# THE SACKIN INDEX OF RANDOM RECURSIVE TREES

Khosro Moradian[1], Ramin Kazemi[2], Mohammad H. Behzadi[3]

*The Sackin index of a tree as of the oldest measures that summarizes the shape of a tree is defined as the sum of the depths of its leaves. In this paper, we study this index in random recursive trees. The mean of this index is given. The lower and upper bounds of the probability generating function are given. Finally, a submartingale on this index is introduced.*

**Keywords:** recursive tree, Sackin index, total path length.

## 1. Introduction

A graph is a collection of points and lines connecting a subset of them [1]. The points and lines of a graph are also called vertices and edges of the graph, respectively. *Trees* are defined as connected graphs without cycles, and their properties are basics of graph theory. For example, a connected graph is a tree, if and only if the number of edges equals the number of nodes minus 1. Furthermore, each pair of nodes is connected by a unique path. A rooted tree is a tree with a countable number of nodes, in which a particular node is distinguished from the others and called the root node. Recursive trees are one of the most natural combinatorial tree models with applications in several fields, e.g., it has been introduced as a model for the spread of epidemics, for pyramid schemes, for the family trees of preserved copies of ancient texts and furthermore it is related to the Bolthausen-Sznitman coalescence model. A recursive tree with $n$ nodes is an unordered rooted tree, where the nodes are labelled by distinct integers from $\{1, 2, 3, ..., n\}$ in such a way that the sequence of labels lying on the unique path from the root node to any node in the tree are always forming an increasing sequence [9]. This implies that the root node is always labelled by 1. It is well known and easy to show by induction that there are $(n-1)!$ different recursive trees with $n$ nodes. It is of particular interest in applications to assume the random recursive tree model and to speak about a random recursive tree with $n$ nodes, which means that one of the $(n-1)!$ possible recursive trees with $n$ nodes is chosen with equal probability, i.e., the probability that a particular tree with $n$ nodes is chosen is always $1/(n-1)!$. Equivalently one may describe random recursive trees via the following tree evolution process, which generates random recursive trees of arbitrary order $n$. At step 1 the process starts with the root labeled by 1. At step $i + 1$ the node with label $i + 1$ is attached to any previous node $v$ of the already grown tree $T$ of order $i$ with probability $p_i(v) = 1/i$. Figure 1 illustrates a recursive tree of order $n = 7$.

The article is organized as follows. In Section 2, we review some previous results on the Sackin index in random trees. In Section 3, first the mean of this index is given. Second, the lower and upper bounds of the probability generating function are given. Finally, a submartingale on this index is introduced.

[1] Department of Statistics, Science and Research Branch, Islamic Azad University, Tehran, Iran.

[2] Department of Statistics, Imam Khomeini International University, Qazvin, Iran, e-mail: `r.kazemi@SCI.ikiu.ac.ir`

[3] Department of Statistics, Science and Research Branch, Islamic Azad University, Tehran, Iran.

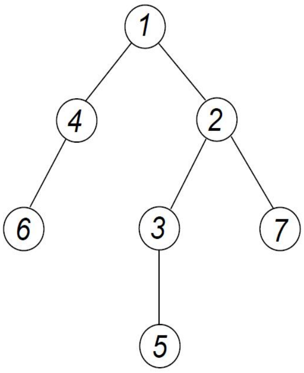

FIGURE 1. A recursive tree of order $n = 7$ with Sackin index $S_7 = 7$ [6].

## 2. Sackin index

The distance $D_k$ between the root and node $k$ in a random recursive tree has been studied by many authors, including Moon [11], Szymański [15]. The total path length of a recursive tree, namely,

$$T_n = \sum_{k=1}^{n} D_k, \tag{1}$$

defined as the sum of all root-to-node distances. This random variable may serve as a global measure of the cost of constructing the tree. The strong dependence among the random variables $D_k$ makes it nontrivial to obtain the exact distribution of $D_n$ be the $n$th harmonic number. Howevere $\mathbb{E}(D_k) = H_{k-1}$. Linearity of expectation gives

