A GREEDY ALGORITHM ESTIMATING THE HEIGHT OF RANDOM TREES

TOMASZ ŁUCZAK†
Mathematical Institute of the Polish Academy of Sciences, Poznań, Poland

Abstract. The behaviour of a greedy algorithm which estimates the height of a random labelled rooted tree is studied. A self-similarity argument is used to characterize the limit distribution of the length H of the path found by such an algorithm in a random rooted tree as the unique solution of an integral equation. Furthermore, it is shown that

$$\lim_{n\to\infty} \frac{\mathrm{E}\, H}{\sqrt{n}} = \frac{\sqrt{2\pi}}{2\sqrt{2} - \ln(3+2\sqrt{2})} = 2.353139\ldots ,$$

i.e. the expected length of the path constructed by the algorithm is roughly 93% of the expected height of a random rooted tree.

1991 Mathematics subject classifications: 68Q25, 05C80, 05C05, 68RXX

Key words: random tree, height, greedy algorithms

1. Introduction.

Let T_n be a random labelled rooted tree on the vertex set $[n] = \{1, 2, \ldots, n\}$ with the root $v_0 \in [n]$ (here and below we assume for convenience that a root is always the vertex number 1). The limit distribution of the height of $\tilde{H} = \tilde{H}(n)$ of T_n was found by Rényi and Szekeres [1] who proved the following result.

THEOREM 0. *For every constant $\beta > 0$*

$$\begin{aligned} \lim_{n\to\infty} (\tilde{H} = \lfloor \sqrt{2n}/\beta \rfloor) &= 2\sqrt{\frac{2\pi}{n}}\beta^2 \sum_{i=1}^{\infty} \left(2i^4\pi^4\beta - 3i^2\pi^2\right) \exp(-\beta\pi^2 i^2) \\ &= \sqrt{\frac{8}{n\beta}} \sum_{i=1}^{\infty} \left(\frac{2i^4}{\beta} - 3i^2\right) \exp\left(-\frac{i^2}{\beta}\right) , \end{aligned} \tag{1}$$

where the convergence is uniform for $\beta \in (c, C)$ for every constants $0 < c < C < \infty$.

Furthermore, they proved that the s-th moment of random variable $h(T_n)/\sqrt{2n}$ tends to $2\Gamma(s/2+1)(s-1)\sum_{i=1}^{\infty} i^{-s}$. In particular, for the expectation and the variance of $h(T_n)$, they got

$$\lim_{n\to\infty} \frac{\mathrm{E}\, h(T_n)}{\sqrt{n}} = \sqrt{2\pi} = 2.50663\ldots \tag{2}$$

and

† On leave from Department of Discrete Mathematics, Adam Mickiewicz University, Poznań, Poland.
Research partially supported by KBN grant 2 1087 91 01.

$$\lim_{n\to\infty} \frac{\operatorname{Var} h(T_n)}{n} = \frac{2\pi(\pi-3)}{3} = 0.29655\ldots .$$

(see also forthcoming paper of Flajolet, Gao, Odlyzko and Richmond [1] for the generalization of these results to other simply generated families of trees).

For a tree T with the root v_0 let $\mathcal{F}(T)$ be the forest of rooted trees obtained from T by removing the root, where as the root of a tree $T' \in \mathcal{F}(T)$ we choose the vertex adjacent to v_0 in T. The height of a tree can be estimated using the following simple greedy algorithm, which constructs a long path starting at the root. In the first step the algorithm removes the root v_0 of $T^{(0)} = T$, chooses the largest tree $T^{(1)}$ from $\mathcal{F}(T^{(0)})$ (if there are more than one of them it picks one with the lexicographically first root) and appends its root to the path. This procedure is repeated until for some h tree $T^{(h)}$ consists only of one vertex.

The purpose of this paper is to apply a simple self-similarity idea to study the length $H = H(n)$ of the path found in a random tree by the above procedure. We characterize the limit distribution of H as the solution of some integral equation and show that the expected value of $H/\sqrt{n}$ tends to an absolute constant C, where

$$C = \frac{\sqrt{2\pi}}{2\sqrt{2} - \ln(3+2\sqrt{2})} = 2.353139\ldots .$$

Thus, in average, the algorithm constructs a path whose length is roughly 93% of the expected height of the tree.

The structure of the paper is the following. In the next section we prove some purely combinatorial facts concerning so called (k,l,m)-decompositions. Based on these results we characterize the limit distribution of H as the solution of some integral equation (section 3) and find the asymptotic value of the expectation of H (section 4). The paper is concluded by some additional remarks and comments on the behaviour of the algorithm.

2. (k,l,m)-decompositions.

A (k,l,m)-*decomposition* (P,F,S,R) of set $[n]$ is a quadruple of graphs, such that:

(i) P is a path $v_0 v_1 \ldots v_{k-1}$ starting at $v_0 = 1$;

(ii) F is a forest of k trees on $n-l-m$ vertices such that vertices $v_0, v_1, \ldots, v_{k-1}$ belong to different trees;

(iii) S is a rooted tree with l vertices and root v_k;

(iv) R is a tree on m vertices rooted at vertex v_{k+1};

(v) $[n] = V(F) \cup V(S) \cup V(R)$ is a partition of set $[n]$.

