

# More combinatorial properties of certain trees

By William C. Lynch\*

A detailed examination of binary search trees reveals that the probability of making precisely  $i$  comparisons in placing the  $(n-1)$ th item in the tree is related to the  $(n-i)$ th symmetric function of the integers  $1, \dots, n$ . A recurrence relation for the moments of this distribution of comparisons is derived, and formulas for the mean number of comparisons and its variance are displayed. These are shown to be in accord with previously published values.

Douglas (1959), Windley (1960), and later Hibbard (1962) introduced binary search trees and gave their applications to sorting, searching, and file maintenance. Windley derives formulas for the mean and variance of the number of comparisons to insert an item, and Hibbard arrives at a similar formula for the mean. Windley remarks that both are difficult to calculate since their defining recurrence relations are unstable.

This paper will show that a deeper analysis of binary search trees reveals some very beautiful mathematical properties. The result of Hibbard will be contained as a special case of the results of this paper. Quantities closely related to the higher moments will be exhibited. The derived recurrence relations will have solutions in integers so that computational difficulties will be lessened.

A binary search tree is constructed in the following way. Items are selected from an infinite pool and inserted one by one into the search tree. These items have sort keys randomly and uniformly distributed in the interval zero to one.

Each node in the tree has potentially two successors, a left successor and a right successor. The first item selected becomes the root of the tree. As a new item is selected its key is compared to the key of the root node. If the selected key is less we pass to the left successor; if greater or equal, to the right successor. The process is then repeated for this new node. If the appropriate successor does not exist, the selected item is inserted as the missing successor, and the process terminates for that item. Fig. 1 gives a sequence of integers and the binary search tree generated by that sequence.

## 1. Construction of the fundamental equation for the number of comparisons

### 1.1. Definitions

We will assume that we are drawing keys from a uniform distribution on  $[0, 1]$ . Let  $P_i(n)$  be the probability of making precisely  $i$  comparisons in placing the randomly drawn  $(n+1)$ th element into a binary search tree containing precisely  $n$  elements. Clearly  $P_0(n) = \delta_{0n}$  and  $P_i(0) = \delta_{i0}$ .

\* Case Institute of Technology, Cleveland, Ohio.

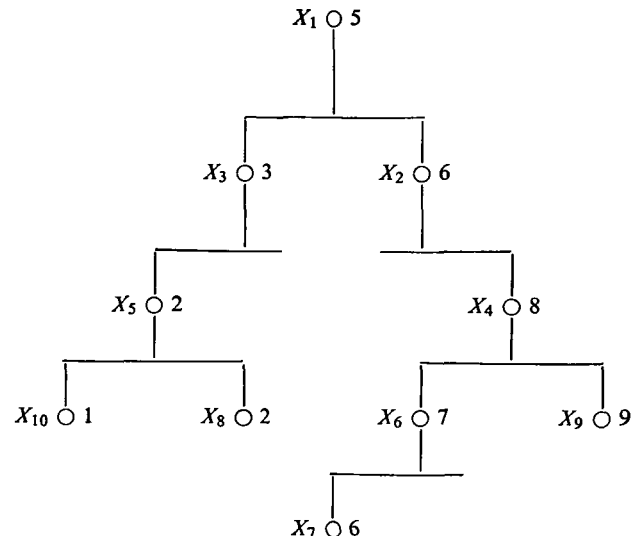


Fig. 1.—Example of a sequence of integers and its associated binary search tree

Let  $P_i(n)|X$  be the probability of making  $i$  comparisons given that the key of the root has value  $X$ .

Let  $P_i(n)|X, k$  be the probability of making  $i$  comparisons given that the root has value  $X$ , and that there are precisely  $k$  nodes to the left of the root (and, of course, precisely  $n - k - 1$  to the right).

We will tacitly assume  $n \geq 1$ . Either  $i \geq 1$  or  $n = i = 0$  or  $P_i(n) = 0$ .

