Applied Mathematics Letters 25 (2012) 2218–2222

Contents lists available at SciVerse ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

Asymptotic distribution of two-protected nodes in random binary search trees

Hosam M. Mahmoud [a], Mark Daniel Ward [b,*]

[a] Department of Statistics, The George Washington University, Washington, DC 20052, USA
[b] Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 27 January 2012
Accepted 6 June 2012

Dedicated to the memory of Philippe Flajolet

Keywords:
Binary search trees
Random structure
Combinatorial probability
Asymptotic analysis

ABSTRACT

We derive exact moments of the number of 2-protected nodes in binary search trees grown from random permutations. Furthermore, we show that a properly normalized version of this tree parameter converges to a Gaussian limit.

1. Introduction

The study of 2-protected nodes in classes of random trees is in the vogue. Cheon and Shapiro [1] investigate the average number of 2-protected nodes in unlabeled, ordered trees and in unary–binary trees (those with 0, 1, or 2 children per node). Mansour [2] considers the average number of 2-protected nodes in k-ary trees. Recently, Du and Prodinger [3] have analyzed the average of this parameter in random digital trees, with a uniform probability model.

In this article, we consider the number of 2-protected nodes in a random binary search tree (BST). These are binary trees, like those in the 2-ary case of Mansour [2], but differ in their underlying probability distribution. Those in [2] are uniformly distributed, i.e., all trees of the same *size* (number of nodes) are equally likely. In contrast, the BST grows from a random permutation that induces a BST probability model, which is *nonuniform*. The BST model is of prime importance in computer science as it represents the backbone of some fundamental algorithms, such as Quicksort (see Knuth [4] or Mahmoud [5]), and are basic efficient data structures in their own right (see Mahmoud [6]).

The BST grows from a uniformly random permutation $(\pi_1, \pi_2, \ldots, \pi_n)$, of $\{1, 2, \ldots, n\}$, as follows. In the computer science jargon, elements of the permutation are often called *keys*. The first key π_1 goes into the root node of a tree, with distinguished left and right subtrees (which are empty as of yet). The second key is guided to the left subtree, if it is smaller than the root key (i.e., if $\pi_2 < \pi_1$), where it is inserted in a node and linked as a left child of the root; otherwise (i.e., if $\pi_2 > \pi_1$) the second key goes into the right subtree, where it is inserted in a node and linked as a right child of the root. Subsequent keys go to the left or right subtrees, according to whether they are smaller than the root key or not, where they are inserted recursively in the subtree by the same algorithm. Note that when the permutations of $\{1, 2, \ldots, n\}$ are equally likely, they give rise to a nonuniform probability distribution on the shapes of BST. We call such distribution the *BST probability model*.

* Corresponding author.
E-mail addresses: hosam@gwu.edu (H.M. Mahmoud), mdw@purdue.edu (M.D. Ward).

0893-9659/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2012.06.005

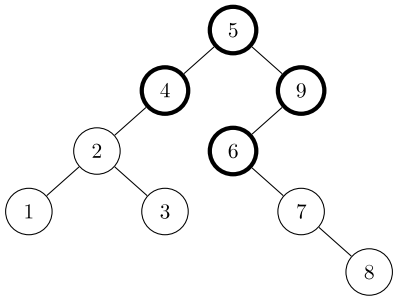

Fig. 1. Example of a binary search tree corresponding to the permutation (5, 9, 6, 4, 7, 2, 3, 1, 8); 2-protected nodes in bold.

This BST probability model is deemed more relevant to computer science applications than the uniform model on binary trees as it conforms more closely to the nature of data arising in sorting and searching applications. For instance, data samples of size n taken from any arbitrary continuous distribution have ranks that are almost surely (since ties occur with probability 0) a random permutation on $\{1, 2, \ldots, n\}$. Such real-numbered data can be assimilated by their ranks to build a binary tree with the aforementioned BST distribution.

A node with no descendants in a BST is a leaf. A node in a BST is said to be a 2-*protected node* if its distance (measured in number of edges) to the nearest descendant leaf is at least 2. Fig. 1 shows a BST of size 9 grown from the permutation (5, 9, 6, 4, 7, 2, 3, 1, 8). The nodes represented by bold circles are 2-protected.

In this note, we investigate the number of 2-protected nodes in a BST. Our program does not stop at the derivation of mean, but continues to find asymptotic distributions.

