THE REGISTER FUNCTION FOR t–ARY TREES

MICHAEL DRMOTA AND HELMUT PRODINGER†

Abstract. For the register function for t–ary trees, recently introduced by Auber et al., we prove that the average is $\log_4 n + O(1)$, if all such trees with n internal nodes are considered to be equally likely.

This result remains true for rooted trees where the set of possible out–degrees is finite. Furthermore we obtain exponential tail estimates for the distribution of the register function. Thus, the distribution is highly concentrated around the mean value.

1. Introduction

The register function of binary trees was introduced by Ershov [3]; the equivalent notion of (Horton–)Strahler numbers was introduced earlier by hydrogeologists Horton [7] and Strahler [14].

This function is recursively defined by $\mathrm{reg}(\mathrm{root}) = 0$, and, if a binary tree T has subtrees T_1 and T_2, then $\mathrm{reg}(T) = \max\{\mathrm{reg}(T_1), \mathrm{reg}(T_2)\}$, provided $\mathrm{reg}(T_1) \neq \mathrm{reg}(T_2)$, otherwise it is $1 + \mathrm{reg}(T_1)$.

Assuming all binary trees with n internal nodes to be equally likely, the average value of the register function was found independently and at the same time [6, 8]; compare also [10]. It is $\log_4 n + O(1)$, and more precision is available and involves complicated (fluctuating) terms. The concept has been extended to unary–binary trees [5]. Various papers about the register function (or Horton–Strahler numbers) have been written; we cite a few here [2, 9, 15, 13, 11].

Recently, Auber et al. [1] have introduced a generalisation to general rooted trees. It is again recursively defined via $\mathrm{reg}(\Box) = 0$ and if the values of the subtrees $\mathrm{reg}(T_1), \dots, \mathrm{reg}(T_t)$ are written in nonincreasing order as $c_1 \geq \cdots \geq c_t$ (where t is the number of descendents) then the register function of the tree T is given by $\mathrm{reg}(T) = \max\{c_1, c_2 + 1, \dots, c_t + t - 1\}$.

The cited paper contains already a few results, but much remains to be done. In this paper we want to investigate the average value of the register function, provided that all trees (with certain degree restrictions) with n nodes are equally likely. We will show that this parameter is $\log_4 n + O(1)$, too, and that the distribution is highly concentrated around the mean. This means that the register function is (with high probability) a "function" of the size of the tree and it "almost" does not depend on the structure of the tree.

2. Results

Let $D \subseteq \{1, 2, 3, \dots\}$ be a finite set that contains at least one element greater than 1 and set $d = \gcd(D)$. For n with $n \equiv 1 \bmod d$ let T_n denote the set of rooted trees of

Date: February 4, 2005.

size n where all nodes have outdegree in D. For example, if $D = \{t\}$ (for some fixed $t \geq 2$) we just get the set of t–ary rooted trees.[1] If we suppose that every tree in T_n is equally likely then every parameter on trees can be interpreted as a random variable.

The main purpose of this paper is to discuss properties of the random variable R_n, the register function on T_n. Our first result is an asymptotic relation for its expected value $\mathbb{E}R_n$.

Theorem 1. *We have, for $n \equiv 1 \bmod d$ as $n \to \infty$*

$$\mathbb{E}R_n = \log_4 n + O(1). \tag{1}$$

We can also show that the register function is highly concentrated around its mean. We obtain exponential tail estimates:

Theorem 2. *We have uniformly for $0 \leq y \leq \left(\frac{1}{2} - \eta\right)\log_4 n$, where $\eta > 0$ is arbitray, and for all n with $n \equiv 1 \bmod d$*

$$\mathbb{P}\{|R_n - \mathbb{E}R_n| \geq y\} = O(2^{-y}). \tag{2}$$

Since $R_n = O(\log n)$ it also follows that all centralized moments are bounded. Unfortunately our methods are not strong enough to get more precise bounds.

The structure of the proof is the following one. First of all we will work out details just for t–ary trees in order to make the presentation more readable. Of course, we will also indicate how the general case of finite D can be treated. In Section 3 we collect some facts on generating functions that encode the distribution of R_n. The main part of the proof is contained in Section 4 where we prove asymptotic relations for these generating functions in order to derive (1). Section 5 is devoted to the proof of Theorem 2. Finally we indicate that a *simplified version* of the generating function leads to the same asymptotic results and sheds some light on the asymptotic structure that is hidden behind the recurrences of the involved generating functions.

3. Generating Functions For t–Ary Trees

The generating function $y = y(z)$ for the number of t–ary trees (where only internal nodes are counted) satisfies the functional equation

$$y = 1 + zy^t.$$

By *Lagrange inversion* we directly obtain the number of t–ary trees with n (internal) nodes:

$$y_n = [z^n]\, y(z) = \frac{1}{(t-1)n+1}\binom{tn}{n} \sim \sqrt{\frac{t}{2\pi(t-1)^3}}\; n^{-3/2}\left(\frac{t^t}{(t-1)^{t-1}}\right)^n.$$

Hence,

$$z_0 = \frac{(t-1)^{t-1}}{t^t}$$

[1]For t–ary trees it is common to count just *internal nodes*. If there are n internal nodes then the total number of nodes equals $n' = tn + 1$. The advantage is that one does not have to care about the restriction $n' \equiv 1 \bmod t$.

is the singularity of $y(z)$ and the local expansion of $y(z)$ around its singularity is given by

$$y(z)=\frac{t}{t-1}-\sqrt{\frac{2t}{(t-1)^3}}\sqrt{1-z\frac{t^t}{(t-1)^{t-1}}}+O\left(1-zt^t(t-1)^{1-t}\right). \qquad (3)$$

In particular, the power series expansion of $y(z)$ is convergent at $z=z_0$ and we have

$$y(z_0)=\frac{t}{t-1}.$$

Furthermore, $z=z_0$ is the only singularity on the circle of convergence $|z|=z_0$ and $y(z)$ can be uniquely analytically continued to a region of the form $|z|<z_0+\varepsilon$, $\arg(z-z_0)\neq 0$, where $\varepsilon>0$.

