1} = 2) = \frac{C^2(\Gamma)}{2\gamma^2}$$

follows.

Equating the mean (given in Proposition 1) to its defining computation, we obtain

$$E[R_{2\gamma+1}] = \frac{(2\gamma+1)C(\Gamma)}{\gamma(\gamma+1)} = \sum_{j=0}^2 j P(R_{2\gamma+1} = j) = P(R_{2\gamma+1} = 1) + 2 \frac{C^2(\Gamma)}{2\gamma^2}.$$

The complete distribution follows.

We can now proceed with the computation

$$E[R_{2\gamma+1}^2] = \sum_{j=0}^2 j^2 P(R_{2\gamma+1} = j) = \frac{\gamma(2\gamma + 1)C(\Gamma) + (\gamma + 1)C^2(\Gamma)}{\gamma^2(\gamma + 1)},$$

needed as a boundary condition in the solution of the recurrence for  $E[R_n^2]$ . We have thus proved the following result.

**Proposition 2.** *For a given rooted tree  $\Gamma$  with size  $\gamma$ , let  $R_n$  be the number of subtrees isomorphic to  $\Gamma$  on the fringe of a random recursive tree of size  $n > 2\gamma$ . Then*

$$\text{var}[R_n] = \frac{(\gamma + 1)(2\gamma + 1) - (3\gamma + 2)C(\Gamma)}{\gamma(\gamma + 1)^2(2\gamma + 1)}C(\Gamma)n.$$

### 6. Limit distributions for varieties of a fixed size

In principle, we can continue pumping higher moments by the recurrence methods utilized for the mean and variance, and attempt to determine limit distributions by a method of recursive moments (see, for example, Chern *et al.* (2002)). However, as already mentioned, the mounting complexity is forbidding. We choose a method that is transparent for limits.

The contraction method was introduced in Rösler (1991) to analyze the quick sort algorithm, and it soon became a popular method because of the transparency of structure that it provides in the limit for processes that have distributional recurrences that may be complicated. Rachev and Rüschemdorf (1995) added several useful extensions. Recently, general contraction theorems and multivariate extensions were added in Neininger (2001), Neininger and Rüschemdorf (2004), and Rösler (2001). Neininger (2002) gave an application of the contraction method to recursive trees. Rösler and Rüschemdorf (2001) provided a valuable survey.

We highlight the rationale behind the contraction method (tailored to our context) in a few paragraphs. An equation like (1) can be normalized in the following manner. Let

$$R_n^* = \frac{1}{\sqrt{n}} \left( R_n - \frac{C(\Gamma)}{\gamma(\gamma + 1)}n. \right)$$

Then,

$$R_n^* = R_{U_n}^* \sqrt{\frac{U_n}{n}} + \tilde{R}_{n-U_n}^* \sqrt{\frac{n-U_n}{n}} - \frac{\mathbf{1}_{\{n-U_n=\gamma\}} \text{Ber}(C(\Gamma))}{\sqrt{n}}.$$

Now, if  $R_n^*$  converges to a limit  $R^*$  then so will  $R_{U_n}^*$  because  $U_n \rightarrow \infty$  almost surely. A symmetric argument applies to the tided random variable in (1), and

$$\sqrt{\frac{U_n}{n}} \xrightarrow{D} \sqrt{U}, \quad \sqrt{\frac{n-U_n}{n}} \xrightarrow{D} \sqrt{1-U}, \quad \frac{\mathbf{1}_{\{n-U_n=\gamma\}} \text{Ber}(C(\Gamma))}{\sqrt{n}} \rightarrow 0.$$

It is plausible to guess that  $R^*$  satisfies the distributional limit equation

$$R^* \stackrel{D}{=} \sqrt{U}R^* + \sqrt{1-U}\tilde{R}^*, \tag{2}$$

where  $\tilde{R}^* \stackrel{D}{=} R^*$  and is independent of it.

The contraction method provides a rigorous proof for this, by viewing the equality in distribution at  $n$  as a contraction in a metric space, the points of which are distribution functions

and distances between distribution functions are measured by some suitable metric, such as Wasserstein's or Zolotarev's. As the  $n$ th distribution function can be represented as a contraction in a metric space, it converges to a unique 'fixed point' (some fixed distribution function), satisfying a distributional equation like (2), which has a normal distribution as a solution.

