ON THE EXTERNAL PATH LENGTH OF RANDOM RECURSIVE k-ARY TREES

Mehri Javanian

Department of Statistics
Faculty of sciences
University of Zanjan
Zanjan
Iran
e-mail: javanian@znu.ac.ir

Abstract. In this paper, we determine the expectation and variance of X_n the external path length in a random recursive k-ary tree of size n.

Key words and phrases: Recursive trees, path length.
2000 Mathematics Subject Classification: 05C05.

1. Introduction

The analysis of the length of paths in tree families has received a lot of attention, see, e.g., [1], [4], [10], [11], [13], often due to their importance in the analysis of algorithms. In [4], [11], [13] 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 recursive k-ary trees. Here we obtain the expectation and variance of the external path length in random recursive k-ary 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.

Note that there is no restriction on the outdegrees of the nodes of a recursive tree. A recursive tree where outdegrees are restricted to k, is called a *recursive k-ary tree*.

The possible insertion positions to join a new node to a tree, are called *external* nodes. In a recursive k-ary tree, the number of nodes can be attached to a node ν of outdegree d_ν is $k - d_\nu$. Therefore the number of all external nodes

in a recursive k-ary tree of size n is,

$$\sum_{\nu=1}^{n}(k-d_\nu) = kn-(n-1) = (k-1)n+1.$$

A *random recursive k-ary tree* of size n is constructed as follows. One starts from a root node holding the label 1; at stage i ($i = 2, 3, ..., n$) a new node holding label i (the ith node) is attached to any previous node ν of outdegree d_ν of the already grown tree T_{i-1} of size $i-1$ with probability $\dfrac{k-d_\nu}{(k-1)(i-1)+1}$ (the number of remaining external nodes for the node ν, $k-d_\nu$, is divided by $(k-1)(i-1)+1$, the number of all external nodes). This function implies that the higher outdegree nodes possess a lower attraction for new neighbors and no node of outdegree greater than k is introduced. These are two properties of interest in chemical applications [2], [5]. The outdegree of a node is investigated in [7] and [9].

A survey of applications and results on recursive trees is given in [12]. These trees are used, e.g., to model chain letters and pyramid schemes [6], and as a simplified growth model of the word wide web [3].

Let D_j be the depth of jth node in a random recursive k-ary 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 $1, 2, \ldots, (k-1)n+1$, and x_j be the depth of the jth external node. The second cumulative random variable is $X_n = \sum_{j=1}^{(k-1)n+1} x_j$. This random variable is called the *external path length*. By the proof of Theorem 1, the relation

$$X_n = (k-1)I_n + nk, \tag{1}$$

and then $X_n = (k-1)\sum_{j=1}^{n} D_j + nk$ can be deduced. The strong dependence between the random variables D_j makes it difficult to compute the exact distribution of X_n.

Throughout this paper we use the term *tree* instead of recursive k-ary tree.

2. Expectation and variance

In this section, the following results for D_n, will be used (see [8]):

$$\mathbf{E}[D_n] = \sum_{x=1}^{n-1} \frac{k}{(k-1)x+1},$$

and

$$\mathbf{Var}[D_n] = \sum_{x=2}^{n-1} \frac{k(k-1)(x-1)}{((k-1)x+1)^2}.$$

Theorem 1

$$\mathbf{E}[X_n] = \sum_{x=1}^{n} \frac{((k-1)n+1)k}{(k-1)x+1},$$

and

$$\mathbf{Var}[X_n] = \sum_{x=1}^{n} \frac{k(k-1)^3(n-x)(x-1)((k-1)n+1)}{((k-1)x+1)^2((k-1)(x+1)+1)}.$$

Proof. By inserting the nth node at level D_n, a tree T_n of size n is obtained from a tree T_{n-1} of size $n-1$. The nth node may replace any of the $(k-1)(n-1)+1$ external nodes of T_{n-1} with probability $1/((k-1)(n-1)+1)$. The new node gives the tree k new external nodes, but one of the external nodes of T_{n-1} is lost in the process. Therefore

$$X_n = X_{n-1} + k(D_n+1) - D_n = X_{n-1} + (k-1)D_n + k.$$

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

$$\begin{aligned}\mathbf{E}[X_n|\mathcal{F}_{n-1}] &= \frac{1}{(k-1)(n-1)+1} \sum_{j=1}^{(k-1)(n-1)+1} (X_{n-1} + (k-1)x_j + k) \\ &= X_{n-1} + k + \frac{k-1}{(k-1)(n-1)+1} \sum_{j=1}^{(k-1)(n-1)+1} x_j.\end{aligned}$$

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

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

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

$$\mathbf{E}[X_n] = \frac{(k-1)n+1}{(k-1)(n-1)+1} \mathbf{E}[X_{n-1}] + k, \tag{3}$$

which can be easily solved under the initial condition $\mathbf{E}[X_1] = k$ to yield the first required result.

