ON PROTECTED NODES IN DIGITAL SEARCH TREES

ROSENA R.X. DU AND HELMUT PRODINGER

Dedicated to Philippe Flajolet (1948–2011)

Abstract. Recently, 2-protected nodes were studied in the context of ordered trees and k-trees. These nodes have a distance of at least 2 to each leaf. Here, we study digital search trees, which are binary trees, but with a different probability distribution underlying. Our result says, that *grosso modo* some 31% of the nodes are 2-protected. Methods include exponential generating functions, contour integration, and some elements from q-analysis.

1. Introduction

Cheon and Shapiro [2] started the study of 2-protected nodes in trees. A node enjoys this property if its distance to any leaf is at least 2. A simpler notion is 1-protected: exactly the nodes that are not leaves are 1-protected. In the cited paper, the family of ordered trees was considered, and it was found that asymptotically a proportion of $\frac{1}{6}$ of the nodes is 2-protected. Recently, Mansour [9] complemented these results by studying k-ary trees.

In the present note, we study the analogous quantity for *Digital Search Trees* (DSTs), a structure that is important in Computer Science [7]. As trees, they are binary trees, but the (probability) distribution is quite different. From a mathematical point of view, they always lead to interesting and nontrivial considerations, with a flair of q-analysis. Here are a few papers of relevance: [3, 10, 6, 8, 5] DSTs are constructed as follows. Given a sequence of binary strings, we place the first in the root node; those starting with "0" ("1") are directed to the left (right) subtree of the root, and are constructed recursively by the same procedure but with the removal of their first bits when comparisons are made. See Figure 1 for an illustration.

In the following section we will show that the proportion of 2-protected nodes in the DST model is about 31%; a more detailed statement will be given later.

We collect here are few notations. These quantities belong to the realm of q-series and can be found in [1], although with a slightly different notation:

$$Q_m=\prod_{k=1}^{m}\Big(1-\frac{1}{2^k}\Big),\quad Q_\infty=\prod_{k=1}^{\infty}\Big(1-\frac{1}{2^k}\Big),\quad Q(x)=\prod_{k=1}^{\infty}\Big(1-\frac{x}{2^k}\Big).$$

There is a formula that is equivalent to one of Euler's partition identities:

$$Q(t)=\sum_{m\ge0}a_{m+1}t^m\qquad\text{with}\qquad a_{m+1}=\frac{(-1)^m2^{-\binom{m+1}{2}}}{Q_m}.$$

Finally, we will use $L=\log 2$.

$A : 1001$
$B : 0110$
$C : 0000$
$D : 1111$
$E : 0100$
$F : 0101$
$G : 1101$
$H : 1110$
$I : 1100$

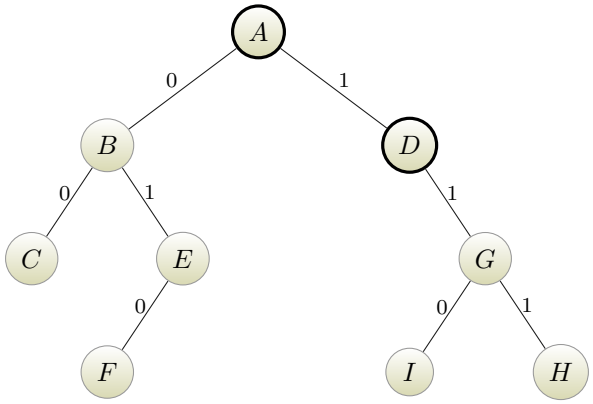

FIGURE 1. *A digital search tree with nine nodes, among which A and D are 2-protected.*

2. Average number of 2-protected nodes

Denote by l_n the average number of 2-protected nodes in a random DST, built from n data. By random we mean that whenever a decision has to be made whether to go down to the left or right, a fair coin is tossed, and a direction is chosen with probability $\frac{1}{2}$.

The following recursion follows from the observation that, provided we have $n+1$ data, k go to the left and $n-k$ go to the right, and such a split happens with probability $\binom{n}{k}2^{-n}$. One node goes to the root and is always 2-protected except in the instances $k = 1$ or $k = n-1$. Therefore

$$\begin{aligned} l_{n+1} &= \sum_{k=0}^{n}\binom{n}{k}2^{-n}\Big(l_k + l_{n-k} + 1\Big) - \sum_{k=1 \text{ or } n-1}\binom{n}{k}2^{-n} \\ &= 1 + 2^{1-n}\sum_{k=0}^{n}\binom{n}{k}l_k - n2^{1-n}. \end{aligned}$$