2. Moments of the number of 2-protected nodes

Let the number of 2-protected nodes in a random BST of size n be X_n. In the tree shown in Fig. 1, $X_9 = 4$. Let U_n be the size in the left subtree of the root, and so $n - 1 - U_n$ is the size of the right subtree. In view of the BST probability model, the root is equally likely to be any of the numbers in the set $\{1, 2, \ldots, n\}$. Thus, U_n is uniformly distributed on the set $\{0, \ldots, n-1\}$, and so symmetrically is $n - 1 - U_n$.

Let $\mathcal{R}_n$ be the event that the root node is *not 2-protected*. Event $\mathcal{R}_n$ occurs if:

- the root is a leaf itself ($n = 1$), or
- both children of the root are leaves (possible when $n = 3$), or
- the root has exactly one child that is a leaf.

For $n \geq 1$, we have a stochastic recurrence for X_n. It is the combined number of 2-protected nodes in the two subtrees of the root, plus 1 (to account for the root being 2-protected) unless $\mathcal{R}_n$ occurs. Thus, we have an equality in distribution, namely,

$$X_n \stackrel{\mathcal{D}}{=} X_{U_n} + \widetilde{X}_{n-1-U_n} + 1 - \mathbf{1}_{\mathcal{R}_n}.$$

(Note: the tilded random variable $\widetilde{X}_{n-1-U_n}$ is conditionally independent of X_{U_n} (given U_n).) The variables X_0, X_1, X_2 are always 0. We are using an *indicator* notation, i.e., $\mathbf{1}_{\mathcal{R}_n} = 1$, if $\mathcal{R}_n$ occurs, and 0 otherwise. Thus, for the moment generating function $\phi_n(t) := \mathbf{E}[e^{X_n t}]$ of X_n, we have

$$\phi_n(t) = \mathbf{E}\left[e^{(\phi_{U_n} + \phi_{n-1-U_n} + 1 - \mathbf{1}_{\mathcal{R}_n})t}\right].$$

When $n \geq 4$, we see that $\mathcal{R}_n$ only occurs if $U_n = 1$ (i.e., the left child of the root is a leaf) or $n - 1 - U_n = 1$ (i.e., the right child of the root is a leaf). (For $n \geq 4$, both children of the root cannot simultaneously be leaves.) Since X_{U_n} and X_{n-1-U_n} are conditionally independent (given U_n), a recurrence ensues by conditioning on U_n. Namely, for $n \geq 4$, we have

$$\begin{aligned}\phi_n(t) &= \frac{e^t}{n} \sum_{\substack{0 \leq k \leq n-1 \\ k \neq 1,\, k \neq n-2}} \phi_k(t)\, \phi_{n-1-k}(t) + \frac{2}{n}\phi_{n-2}(t) \\ &= \frac{e^t}{n} \sum_{k=0}^{n-1} \phi_k(t)\, \phi_{n-1-k}(t) + \frac{2}{n}\phi_{n-2}(t)(1 - e^t). \end{aligned} \tag{1}$$

Differentiating r times with respect to t, then setting $t = 0$, gives a recursion for $\mathbf{E}[X_n^r]$. As r increases, the recurrence equations quickly become more complicated, a phenomenon commonly called the combinatorial explosion. It is sufficient for our purpose to get an exact solution for the recurrence relations for the first two moments, and from there we shall manage to get a shortcut to the higher asymptotic moments.

For $r = 1$ and $n \geq 4$, we obtain a recurrence for the mean, namely,

$$\mathbf{E}[X_n] = \frac{2}{n}\sum_{k=0}^{n-1}\mathbf{E}[X_k] + 1 - \frac{2}{n}. \tag{2}$$

The recurrence for $n\mathbf{E}[X_n]$ can be solved by standard methods such as differencing, for example. If we denote Eq. (2) as $\mathcal{S}(n)$, then for $n \geq 5$, we see $\mathcal{S}(n) - \mathcal{S}(n-1)$ has telescoping sums that disappear, and the resulting linear recurrence can be easily solved, with boundary conditions $\mathbf{E}[X_0] = \mathbf{E}[X_1] = \mathbf{E}[X_2] = 0$, $\mathbf{E}[X_3] = 2/3$, and $\mathbf{E}[X_4] = 5/6$. The boundary condition $n = 4$ agrees with the general form, and we get a simple solution for the mean.