Note further, that $1-tzy^{t-1}$ has the local expansion

$$1-tzy(z)^{t-1}=\sqrt{\frac{2(t-1)}{t}}\sqrt{1-z/z_0}+O\left(1-z/z_0\right). \qquad (4)$$

In what follows we will make use of the abbreviation

$$V:=\frac{1-tzy(z)^{t-1}}{z\binom{t}{2}y(z)^{t-2}}=\sqrt{\frac{8t}{(t-1)^3}}\sqrt{1-z/z_0}+O\left(1-z/z_0\right). \qquad (5)$$

Since

$$y(z)=\frac{t}{t-1}-\frac{V}{2}+O(|V|^2)$$

"everything" can be expressed in terms of V. In particular, a local expansion in terms of V translates into a local expansion around the singularity z_0. In this type of tree enumeration problems, one can always decide whether z or y is the independent variable. It is natural to take z, but usually it is easier to work with y as independent variable. However, one can always "translate."

It seems to be natural to work with the generating function

$$R_p(z)=\sum_{n\geq 0}[\text{number of } t\text{–ary trees with } n \text{ internal nodes and register function } p]\cdot z^n,$$

but, as noticed already in the (classical) binary case [12], it is more convenient to work with

$$S_p(z)=\sum_{n\geq 0}[\text{number of } t\text{–ary trees with } n \text{ internal nodes and register function} \geq p]\cdot z^n.$$

Of course, $R_p=S_p-S_{p+1}$, and $S_0=y$.

The paper [1] has already the recursion for these functions, if one makes the proper adjustments (as already mentioned, they count the leaves also as internal nodes, which amount to the generating function $y=z(1+y^t)$, but we decided to study the more common version given by $y=1+zy^t$):

One sets $P_1(y,g_0)=g_0$, and recursively

$$P_t(y,g_0,\dots,g_{t-1})=g_{t-1}^t+t\int_{g_{t-1}}^{y}P_{t-1}(\tau,g_0,\dots,g_{t-2})d\tau.$$

For example, one gets

$$P_2(y, g_0, g_1) = g_1^2 - 2g_0g_1 + 2g_0y,$$
$$P_3(y, g_0, g_1, g_2) = g_2^3 - 3g_1^2g_2 + 3g_0g_1g_2 - 3g_0g_2^2 + 3g_1^2y - 6g_0g_1y + 3g_0y^2.$$

Then $S_p = S_p(z)$ satisfies a recurrence relation of the form

$$S_p = zP_t(y, S_p, S_{p-1}, \dots, S_{p-t+1}),$$

that can be made explicit since g_0 (resp. S_p) occurs in P_t at most in first order, in particular define D_t and N_t by $P_t = g_0D_t + N_t$ then we have

$$S_p = \frac{zN_t(y, S_{p-1}, \dots, S_{p-t+1})}{1 - zD_t(y, S_{p-1}, \dots, S_{p-t+1})}. \tag{6}$$

Here are the first few instances (for $t = 2, 3, 4$):

$$S_p = \frac{zS_{p-1}^2}{1 - 2zy + 2zS_{p-1}},$$

$$S_p = \frac{z\left(3S_{p-1}^2y + S_{p-2}^3 - 3S_{p-1}^2S_{p-2}\right)}{1 - 3zy^2 + 6zS_{p-1}y + 3zS_{p-2}^2 - 6zS_{p-1}S_{p-2}},$$

$$S_p = \frac{z(6S_{p-1}^2y^2+S_{p-3}^4-6S_{p-1}^2S_{p-3}^2+4S_{p-2}^3y-4S_{p-2}^3S_{p-3}-12S_{p-1}^2S_{p-2}y+12S_{p-1}^2S_{p-2}S_{p-3})}{1-4zy^3+12zS_{p-1}y^2+4zS_{p-3}^3-12zS_{p-1}S_{p-3}^2+12zS_{p-2}^2y-12zS_{p-2}^2S_{p-3}-24zS_{p-1}S_{p-2}y+24zS_{p-1}S_{p-2}S_{p-3}}.$$

Note further that the recursion holds for $p \ge 1$, if one chooses initial conditions $S_0 = S_{-1} = S_{-2} = \cdots = y$.

It is clear that N_t "starts" with $N_t = \binom{t}{2}S_{p-1}^2y^{t-2} + \cdots$ and D_t with $D_t = y^{t-1} - 2\binom{t}{2}S_{p-1}y^{t-2} + \cdots$, where we only encounter the "leading terms" with respect to y. This means that we can rewrite (6) to

$$S_p = \frac{z\left[\binom{t}{2}S_{p-1}^2y^{t-2} + A_p\right]}{1 - tzy^{t-1} + z\left[2\binom{t}{2}S_{p-1}y^{t-2} + B_p\right]}, \tag{7}$$

where A_p is a homogeneous polynomial of degree t in $y, S_{p-1}, S_{p-2}, \dots, S_{p-t+1}$, B_p is a homogeneous polynomial of degree $t-1$ in $y, S_{p-1}, S_{p-2}, \dots, S_{p-t+1}$, and the degrees of y are all smaller than $t-2$.