We say that a (k,l,m)-decomposition (P,F,S,R) is *contained* in T if T is the rooted tree with vertex set $[n]$, root v_0 and edges $E(T) = E(P) \cup E(F) \cup E(S) \cup E(R) \cup \{v_{k-1}v_k, v_k v_{k+1}\}$. A (k,l,m)-decomposition (P,F,S,R) is *sub-decomposition* of (k',l',m')-decomposition if both (P,F,S,R) and (P',F',S',R') are contained in the same tree and $P \subseteq P'$. We call a

(k,l,m)-decomposition (P,F,S,R) *bad* if $\mathcal{F}(S)$ contains a tree with more than l vertices (or a tree with precisely l vertices with the label of the root smaller than the label of the root of R) and *good* otherwise. Finally, we say that a decomposition *proper* if all its sub-decompositions are good.

It is not hard to see that the notion of decomposition emerges naturally in the analysis of the algorithm. Indeed, suppose that the algorithm run on a tree $T = T^{(0)}$ results in getting a path $P_h = v_0 v_1 \dots v_h$. Then, for every $k < h$, deleting from T edges $\{v_{k-1}, v_k\}$ $\{v_k, v_{k+1}\}$ splits the tree into three parts F_k, S_k and R_k which, together with the path $P_k = v_0 v_1 \dots v_{k-1}$, form a proper decomposition contained in T. On the other hand, if a proper decomposition (P,F,S,R) is contained in T, the algorithm run on T generates path P in the $(k-1)$-th step. Thus, $|T^{(k)}| = m$ if and only if $T = T^{(0)}$ contains some proper (k,l,m)-decomposition, and, since for given k (or given m) every tree contains at most one (k,l,m)-decomposition, studying the behaviour of the algorithm on random rooted trees can be reduced to an analysis of the probability that a random tree contains some proper (k,l,m)-decomposition of $[n]$.

Let $a(n;k,l,m)$ be the number of all (k,l,m)-decompositions of $[n]$ and let $\hat{a}(n;k,l,m)$ denote the probability that a random tree contains a (k,l,m)-decomposition, i.e. $\hat{a}(n;k,l,m) = a(n;k,l,m)/n^{n-2}$.

FACT 1. *For $k \geq 1$ we have*

$$\begin{aligned}\hat{a}(n;k,l,m) &= \frac{n!}{n^{n-1}}\frac{l^{l-1}}{l!}\frac{m^{m-1}}{m!}k\frac{(n-l-m)^{n-l-m-k-1}}{(n-m-l-k)!} \\ &= \frac{1}{2\pi}\frac{n^{3/2}}{l^{3/2}m^{3/2}}\frac{k}{(n-m-l)^{3/2}}\exp\left(-\frac{k^2}{2(n-m-l)}\right) \\ &\qquad \times \exp\left(O\left(\frac{k^3}{(n-m-l)^2}+\frac{k}{n-m-l}+\frac{1}{k}+\frac{1}{l}+\frac{1}{m}\right)\right),\end{aligned}$$

whereas

$$\hat{a}(n;0,l,m) = \begin{cases} \dfrac{n!}{n^{n-1}}\dfrac{l^{l-1}}{l!}\dfrac{(n-l)^{n-l-1}}{(n-l)!} & \text{if} \quad l+m=n \\ 0 & \text{if} \quad l+m \neq n. \end{cases}$$

Proof. To build a (k,l,m)-decomposition we must divide the set $\{2,3,\dots,n\}$ into four parts of $k-1$, $n-m-l-k$, l and m elements respectively, arrange the vertices of the first set in a path P in one of $(k-1)!$ ways, construct a rooted forest on the first two sets and vertex $v_0 = 1$ in one of $k(n-l-m)^{n-l-m-k-1}$ ways, and build rooted trees on each

from the remaining two sets. Thus, using Stirling's formula, we get

$$\begin{aligned}
\hat{a}(n;k,l,m) &= \frac{(n-1)!}{(k-1)!\,(n-m-l-k)!\,l!\,m!}(k-1)!k(n-l-m)^{n-l-m-k-1}\frac{l^{l-1}m^{m-1}}{n^{n-2}} \\
&= \frac{n!}{n^{n-1}}\frac{l^{l-1}}{l!}\frac{m^{m-1}}{m!}k\frac{(n-l-m)^{n-l-m-k-1}}{(n-m-l-k)!} \\
&= \frac{1}{2\pi}\frac{kn^{3/2}}{l^{3/2}m^{3/2}}\frac{(n-l-m)^{n-l-m-k-1}}{(n-m-l-k)^{n-m-l-k+1/2}} \\
&\qquad \times \exp(k-l+O(1/(n-m-l)+1/k+1/l+1/m)) \\
&= \frac{1}{2\pi}\frac{n^{3/2}}{l^{3/2}m^{3/2}}\frac{k}{(n-m-l)^{3/2}}\exp\left(-\frac{k^2}{2(n-m-l)}\right) \\
&\qquad \times \exp\left(O\left(\frac{k^3}{(n-m-l)^2}+\frac{k}{n-m-l}+\frac{1}{k}+\frac{1}{l}+\frac{1}{m}\right)\right)\ .
\end{aligned}$$