Theorem 2.1. *Let X_n denote the number of 2-protected nodes in a binary search tree grown from a uniformly chosen random permutation of $\{1, \ldots, n\}$. Then we have*

$$\mathbf{E}[X_n] = \frac{11}{30}\,n - \frac{19}{30}, \quad \text{for } n \geq 4.$$

For $r = 2$, we use (1) to develop a recurrence for the second moment:

$$\mathbf{E}[X_n^2] = \frac{2}{n}\sum_{k=0}^{n-1}\mathbf{E}[X_k^2] + \frac{4}{n}\sum_{k=0}^{n-1}\mathbf{E}[X_k] + \frac{2}{n}\sum_{k=0}^{n-1}\mathbf{E}[X_k]\,\mathbf{E}[X_{n-1-k}] - \frac{4}{n}\mathbf{E}[X_{n-2}] + \frac{n-2}{n},$$

valid for $n \geq 4$. With $\mathbf{E}[X_k]$ now determined, we can solve the recurrence for $\mathbf{E}[X_k^2]$. Solving the recurrence for the second moment is not quite as simple as solving that for the first moment. For instance, differencing does not shave off the sums. A more straightforward strategy is to guess a solution, then prove it by induction. This procedure yields the following result.

Theorem 2.2. *Let X_n denote the number of 2-protected nodes in a binary search tree grown from a uniformly chosen random permutation of $\{1, \ldots, n\}$. Then we have*

$$\mathbf{E}[X_n^2] = \frac{121}{900}\,n^2 - \frac{151}{450}\,n + \frac{53}{100}, \quad \text{for } n \geq 8,$$

and the variance follows:

$$\mathbf{Var}[X_n] = \mathbf{E}[X_n^2] - (\mathbf{E}[X_n])^2 = \frac{29}{225}\,n + \frac{29}{225}, \quad \text{for } n \geq 8.$$

Note the exact cancellation of the quadratic terms, leaving only a linear variance, which gives a chance for asymptotic normality to hold, as it fits nicely into the "two moments and a recurrence paradigm" given by Pittel [7].

Moments of arbitrarily high degree can be found similarly. For instance, we have

$$\mathbf{E}[X_n^3] = \frac{1331}{27000}n^3 - \frac{341}{3000}n^2 + \frac{10055641}{27027000}n - \frac{12566959}{27027000}, \quad \text{for } n \geq 12,$$

$$\mathbf{E}[X_n^4] = \frac{14641}{810000}n^4 - \frac{847}{40500}n^3 + \frac{1238257}{7371000}n^2 - \frac{2515391}{6563700}n + \frac{5648494433}{13783770000}, \quad \text{for } n \geq 16,$$

$$\mathbf{E}[X_n^5] = \frac{161051}{24300000}n^5 + \frac{30613}{4860000}n^4 + \frac{15266801}{221130000}n^3 - \frac{527943277}{2756754000}n^2 + \frac{2207719797571}{6110804700000}n - \frac{319695619487}{846111420000}, \quad \text{for } n \geq 20,$$

etc., and in general, we conjecture the following.

Conjecture 2.1. *Let X_n denote the number of 2-protected nodes in a binary search tree grown from a uniformly chosen random permutation of $\{1, \ldots, n\}$. For each fixed integer $k \geq 1$, there exists a polynomial $p_k(n)$ of degree k, the leading term of which is $(11/30)^k$, such that $\mathbf{E}[X_n^k] = p_k(n)$, for all $n \geq 4k$.*

3. Asymptotic normality

The main result of this note is the following.

Theorem 3.1. *Let X_n be the number of 2-protected nodes in a random binary search tree grown from a uniformly chosen random permutation of $\{1, \ldots, n\}$. Then X_n, properly normalized, converges in distribution, namely,*

$$\frac{X_n - \frac{11}{30}\,n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{29}{225}\right).$$

Proof. Let $X_n^* = n^{-1/2}(X_n - \frac{11}{30}n)$, and

$$\phi_{X_n^*}(t) = \mathbf{E}\left[\exp\left(\frac{X_n - \frac{11}{30}n}{\sqrt{n}}t\right)\right] = \phi_{X_n}\left(\frac{t}{\sqrt{n}}\right)\exp\left(-\frac{11}{30}t\sqrt{n}\right),$$

be its moment generating function. The recurrence (1) can be "normalized" in the form