$$\mathbb{E}(T_n) = n(H_n - 1), \tag{2}$$

which is asymptotically equivalent to $n \ln n$. Sackin index is one of the oldest measure that summarizes the shape of a tree [12, 13]. It adds the number of internal nodes between each leaf of the tree and the root to form the following index $S_n = \sum_{i=1}^{n} N_i$ , where the sum runs over the $n$ leaves of the tree and $N_i$ is the number of internal nodes crossed in the path from $i$ to the root (including the root). An equivalent formulation of $S_n$ is by counting the number of leaves under each internal nodes $S_n = \sum_{j=1}^{n-1} \overline{N}_j$, where $\overline{N}_j$ is the number of leaves that descend from the ancestor $j$. This is a well-known result in systematic biology that the expectation of $S_n$ under the Yule model is of order $2n \ln n$ [7]:

$$\mathbb{E}(S_n) = 2n(H_n - 3/2).$$

The variance is more complex, but it can be estimated by noticing the analogy with a classical problem in theoretical computer science. Let $\mathcal{T}_n$ be the set of isomorphism classes of phylogenetic trees with $n$ leaves. Then under the uniform model,

$$\mathbb{E}(S_n) = \frac{n}{2n-3} \sum_{k \geq 0} \frac{(2)_k (2)_k (2-n)_k}{(1)_k (4-2n)_k} \frac{2^k}{k!},$$

where $(a)_k := a(a+1)...(a+k-1)$ [10].

## 3. The Main Resluts

For a (rooted) path $P_n$, $S_{P_n} = 1 \times (n-1) = n-1$. For a star $S_n$, $S_{S_n} = (n-1) \times 1 = n-1$. Thus $S_{P_n} = S_{S_n} = (n-1)$.

**Theorem 3.1.** *Let $S_n$ be the Sackin index of the random recursive tree of order $n$, Then for $n \geq 3$*

$$\mathbb{E}(S_n) = \frac{n}{2}\Big(H_n - \frac{1}{2}\Big).$$

*Proof.* Let $U_n$ be a randomly chosen node belong to $T$ of order $n$ and $\mathcal{F}_n$ be the sigma-field generated by the first $n$ stages of the recursive trees. Also, let

$$\mathbb{I}(D_k) = \begin{cases} 0, & \text{if node } k \text{ is a leaf or root in } T \text{ of order } n-1 \\ D_k, & \text{if node } k \text{ is a non-leaf in } T \text{ of order } n-1. \end{cases}$$

By stochastic growth rule of the random recursive trees and definition of $S_n$,

$$S_n = S_{n-1} + \mathbb{I}(D_{U_{n-1}}) + 1. \tag{3}$$

From (3),

$$\begin{aligned} \mathbb{E}(S_n|\mathcal{F}_{n-1}) &= S_{n-1} + \mathbb{E}(\mathbb{I}(D_{U_{n-1}})|\mathcal{F}_{n-1}) + 1 \\ &= S_{n-1} + \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbb{I}(D_i) + 1, \end{aligned} \tag{4}$$

since $S_{n-1}$ is $\mathcal{F}_{n-1}$-measurable [2] and the label $n$ is attached to any node $v$ of the already grown tree $T$ of order $n-1$ with probability $\frac{1}{n-1}$. But

$$\sum_{i=1}^{n-1}\mathbb{I}(D_i) = T_{n-1} - S_{n-1}.$$

Thus

$$\mathbb{E}(S_n|\mathcal{F}_{n-1}) = \frac{n-2}{n-1}S_{n-1} + \frac{1}{n-1}T_{n-1} + 1. \tag{5}$$

Taking expectation of the relation (5):

$$\begin{aligned} \mathbb{E}(S_n) &= \frac{n-2}{n-1}\mathbb{E}(S_{n-1}) + \frac{1}{n-1}\mathbb{E}(T_{n-1}) + 1 \\ &= \frac{n-2}{n-1}\mathbb{E}(S_{n-1}) + H_{n-1}. \end{aligned} \tag{6}$$

The recurrence equation (6) leads to

$$\begin{aligned} \mathbb{E}(S_n) &= \frac{1}{n-1} + \frac{1}{n-1}\sum_{j=2}^{n-2} jH_j + H_{n-1} \\ &= \frac{1}{n-1} + \frac{1}{n-1}\Big(\frac{(n-2)(n-1)}{2}H_{n-1} - \frac{(n-2)(n-1)}{4} - 1\Big) \\ &+ \ H_{n-1} \\ &= \frac{n}{2}\Big(H_n - \frac{1}{2}\Big). \end{aligned}$$