Finally we want to state the differences if we consider finite outdegree sets D of cardinality greater than 1. Here the generating function $y(z)$ of the numbers $y_n = |T_n|$ (where all nodes are counted) is given by

$$y(z) = z\sum_{t\in D} y(z)^t.$$

For convenience, set $\Phi(y) = \sum_{t\in D} y^t$, so that we have $y(z) = z\Phi(y(z))$. It is well known (see [4]) that the series $y(z)$ converges as an analytic function inside the complex disc $|z| < z_0$, where $z_0 = \tau/\Phi(\tau)$ and τ is the unique positive real solution of the equation $\Phi(\tau) = \tau\Phi'(\tau)$. Furthermore, $y(z)$ has dominant singularities of square–root type (3):

$$y(z) = \tau - \sqrt{\frac{2\Phi(\tau)}{\Phi''(\tau)}}\sqrt{1 - z/z_0} + O(|1 - z/z_0|).$$

If $d = \gcd(D) = 1$ then $z = z_0$ is the only singularity on the circle $|z| = z_0$. If $d > 1$ then the points $z_0 e^{2\pi ik/d}$, $k = 1, \ldots, d-1$, are also square–root singularities of the same type. Eventually this leads to the asymptotic expansion

$$y_n = d\sqrt{\frac{\Phi(\tau)}{2\pi\Phi''(\tau)}}n^{-3/2}z_0^n\left(1+O\Big(\frac{1}{n}\Big)\right)$$

if $n \equiv 1 \bmod d$ (and $y_n = 0$ otherwise). In what follows we will always assume that $d = 1$. The case $d > 1$ can be treated in a completely similar way. Note that we also have the local expansion

$$z\Phi'(y(z)) = 1 - \sqrt{\frac{2\tau^2\Phi''(\tau)}{\Phi(\tau)}}\sqrt{1-z/z_0} + O(|1-z/z_0|).$$

Thus, with

$$V = \frac{1-z\Phi'(y(z))}{z\Phi''(y(z))/2} = \sqrt{\frac{8\Phi(\tau)}{\Phi''(\tau)}}\sqrt{1-z/z_0} + O(|1-z/z_0|)$$

we get $y(z) = \tau - V/2 + O(|V|^2)$.

Furthermore, since $S_p = S_p(z)$ satisfies the recurrence relation

$$S_p = z\sum_{t\in D} P_t(y, S_p, S_{p-1}, \ldots, S_{p-t+1})$$

we thus get

$$S_p = \frac{z\left[\frac{1}{2}\Phi''(y)S_{p-1}^2 + A_p\right]}{1 - z\Phi'(y) + z\left[\Phi''(y)S_{p-1} + B_p\right]}, \tag{8}$$

where A_p is (now) a polynomial of degree $t_{\max} = \max(D)$ in $y, S_{p-1}, S_{p-2}, \ldots, S_{p-t+1}$, B_p is a of degree $t_{\max}-1$ in $y, S_{p-1}, S_{p-2}, \ldots, S_{p-t+1}$, and the degrees of y are all smaller than $t_{\max}-2$. This means that the "general case" follows completely the same pattern as the t–ary case.

4. Asymptotic Properties For The Expected Value

In order to prove an asymptotic expansion for the expected values $\mathbb{E}R_n$ for the register function of t–ary trees we consider the generating function

$$E(z) = \sum_{n\geq 0}\mathbb{E}R_n\, y_n z^n = \sum_{p\geq 1} S_p(z). \tag{9}$$

We will show that the behaviour of $E(z)$ around its singularity is of the following form.

Proposition 1. *There exists a constant $D > 0$ such that*

$$E(z) = D - \sqrt{\frac{t}{2(t-1)^3}}\,\sqrt{1-z/z_0}\,\log\frac{1}{1-z/z_0} + O\Big(\sqrt{1-z/z_0}\Big)$$

for $|z-z_0| \leq \varepsilon$ and $|\arg(z-z_0)| \geq \frac{\pi}{2}-\eta$, where $\varepsilon > 0$ and $\eta > 0$ are sufficiently small constants. Furthermore, $E(z)$ is analytic and uniformly bounded in a range of the form $|z| < z_0+\varepsilon_2$, $|z-z_0| > \varepsilon$, where $\varepsilon_2 > 0$ is another sufficiently small constant.

Theorem 1 is an immediate consequence of Proposition 1 by a *transfer lemma* of Flajolet and Odlyzko, see [4].

We start our analysis with the asymptotic behaviour of $S_p(z_0)$, where $z_0 = (t-1)^{t-1}t^{-t}$ is the singularity (and also the radius of convergence) of $y(z)$.

Lemma 1. *The sequence $S_p(z_0)$ satisfies $S_p(z_0)/S_{p-1}(z_0) \ge 1/t$ and $S_p(z_0)/S_{p-1}(z_0) \to 1/2$ as $p \to \infty$. In particular, we have*

$$S_p(z_0) = C_p 2^{-p}, \tag{10}$$

where the sequence C_p satisfies

$$C_p = C_\infty + O(2^{-p})$$

with a positive constant C_∞.

Note that Lemma 1 can be restated as $S_p(z_0) = C_\infty 2^{-p} + O(4^{-p})$ but (10) is more useful for our purpose, see Lemma 2.

Proof. Most parts of Lemma 1 are contained in [1]. However, since some of the ideas of the proof will be used in the sequel we provide a complete proof.

First observe that $S_p(z_0) \to 0$ monotonically. In fact, from the combinatorial interpretation it directly follows that $S_p(z_0) \le S_{p-1}(z_0)$. Further, if the register function of a tree T is at least p then T must have at least $2^p - 1$ nodes. Hence

$$S_p(z_0) \le \sum_{n \ge 2^p - 1} y_n z_0^n \to 0,$$

where we have also used the fact that $y(z_0)$ is finite.