Similarly, for $l+m=n$ we have

$$\hat{a}(n;0,l,m)=\binom{n-1}{l-1}\frac{l^{l-2}(n-l)^{n-l-1}}{n^{n-2}}=\frac{n!}{n^{n-1}}\frac{l^{l-1}}{l!}\frac{(n-l)^{n-l-1}}{(n-l)!}\ . \quad \square$$

Although Fact 1 provides quite precise value of $a(n;k,l,m)$ we need to find a way to distinguish those decompositions which are proper. Clearly, if $m+l \geq n/2$, then $m \geq l$ is a sufficient condition for (k,l,m)-decomposition to be good, and thus, to be proper. Let us call a tree T *simple* if for T this condition turns out to be also necessary, i.e. if for every proper (k,l,m)-decomposition contained in T $l+m \geq n/2$ implies $m \geq l$. We shall show that almost every rooted tree is simple and, consequently, if $l+m \geq n/2$ and $m \geq l$ then we can approximate the probability that a random tree contains a proper (k,l,m)-decomposition by $\hat{a}(n;k,l,m)$. Let us start with the following simple fact on the structure of a random tree (we should mention that Meir and Moon [4] proved a stronger result using generating functions – we give here a proof for the completeness of the argument).

FACT 2. *If T_n is a random rooted tree then the probability that all components of $\mathcal{F}(T_n)$ are smaller than $\lfloor (n-1)/2 \rfloor + 1$ is of an order at most $O(n^{-1})$.*

Furthermore, the probability that $\mathcal{F}(T_n)$ consists of two components of equal sizes is at most $O(n^{-3/2})$.

Proof. We prove first the second part of the assertion. Obviously, we need to show it only for odd n. Let $n=2k+1$. Then the probability that $\mathcal{F}(T_n)$ consists of two components each of size k can be bounded from above by

$$\binom{2k}{k}\frac{k^{k-1}k^{k-1}}{(2k+1)^{2k-1}}=\frac{1+o(1)}{\sqrt{\pi}}\frac{2^{2k}}{\sqrt{k}}\frac{k^{2k-2}}{(2k+1)^{2k-1}}=O(k^{-3/2})=O(n^{-3/2})\,. \tag{3}$$

Now note that every tree T with the root v_0 for which one component of $\mathcal{F}(T)$ is larger than $\lfloor (n-1)/2 \rfloor$ can be uniquely decomposed into two rooted trees – a small one rooted at v_0 and a large one whose root is adjacent to v_0. Thus, the number of such trees is equal to

$$\sum_{k=\lfloor (n-1)/2 \rfloor + 1}^{n-1} \binom{n-1}{k} k^{k-1}(n-k)^{n-k-2} = \frac{1}{n} \sum_{k=\lfloor (n-1)/2 \rfloor + 1}^{n-1} \binom{n}{k} k^{k-1}(n-k)^{n-k-1}$$

$$= \frac{1}{2n} \sum_{k=1}^{n-1} \binom{n}{k} k^{k-1}(n-k)^{n-k-1} - \frac{1}{2n} \binom{n}{\lfloor (n-1)/2 \rfloor} (\lfloor (n-1)/2 \rfloor)^{2\lfloor (n-1)/2 \rfloor - 2} ,$$

where the last term appears only in the case when n is odd and, by elementary calculation very similar to those of (3), can be shown to be of an order $O(n^{n-7/2})$. Furthermore, observe that

$$\frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} k^{k-1}(n-k)^{n-k-1} = n^{n-2}(n-1) ,$$

since both sizes of the equality count the number of trees with one distinguished edge. Thus, for the probability that $\mathcal{F}(T_n)$ contains no tree of size at least $\lfloor n/2 \rfloor + 1$ we get

$$1 - \frac{1}{2n} \sum_{k=1}^{n-1} \binom{n}{k} \frac{k^{k-1}(n-k)^{n-k-1}}{n^{n-2}} + O(n^{-3/2}) = 1 - \frac{n-1}{n} + O(n^{-3/2}) = O(n^{-1}) . \quad \square$$

FACT 3. *A random tree T_n is simple with probability $1 - O(n^{-1/4})$.*

Proof. Let us suppose that the algorithm found a path $v_0 v_1 \ldots v_{k-1}$ in T_n and identified the vertex set $V^{(k)}$ of a tree $T_n^{(k)}$, where $|V_k| \geq n/2$. Since T_n is a random tree all rooted trees with vertex set $|V_k| = l + m$ are equally likely to appear as $T_n^{(k)}$, so with probability at least $1 - O(1/(l+m)) = 1 - O(n^{-1})$ the size m of the largest tree in $\mathcal{F}(T_n^{(k)})$ is larger than $\lfloor (m + l - 1)/2 \rfloor + 1$, i.e. $m \geq l$. Furthermore, (2) together with Markov's inequality implies that with probability $1 - O(n^{-1/4})$ the height of a random tree is less than $n^{3/4}$. Thus, the number of steps in the algorithm is at most $n^{3/4}$ and the probability that in some of them we generate a good (k,l,m)-decomposition with $l > m$ and $l + m \geq n/2$ is $O(n^{3/4-1}) = O(n^{-1/4})$. $\square$