$$\begin{aligned}\phi_{X_n}\left(\frac{u}{\sqrt{n}}\right)\exp\left(-\frac{11}{30}u\sqrt{n}\right) = {} & \frac{\exp\left(\frac{u}{\sqrt{n}}\right)\exp\left(-\frac{11u}{30\sqrt{n}}\right)}{n} \\ & \times \sum_{k=0}^{n-1}\phi_{X_k}\left(\frac{u}{\sqrt{n}}\right)\exp\left(-\frac{11ku}{30\sqrt{n}}\right)\phi_{\widetilde{X}_{n-1-k}}\left(\frac{u}{\sqrt{n}}\right)\exp\left(-\frac{11(n-1-k)u}{30\sqrt{n}}\right) \\ & + \frac{2}{n}\phi_{X_{n-2}}\left(\frac{u}{\sqrt{n}}\right)\exp\left(-\frac{11}{30}u\sqrt{n}\right)\left(1-\exp\left(\frac{u}{\sqrt{n}}\right)\right),\end{aligned}$$

which we can write as

$$\begin{aligned}\phi_{X_n^*}(u) = {} & \frac{\exp\left(\frac{u}{\sqrt{n}}\right)\exp\left(-\frac{11u}{30\sqrt{n}}\right)}{n}\sum_{k=0}^{n-1}\phi_{X_k^*}\left(u\sqrt{\frac{k}{n}}\right)\phi_{X_{n-1-k}^*}\left(u\sqrt{\frac{n-1-k}{n}}\right) \\ & + \frac{2}{n}\phi_{X_{n-2}^*}\left(u\sqrt{\frac{n-2}{n}}\right)\exp\left(\frac{11}{30}u\sqrt{n-2} - \frac{11}{30}u\sqrt{n}\right)\left(1-\exp\left(\frac{u}{\sqrt{n}}\right)\right).\end{aligned}$$

In view of Pittel's paradigm [7], a limit $\phi_{X^*}(u)$ (the moment generating function of a limiting random variable X^*) exists, as $n \to \infty$. Passage to the limit in the latter relation yields

$$\phi_{X^*}(u) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\phi_{X^*}\left(u\sqrt{\frac{k}{n}}c\right)\phi_{X^*}\left(u\sqrt{\frac{n-1-k}{n}}\right) + \lim_{n\to\infty}O(n^{-3/2}).$$

Put $k/n = x_{k,n}$ to represent the last relation as

$$\phi_{X^*}(u) = \lim_{n\to\infty}\sum_{x_{k,n}=0}^{(n-1)/n}\phi_{X^*}(u\sqrt{x_{k,n}})\phi_{X^*}(u\sqrt{x_{n-1-k,n}})\Delta x_{k,n},$$

where $\Delta x_{k,n} = x_{k,n} - x_{k-1,n}$ is the difference operator, and the summation index $x_{k,n}$ moves up in increments of size $1/n$. By the usual interpretation of Riemann integrals, we finally write

$$\phi_{X^*}(u) = \int_{y=0}^{1}\phi_{X^*}(u\sqrt{y})\,\phi_{X^*}(u\sqrt{1-y})\,dy.$$

This integral functional equation has the function $e^{c^2u^2/2}$ as a solution. This function is the moment generating function of the normal $\mathcal{N}(0, c^2)$ random variable. By Lévy's continuity theorem we get the desired convergence in distribution:

$$\frac{X_n - \frac{11}{30}n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, c^2),$$

for an appropriate value of c^2. Of course, it must be $\frac{29}{225}$, the coefficient of the leading asymptotic term in the variance. □

4. Extended binary search trees

BSTs are often extended by adding special *external* nodes as children. A sufficient number of these external nodes are supplied to each original node (now thought of as internal) to make its outdegree equal to two. In this variant, *the* 2*-protected nodes are cushioned from the external nodes by at least one internal node*. As an example, see Fig. 2, in which we have added the external nodes to the tree in Fig. 1, and we have again noted the 2-protected nodes for this modified model in bold.

If X_n denotes the number of 2-protected nodes in extended binary trees, then we have

$$\mathbf{E}[X_n] = \frac{1}{3}n - \frac{2}{3}, \quad \text{for } n \geq 2,$$

and

$$\mathbf{Var}[X_n] = \frac{2}{45}n + \frac{2}{45}, \quad \text{for } n \geq 4.$$

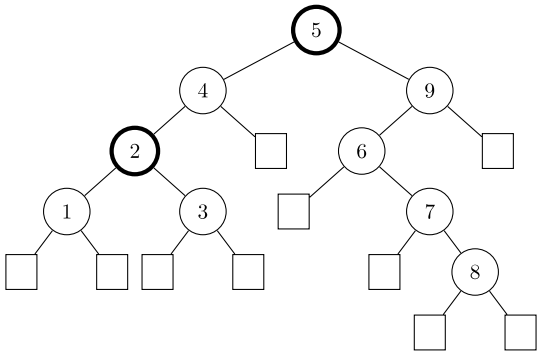

Fig. 2. Example of an *extended* binary search tree for the permutation (5, 9, 6, 4, 7, 2, 3, 1, 8); 2-protected nodes in bold.