□

Theorem 3.1 shows that $\mathbb{E}(S_n)$ is asymptotically equivalent to $\frac{n}{2}\ln n$. Stanley [14] gives the following mapping. Let $\sigma = (\sigma_1, ..., \sigma_{n-1})$ be a permutation on $\{1, 2, ..., n-1\}$. Construct a recursive tree with nodes 0,1,...,$n-1$ by making 0 the root and defining the parent of node $i$ to be the rightmost element $j$ of $\sigma$ which both precedes $i$ and is less than $i$. If there is no such element $j$, then define the parent of $i$ to be the root 0. Finally, to convert to a recursive tree on nodes $\{1, 2, ..., n\}$, simply add 1 to each label. For example, the permutation $(1, 2, 3)$ corresponds to the linear tree of order 4 where $i$ is the parent of $i+1$ for $i = 1, 2, 3$; the permutation $(3, 2, 1)$ corresponds to the tree where nodes 2, 3, and 4 are each children of the root 1. This mapping is bijective between permutations of $\{1, ..., n-1\}$ and recursive trees with label set $\{1, ..., n\}$. Note that in this correspondence the order of the subtree rooted at node 2 is one greater than the number of elements in the corresponding permutation of order $n-1$ that succeed 1. This number, in turn, is just $n$ minus the position of 1. The position of 1 is, of course, distributed uniformly on $\{1, ..., n-1\}$.

**Theorem 3.2.** *For $n \geq 3$,*

$$S_n \overset{d}{=} M + S_M + S^*_{n-M},$$

*where $M$ is the number of leaves in the subtree rooted at node 2 and the random variables $M, S_1, ...S_M, S^*_1, ..., S^*_{n-M}$ are all mutually independent.*

*Proof.* Let $M$ be the number of leaves in the subtree rooted at node 2. Then $M + S_M$ accounts for the contribution to Sackin index in the subtree rooted at node 2, and $S^*_{n-M}$ accounts for the contribution to Sackin index from all the remaining subtrees. The theorem will follow from the fact that in a random recursive tree of order $n$ the order of the subtree rooted at node 2 is distributed uniformly on $\{1, ..., n-1\}$. □

**Theorem 3.3.** *Let $\phi_n(t) = \sum_k t^k P(S_n = k)$ be the probability generating function of $S_n$. Then for $n > 3$,*

$$\frac{t^{n-1}}{(n-1)!} \leq \phi_n(t) \leq \frac{t^{n-1}}{n-1} + \frac{1}{1-t}.$$

*Proof.* Given tree $T$ of order $n-1$, pick a node uniformly at random. If leaf $v$ is picked, then tree $T$ of order $n$ is formed by making node $n$ as a child of $v$ and $S_n = S_{n-1} + 1$. If non-leaf $v$ is picked, then $S_n = S_{n-1} + D_n$. By this constraction, conditional on the tree $T$ of order $n-1$,

$$P(S_n = k) = \frac{L_{n-1}}{n-1} P(S_{n-1} = k-1) + \frac{n-1-L_{n-1}}{n-1} P(S_{n-1} = k - D_n),$$

where $L_n$ is the number of leaves in $T$ of order $n$. It is obvious that $\max L_{n-1} \leq n-2$. Thus

$$P(S_n = k) \leq \frac{n-2}{n-1} P(S_{n-1} = k-1) + \frac{1}{n-1}. \tag{7}$$

Multiplying (7) by $t^k$ and summing over $k$,

$$\phi_n(t) \leq \frac{n-2}{n-1} t\phi_{n-1}(t) + \frac{1}{n-1}\frac{1}{1-t}. \tag{8}$$

The recurrence (8) leads to

$$\phi_n(t) \leq \frac{1}{n-1} t^{n-2} \phi_2(t) + \frac{1}{1-t}.$$

But

$$\phi_2(t) = \sum_{k=1}^{1} t^k P(S_2 = k) = t.$$

Thus

$$\phi_n(t) \leq \frac{1}{n-1} t^{n-1} + \frac{1}{1-t}.$$

We can obtain the lower bound similarly, since $\min L_{n-1} \geq 1$. □

**Definition 3.1.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\{X_1, X_2, ...\}$ a sequence of integrable random variables on $(\Omega, \mathcal{F}, P)$, and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ an increasing sequence of sub sigma-fields of $\mathcal{F}$; $X_n$ is assumed $\mathcal{F}_n$-measurable. The sequence $(X_n)_{n\geq 1}$ is said to be a submartingale relative to the $\mathcal{F}_n$ iff for all $n = 1, 2, ...$, $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$ (a.e.).*