### 1.2. Derivation of the recurrence equations for $P_i(n)$

By independence and mutual exclusion

$$P_i(n)|X, k = X \cdot P_{i-1}(k) + (1 - X) \cdot P_{i-1}(n - k - 1).$$

Again by mutual exclusion

$$P_i(n)|X = \sum_{k=0}^{n-1} \binom{n-1}{k} X^k (1 - X)^{n-1-k} \cdot P_i(n)|X, k.$$

Sequence  $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$   
5, 6, 3, 8, 2, 7, 6, 2, 9, 1

And again by mutual exclusion, independence, and the uniform distribution

$$P_i(n) = \int_0^1 P_i(n) |X| \cdot dX.$$

Putting these together we have

$$P_i(n) = \int_0^1 \left( \sum_{k=0}^{n-1} \binom{n-1}{k} X^k (1-X)^{n-1-k} \cdot [X \cdot P_{i-1}(k) + (1-X) \cdot P_{i-1}(n-k-1)] \right) dX.$$

Rearranging,

$$P_i(n) = \sum_{k=0}^{n-1} P_{i-1}(k) \cdot \binom{n-1}{k} \int_0^1 X^{k+1} (1-X)^{n-1-k} dX + \sum_{k=0}^{n-1} P_{i-1}(n-k-1) \binom{n-1}{k} \int_0^1 X^k (1-X)^{n-k} dX.$$

Since

$$\int_0^1 X^k (1-X)^{n-k} dX = 1 / \left[ (n+1) \binom{n}{k} \right]$$

we have

$$\begin{aligned} P_i(n) &= \sum_{k=0}^{n-1} \frac{k+1}{n(n+1)} \cdot P_{i-1}(k) \\ &\quad + \sum_{k=0}^{n-1} \frac{n-k}{n(n+1)} P_{i-1}(n-1-k) \\ &= \frac{2}{n(n+1)} \sum_{k=0}^{n-1} (k+1) P_{i-1}(k). \end{aligned}$$

Reducing  $n$  by 1 we have

$$P_i(n-1) = \frac{2}{n(n-1)} \sum_{k=0}^{n-2} (k+1) P_{i-1}(k)$$

so that

$$P_i(n) = \frac{2}{n+1} P_{i-1}(n-1) + \frac{n-1}{n+1} P_i(n-1).$$

This is our desired result.

### 1.3. Rearrangement of the equation for $P_i(n)$

Let

$$S_i(n) = \frac{(n+2)!}{2^{n+1-i}} P_{n+1-i}(n+1).$$

Hence

$$P_i(n) = \frac{2^i}{(n+1)!} S_{n-i}(n-1).$$

This and equation (1) give us

$$S_i(n) = S_i(n-1) + n \cdot S_{i-1}(n-1). \quad (2)$$

Now we wish to find boundary conditions.

$$S_i(0) = \frac{2!}{2^{1-i}} P_{1-i}(1) = 2^i P_{1-i}(1),$$

$$\begin{aligned} P_{1-i}(1) &= \frac{2}{2} P_{-i}(0) + \frac{0}{2} P_{1-i}(0) \quad \text{by (1)} \\ &= P_{-i}(0) = \delta_{-i0} = \delta_{i0}. \end{aligned}$$

Hence

$$S_i(0) = 2^i \delta_{i0} = \delta_{i0}.$$

For the other boundary condition

$$S_0(n) = \frac{(n+2)!}{2^{n+1}} P_{n+1}(n+1),$$

$$P_{n+1}(n+1) = \frac{2}{n+2} P_n(n) + \frac{n}{n+2} P_{n+1}(n) \quad \text{by (1).}$$

But we cannot make  $n+1$  comparisons in a tree that contains only  $n$  elements. Hence  $P_{n+1}(n) = 0$ .

Hence

$$P_{n+1}(n+1) = \frac{2}{(n+2)} P_n(n).$$

Since  $P_0(0) = 1$  clearly we have

$$P_{n+1}(n+1) = \frac{2^{n+1}}{(n+2)!}$$

so that

$$S_0(n) = \frac{(n+2)!}{2^{n+1}} \cdot \frac{2^{n+1}}{(n+2)!} = 1.$$

### 1.4. Interpretation of equation (2)

We will now try to interpret equation (2).

Consider the following derivation of the Pascal triangle identity.

$$\binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1}. \quad (3)$$

To find  $\binom{n}{i}$ , the number of combinations of  $n$  things taken  $i$  at a time we partition the combinations into two classes. Class 1 will be all combinations which do not contain the  $n$ th object. Class 2 will be all that do contain the  $n$ th object. Since we must choose  $i$  things from  $n-1$  objects to get a combination in class 1, class 1 contains  $\binom{n-1}{i}$  objects. Each combination in class 2 has already chosen the  $n$ th object. We must now choose  $i-1$  from the remaining  $n-1$ . Hence class 2 contains  $\binom{n-1}{i-1}$  combinations and (3) is proven.