FACT 4. *Let T_n be a random tree and $r \geq n/2$. Then, for every $r/2 \leq m \leq r$,*

$$p_n(r,k,m) = \mathrm{Prob}(\min_i \{|T_n^{(i)}| \leq r\} = k+1, |T_n^{(k+1)}| = m)$$

$$= (1 + O(n^{-1/4})) \sum_{l=r-m+1}^{\min\{m, n-m-k\}} \hat{a}(n;k,l,m)$$

so, for $k \geq 1$,

$$p_n(r,k,m) = \frac{1+O(n^{-1/4})}{2\pi} \sum_{l=r-m+1}^{\min\{m,n-m-k\}} \frac{n^{3/2}}{l^{3/2}m^{3/2}} \frac{k}{(n-m-l)^{3/2}} \exp\left(-\frac{k^2}{2(n-m-l)}\right)$$
$$\times \exp\left(O\left(\frac{k^3}{i^2} + \frac{1}{n-m-i} + \frac{1}{i} + \frac{1}{k} + \frac{1}{m}\right)\right).$$

Proof. If $|T_n^{(k)}| > r \geq n/2$ and $|T_n^{(k+1)}| = m \leq r$ then T_n contains a proper (k,l,m)-decomposition, for which $m \leq r$ but $m+l > r$. Since for a given k, a tree contains at most one (k,l,m)-decomposition, the assertion follows immediately from Facts 1 and 3. □

The case when $r = \lfloor n/2 \rfloor$ will be particularly important for our further considerations, so we state it in a bit stronger form as a separate lemma.

LEMMA 5. *Let f be a a function defined as*

$$f(x,y) = \frac{1}{2\pi} \int_{1/2-y}^{y} \frac{x}{t^{3/2}y^{3/2}(1-t-y)^{3/2}} \exp\left(-\frac{x^2}{2(1-y-t)}\right) dt, \tag{4}$$

where $x > 0$ and $y \in (1/4, 1/2)$. Then the probability $p_n(\lfloor n/2 \rfloor, \lfloor x\sqrt{n} \rfloor, \lfloor yn \rfloor)$ that for $k = \lfloor x\sqrt{n} \rfloor$ we have $|T_n^{(k)}| > \lfloor n/2 \rfloor$ and $|T_n^{(k+1)}| = \lfloor yn \rfloor \leq \lfloor n/2 \rfloor$ is given by

$$p_n(\lfloor n/2 \rfloor, \lfloor \alpha\sqrt{n} \rfloor, \lfloor \beta n \rfloor) = \frac{1+o(1)}{n^{3/2}} f(x,y)$$

where for every $\epsilon > 0$ the quantity $o(1)$ tends to 0 uniformly for every $x > \epsilon$ and $1/4 + \epsilon < y < 1/2 - \epsilon$. □

Proof. Similarly as Fact 4 the Lemma 5 follows from Facts 1 and 3. □

3. The limit distribution of H.

In this section we shall use a kind of rescaling argument to find the limit distribution of H, the length of the path constructed by the algorithm. Let us define recursively the two sequences of random variables $\{\hat{H}_i\}$ and $\{W_i\}$ setting $\hat{H}_0 = \min_j\{|T_n^{(j)}| \leq n/2\}$, $W_0 = |T_n^{(\hat{H}_0)}|$ whereas for $i \geq 1$ let

$$\hat{H}_i = \min_j\{|T_n^{(j)}| \leq W_{i-1}/2\}$$

and $W_i = |T_n^{(\hat{H}_i)}|$. Furthermore, set $H_0 = \hat{H}_0$ and $H_i = \hat{H}_i - \hat{H}_{i-1}$ for $1 \leq r \leq n-1$. Thus, W_i denotes the size of the tree $T_n^{(k)}$ when it first drops under $W_{i-1}/2$ and H_i is the number of steps of the algorithm between two such moments. Note that for every $i \geq 0$ we have $W_i \leq 2^{-i-1}n$.

Thus, the length of the path found by the algorithm can be written as a sum of H_i's, so

$$\begin{aligned} \text{Prob}(H > k) &= \text{Prob}(\sum_{i\geq 0} H_i > k) \\ &= \text{Prob}(H_0 > k) + \sum_{j\geq 1} \text{Prob}(\sum_{i=0}^{j} H_i > k \wedge \sum_{i=0}^{j-1} H_i \leq k) \ . \end{aligned} \tag{5}$$

In order to characterize the behaviour of the probabilities $\text{Prob}(\sum_{i=0}^{j} H_i > k \wedge \sum_{i=0}^{j-1} \leq k)$ we introduce an integral operator A setting

$$(Ag)(x) = \int_0^x \int_{1/4}^{1/2} f(z,y) g((x-z)/\sqrt{y}) dy dz \ ,$$

where f is the function defined by (4). Furthermore, let

$$g_0(x) = \int_x^\infty \int_{1/4}^{1/2} f(z,y) dy dz \tag{6}$$

and for $j \geq 1$

$$g_j = A g_{j-1} = A^j g_0 \ . \tag{7}$$

Note that the function g_j is well defined, i.e. all integrals which appear in its definition converge. As a matter of fact, our next result gives an explicit upper bound for the value of g_j.