**Theorem 3.4.** *Let $S_n$ be the Sackin index of the random recursive tree of order $n$. Then the sequence $(Z_n)_{n\geq 0}$ with*

$$Z_n = S_n - \frac{\mathbb{E}(S_n)}{n},$$

*iz a submartingale.*

*Proof.* It is obvious that $\mathbb{I}(D_{U_{n-1}}) \geq 0$, then for all $n \geq 2$, $S_n \geq S_{n-1}+1$. Given the history of insertions $D_1, ..., D_{n-1}$, the values of $Z_1, ..., Z_{n-1}$ are completely determined. Therefore, we can equivalently condition on $D_1, ..., D_{n-1}$. From Theorem 3.1,

$$\begin{aligned}
\mathbb{E}(Z_n|Z_1, ..., Z_{n-1}) &= \mathbb{E}(Z_n|D_1, ..., D_{n-1}) \\
&\geq \mathbb{E}(S_{n-1}+1|D_1, ..., D_{n-1}) - \frac{1}{n}\frac{n}{2}\Big(H_n - \frac{1}{2}\Big) \\
&= S_{n-1} + 1 - \frac{1}{2}\Big(H_{n-1} + \frac{1}{n} - \frac{1}{2}\Big) \\
&\geq S_{n-1} - \frac{1}{2}\Big(H_{n-1} - \frac{1}{2}\Big) \\
&= Z_{n-1}.
\end{aligned}$$

Also $\mathbb{E}(|Z_n|) < \infty$ exists for each $n$ and proof is completed. □

By Theorem 3.4 and the submartingale convergence theorem [2], there is an integrable random variable $Z_\infty$ such that $Z_n \to Z_\infty$ almost everywhere. Let $T_1, T_2, ...$ be an increasing sequence of finite stopping times for $(Z_n)_{n\geq 1}$ and $Y_n = Z_{T_n}$, $n = 1, 2, 3, ....$ Also, let
1- $\mathbb{E}(Y_n) < \infty$ for all $n$ and
2- $\liminf_{k\to\infty} \int_{\{T_n>k\}} |Z_n| dP = 0$ for all $n$.
By optimal sampling theorem, $(Y_n)_{n\geq 1}$ is a submartingale relative to the $\mathcal{F}_{T_n}$. Let $T$ be a stopping time for $(Z_n)_{n\geq 1}$. Then

$$\mathbb{E}(|Z_T|) \leq 2\mathbb{E}(Z_n^+),$$

since $Z_1 = 0$. If $T$ is a finite stopping time for $(Z_n)_{n\geq 1}$, then

$$\mathbb{E}(|Z_T|) \leq 2 \sup_n \mathbb{E}(Z_n^+).$$

**Theorem 3.5.** *If $\lambda \geq 0$, then $P(\max_{1\leq i\leq n} Z_i \geq \lambda) \leq \frac{\mathbb{E}(Z_n^+)}{\lambda}$. Also*

$$P(\sup_n Z_n > \lambda) \leq \frac{\sup_n \mathbb{E}(Z_n^+)}{\lambda}.$$

*Proof.* This is an immediate consequence of Doob's supermartingale inequality [2]. □

## 4. Conclusion and open problems

In this paper, we studied the Sackin index in random recursive trees. The mean of this index was given. The lower and upper bounds of the probability generating function were given. Finally, a submartingale on this index was introduced.

An interesting and natural generalization of random recursive trees has been introduced in [8] by Mahmoud and Smythe, which are called bucket recursive trees. In this model the nodes of a bucket recursive tree are buckets, which can contain up to a fixed integer amount of $b \geq 1$ labels. A probabilistic description of random bucket recursive trees is given by a generalization of the stochastic growth rule for ordinary random recursive trees (which are the special instance $b = 1$), where a tree grows by progressive attraction of increasing integer labels: when inserting label $n+1$ into an existing bucket recursive tree containing $n$ labels (i.e., containing the labels $\{1, 2, ..., n\}$) all $n$ existing labels in the tree compete to attract the label $n+1$, where all existing labels have equal chance to recruit the new label. If the label winning this competition is contained in a node with less than $b$ labels (an unsaturated bucket or node), label $n+1$ is added to this node, otherwise if the winning label is contained in a node with already $b$ labels (a saturated bucket or node), label $n+1$ is attached to this node as a new bucket containing only the label $n+1$. Starting with a single bucket as root node containing only label 1 leads after $n-1$ insertion steps, where the labels $2, 3, ..., n$ are successively inserted according to this growth rule, to a so called random bucket recursive tree with $n$ labels and maximal bucket size $b$. Of course, the above growth rule for inserting the label $n+1$ could also be formulated by saying that, for an existing bucket recursive tree $T$ with $n$ labels, the probability that a certain node $v \in T$ with capacity $1 \leq c(v) \leq b$ attracts the new label $n+1$ is proportional to the number of labels contained in $v$, i.e., $\frac{c(v)}{n}$.

