EXTERNAL PATH LENGTH OF RANDOM

*m*-ORIENTED RECURSIVE

M. JAVANIAN* AND M. G. VAHIDI-ASL**

*Department of Statistics, Faculty of Science, Zanjan University, Zanjan, Iran
**Department of Statistics, Faculty of Mathematical Science, Shahid Beheshti University, Tehran, Iran
E-mail Addresses: m-vahidi@sbu.ac.ir

(Received: June 17, 2011)

ABSTRACT: The random recursive tree is a combinatorial structure used to model a variety of applications such as contagion, chain letters, philology, etc. In this paper, we determine the expectation and variance of \( X_n \), the external path length in a random *m*-oriented recursive tree of size \( n \).

Keywords: Random recursive trees, path length.

1. INTRODUCTION

The analysis of the length of paths in tree families has received a lot of attention, see, e.g., [1,3,6-8], often due to their importance in the analysis of algorithms. In [3,6,7] the total path length is investigated in random recursive trees. However, up to now there is no result about the external path length of random *m*-oriented recursive trees. Here we obtain the expectation and variance of the external path length in random *m*-oriented recursive trees.

By a recursive tree we mean a labeled rooted tree such that each path from the root to any node of the tree is labeled with an increasing sequence of labels.

A survey of applications and results on recursive trees is given in [10]. These trees are used, e.g., to model chain letters and pyramid schemes [5], and as a simplified growth model of the World Wide Web [2].
In working with recursive trees it is convenient to consider an extension of these trees obtained by adding a different type of node called external at each possible insertion position.

Orientation in the plane was not taken into account in the definition of recursive trees. The two labeled trees in Fig. 2 are only two drawings of the same recursive tree. If different orientations are taken to represent different trees, we arrive at a definition of a plane-oriented recursive tree. In such a tree, if a node has outdegree \( d \); there are \( d \) children under it, with \( d+1 \) external nodes.

Figure 2 shows one of the plane-oriented recursive trees of Fig. 1 after it has been extended; the external nodes are shown as squares in Fig. 2.
In this paper we consider m-oriented recursive trees the generalization of random plane-oriented recursive trees where if a node has outdegree \( d \); there are \( d \) children under it, with \((m-1)d+1\) external nodes, (this generalization first defined in [9]). If \( d(i) \) denotes the outdegree of the \( i \)th node then the total number of external nodes in an extended \( m \)-oriented recursive tree of size \( n \) is

\[
\sum_{i=1}^{n} ((m-1)d(i) + 1) = (m-1)\sum_{i=1}^{n} d(i) + n = (m-1)(n-1) + n = m(n-1)+1.
\]

A random \( m \)-oriented recursive tree of size \( n \) is constructed as follows. One starts from a root node holding the label 1; at stage \( i \) (\( i = 2, 3, \ldots, n \)) a new node holding label \( i \) (the \( i \)th node) is attached to any previous node \( j \) of outdegree \( d(j) \) of the already grown tree \( T_{i-1} \) of size \( i-1 \) with probability

\[
\frac{(m-1)d(j) + 1}{m(i-2) + 1}
\]

(the number of remaining external nodes for the node \( j \), \((m-1)d(j)+1\), is divided by \( m(i-2)+1 \), the number of all external nodes). This function implies that the higher outdegree nodes possess a higher attraction for new neighbors.

As the tree grows by the progressive insertion of nodes, two cumulative random variables may serve as measures of the overall cost of construction of the tree, or the cost of later processing of the whole tree if each internal or external node is to be accessed equally often.

Let \( D_j \) be the depth of \( j \)th node in a random \( m \)-oriented recursive tree of size \( n \). The first cumulative random variable is the internal path length

\[
I_n = \sum_{j=1}^{n} D_j.
\]

Suppose the external nodes are indexed by 1, 2, \ldots, \( m(n-1)+1 \), and \( x_j \) be the depth of the \( j \)th external node. The second cumulative random variable is

\[
X_n = \sum_{j=1}^{m(n-1)+1} x_j.
\]

This random variable is called the external path length. By the proof of Theorem 1, the relation
\( X_n = mI_n + n \) (1)

and then \( X_n = m \sum_{j=1}^{n} D_j + n \) can be deduced. The strong dependence between the random variables \( D_j \) makes it difficult to compute the exact distribution of \( X_n \). In Section 2, we compute the expectation and variance of \( X_n \).

In the following, the term random tree without qualification will refer to a random \( m \)-oriented recursive tree.

2. EXPERIMENTAL PROCEDURES

The total path length in random recursive trees has been investigated in [3]. Expectation and variance of the external path length of plane-oriented recursive trees are derived in [8]. In this section, the following results for \( D_n \), will be used (see [4]):

\[
E[D_n] = \sum_{j=1}^{n-1} \frac{1}{m(j-1)+1} \quad \text{and} \quad Var[D_n] = \sum_{j=1}^{n-1} \frac{m(j-1)}{(m(j-1)+1)^2}.
\]

**THEOREM 1:**

\[
E[X_n] = \sum_{j=1}^{n} \frac{m(n-1)+1}{m(j-1)+1},
\]

and

\[
Var[X_n] = \sum_{j=1}^{n} \frac{m^3(n-j)(j-1)(m(n-1)+1)}{(m(j-1)+1)^2(mj+1)}.
\]

**Proof:** Observe that a tree \( T_n \) of size \( n \) is obtained algorithmically from a tree \( T_{n-1} \) of size \( n-1 \) by inserting the \( n \)th node at level \( D_n \). The \( n \)th node may replace any of the \( m(n-2)+1 \) external nodes of \( T_{n-1} \) with probability
The new node gives the tree $m+1$ new external nodes, but one of the external nodes of $T_{n-1}$ is lost in the process. The net gain in the external path length is therefore $mD_n + (D_n + 1) - D_n = mD_n + 1$, i.e., $X_n = X_{n-1} + mD_n + 1$.