FACT 6. *For every $j \geq 0$, g_j is a non-negative function, bounded above by 1, such that*

$$\int_0^\infty g_j(x) dx \leq \Big(\int_0^\infty \int_{1/4}^{1/2} \sqrt{y} f(x,y) dy dx \Big)^j < 2^{-j/2} \ .$$

Proof. Note first that $f(x,y)$ is related to the density of a random variable so

$$\int_0^\infty \int_{1/4}^{1/2} f(x,y) dy dx = 1 \tag{8}$$

(which, of course, can be verified by direct computation) and the assertion follows for g_0. Now assume it holds for g_{j-1}. Then

$$g_j(x) = \int_0^x \int_{1/4}^{1/2} f(z,y) g_{j-1}((x-z)/\sqrt{y}) dy dz \leq \int_0^x \int_{1/4}^{1/2} f(z,y) dy dz < 1 \ .$$

Similarly,

$$\begin{aligned}\int_0^\infty g_j(x)dx &= \int_0^\infty \int_0^x \int_{1/4}^{1/2} f(z,y)g_{j-1}((x-z)/\sqrt{y})dydzdx \\ &= \int_0^\infty \int_0^\infty \int_{1/4}^{1/2} \sqrt{y}f(z,y)g_{j-1}(u)dydudz \\ &= \int_0^\infty \int_{1/4}^{1/2} \sqrt{y}f(z,y)dydz \int_0^\infty g_{j-1}(u)du \\ &< \sqrt{2}/2 \int_0^\infty g_{j-1}(u)du \le 2^{-j/2} \ . \ \square\end{aligned}$$

Our next result shows that functions g_j are closely related to our problem.

LEMMA 7. *For every $x > 0$ we have*

$$\mathrm{Prob}(H_0 > \lfloor x\sqrt{n} \rfloor) = (1 + o(1))g_0 \,,$$

and for $j \ge 1$

$$\mathrm{Prob}(\sum_{i=0}^{j} H_i > \lfloor x\sqrt{n} \rfloor \wedge \sum_{i=0}^{j-1} H_i \le \lfloor x\sqrt{n} \rfloor) = (1 + o(1))g_i(x) \,,$$

where, for given positive constants c, C, the quantity $o(1)$ tends to 0 uniformly for every $x \in (c, C)$.

Proof. We shall use the induction over j. The formula for $\mathrm{Prob}(H_0 > \lfloor x\sqrt{n} \rfloor)$ is a straightforward consequence of Lemma 5. On the other hand, for every k and $j \ge 1$

$$\mathrm{Prob}(\sum_{i=0}^{j} H_i > k \wedge \sum_{i=0}^{j-1} H_i \le k) = \sum_{l=1}^{k-j} \mathrm{Prob}(H_0 = l \wedge \sum_{i=1}^{j} H_i > k-l \wedge \sum_{i=1}^{j-1} H_i \le k-l)$$

(9)

$$= \sum_m \sum_l \mathrm{Prob}(\sum_{i=1}^{j} H_i > k-l \wedge \sum_{i=1}^{j-1} H_i \le k-l | H_0 = l \wedge W_0 = m) \, \mathrm{Prob}(H_0 = l \wedge W_0 = m).$$

Now we shall use a kind of 'rescaling' idea. As we have already noticed in the proof of Fact 3, in the first k steps the algorithm uses no information about tree $T^{(k)}$ except for its size. Thus, when the algorithm is run on a random tree, in each step we can treat $T_n^{(k)}$ as a random tree on $|T_n^{(k)}|$ vertices. Hence the probability $\mathrm{Prob}(\sum_{i=1}^{j} H_i > k-l \wedge \sum_{i=0}^{j-1} H_i \le k-l | H_0 = l \wedge W_0 = m)$ is precisely the probability that if we run the algorithm on a random tree with m vertices then $\sum_{i=0}^{j-1} H_i > k-l$ and $\sum_{i=0}^{j-2} H_i \le k-l$. Since due to Lemma 5 the function $f(x,y)$ determines the joint distribution of (H_0, W_0), the assertion follows from (9) and the definition of A. $\square$

As a consequence of Lemma 7 we get the limit distribution of H.