Next we show that $S_p(z_0) \ge S_{p-1}(z_0)/t$ for all $p \ge 1$ (compare with [1]). Set (as above) $P_t = g_0 D_t + N_t$ and $E_t = ty^{t-1} - D_t$. Since $1 - tz_0 y(z_0)^{t-1} = 0$ we get for $z = z_0$

$$S_p = \frac{N_t(y, S_{p-1}, \ldots, S_{p-t+1})}{E_t(y, S_{p-1}, \ldots, S_{p-t+1})}$$

and N_t and E_t satisfy the recurrences

$$N_t(y, g_1, \ldots, g_{t-1}) = g_{t-1}^t + t\int_{g_{t-1}}^{y} N_{t-1}(\tau, g_1, \ldots, g_{t-2})\, d\tau$$

and

$$E_t(y, g_1, \ldots, g_{t-1}) = t g_{t-1}^{t-1} + t\int_{g_{t-1}}^{y} E_{t-1}(\tau, g_1, \ldots, g_{t-2})\, d\tau.$$

For example, for $t = 3$ we have

$$S_p = \frac{3S_{p-1}^2 y - 3S_{p-1}^2 S_{p-2} + S_{p-2}^3}{6S_{p-1}y - 6S_{p-1}S_{p-2} + 3S_{p-2}^2}.$$

The recurrence relations for N_t and E_t also imply

$$N_t - \frac{g_1}{t}E_t = (g_{t-1} - g_1)g_{t-1}^{t-1} + t\int_{g_{t-1}}^{y} (N_{t-1} - g_1 E_{t-1})\, d\tau.$$

Assume that $0 \le g_1 \le g_2 \le \cdots \le g_{t-1} \le y$. Then we have $N_2 - (g_1/t)E_2 = g_1^2(1-2/t) \ge 0$ and by induction

$$N_t(y, g_1, \ldots, g_{t-1}) \ge \frac{g_{t-1}}{t} E_t(y, g_1 \ldots, g_{t-1}) \ge 0.$$

Consequently

$$S_p(z_0) = \frac{N_t(y(z_0), S_{p-1}(z_0), \ldots, S_{p-t+1}(z_0))}{E_t(y(z_0), S_{p-1}(z_0), \ldots, S_{p-t+1}(z_0))} \ge \frac{S_{p-1}(z_0)}{t}.$$

We now use (7) to represent the ratio

$$\frac{S_p}{S_{p-1}} = \frac{1+A'_p}{2+B'_p},$$

where

$$A'_p = \frac{A_p}{\binom{t}{2} S_{p-1}^2 y^{t-2}}, \qquad B'_p = \frac{B_p}{\binom{t}{2} S_{p-1} y^{t-2}}.$$

For example, for $t = 3$ we have

$$A'_p = \frac{S_{p-2}^3}{3yS_{p-1}^2} - \frac{S_{p-2}}{y} \tag{11}$$

and

$$B'_p = \frac{S_{p-2}^2}{yS_{p-1}} - \frac{2S_{p-2}}{y}. \tag{12}$$

Since we know that S_{p-j}/S_{p-1} is bounded for each fixed $j \ge 1$ and that $S_p \to 0$ (as $p \to \infty$) we also get that $A'_p \to 0$ and $B'_p \to 0$. Consequently

$$\frac{S_p}{S_{p-1}} \to \frac{1}{2} \qquad (p \to \infty).$$

In particular we have $S_p \le \frac{3}{4} S_{p-1}$ for sufficiently large $p \ge p_0$ and, thus, $S_p = O((3/4)^p)$. This also implies $A'_p = O((3/4)^p)$ and $B' = O((3/4)^p)$ which gives

$$\frac{S_p}{S_{p-1}} = \frac{1}{2}\big(1 + O((3/4)^p)\big).$$

Hence,

$$S_p = S_0 \frac{S_1}{S_0} \cdots \frac{S_p}{S_{p-1}} = O(2^{-p})$$

and consequently $A'_p = O(2^{-p})$ and $B'_p = O(2^{-p})$. The proof is now completed by setting

$$C_p = S_0 \prod_{j<p} \frac{1+A'_p}{1+B'_p/2}.$$

Obviously we also have $C_p = C_\infty + O(2^{-p})$, where

$$C_\infty = S_0 \prod_{j\ge 1} \frac{1+A'_p}{1+B'_p/2}.$$

□

Next we consider $S_p(z)$ when z is close to z_0. We will state all properties is terms of $V = (1 - tzy(z)^{t-1})/(z\binom{t}{2}y(z)^{t-2})$, see (5). For the sake of transparency we will split up our considerations into several lemmata.

Lemma 2. *Suppose that* $|z - z_0| < \varepsilon$ *and* $|\arg(z - z_0)| \ge \frac{\pi}{2} - \eta$, *that is,* $|V| < \varepsilon'$ *and* $|\arg(V)| \le \frac{\pi}{4} + \eta'$, *where* $\varepsilon, \eta > 0$ *resp.* $\varepsilon', \eta' > 0$ *are sufficiently small constants. Then there exists* $\delta > 0$ *and constants* $C', C'' > 0$ *such that*

$$C' 2^{-p} \le |S_p(z)| \le C'' 2^{-p}$$

for all complex $p \le -\log_2 |V| - \delta$.

Proof. In order to simplify the notation we restrict ourselves to the case $t = 3$. The general case runs along the same lines. Further, we again use the recurrence for S_p of the form

$$S_p = \frac{S_{p-1}(1 + A_p')}{2 + V/S_{p-1} + B_p'}, \tag{13}$$

where $A_p' = A_p/\left(\binom{t}{2} S_{p-1}^2 y^{t-2}\right)$ and $B_p' = B_p/\left(\binom{t}{2} S_{p-1} y^{t-2}\right)$. In particular, for $t = 3$ we have (11) and (12).