This recursion is true for $n \geq 3$, with initial conditions $l_0 = l_1 = l_2 = 0$, $l_3 = \frac{1}{2}$. Our treatment follows [3]. We introduce the exponential generating function $L(z) = \sum_{n\geq 0} l_n z^n/n!$ and translate the recursion:

$$\sum_{n\geq 3} l_{n+1}\frac{z^n}{n!} = \sum_{n\geq 3}\frac{z^n}{n!} + \sum_{n\geq 3}\frac{z^n}{n!}\sum_{k=0}^{n}\binom{n}{k}2^{1-n}l_k - \sum_{n\geq 3} n2^{1-n}\frac{z^n}{n!}$$

or

$$\sum_{n\geq 0} l_{n+1}\frac{z^n}{n!} - l_3\frac{z^2}{2!} = \sum_{n\geq 3}\frac{z^n}{n!} + \sum_{n\geq 0}\frac{z^n}{n!}\sum_{k=0}^{n}\binom{n}{k}2^{1-n}l_k - \sum_{n\geq 3} n2^{1-n}\frac{z^n}{n!},$$

which leads after some simple manipulations to

$$L'(z) = e^z - ze^{z/2} - 1 + \frac{z^2}{4} + 2e^{z/2}L(\frac{z}{2}).$$

Now we introduce the *Poisson generating function* $M(z) = e^{-z}L(z) = \sum_{n\geq 0} m_n z^n/n!$ and rewrite the equation:

$$M'(z) + M(z) = 1 - ze^{-z/2} - e^{-z} + \frac{z^2}{4}e^{-z} + 2M(\frac{z}{2}).$$

For $n \geq 1$, we can read off the coefficients of $z^n/n!$:

$$m_{n+1} = -(1-2^{1-n})m_n + n(-1)^n 2^{1-n} - (-1)^n + \frac{n(n-1)}{4}(-1)^n.$$

In order to solve it, we rewrite it as

$$\frac{m_{n+1}(-1)^n}{Q_{n-1}} = \frac{m_n(-1)^{n-1}}{Q_{n-2}} + \frac{n2^{1-n} - 1 + \frac{n(n-1)}{4}}{Q_{n-1}},$$

which can be summed and leads to

$$\frac{m_{N+1}(-1)^N}{Q_{N-1}} = \sum_{n=2}^{N} \frac{n2^{1-n} - 1 + \frac{n(n-1)}{4}}{Q_{n-1}}$$

and eventually to

$$m_N = Q_{N-2}(-1)^N \sum_{n=1}^{N-2} \frac{1 - (n+1)2^{-n} - \frac{n(n+1)}{4}}{Q_n}.$$

Since

$$l_n = \sum_{k=2}^{n} \binom{n}{k} m_k$$

we found the following explicit formula that we formulate as a theorem.

Theorem 1. *The average number of 2-protected nodes in random DSTs of size $N \geq 1$ is exactly given by*

$$l_N = \sum_{k=2}^{N} \binom{N}{k} (-1)^k Q_{k-2} \sum_{n=1}^{k-2} \frac{1 - (n+1)2^{-n} - \frac{n(n+1)}{4}}{Q_n}.$$

Now we turn to the asymptotic evaluation of l_N as $N \to \infty$. Again, we follow the approach in [3] and use Rice's integrals, which means that we are able to rewrite l_N as a contour integral. Changing the contour of integration and collecting residues produces the asymptotic expansion of interest. Many examples have been described in [4]. In order to do so, one must extend the function

$$Q_{k-2} \sum_{n=1}^{k-2} \frac{1 - (n+1)2^{-n} - \frac{n(n+1)}{4}}{Q_n}$$

so that is makes sense for any complex k, not just integers. This will be discussed now.

We have $Q_{k-2} = Q_\infty/Q(2^{2-k})$, and this makes sense for any k. Now we have, using Euler's identity mentioned in the Introduction,

$$\frac{1}{Q_n} = \frac{Q(2^{-n})}{Q_\infty} = \frac{1}{Q_\infty} \sum_{m\geq 0} a_{m+1} 2^{-nm},$$

and this makes sense for any n, since the smallness of the a_m's handles all convergence issues. Therefore

$$\sum_{n=1}^{k-2} \frac{1-(n+1)2^{-n}-\frac{n(n+1)}{4}}{Q_n} = \frac{1}{Q_\infty}\sum_{m\ge 0} a_{m+1} \sum_{n=1}^{k-2} \Big[1-(n+1)2^{-n}-\frac{n(n+1)}{4}\Big]2^{-nm}.$$

The inner sum (on n) can be explicitly evaluated, but since it is long and ugly, we don't display it here, but the resulting form (that we keep in our Maple calculation) can be used for any $k \in \mathbb{C}$.