Kazemi [4, 5] introduced a new version of bucket recursive trees where the nodes are buckets with variable capacities labelled with integers $1, 2, ..., n$. In fact, the capacity of buckets is a random variable in these models. An order-$n$ bucket recursive tree $T$ with variable bucket capacities and maximal bucket size $b$ starts with the root labelled by 1. The tree grows by progressive attraction of increasing integer labels: when inserting label $j+1$ into an existing bucket recursive tree $T$ of order $j$, except the labels in the non-leaf nodes with capacity $< b$ all labels in the tree (containing label 1) compete to attract the label $j+1$. For the root node and nodes with capacity $b$, we always produce a new node $j+1$. But for a leaf with capacity $c < b$, either the label $j+1$ is attached to this leaf as a new bucket containing only the label $j+1$ or is added to that leaf and make a node with capacity $c+1$. This process ends with inserting the label $n$ (i.e., the largest label) in the tree. Let $|.|$ denotes the size of sets. The probability $p$, which gives the probability that label $n$ is attracted by node $v$ in the tree of order $n-1$ is: $p = \frac{c(v)}{n-1-|\gamma|}$, where $\gamma = \{v \in T;\ c = c(v) < b,$ and $v$ is a non-leaf$\}$.

As an open problem, it would be interesting to consider the Sackin index of random bucket recursive trees. Furthermore, it is challenging to determine extremal values of the Sackin index among all trees with $n$ vertices.

## REFERENCES

[1] *Azari, M. Iranmanesh, A.* Edge-Wiener type invariants of splices and links of graphs, U.P.B. Sci. Bull., Series A, 77(3) (2015), 143-154.

[2] *Billingsley, P.* Probability and Measure, John Wiley and Sons, New York, 1995.

[3] *Borovićanin, B., Furtula, B.* On extremal Zagreb indices of trees with given domination number, Appl. Math. Comput., 279(10) (2016), 208-218.

[4] *Kazemi, R.* Depth in bucket recursive trees with variable capacities of buckets, Acta Math. Sin, English Series, 30(2) (2014), 305-310.

[5] *Kazemi, R.* Branches in bucket recursive trees with variable capacities of buckets, U.P.B. Sci. Bull., Series A, 77(1) (2015), 109-114.

[6] *Kazemi, R., Meimondari, L. K.* Degree distance and Gutman index of increasing trees. Trans. Comb, 5(2) (2016), 23-31.

[7] *Kirkpatrick, M., Slatkin, M.* Searching for evolutionary patterns in the shape of a phylogenetic tree, Evolution, 47(4) (1993) 11-71.

[8] *Mahmoud, H., Smythe, R.* Probabilistic analysis of bucket recursive trees. Theor. Comput. Sci. 144 (1995), 221-249.

[9] *Meir, A. Moon, J. W.* Recursive trees with no nodes of out-degree one, Congressus Numerantium, 66 (1988), 49-62.

[10] *Mir, A., Rosselló, F., Rotger, L.* A new balance index for phylogenetic trees, Math. Biosci, 241(1) (2013), 125-136.

[11] *Moon, J. W.* The distance between nodes in recursive trees. London Mathematics Society Lecture Notes Series, No. 13 London: Cambridge University Press, (1974) 125-132.

[12] *Sackin, M. J.* Good and bad phenograms, Syst. Zool. 21 (1972) 225.

[13] *Shao, K. Sokal, R. R.* Tree balance, Syst. Zool. 39 (1990), 266.

[14] *Stanley, R. P.* Enumerative Combinatorics, Vol. I. Wadsworth and Brooks/Cole, Monterey, Calif, 1986.

[15] *Szymański, J.* On the complexity of algorithms on recursive trees. Theo. Comp. Sci. 74 (1990), 355-361