First, we want to show that we have

$$\frac{3}{10} \le \frac{|S_p(z)|}{|S_{p-1}(z)|} \le \frac{5}{6} \tag{14}$$

for $p \ge p_1$ (where p_1 has to be chosen appropriately) and for all p that satisfy $|S_{p-1}| \ge 4|V|$ (where z is close to z_0 according to the assumptions of Lemma 2).

Since $S_p(z_0) \to 0$ and $S_p(z_0)/S_{p-1}(z_0) \to 1/2$ as $p \to \infty$ and since all functions $z \to S_p(z)$ are continuous it follows that there exist p_1 and $\varepsilon > 0$ such that $|S_{p_1-1}(z)|/|S_{p_1-2}(z)| \ge 3/10$, $|S_{p_1-2}(z)| \le 1/30$, and $|y(z)| \ge 1$ for all z with $|z - z_0| < \varepsilon$ and $\arg(z - z_0) \ne 0$. We now show by induction that this will be then satisfied for all $p \ge p_1$ as long $|S_{p-1}| \ge 4|V|$. First we get

$$\begin{aligned} |A_{p_1}'| &\le \frac{|S_{p_1-2}|^3}{3|yS_{p_1-1}|^2} + \frac{|S_{p_1-2}|}{|y|} \\ &\le \left(\frac{10}{3}\right)^2 \frac{1}{3}\frac{1}{30} + \frac{1}{30} \\ &\le \frac{1}{4} \end{aligned}$$

and similarly

$$\begin{aligned} |B_{p_1}'| &\le \frac{|S_{p_1-2}|^2}{|yS_{p_1-1}|} + \frac{2|S_{p-2}|}{|y|} \\ &\le \frac{10}{3}\frac{1}{30} + \frac{2}{30} \\ &\le \frac{1}{4}. \end{aligned}$$

Since we also assume that $|V/S_{p_1-1}| \le \frac{1}{4}$ we thus obtain

$$\frac{1 - \frac{1}{4}}{2 + \frac{1}{2}} \le \frac{|1 + A_{p_1}'|}{|2 + V/S_{p_1-1} + B_{p_1}'|} \le \frac{1 + \frac{1}{4}}{2 - \frac{1}{2}}$$

or

$$\frac{3}{10} \le \frac{|S_{p_1}|}{|S_{p_1-1}|} \le \frac{5}{6}.$$

Furthermore, we have $|S_{p_1}| \le \frac{5}{6}|S_{p_1-1}| \le |S_{p_1-1}| \le 1/30$ so that we can proceed by induction. This proves (14).

These considerations also prove $S_p = O((5/6)^p)$ and consequently $A'_p = O((5/6)^p)$ and $B'_p = O((5/6)^p)$. Set $\tilde p = \tilde p(z) = \max\{p : |S_{p-1}(z)| \ge 4|V|\}$. Then we also have $|V/S_{p-1}| = O((5/6)^{\tilde p - p})$. Consequently

$$\left|\frac{S_p}{S_{p-1}}\right| = \frac{1}{2} + O\left(\left(\frac{5}{6}\right)^p + \left(\frac{5}{6}\right)^{\tilde p - p}\right)$$

and also

$$C' 2^{-p} \le |S_p(z)| \le C'' 2^{-p}$$

for $p \le \tilde p$. Finally with help of this estimate it also follows that $\tilde p = -\log|V| + O(1)$. This completes the proof of Lemma 2. □

Lemma 3. *Suppose that z satisfies the same assumptions as in Lemma 2. Then there exist $\eta > 0$ such that*

$$S_p(z) = C_p 2^{-p} - \frac{V}{2} + O\left(p\, 2^{-p}|V|\right) + O\left(2^p |V|^2\right) \tag{15}$$

for $p \le -\log_2|V|$ (where the constants $C_p = 2^p S_p(z_0)$ are from Lemma 1)

Proof. We have to be a little bit more precise than before. From (13) we get

$$S_p = S_{p-1}\frac{1 + A'_p}{(1 + B'_p/2)} - \frac{V}{4}\frac{1 + A'_p}{(1 + B'_p/2)^2} + O(2^p|V|^2)$$

which yields

$$\begin{aligned} S_p &= S_0 \prod_{j \le p} \frac{1 + A'_j}{1 + B'_j/2}\,\frac{1}{2^p} - \frac{V}{4}\sum_{j\le p}\frac{1 + A'_j}{(1 + B'_j/2)^2} 2^{j-p} + O(2^p|V|^2) \\ &= C_p(V) 2^{-p} - \frac{V}{2}\left(1 + O(p\, 2^{-p})\right) + O(2^p|V|^2), \end{aligned}$$

where

$$C_p(V) = S_0 \prod_{j\le p} \frac{1 + A'_j}{1 + B'_j/2}.$$

We will now show by induction that $C_p(V) = C_p + O(p|V|)$. Of course, (15) is then immediate.

Suppose that we already know that $|C_j(V) - C_j| \le Cj|V|$ for $j < p$ (with some $C \ge 1$ that will be fixed in the sequel and for some sufficiently large p that will be also specified). This assumption also gives

$$S_j = C_j 2^{-p} + O(Cp\, 2^{-p}|V|) - \frac{V}{2} + O(2^p|V|^2)$$

for $j < p$ and also

$$A'_j = a_j 2^{-j} + O(Cj\, 2^{-j}|V|) + O(|V|) \quad \text{and} \quad B'_j = b_j 2^{-j} + O(Cj\, 2^{-j}|V|) + O(|V|)$$

for $j \le p$, where a_j, b_j are proper constants that satisfy $a_j = a_\infty + O(2^{-j})$ and $b_j = b_\infty + O(2^{-j})$. (Note that $2^p|V| \le 1$ for $p \le -\log_2|V|$ and that we can always assume that $|V|$ is sufficiently small.) Hence,

$$\begin{aligned} C_p(V) &= (y(z_0)+O(V))\prod_{j\le p}\frac{1+a_j2^{-j}+O(Cj\,2^{-j}|V|)+O(|V|)}{1+b_j2^{-j-1}+O(Cj2^{-j}|V|)+O(|V|)} \\ &= y(z_0)\prod_{j\le p}\frac{1+a_j2^{-j}}{1+b_j2^{-j-1}}\prod_{j\le p}\frac{1+O(C\,j2^{-j}|V|)+O(|V|)}{1+O(Cj\,2^{-j}|V|)+O(|V|)} \\ &= C_p\,(1+O(C|V|)+O(p|V|)) \\ &= C_p + O(C|V|+p|V|). \end{aligned}$$