The integral expression is

$$l_N = -\frac{1}{2\pi i}\int_{\mathcal{C}} \frac{\Gamma(N+1)\Gamma(-z)}{\Gamma(N+1-z)}\psi(z)dz,$$

where $\mathcal{C}$ encircles the poles $2, 3, \dots, N$ and no others. The function $\psi(z)$ is the extension of

$$Q_{k-2}\sum_{n=1}^{k-2} \frac{1-(n+1)2^{-n}-\frac{n(n+1)}{4}}{Q_n}$$

as just discussed. Changing the contour, one encounters other poles. They must be subtracted and produce the asymptotic expansion that we need. The main contribution comes from $z = 1$. There are also poles at $z = 1 + \chi_k$, with $\chi_k = \frac{2\pi i k}{L}$, and they contribute a tiny oscillating function $N \cdot \delta(\log_2 N)$, where the amplitude of $\delta(x)$ is typically smaller than 10^{-5}. In order to keep this note short and crisp, we refrain from computing this function explicitly. It is not difficult, and there are many similar examples in the literature. So we concentrate now on $z = 1$, and we will find a simple pole. As a first step, we consider

$$\lim_{k\to 1} \frac{Q_{k-1}}{1-2^{1-k}} \sum_{n=1}^{k-2} \Big[1-(n+1)2^{-n}-\frac{n(n+1)}{4}\Big]2^{-nm}.$$

This limit can be computed by Maple, with the result

$$b_m := \frac{1}{4L}\frac{B(2^{-m})}{(2^{-m}-1)^3(2^{-m}-2)^2}$$

and $B(x) := 16L - 48xL + 48x^2L - 16x^3L - 20 + 60x - 69x^2 + 36x^3 - 7x^4 - 8x\log(x) + 12x^2\log(x) - 10x^3\log(x) + 4x^4\log(x)$. Note that b_0 is interpreted as a limit:

$$b_0 = \frac{37}{12L} - 4.$$

So we are left with the negative residue of

$$-\frac{\Gamma(N+1)\Gamma(-z)}{\Gamma(N+1-z)}$$

at $z = 1$, which is just N. Summarizing, we found the asymptotic behaviour.

Theorem 2. *The average number l_N of 2-protected nodes in random DSTs of size N admits the asymptotic expansion*

$$l_N = N \cdot \frac{1}{Q_\infty} \sum_{m \geq 0} a_{m+1} b_m + N \cdot \delta(\log_2 N) + O(1),$$

where the numerical constant evaluates to $0.30707981393605921828549\ldots$. *The tiny periodic function $\delta(x)$ has a Fourier expansion that could be computed in principle. The remainder term $O(1)$ stems from the next pole at $z = 0$.*

For example, $l_{500}/500 = 0.305710\ldots$.

Remark. Flajolet and Sedgewick in [3] solved an open problem of Knuth [7], and considered the number of endnodes. They found this to be on average as $\beta \cdot N$, with $\beta = 0.372046812\ldots$. Again, there are tiny oscillations. The quantity $(1-\beta)N$ is (asymptotically) the number of 1-protected nodes. So, there are roughly 63% 1-protected nodes, and our new results say that there are about 31% 2-protected nodes.

Acknowledgement. The first author is partially supported by the National Science Foundation of China under Grant 10801053, and the Shanghai Rising-Star Program (No. 10QA1401900). The second author is supported by an International Science and Technology Agreement (Grant 67215) from the NRF (South Africa).

References

[1] George E. Andrews. *The theory of partitions*. Addison-Wesley Publishing Co., Reading, Mass.-London-Amsterdam, 1976. Encyclopedia of Mathematics and its Applications, Vol. 2.

[2] Gi-Sang Cheon and Louis W. Shapiro. Protected points in ordered trees. *Appl. Math. Lett.*, 21(5):516–520, 2008.

[3] Philippe Flajolet and Robert Sedgewick. Digital search trees revisited. *SIAM J. Comput.*, 15(3):748–767, 1986.

[4] Philippe Flajolet and Robert Sedgewick. Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoret. Comput. Sci.*, 144(1-2):101–124, 1995. Special volume on mathematical analysis of algorithms.

[5] Hsien-Kuei Hwang, Michael Fuchs, and Vytas Zacharovas. Asymptotic variance of random symmetric digital search trees. *Discrete Math. Theor. Comput. Sci.*, 12(2):103–165, 2010.

[6] Peter Kirschenhofer and Helmut Prodinger. Eine Anwendung der Theorie der Modulfunktionen in der Informatik. *Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II*, 197(4-7):339–366, 1988.

[7] Donald E. Knuth. *The Art of Computer Programming*, volume 3: Sorting and Searching. Addison-Wesley, 1973. Second edition, 1998.

[8] Guy Louchard and Helmut Prodinger. Asymptotics of the moments of extreme-value related distribution functions. *Algorithmica*, 46(3-4):431–467, 2006.

[9] Toufik Mansour. Protected points in k-ary trees. *Appl. Math. Lett.*, 24(4):478–480, 2011.

[10] Helmut Prodinger. External internal nodes in digital search trees via Mellin transforms. *SIAM J. Comput.*, 21(6):1180–1183, 1992.

Rosena R.X. Du, Department of Mathematics, East China Normal University, 500 Dongchuan Road, Shanghai, 200241, P. R. China.
E-mail address: rxdu@math.ecnu.edu.cn

Helmut Prodinger, Mathematics Department, Stellenbosch University, 7602 Stellenbosch, South Africa.
E-mail address: hproding@sun.ac.za