Theorem 1 is an immediate consequence of Propositions 1 and 2 and Corollary 5.2 of Neininger and Rüschemdorf (2004). Expressed explicitly, the statement reads as follows. In a random recursive tree of size  $n$  the number  $R_n = R(n, \Gamma)$  of subtrees on the fringe that are isomorphic to a given fixed rooted tree  $\Gamma$  (of size  $\gamma$ ) satisfies

$$\frac{R_n - C(\Gamma)n/(\gamma(\gamma + 1))}{\sqrt{n}} \xrightarrow{D} \mathcal{N}\left(0, \frac{(\gamma + 1)(2\gamma + 1) - (3\gamma + 2)C(\Gamma)}{\gamma(\gamma + 1)^2(2\gamma + 1)} C(\Gamma)\right).$$

### Acknowledgements

The work of the first author was supported by the National Natural Science Foundation of China (grant number 10771024) and the Special Foundation of USTC.

### References

- BERGERON, F., FLAJOLET, P. AND SALVY, B. (1992). Varieties of increasing trees. In *CAAP '92* (Lecture Notes Comput. **581**), Springer, Berlin, pp. 24–48..
- CHERN, H.-H., HWANG, H.-K. AND TSAI, T.-H. (2002). An asymptotic theory for Cauchy–Euler differential equations with applications to the analysis of algorithms. *J. Algorithms* **44**, 177–225.
- CHYZAK, F., DRMOTA, M., KLAUSNER, T. AND KOK, G. (2008). The distribution of patterns in random trees. *Combinatorics Prob. Comput.* **17**, 21–59.
- DOBROW, R. P. AND FILL, J. A. (1996). Multiway trees of maximum and minimum probability under the random permutation model. *Combinatorics Prob. Comput.* **5**, 351–371.
- FENG, Q., MAHMOUD, H. M. AND PANHOLZER, A. (2008). Phase changes in subtree varieties in random recursive and binary search trees. *SIAM J. Discrete Math.* **22**, 160–184.
- FILL, J. A. (1996). On the distribution of binary search trees under the random permutation model. *Random Structures Algorithms* **8**, 1–25.
- FILL, J. A. AND KAPUR, N. (2005). Transfer theorems and asymptotic distributional results for  $m$ -ary search trees. *Random Structures Algorithms* **26**, 359–391.
- FLAJOLET, P., GOURDON, X. AND MARTÍNEZ, C. (1997). Patterns in random binary search trees. *Random Structures Algorithms* **11**, 223–244.
- KNUTH, D. E. (1998). *The Art of Computer Programming III. Sorting and Searching*, 2nd edn. Addison-Wesley, Reading, MA.
- MAHMOUD, H. M. (1992). *Evolution of Random Search Trees*. John Wiley, New York.
- MAHMOUD, H. M., SMYTHE, R. T. AND RÉGNIER, M. (1997). Analysis of Boyer-Moore-Horspool string-matching heuristic. *Random Structures Algorithms* **10**, 169–186.
- NEININGER, R. (2001). On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Structures Algorithms* **19**, 498–524.
- NEININGER, R. (2002). The Wiener index of random trees. *Combinatorics Prob. Comput.* **11**, 587–597.
- NEININGER, R. AND RÜSCHENDORF, L. (2004). A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Prob.* **14**, 378–418.
- RACHEV, S. T. AND RÜSCHENDORF, L. (1995). Probability metrics and recursive algorithms. *Adv. Appl. Prob.* **27**, 770–799.
- RÉGNIER, M. AND SZPANKOWSKI, W. (1998) On pattern frequency occurrences in a Markovian sequence. Average-case analysis of algorithms. *Algorithmica* **22**, 631–649.
- RÖSLER, U. (1991). A limit theorem for “Quicksort”. *RAIRO Inform. Théor. Appl.* **25**, 85–100.
- RÖSLER, U. (2001). On the analysis of stochastic divide and conquer algorithms. *Algorithmica* **29**, 238–261.
- RÖSLER, U. AND RÜSCHENDORF, L. (2001). The contraction method for recursive algorithms. *Algorithmica* **29**, 3–33.
- SMYTHE, R. T. AND MAHMOUD, H. M. (1995). A survey of recursive trees. *Theory Prob. Math. Statist.* **51**, 1–27.
- VAN DER HOFSTAD, R., HOOGHIEMSTRA, G. AND VAN MIEGHEM, P. (2002). On the covariance of the level sizes in random recursive trees. *Random Structures Algorithms* **20**, 519–539.