Let $\mathcal{F}_n$ denote the sigma field generated by tree $T_n$. When the shape of the tree $T_{n-1}$ is available, the levels $x_1, \ldots, x_{m(n-2)+1}$ of the external nodes are completely determined. Thus $D_n$ may assume any of the values $x_1, \ldots, x_{m(n-2)+1}$ with equal probability $1/(m(n-2)+1)$. We can now formulate a conditional expectation,

$$E[X_n | \mathcal{F}_{n-1}] = \frac{1}{m(n-2)+1} \sum_{j=1}^{m(n-2)+1} (X_{n-1} + mx_j + 1)$$

$$= X_{n-1} + 1 + \frac{m}{m(n-2)+1} \sum_{j=1}^{m(n-2)+1} x_j.$$ 

But the remaining sum is the external path length of $T_{n-1}$, i.e.,

$$E[X_n | \mathcal{F}_{n-1}] = \frac{m(n-1)+1}{m(n-2)+1} X_{n-1} + 1.$$ 

Taking expectations of the last relation we get the following recurrence on expected external path length

$$E[X_n] = \frac{m(n-1)+1}{m(n-2)+1} E[X_{n-1}] + 1,$$

which can be easily solved under the initial condition $E[X_1] = 1$ to yield the first required result.

To compute the variance of $X_n$ we formulate a recurrence for
\[ Q_n := \frac{\text{Var}[X_n]}{(mn+1)(m(n-1)+1)} \]

as follows. Let \( Z_n = \frac{X_n - E[X_n]}{m(n-1)+1} \). Replace \( X_n \) by \( X_{n-1} + mD_n + 1 \) in the definition of \( Z_n \) and write

\[ Z_n = \frac{X_{n-1} + mD_n + 1 - E[X_{n-1} + mD_n + 1]}{m(n-1)+1} \]

\[ = \frac{m(n-2)+1}{m(n-1)+1} Z_{n-1} + \frac{m}{m(n-1)+1} (D_n - E[D_n]). \]

Upon squaring the latter relation and taking expectations we get

\[
E[Z_n^2] = \left( \frac{m(n-2)+1}{m(n-1)+1} \right)^2 E[Z_{n-1}^2] + \left( \frac{m}{m(n-1)+1} \right)^2 \text{Var}[D_n] \\
+ \frac{2m(m(n-2)+1)}{(m(n-1)+1)^2} E[Z_{n-1} (D_n - E[D_n])].
\]

(2)

Since the component \( E[Z_{n-1}E[D_n]] \) is zero, in the last term we need only to find \( E[Z_{n-1}D_n] \). For the required term we compute

\[ E[Z_{n-1}D_n] = E[E[Z_{n-1}D_n | \mathcal{F}_{n-1}]] = E[Z_{n-1}E[D_n | \mathcal{F}_{n-1}]]. \]

But according to the algorithmic development,

\[ E[D_n | \mathcal{F}_{n-1}] = \sum_{j=1}^{m(n-2)+1} \frac{x_j}{m(n-2)+1} = \frac{X_{n-1}}{m(n-2)+1}. \]

So,

\[ E[Z_{n-1}D_n] = E[Z_{n-1}^2]. \]
Plugging this relation into (2) we arrive at the recurrence

\[
E[Z_n^2] = \frac{(mn+1)(m(n-2)+1)}{(m(n-1)+1)^2} E[Z_{n-1}^2]
\]

\[
+ \frac{m^2}{(m(n-1)+1)^2} \text{Var}[D_n].
\]  

(3)

The substitution \(Q_n\) linearizes the recurrence (3) into the simple recurrence

\[
Q_n = Q_{n-1} + \frac{m^2}{(m(n-1)+1)(mn+1)} \text{Var}[D_n].
\]

By the relation for the variance of \(D_n\), the solution to the last recurrence gives

\[
Q_n = \sum_{j=3}^{n} \frac{m^2}{(m(j-1)+1)(mj+1)} \sum_{k=1}^{j-1} \frac{m(k-1)}{(m(k-1)+1)^2}.
\]

Expanding \(1/(m(j-1)+1)(mj+1)\) by partial fractions and collapsing the resulting telescopic sums, we have

\[
Q_n = \sum_{j=2}^{n-1} \frac{m^3(n-j)(j-1)}{(m(j-1)+1)^2(mj+1)(mn+1)}.
\]

So by definition of \(Q_n\), the proof is complete.

**REMARK:** By (1) the expectation of internal path length \(I_n\) is

\[
E[I_n] = \frac{1}{m} \sum_{j=1}^{n} \frac{m(n-1)+1}{m(j-1)+1} - \frac{n}{m}.
\]
So the average external path length is asymptotically $m$ times as much as the average internal path length $E[T_n] = \frac{n}{m} \ln n$.

The standard deviation of the external path length of a random $m$-oriented recursive tree is relatively small compared to the mean value, since
\[
\text{Var}[X_n] = (mn+1)(m(n-1)+1)Q_n \quad \text{and} \quad Q_n = O(1)
\]
then
\[
\sqrt{\text{Var}[X_n]} = O(n), \quad \text{while} \quad E[X_n] = n \ln n, \quad \text{as} \quad n \to \infty.
\]
From an application of Chebychev's inequality we can conclude that
\[
\frac{X_n}{n \ln n} \to 1, \quad \text{in probability.}
\]

REFERENCES