Now consider  $Q_i(n)$ , the  $i$ th symmetric function on  $X_1, X_2, \dots, X_n$ . This function is constructed by selecting  $i$  numbers from this list, multiplying them together, and summing over the  $\binom{n}{i}$  possible distinct products. The terms of this summation may be partitioned into two classes. Class 1 will be those that do not contain  $X_n$  as a factor. Class 2 will be those terms that do contain  $X_n$  as a factor.

The terms in class 1 and the terms of  $Q_i(n-1)$  are identical. If we delete  $X_n$  from each term of class 2, these new terms are identical to those of  $Q_{i-1}(n-1)$ . Hence

$$Q_i(n) = Q_i(n-1) + X_n Q_{i-1}(n-1).$$

Clearly  $Q_0(n) = 1$  since there is only one way to select no terms. The product over the empty set is 1. It is also clear that  $Q_i(0) = \delta_{i0}$ .

Since the solution of (2) under the given boundary conditions is clearly unique we have that  $S_i(n)$  is the  $i$ th symmetric function of the integers  $1, 2, 3, \dots, n$ .

It is well known that if

$$\sum_{i=0}^n a_i x^i = (x + \rho_1)(x + \rho_2)(\dots)(x + \rho_n),$$

then  $a_i$  is the  $(n-1)$ th symmetric function on  $\rho_1, \rho_2, \dots, \rho_n$ . The  $\rho$ 's are the negatives of the roots. It is clear that  $(x+1)(x+2)(\dots)(x+n)$  is the generating function for  $S_{n-i}(n)$ .

Also

$$\begin{aligned} \sum_{i=0}^n P_i(n) &= \frac{1}{(n+1)!} \cdot \sum_{i=1}^n S_{n-i}(n-1) \cdot 2^i + \frac{P_0(n)}{(n+1)!} \\ &= \frac{2}{(n+1)!} \cdot \sum_{i=0}^{n-1} S_{(n-1)-i}(n-1) \cdot 2^i + \frac{0}{(n+1)!} \\ &= \frac{2}{(n+1)!} \cdot (2+1)(2+2)(\dots)(2+n-1) \\ &= \frac{(n+1)!}{(n+1)!} = 1 \end{aligned} \quad (4)$$

as it should be.

## 2. Construction of the moments $P_i(n)$

We observe that  $\binom{i}{j}$  is a  $j$ th-degree polynomial in  $i$  and that the set of polynomials  $\left\{ \binom{i}{j} \mid 0 \leq j \leq i \right\}$  are linearly independent. They are transformed into the linearly independent set of polynomials  $\{i^j \mid 0 \leq j \leq i\}$  by means of the Stirling numbers.

Instead of computing the  $j$ th moment,

$$\mu_j(n) = \sum_{i=0}^n i^j P_i(n),$$

we will compute

$$\mu'_j(n) = \sum_{i=0}^n \binom{i}{j} P_i(n).$$

The  $\mu_j$ 's are then connected to the  $\mu'_j$ 's by means of the Stirling numbers.

Applying (1)

$$\begin{aligned} \mu'_j(n) &= \sum_{i=0}^n \binom{i}{j} P_i(n) = \frac{2}{n+1} \sum_{i=0}^n \binom{i}{j} P_{i-1}(n-1) \\ &\quad + \frac{n-1}{n+1} \sum_{i=0}^n \binom{i}{j} P_i(n-1). \end{aligned}$$

Since  $P_n(n-1) = 0$ ,

$$\frac{n-1}{n+1} \sum_{i=0}^n \binom{i}{j} P_i(n-1) = \frac{n-1}{n+1} \sum_{i=0}^{n-1} \binom{i}{j} P_i(n-1)$$

$$= \frac{n-1}{n+1} \mu'_j(n-1).$$

On the other hand,

$$\begin{aligned} \sum_{i=0}^n \binom{i}{j} P_{i-1}(n-1) &= \sum_{i=-1}^{n-1} \binom{i+1}{j} P_i(n-1) \\ &= \sum_{i=0}^{n-1} \binom{i+1}{j} P_i(n-1) \text{ since } P_{-1}(n-1) = 0 \\ &= \sum_{i=0}^{n-1} \binom{i}{j} P_i(n-1) + \sum_{i=0}^{n-1} \binom{i}{j-1} P_i(n-1) \\ &= \mu'_j(n-1) + \mu'_{j-1}(n-1). \end{aligned}$$