THEOREM 8. *For every constant* $x \geq 0$

$$\lim_{n\to\infty} \mathrm{Prob}(H > x\sqrt{n}) = h(x),$$

where

$$h(x) = \sum_{j=0}^{\infty} g_j(x) = \sum_{j=0}^{\infty} (A^j g_0)(x)$$

and functions g_j *are defined by (6) and (7). Equivalently, the function* h *is the only continuous solution of the integral equation*

$$\begin{aligned} h(x) &= g_0(x) + (Ah)(x) \\ &= \int_x^{\infty} \int_{1/4}^{1/2} f(z,y)dydz + \int_0^x \int_{1/4}^{1/2} f(z,y)h((x-z)/\sqrt{y})dydz \ , \end{aligned} \tag{10}$$

where the function f *is given by (4).*

Proof. Fix constants $\delta > 0$ and $\epsilon > 0$ and pick J and N large enough so that the probability that the actual height of a random tree with less than $2^{-J-1}n$ vertices is larger than $\delta\sqrt{n}$ is smaller than $\epsilon/2$ for every $n \geq N$ (the existence of such constants follows from Theorem 0). From Lemma 7 and (5) it follows that we may approximate the probability $\mathrm{Prob}(\sum_{j=0}^{J-1} H_j \geq x\sqrt{n})$ by $\sum_{j=0}^{J-1} g_j(x)$. But $P(\sum_{j\geq J} H_j \geq \delta\sqrt{n})$ is the probability that the algorithm will find a path of length larger than $\delta\sqrt{n}$ in a tree W_n of less than $\delta\sqrt{n}$ vertices, and, according to our assumption, is smaller than $\epsilon/2$. Thus, for every $\delta, \epsilon > 0$ we can choose n large enough, so that

$$h(x) - \epsilon \leq \mathrm{Prob}(H > (x+\delta)\sqrt{n}) \leq \mathrm{Prob}(H > x\sqrt{n}) \leq h(x) + \epsilon \ .$$

This proves the first part of the assertion.

The second part of Theorem 8 follows from the fact that, due to Fact 6, the series $\sum_j g_j$ converges in the L_1-norm. Thus, a function h is uniquely determined up to the set of measure zero, and since the kernel of the integral equation is absolutely continuous in the whole range of the integration, h can be chosen to be continuous. □

Remark. Let us note that once we know that the limit $\lim_{n\to\infty} \mathrm{Prob}(H > x\sqrt{n})$ exists for each $x \geq 0$ the fact that h fulfills the integral equation $h = g_0 + Ah$ is quite natural and follows immediately from the fact that $H = H_0 + \lambda H$, where λ is a random variable which plays the role of a 'scaling factor'.

4. The expectation of H.

Having computed the distribution of H it is not hard to guess the value of its mean. Clearly, $EH/\sqrt{n}$ should converge to the expected value of the random variable Z, where $P(Z > x) = h(x)$ and $h(x)$ is given by Theorem 8. But $xh(x) \to 0$ as $x \to \infty$ (as a matter

of fact Theorem 0 says that the probability that the actual height of a random tree is larger than x decreases exponentially with x), so

$$\mu = \mathrm{E}\, Z = \int_0^\infty h(x)dx.$$

Now if we integrate both sides of (9), after elementary calculations we arrive at

$$\mu = \int_0^\infty \int_{1/4}^{1/2} xf(x,y)dydx + \mu \int_0^\infty \int_{1/4}^{1/2} \sqrt{y} f(x,y)dydx \tag{11}$$

so, consequently,

$$\mu = \frac{\int_0^\infty \int_{1/4}^{1/2} xf(x,y)dydx}{1 - \int_0^\infty \int_{1/4}^{1/2} \sqrt{y} f(x,y)dydx} \ .$$

Note furthermore that, similarly as (10), the equation (11) can easily be deduced from the 'scaling' relation $H = H_0 + \lambda H$, once we know that the expectation of $H/\sqrt{n}$ exists. Unfortunately, the existence of the limit $\lim_{n\to\infty} \mathrm{E}\, H/\sqrt{n}$ is *not* implied by the existence of the limit distribution $h(x)$ (even if one can prove that the convergence is uniform for every $x \in (0,\infty)$ which indeed is the case – see section 5). Hence, we shall deduce (11) from Lemma 5, following the way which led us to Theorem 8.

We find first the limit distributions of random variables H_i. Not surprisingly, we shall do it recursively, using a kind of integral operator.

Thus, let B be an operator which maps an integrable function r into the function Br such that

$$(Br)(x) = \int_0^\infty \int_{1/4}^{1/2} \frac{f(z,y)}{\sqrt{y}} r\left(\frac{x}{\sqrt{y}}\right) dydz \ ,$$

where function f is given by (4).

Let us note two simple properties of B.

FACT 9. *For every non-negative, integrable function r*

$$\int_0^\infty (Br)(x)dx = \int_0^\infty r(x)dx$$

and, if $m = \int_0^\infty r(x)dx < \infty$, then

$$\int_0^\infty x(Br)(x)dx = m \int_0^\infty \int_{1/4}^{1/2} \sqrt{y} f(z,y)dydz < \frac{\sqrt{2}}{2} m < \infty \ . \tag{12}$$

Proof. Both equalities follows immediately from (8) and the definition of B. □

Now, for $x \geq 0$, let

$$r_0(x) = \int_{1/4}^{1/2} f(x,y)dy$$

and for $j \geq 1$

$$r_j = Br_{j-1} \ .$$

Then the distribution of H_j is characterized by the following local limit theorem.