To compute the variance of X_n we formulate a recurrence for

$$Q_n := \frac{\mathbf{Var}[X_n]}{((k-1)n+1)((k-1)(n+1)+1)}$$

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

$$\begin{aligned} Z_n &= \frac{X_{n-1} + (k-1)D_n + k - \mathbf{E}[X_{n-1} + (k-1)D_n + k]}{(k-1)n+1} \\ &= \frac{(k-1)(n-1)+1}{(k-1)n+1} Z_{n-1} + \frac{k-1}{(k-1)n+1}(D_n - \mathbf{E}[D_n]). \end{aligned}$$

By squaring the latter relation and taking expectations we get

$$\begin{aligned} \mathbf{E}[Z_n^2] &= \Big(\frac{(k-1)(n-1)+1}{(k-1)n+1}\Big)^2 \mathbf{E}[Z_{n-1}^2] + \Big(\frac{k-1}{(k-1)n+1}\Big)^2 \mathbf{Var}[D_n] \\ &+ \frac{2(k-1)((k-1)(n-1)+1)}{((k-1)n+1)^2} \mathbf{E}[Z_{n-1}(D_n - \mathbf{E}[D_n])]. \end{aligned} \tag{4}$$

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

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

But according to the algorithmic development,

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

So,

$$\mathbf{E}[Z_{n-1}D_n] = \mathbf{E}[Z_{n-1}^2].$$

Put this relation into (4) we arrive at the recurrence

$$\begin{aligned} \mathbf{E}[Z_n^2] &= \frac{((k-1)(n+1)+1)((k-1)(n-1)+1)}{((k-1)n+1)^2} \mathbf{E}[Z_{n-1}^2] \\ &+ \frac{(k-1)^2}{((k-1)n+1)^2} \mathbf{Var}[D_n]. \end{aligned} \tag{5}$$

The substitution Q_n linearizes the recurrence (5) into the simple recurrence

$$Q_n = Q_{n-1} + \frac{(k-1)^2}{((k-1)n+1)((k-1)(n+1)+1)} \mathbf{Var}[D_n].$$

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

$$Q_n = \sum_{m=3}^{n} \frac{(k-1)^2}{((k-1)m+1)((k-1)(m+1)+1)} \sum_{x=2}^{m-1} \frac{k(k-1)(x-1)}{((k-1)x+1)^2}.$$

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

$$Q_n = \sum_{x=2}^{n-1} \frac{k(k-1)^3(n-x)(x-1)}{((k-1)x+1)^2((k-1)(x+1)+1)((k-1)(n+1)+1)}. \tag{6}$$

So, by definition of Q_n, the proof is complete. ∎

Remark. By (1) the expectation of internal path length I_n is

$$\mathbf{E}[I_n] = \frac{k}{k-1} \sum_{j=2}^{n-1} \frac{(k-1)n+1}{(k-1)j+1} - \frac{n-1}{k-1}.$$

So the average external path length is asymptotically $k-1$ times as much as the average internal path length $\mathbf{E}[I_n] \sim \frac{k}{k-1} n \ln n$.

References

[1] Aguech, R., Lasmar, N., Mahmoud, H., *Extremal weighted path lengths in random binary search trees*, Probab. Engrg. Inform. Sci., 21 (1) (2007), 133–141.

[2] Balińska, K.T., Quintas, L.V., *The sequential generation of random f-graphs. Line maximal 2-, 3-, and 4-graphs*, Computers Chemistry, 14 (1990), 323–328.

[3] Chan, D.Y.C., Hughes, B.D., Leong, A.S., Reed, W.J., *Stochasticaly evolving networks*, Phys. Rev. E, 68 (066124) (2003), 24.

[4] Dobrow, R.P., Fill, J., *Total path length for random recursive trees. Random graphs and combinatorial structures*, Combin. Probab. Comput., 8 (4) (1999), 317–333.

[5] Galina, H., Szustalewicz, A., *A klinic theory of stepwise crosslinking polymerization with substitution effect*, Macromolecules, 22 (1989), 3124–3129.

[6] Gastwirth, J.L., Bhattacharya, P.K., *Two probability models of pyramids or chain letter schemes demonstrating that their promotional claims are unreliable*, Operations Research, 32 (1984), 527–536.

[7] Janson, S., *Asymptotic degree distribution in random recursive trees*, Random Structures and Algorithms, 26 (2005), 69–83.

[8] Javanian, M., Vahidi-Asl, M.Q., *Depth of nodes in random recursive k-ary trees*, Inform. Process. Lett., 98 (3) (2006), 115–118.

[9] Kuba, M., Panholzer, A., *On the degree distribution of the nodes in increasing trees*, Journal of Combinatorial Theory, Series A, 114 (2007), 597–618.

[10] Kuba, M., Panholzer, A., *On weighted path lengths and distances in increasing trees*, Probab. Engrg. Inform. Sci., 21 (3) (2007), 419–433.

[11] Mahmoud, H., *Limiting distributions for path lengths in recursive trees*, Probab. Engrg. Inform. Sci., 5 (1) (1991), 53–59.

[12] Mahmoud, H., Smythe, R.T., *A survey of recursive trees*, Theory Probab. Math. Statist., 51 (1995), 1–27.

[13] Szymanski, J., *On the complexity of algorithms on recursive trees*, Theoretical Computer Science, 74 (3) (1990), 355-361.

Accepted: 16.04.2010