We then have

$$\begin{aligned} \mu'_j(n) &= \frac{2}{n+1} \mu'_j(n-1) + \frac{2}{n+1} \mu'_{j-1}(n-1) \\ &\quad + \frac{n-1}{n+1} \mu'_j(n-1) \\ &= \mu'_j(n-1) + \frac{2}{n+1} \mu'_{j-1}(n-1). \end{aligned}$$

From (4) we have

$$\begin{aligned} \mu'_0(n) &= \sum_{i=0}^n \binom{i}{0} P_i(n) = \sum_{i=0}^n P_i(n) = 1, \\ \mu'_j(0) &= \sum_{i=0}^0 \binom{0}{j} P_i(0) = \binom{0}{j} P_0(0) = \binom{0}{j} = \delta_{j0}. \end{aligned}$$

We then conclude that  $\mu'_j(n)$  is then the  $j$ th symmetric function of  $\frac{2}{2}, \frac{2}{3}, \dots, \frac{2}{n+1}$ .

We may now look into  $\mu_1(n)$ ,  $\mu_2(n)$  and  $\text{var}(n)$ . We first observe that

$$(\mu'_1(n))^2 = 2\mu'_2(n) + \sum_{i=2}^{n+1} \frac{4}{i^2}.$$

Also

$$\mu'_2(n) = \sum_{i=0}^n \frac{i(i-1)}{2} P_i(n) = \frac{1}{2} \mu_2(n) - \frac{1}{2} \mu_1(n)$$

and

$$\mu'_1(n) = \sum_{i=0}^n i P_i(n) = \mu_1(n) = \sum_{i=2}^{n+1} \frac{1}{i}$$

so that

$$\begin{aligned} \mu_1(n) &= \mu'_1(n), \\ \mu_2(n) &= 2\mu'_2(n) + \mu'_1(n), \\ \text{var}(n) &= \sum_{i=0}^n (i - \mu_1(n))^2 P_i(n) \\ &= \mu_2(n) - 2(\mu_1(n))^2 + (\mu_1(n))^2 \\ &= \mu_2(n) - (\mu_1(n))^2 \\ &= 2\mu'_2(n) + \mu'_1(n) - (\mu'_1(n))^2 \\ &= 2\mu'_2(n) + \mu'_1(n) - 2\mu'_2(n) - \sum_{i=2}^{n+1} \frac{4}{i^2}, \end{aligned}$$

$$\text{var}(n) = \mu'_1(n) - 4 \sum_{i=2}^{n+1} i^{-2}.$$

$$\text{var}(n) \sim 2(\gamma - 1 + \ln(n+1)) - 4\left(\frac{\pi^2}{6} - 1\right)$$

so that

$$\mu_1(n) \sim \ln(n+1)^2 - 0.845 \dots,$$

$$\text{var}(n) \sim \ln(n+1)^2 - 3.425 \dots$$

For large  $n$ ,  $\mu_1(n) \sim 2(\gamma - 1 + \ln(n+1))$  where  $\gamma$  is Euler's constant

## References

- DOUGLAS, A. S. (1959). "Techniques for the Recording of, and Reference to data in a Computer", *The Computer Journal*, Vol. 2, p. 1.
- HIBBARD, T. (1962). "Some Combinatorial Properties of Certain Trees," *J. Assoc. Comp. Mach.*, Vol. 9, p. 13.
- WINDLEY, P. F. (1960). "Trees, Forests, and Rearranging," *The Computer Journal*, Vol. 3, p. 84.

## Data Transmission Handbook

The *Data Transmission Handbook 1964*, produced by the Data Transmission Committee of The British Computer Society, will be sent free to members early in 1965.

Additional copies may be purchased at 13s. 6d. per copy post free; this offer also applies to non-society members. As the number of copies from the first printing is limited, those requiring extra copies are advised to place their orders now with the Assistant Secretary, The British Computer Society, Finsbury Court, Finsbury Pavement, London, E.C.2., accompanied by a remittance.

## THE COMPUTING AND DATA PROCESSING SOCIETY OF CANADA

Mr. R. L. Sutton, Editor-in-Chief, reports that the *Quarterly Bulletin* of the Canadian Society has now reached its fifth volume. Members of The British Computer Society may subscribe to the publication at the special price of £1 per annum. Orders (with remittance), should be sent to The Assistant Secretary, The British Computer Society, Finsbury Court, Finsbury Pavement, London, E.C.2.

Readers in North America who wish for particulars, should write direct to The Computing and Data Processing Society of Canada, c/o Mr. R. L. Sutton, Confederation Life Association, 321 Bloor Street East, Toronto 5, Ontario.