LEMMA 10. *For every $j \geq 0$ and $x > 0$*

$$\lim_{n\to\infty} \sqrt{n}\,\mathrm{Prob}(H_j = \lfloor xn \rfloor) = (1 + o(1))r_j(x) \ ,$$

where for every given constants $0 < c < C < \infty$ the quantity $o(1)$ tends to 0 uniformly for $x \in (c, C)$.

Proof. For $j = 0$ the assertion follows straightforwardly from Lemma 5. Now note that for $j \geq 1$ we have

$$\mathrm{Prob}(H_j = k_j) = \sum_k \sum_m \mathrm{Prob}(H_j = k_j | H_0 = k \wedge W_0 = m)\,\mathrm{Prob}(H_0 = k \wedge W_0 = m) \ .$$

But, similarly as in the proof of Lemma 6, the probability $\mathrm{Prob}(H_j = k_j | H_0 = k \wedge W_0 = m)$ is just the probability that $H_{j-1} = k_j$ in a random tree on m vertices. Thus, the assertion follows from Lemma 5. □

THEOREM 11.

$$\lim_{n\to\infty} \frac{\mathrm{E}\,H}{\sqrt{n}} = \mu \ ,$$

where

$$\mu = \sum_{j=0}^{\infty} \int_0^\infty x r_j(x)dx = \frac{\int_0^\infty \int_{1/4}^{1/2} x f(x,y)dydx}{1 - \int_0^\infty \int_{1/4}^{1/2} \sqrt{y} f(x,y)dydx} \ .$$

Proof. Fix $\epsilon > 0$. Note first that, since $\sum_{j\geq J} H_j$ is the length of the path found by the algorithm in a random tree of size $W_j \leq 2^{-J-1}$, choosing J large enough we can made the expectation of $\sum_{j\geq J} H_j$ smaller than $0.1\epsilon\sqrt{n}$ for large n (by (2) we can take any J such that $\sqrt{2\pi}2^{-J/2-1/2} \leq 0.1\epsilon$). Now, for $j \geq 0$ and $a < b$ define a random variable $H_j(a,b)$ setting

$$H_j(a,b) = \begin{cases} H_j & \text{if} \quad a\sqrt{n} < X < b\sqrt{n} \\ 0 & \text{otherwise.} \end{cases}$$

Choose a constant C such that for all j, $0 \leq j \leq J$, and n large enough

(i) $\mathrm{E}\,H_j(C,\infty)/\sqrt{n} \leq 0.1\epsilon/J$,

(ii) $\int_C^\infty x r_j(x)dx \leq 0.1\epsilon$.

Note that such a constant C exists since, by (2),

$$\mathrm{E}\,H_j/\sqrt{n} \leq EH/\sqrt{n} \leq 3$$

and Fact 9 implies that $\int_0^\infty x r_j(x)dx < 2^{-j/2} < \infty$. Due to Lemma 10 the function r_j approximates uniformly the distribution of H_j in every finite interval, so, for every j, $0 \leq j \leq J$, and n large enough we have

$$\left| \mathrm{E}\,H_j(0.1\epsilon/J, C)/\sqrt{n} - \int_{0.1\epsilon/J}^{C} x r_j(x)dx \right| \leq 0.3\epsilon/J \ .$$

Thus,

$$\begin{aligned}
\Big|\,\mathrm{E}\,H/\sqrt{n} - \sum_{j\ge 0}\int_0^\infty x r_j(x)dx\Big| &\le \Big|\sum_{j\ge 0}\mathrm{E}\,H_j/\sqrt{n} - \sum_{j\ge 0}\int_0^\infty x r_j(x)dx\Big| \\
&\le \Big|\sum_{j=0}^{J}\mathrm{E}\,H_j/\sqrt{n} - \sum_{j=0}^{J}\int_0^\infty x r_j(x)dx\Big| + 0.2\epsilon \\
&\le \sum_{j=0}^{J}\Big|\,\mathrm{E}\,H_j(0.1\epsilon/J, C)/\sqrt{n} - \int_{0.1\epsilon/J}^{C} x r_j(x)dx\Big| + \sum_{j=0}^{J}\mathrm{E}\,H(0, 0.1\epsilon/J)/\sqrt{n} \\
&\quad + \sum_{j=0}^{J}\Big(\mathrm{E}\,H(C,\infty)/\sqrt{n} + \int_0^{0.1\epsilon/J} x r_j(x)dx + \int_C^\infty x r_j(x)dx\Big) + 0.2\epsilon < \epsilon. \;\square
\end{aligned}$$

COROLLARY 12.

$$\lim_{n\to\infty}\frac{\mathrm{E}\,H}{\sqrt{n}} = \frac{\sqrt{2\pi}}{2\sqrt{2} - \ln(3+2\sqrt{2})} = 2.353139\ldots .$$