The corresponding central limit result is

$$\frac{X_n - \frac{1}{3}n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{2}{45}\right).$$

These results can be obtained by very similar methods as those we applied to the unextended BST. However, most of these results for extended BSTs are already implied in the published literature. For instance, in the extended BST the 2-protected nodes are the nodes of outdegree 2 in the tree before it got extended. The exact average of these appears in [8]. The asymptotic distribution appears in [9], where he uses an m-dependent central limit theorem for stationary random variables, due to Hoeffding and Robbins [10]. Mahmoud [11] gives an account of a proof based on modeling by Pólya urn models. The only thing new here is the *exact* variance, which we get via the exact second moment,

$$\mathbf{E}[X_n^2] = \frac{1}{9}n^2 - \frac{2}{5}n + \frac{22}{45}, \quad \text{for } n \geq 4.$$

Another new aspect is that we can again use our recursive methods to develop exact higher moments for the number of 2-protected nodes in an extended BST, e.g.,

$$\mathbf{E}[X_n^3] = \frac{1}{27}n^3 - \frac{8}{45}n^2 + +\frac{376}{945}n - \frac{122}{315}, \quad \text{for } n \geq 6,$$
$$\mathbf{E}[X_n^4] = \frac{1}{81}n^4 - \frac{28}{405}n^3 + \frac{2984}{14175}n^2 - \frac{5458}{14175}n + \frac{218}{675}, \quad \text{for } n \geq 8,$$
$$\mathbf{E}[X_n^5] = \frac{1}{243}n^5 - \frac{2}{81}n^4 + \frac{764}{8505}n^3 - \frac{94}{405}n^2 + \frac{6956}{18711}n - \frac{8654}{31185}, \quad \text{for } n \geq 10,$$

etc. In general, we conjecture the following.

Conjecture 4.1. *Let X_n denote the number of 2-protected nodes in an extended binary search tree grown from a uniformly chosen random permutation of $\{1, \ldots, n\}$. For each fixed integer $k \geq 1$, there exists a polynomial $p_k(n)$ of degree k, the leading term of which is $1/3^k$, such that $\mathbf{E}[X_n^k] = p_k(n)$, for all $n \geq 2k$.*

Acknowledgments

This research was done while the first author was visiting Purdue University. The support the first author received from Purdue's Department of Statistics is greatly appreciated. The second author was supported by NSF Science & Technology Center for Science of Information Grant CCF-0939370.

References

[1] G.-S. Cheon, L.W. Shapiro, Protected points in ordered trees, Applied Mathematics Letters 21 (2008) 516–520.
[2] T. Mansour, Protected points in k-ary trees, Applied Mathematics Letters 24 (2011) 478–480.
[3] R.R. Du, H. Prodinger, Notes on protected nodes in digital search trees, Applied Mathematics Letters 25 (2012) 1025–1028.
[4] D.E. Knuth, The Art of Computer Programming, second ed., in: Sorting and Searching, vol. 3, Addison-Wesley, Reading, MA, 1998, Originally published in 1973.
[5] H.M. Mahmoud, Sorting: A Distribution Theory, Wiley, New York, NY, 2000.
[6] H.M. Mahmoud, Evolution of Random Search Trees, Wiley, New York, NY, 1992.
[7] B. Pittel, Normal convergence problem? two moments and a recurrence may be the clues, The Annals of Applied Probability 9 (1999) 1260–1302.
[8] H.M. Mahmoud, The expected distribution of degrees in random binary search trees, The Computer Journal 29 (1986) 36–37.
[9] L. Devroye, Limit laws for local counters in random binary search trees, Random Structures and Algorithms 2 (1991) 303–315.
[10] W. Hoeffding, H. Robbins, The central limit theorem for dependent random variables, Duke Mathematical Journal 15 (1948) 773–780.
[11] H.M. Mahmoud, Pólya Urn Models, Chapman, Orlando, FL, 2008.