This means that there exists a univeral constant $c > 0$ such that $|C_p(V) - C_p| \le c(C|V| + p|V|)$ if $|V|$ is sufficiently small. We can now assume that our induction has started for some $p \ge 2c$ and that $C \ge 2c$ was chosen appropriately. Then $cC + cp \le Cp$ and consequently $|C_p(V) - C_p| \le Cp|V|$. This completes the proof of $C_p(V) = C_p + O(p|V|)$ for $p \le p_0 = \lfloor -\log_2 V\rfloor$. □

Up to $p \le \lfloor -\log_2|V| - \delta\rfloor$ the behaviour of S_p is very regular. The reason is that S_p is *large* compared to V. This means that the denominator $V + 2S_{p-1} + B''_p$ of

$$S_p = \frac{S_{p-1}^2 + A''_p}{V + 2S_{p-1} + B''_p}, \tag{16}$$

is dominated by the behaviour of S_{p-1} and V has only a minor influence; here $A''_p = S_{p-1}^2 A'_p$ and $B''_p = S_{p-1}B'_p$ are polynomials in $S_{p-1}, \ldots, S_{p-t+1}$. However, if $2S_{p-1}$ is of order V, in particular if $2S_{p-1}$ is close to $-V$ then it might occur that S_p gets arbitrarily large and we are confronted with an *chaotic* behaviour. On the other hand, if S_p (or S_{p-1}) is small compared to V then the denominator of (16) is dominated by the V and $S_p \approx S_{p-1}^2/V$. This means that S_p will converge to 0 very rapidly. This means that one has to manage the *gap* between $p \approx -\log_2|V| - \delta$ where $S_p \approx 2^\delta|V|$ and $p \approx -\log_2|V| + \delta'$ where $|S_p|$ should be small compared to $|V|$. Fortunately this *phase transition* of finitely many steps can be managed in the following way.

Lemma 4. *Suppose that z satisfies the same assumptions as in Lemma 2 and set $p = \lfloor -\log_2|V| - \delta\rfloor$, where $\delta = O(1)$ is chosen in a way that $|S_p(z)| \ge 2|V|$. Then for every $\delta' > 0$ we have uniformly for all z (that satisfy the same assumptions as in Lemma 2) and for $0 \le \ell \le \delta + \delta'$*

$$S_{p+\ell}(z) = \frac{V}{\left(\frac{V}{S_p}+1\right)^{2^\ell} - 1} + O(|V|^2). \tag{17}$$

Proof. For a moment let us assume that $A''_p = B''_p = 0$, that is, we consider the recurrence

$$T_p = \frac{T_{p-1}^2}{V + 2T_{p-1}} \tag{18}$$

instead of (16). This recurrence can be explicitly solved since it is equivalent to

$$\frac{V}{T_p} + 1 = \left(\frac{V}{T_{p-1}} + 1\right)^2.$$

Hence

$$T_{p+\ell} = \frac{V}{\left(\frac{V}{T_p}+1\right)^{2^\ell} - 1}.$$

Here we have $\lim\limits_{\ell\to\infty} T_{p+\ell} = 0$ if and only if

$$\left|\frac{V}{T_p}+1\right| > 1.$$

In particular, this is satisfied if $\Re(V/T_p) > 0$ or if $|\arg(T_p)| < \frac{\pi}{4} - \eta_2'$ (since we have assumed that $|\arg(V)| < \frac{\pi}{4} + \eta_2'$ for some small constant η_2').

Since S_p is asymptotically given by (17) it follows that $\arg(S_p) = O(2^{-\delta})$ if $p \le -\log_2|V| - \delta$. Thus we can start with a *save starting point* $T_p \approx S_p$.

The crucial step is now to observe that the explicit solution T_p of (18) is a good approximation for S_p. Assume that $T_p = S_p$ for $p = \lfloor -\log_2|V| - \delta\rfloor$. Then T_p (and S_p) are of order $|V|$, A_p'' is of order $|V|^3$ and and B_p'' of order $|V|^2$.

Set

$$f(x,\varepsilon,\eta) = \frac{x^2+\varepsilon}{1+2x+\eta}.$$

Then we uniformly have $f(x,\varepsilon,\eta) = f(x,0,0) + O(\max\{|\varepsilon|,|\eta|\})$ uniformly if x varies in a compact set that avoids $x = -1/2$ as $\max\{|\varepsilon|,|\eta|\} \to 0$. Consequently it follows by induction that

$$f^\ell(x,\varepsilon_1,\ldots,\varepsilon_\ell,\eta_1,\ldots,\eta_\ell) = f^k(x,0,0) + O\left(\max_{1\le j\le \ell}\{|\varepsilon_j|,|\eta_j|\}\right)$$

for every fixed ℓ (where f^ℓ denotes the ℓ-th iterate of f, for example $f^2(x,\varepsilon_1,\varepsilon_2,\eta_1,\eta_2) = f(f(x,\varepsilon_1,\eta_1),\varepsilon_2,\eta_2)$).