Proof. In order to verify the assertion one needs only to compute integrals given by Theorem 11. It is elementary, but tedious and not very exciting task, so we shall leave it to the favourite symbolic mathematical computation program of the reader. $\square$

5. Final remarks and comments.

The main purpose of this note was to present a simple rescaling idea which allows us to 'guess' and verify the asymptotic behaviour of the algorithm without invoking generating functions. Thus, for the sake of simplicity, we have not stated our results in the strongest possible form. One can easily see that, due to the self-similarity argument, the local limit distribution of H, defined as

$$\hat{h}(x) = \lim_{n\to\infty}\sqrt{n}\,\mathrm{Prob}(H = \lfloor x\sqrt{n}\rfloor)\;,$$

is the only solution of the integral equation

$$\hat{h}(x) = \int_{1/4}^{1/2} f(x,y)dy + \int_0^x\int_{1/4}^{1/2}\frac{f(z,y)}{\sqrt{y}}\hat{h}\left(\frac{x-z}{\sqrt{y}}\right)dydz\;,$$

but the proof of the existence of $\hat{h}$ is slightly more involved than that for $h(x)$. One can also show that the convergence of $\mathrm{Prob}(H > x\sqrt{n})$ is uniform for every $x \in (0,\infty)$, and that $\sqrt{n}\,\mathrm{Prob}(H = \lfloor x\sqrt{n}\rfloor)$ tends to $\hat{h}$ uniformly for $x \in (c,\infty)$, for every $c > 0$.

It is not very hard to investigate how many vertices are 'used' by the algorithm to generate a path of length $k = k(n)$. One can study two natural random variables related

to this problem: $X^{(k)} = n - |T^{(k)}|$ which counts vertices 'rejected' after first k steps, and $Y^{(m)} = \max_k\{|T^{(k)}| \geq n-m\}$ which tells us about the length of the path generated before rejecting m vertices. (Let us note that the second choice seems to be a bit better: both medians and expectation of $Y^{(k)}$ are of the order $\sqrt{m}$, whereas the median of $X^{(k)}$ is of the order k^2 but its expectation grows like $k\sqrt{n}$. This rather unpleasant behaviour of $X^{(k)}$ is not unexpected and for $k = 1$ has been well known and studied (see Meir and Moon [4]).) The limit distribution of $X^{(k)}$ for $k = o(\sqrt{n})$ and $Y^{(m)}$ for $m \leq n/2$ can be immediately derived from Fact 4. However in order to investigate the asymptotic behaviour of $X^{(k)}$ and $Y^{(m)}$ for large k and m, one should use an argument similar to that we use to prove Theorem 8.

Corollary 12 states that the expectation of the length H of the path found by the algorithm is more than 93% of the expectation of the actual height $\tilde{H}$ of a random tree. The greedy procedure is also quite quick: if a random tree is given in, say, preordered form it finds a path of length H in the expected time $O(H)$. On the other hand, one should keep in mind that the height of a vertex chosen at random from a random tree is $\sum_{j=2}^{n}(n)_j/n^j$ (Meir and Moon [3]) which, for large n, is just a half of the expectation of the height of the tree $\tilde{H}$. Note also that the ratio $\mathrm{E}\,\tilde{H}/\mathrm{E}\,H$ is a rather crude measure of the efficiency of the algorithm. It would be also interesting (and possibly even more adequate) to study the behaviour of the random variable $\tilde{H}/H$. More specifically, one can ask what is the length of the path found by the algorithm in a tree $T_{n,k}$, chosen at random from the family of all rooted trees of n vertices and height $k = k(n)$. For large k, i.e. when $k/\sqrt{n} \to \infty$, the structure of $T_{n,k}$ was studied in [2]. One can show that in this case with probability tending to 1 as $n \to \infty$ the algorithm constructs a path of length $k - O(n/k)$. Nevertheless, the case when k is small, and, in particular, when k is of the order of $\sqrt{n}$, seems to be much harder to handle.

Finally, let us mention that an analogous rescaling idea can be applied to study the behaviour of the algorithm run on other probabilistic models of random trees, provided that

- every branch of a random tree of size m can be treated as a random tree of size m (so we can use self-similarity argument)
- we know the formula for the number of forests which consists of such a trees (so the scaling function f can be effectively computed).

Acknowledgement. I would like to thank Boris Pittel who introduced this problem to me.

REFERENCES

[1] P. Flajolet, J. Gao, A. M. Odlyzko and B. Richmond, *The distribution of heights of binary trees and other simple trees*, INRIA TR 1749, 1992.

[2] T. Łuczak, *The number of trees with a large diameter*, J. Austral. Math. Soc.. to appear

[3] A. Meir and J. W. Moon, *On the altitude of nodes in random trees*, Can. J. Math., **30** (1978), pp. 997–1015.
[4] A. Meir and J. W. Moon, *On major and minor branches of rooted trees*, Can. J. Math., **39** (1987), pp. 673–693.
[5] A. Rényi and G. Szekeres, *On the height of trees*, J. Austral. Math. Soc., **7** (1967), pp. 497–507.