Now we have

$$S_p/V = f(S_{p-1}/V, A_p''/V^2, B_p''/V) \quad \text{and} \quad T_p/V = f(T_{p-1}/V, 0, 0),$$

compare with (16) and (18). Hence, if we now consider $\delta+\delta'$ steps (where δ' is any fixed number) then we have $S_{p+\ell} = T_{p+\ell} + O(|V|^2)$ for $0 \le \ell \le \delta+\delta'$. This completes the proof of the lemma. ☐

The final case $p \ge -\log_2|V| + \delta'$ can be managed in a quite easy way.

Lemma 5. *Suppose that z satisfies the same assumptions as in Lemma 2 and that $p \ge -\log_2|V| + \delta'$. Then*

$$|S_p(z)| = O\left(|V|e^{-\eta 2^p|V|}\right). \tag{19}$$

Proof. For the range $p \ge -\log_2|V| + \delta'$ we proceed as follows. We estimate $S_{p-1}^2 + A_p''$ in a crude way and get $|S_{p-1}^2 + A_p''| \le c'|S_{p-t+1}|^2$ for some constant $c' > 0$. Furthermore, there exists $\kappa > 0$ with $|2S_{p-1} + B_p''| \le |V|/2$ and $|S_{p-t+1}|^2 \le \frac{1}{2}|V|/(2c')$ for $p \ge p_0 + \kappa$. Consequently

$$|S_p| \le \frac{|S_{p-t+1}|^2}{|V|/(2c')}$$

for $p \ge p_0 + \kappa$ and by induction we get

$$|S_p| \le \frac{|V|}{2c'} \left(\frac{|S_{p_0 - t+1+\kappa}|}{|V|/(2c')} \right)^{2^{p-p_0-\kappa}}$$
$$= O\left(|V| e^{-\eta 2^p |V|}\right),$$

where $\eta > 0$ is sufficiently small. □

Lemma 6. *Let $\varepsilon > 0$ be from Lemma 2. Then There exists $\varepsilon_2 > 0$ such that for all z with $z_0 \le |z| \le z_0 + \varepsilon_2$ and $|z - z_0| \ge \varepsilon$ we have*

$$|S_p(z)| \le 2^{-p}. \tag{20}$$

Proof. If we restrict z in the range $z_0 \le |z| \le z_0 + \varepsilon/2$ and $|z - z_0| \ge \varepsilon$ then $|V|$ is uniformly bounded below by $|V| \ge \varepsilon'$ for some $\varepsilon' > 0$. Furthermore, we also have $|y(z)| \ge \varepsilon''$ for some $\varepsilon'' > 0$ since $y = 1 + zy^t$.

The idea of the proof is to show that (for sufficiently large p) $|S_{p-1}^2 + A_p''| \le 2^{-(p-1)}\varepsilon'$ and $|2S_{p-1} + B_p''| \le \frac{1}{2}\varepsilon'$. By assuming that we get

$$|S_p| \le \frac{2^{-(p-1)}\varepsilon'}{|V| - \frac{1}{2}\varepsilon'} \le 2^{-p}$$

and the result follows by induction.

In order to simplify notation we just consider the case $t = 3$. Here we have

$$A_p'' = \frac{S_{p-2}^3}{3y} - \frac{S_{p-2}S_{p-1}^2}{y}$$

and

$$B_p'' = \frac{S_{p-2}^2}{y} - \frac{2S_{p-2}S_{p-1}}{y}.$$

If $|z| = z_0$ we have $|S_p(z)| \le S_p(z_0) \le C' 2^{-p}$, where we can assume that $C' \ge 1$. We now fix some "starting" p. By continuity there exists $0 < \varepsilon_2 \le \varepsilon/2$ such that

$$|S_{p-1}(z)| \le 2C' 2^{-p+1} \quad \text{and} \quad |S_{p-2}(z)| \le 2C' 2^{-p+2}$$

for all z with $z_0 \le |z| \le z_0 + \varepsilon_2$ and $|z - z_0| \ge \varepsilon$. Hence

$$|S_{p-1}^2 + A_p''| \le 4C'^2 2^{-2p+4} + \frac{8C'^3 2^{-3p+6}}{3\varepsilon''} + \frac{8C'^3 2^{-3p+5}}{\varepsilon''}$$

and

$$|2S_{p-1} + B_p''| \le 4C' 2^{-p+1} + \frac{4C'^2 2^{-2p+4}}{\varepsilon''} + \frac{8C'^2 2^{-2p+3}}{\varepsilon''}.$$

If we choose the "starting" p sufficiently large then we surely get

$$|S_{p-1}^2 + A_p''| \le 2^{-(p-1)}\varepsilon' \quad \text{and} \quad |2S_{p-1} + B_p' S_{p-1}| \le \frac{1}{2}\varepsilon'$$

and (as noted above) $|S_p| \le 2^{-p} \le C' 2^{-p}$. This completes the proof of Lemma 6 (by induction). □

Now it is easy to complete the proof of Proposition 1.

Proof. We first consider the range $|z - z_0| < \varepsilon$ and $|\arg(z - z_0)| \ge \frac{\pi}{2} - \eta$ and split up the sum (9) into two parts:

$$E(z) = \sum_{p \le -\log_2 |V|} S_p(z) + \sum_{p > -\log_2 |V|} S_p(z) = E_1 + E_2.$$

By Lemma 3 we have

$$\begin{aligned} E_1 &= \sum_{p \le -\log_2 |V|} C_p 2^{-p} + \frac{V}{2} \log_2 |V| + O(|V|) \\ &= \sum_{p \ge 0} C_p 2^{-p} + \frac{V}{2} \log_2 V + O(|V|). \end{aligned}$$

and

$$E_2 = O(|V|).$$

By (5) this directly translates to (9).

Finally, Lemma 6 implies that $E(z)$ is bounded in the range $z_0 \le |z| \le z_0 + \varepsilon_2$ and $|z - z_0| \ge \varepsilon$. □

5. Tail Estimates

In this section we shortly comment on the proof of Theorem 2. We can use the estimates of Lemma 3 and Lemma 4 to get approximations for

$$\mathbb{P}\{R_n \ge p\} = \frac{[z^n]\, S_p(z)}{y_n}.$$

In order to extract the coefficient of $[z^n]\, S_p(z)$ we *copy* the methods of [4], that is we use Cauchy's formula and integrate *around the singularity* with distance $|1 - z/z_0| = \frac{1}{n}$.

In particular, if $p \le \log_4 n$ then we have to use (15) and (20) and we get

$$\mathbb{P}\{R_n \ge p\} = 1 + O(2^p/\sqrt{n}\,) + O(p\, 2^{-p}) + O(e^{\eta'' n})$$

For the case $p \ge \log_4 n$ we apply (19) and (20) and derive

$$\mathbb{P}\{R_n \ge p\} = O(e^{-\eta' 2^p/\sqrt{n}}\,) + O(e^{\eta'' n}),$$

where η' and η'' are positive constants. Of course, these two estimates imply Theorem 2.

6. The approximate recursion

Let us consider the simplified recursion

$$S_p = \frac{\binom{t}{2} z S_{p-1}^2 y^{t-2}}{1 - tzy^{t-1} + 2\binom{t}{2} z S_{p-1} y^{t-2}}, \qquad S_0 = y,$$

which is obtained from the original one by discarding the less important terms. It is exact for the classical case $t = 2$. Since one can say a lot more in the binary case, we will sketch that this is also the case for this simplified recursion, which has, as demonstrated before, the explicit solution

$$\frac{V}{S_p} + 1 = \Big(\frac{V}{y} + 1\Big)^{2^p},$$

or

$$S_p = V\frac{\left(\frac{V}{y}+1\right)^{-2^p}}{1-\left(\frac{V}{y}+1\right)^{-2^p}}.$$

Set $\frac{V}{y}+1=e^\tau$, then

$$E_p(z)=\sum_{p\ge1}S_p(z)\sim\frac{t}{t-1}\tau\sum_{p\ge1}\frac{e^{-2^p\tau}}{1-e^{-2^p\tau}}.$$

But this series is well understood [12], with the result

$$E_p(z)\sim D-\frac{t}{t-1}\tau\log_4\tau,$$

with more terms being available. Rewriting,

$$\tau\sim\frac{t-1}{t}\sqrt{\frac{8t}{(t-1)^3}}\sqrt{1-z/z_0},$$

we are completely in the same situation as in Proposition 1. We find that the average value (related to the simplified recursion) satisfies

$$\frac{1}{y_n}[z^n]\sum_{p\ge1}S_p\sim\log_4 n,$$

with more terms available (including periodic functions of $\log_4 n$). However, since we do not fully understand how well the original recursion is approximated by the simplified recursion, we do not pursue this any further.

References

[1] D. Auber, J.-P. Domenger, M. Delest, P. Duchon, and J.-M. Fédou. New Strahler numbers for rooted plane trees. In *Mathematics and Computer Science III*, pages 203–215. Birkäuser, Basel, 2004.

[2] L. Devroye and P. Kruszewski. On the Horton–Strahler number for random tries. *RAIRO Inform. Théor. Appl.*, 30(5):443–456, 1996.

[3] A. P. Ershov. On programming of arithmetic operations. *Communications of the ACM*, 1:3–6, 1958.

[4] P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, 3:216–240, 1990.

[5] P. Flajolet and H. Prodinger. Register allocation for unary-binary trees. *SIAM J. Comput.*, 15:629–640, 1986.

[6] P. Flajolet, J.–C. Raoult, and J. Vuillemin. The number of registers required for evaluating arithmetic expressions. *Theoretical Computer Science*, 9(1):99–125, 1979.

[7] R. E. Horton. Erosioned development of systems and their drainage basins. *Bulletin geological society of America*, 56:275–370, 1945.

[8] R. Kemp. The average number of registers needed to evaluate a binary tree optimally. *Acta Inform.*, 11(4):363–372, 1978/79.

[9] P. Kruszewski. A note on the Horton–Strahler number for random binary search trees. *Inform. Process. Lett.*, 69(1):47–51, 1999.

[10] A. Meir, J. W. Moon, and J. R. Pounder. On the order of random channel networks. *SIAM J. Algebraic Discrete Methods*, 1(1):25–33, 1980.

[11] M. Nebel. A unified approach to the analysis of Horton–Strahler parameters of binary tree structures. *Random Structures Algorithms*, 21(3-4):252–277, 2002. Random structures and algorithms (Poznan, 2001).

[12] H. Prodinger. Some analytic techniques for the investigation of the asymptotic behaviour of tree parameters. *EATCS Bulletin*, 47:180–199, 1992.
[13] H. Prodinger. On a problem of Yekutieli and Mandelbrot about the bifurcation ratio of binary trees. *Theoretical Computer Science*, 181:181–194, 1997.
[14] A. N. Strahler. Hypsomic analysis of erosional topography. *Bulletin geological society of America*, 63:1117–1142, 1952.
[15] I. Yekutieli and B. Mandelbrot. Horton–Strahler ordering of random binary trees. *J. Phys. A*, 27(2):285–293, 1994.

Michael Drmota, Department of Discrete Mathematics and Geometry, Wiedner Hauptstr. 8–10/104, A-1040 Wien, Austria

E-mail address: michael.drmota@tuwien.ac.at

Helmut Prodinger, Mathematics Department, University of Stellenbosch, 7602 Stellenbosch, South Africa

E-mail address: hproding@sun.ac.za