On the distribution of random walk hitting times in random trees.*

Joubert Oosthuizen and Stephan Wagner†

Abstract

The hitting time H_{xy} between two vertices x and y of a graph is the average time that the standard simple random walk takes to get from x to y. In this paper, we study the distribution of the hitting time between two randomly chosen vertices of a random tree. We consider both uniformly random labelled trees and a more general model with vertex weights akin to simply generated trees. We show that the r-th moment of the hitting time is of asymptotic order $n^{3r/2}$ in trees of order n, and we describe the limiting distribution upon normalisation by means of its moments. Moreover, we also obtain joint moments with the distance between the two selected vertices.

Finally, we discuss a somewhat different model of randomness, namely random recursive trees. In this setup, the root is of special importance, and so we study the hitting time from the root to a random vertex or from a random vertex to the root. Interestingly, the hitting time from the root is of order $n \log n$, with a normal limit law, while the hitting time to the root is only of linear order and has a non-Gaussian limit law.

1 Introduction

We consider the standard simple random walk on a finite graph, which starts at a vertex x and moves at each step to one of the neighbours, chosen uniformly at random. This is a standard example of a finite Markov chain, and it is well known that every vertex is eventually reached with probability 1. The average time it takes the random walk to reach y from x is called the *hitting time* (or *first passage time*) and denoted by H_{xy}.

The hitting time is a very natural parameter associated with a random walk and as such, it has been studied quite thoroughly. The first results on the distribution of the hitting time in random trees are due to Moon [14], who obtained the asymptotic behaviour of the first two moments of the hitting time between two randomly selected nodes of a uniformly random labelled tree. Aldous [1] describes a way to obtain the mean in the setting of the continuum random tree (the continuous limit of uniformly random labelled trees and other random tree families) and mentions that the variance could be treated in a similar way, but that "the computations look messy". We remark that the random walk on random trees has been shown to converge to Brownian motion on the continuum random tree in great generality [4]. However, interchangeability of the various limiting processes in this context is far from trivial.

Aldous also mentions the *cover time* (the time that the random walk needs to visit all vertices) in [1]. This parameter is much harder to analyse. Not even the asymptotic behaviour of the mean is known, although a heuristic argument is provided in [2]. A bound for the mean is given in [10].

In [12], Löwe and Torres consider the average hitting time from the root of a subcritical Galton-Watson tree to a random vertex and determine its order of magnitude in dependence of a parameter. An alternative approach that leads to essentially the same result was described in the recent paper [10] on the *cover cost* of a vertex in a graph. This quantity $CC(x)$ is simply defined as the sum of all hitting times from x:

$$CC(x) = \sum_y H_{xy}.$$

Making use of a connection to electrical networks, it can be shown that the cover cost of a vertex

*This material is based upon work supported by the National Research Foundation of South Africa under grant number 96236.

†Department of Mathematical Sciences, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

in a tree can be expressed in terms of distances. Specifically, if we let $D(x) = \sum_w d(x,w)$ be the (total) *distance* from x (i.e., the sum of all distances from x) and $W(T) = \sum_{\{x,y\}} d(x,y) = \frac{1}{2}\sum_x D(x)$ the *Wiener index* of a tree (i.e., the sum of all distances between pairs of vertices in T), then we have

$$CC(x) = 2W(T) - D(x).$$

This formula, combined with results of Janson [11] on the Wiener index, makes it possible to derive the asymptotic behaviour of the mean cover cost of a random vertex in a random labelled tree of order n, and even to characterise the limiting distribution. Up to a factor n, the mean cover cost is the same as the average hitting time from a random vertex to another random vertex, which is also known as the *mean first passage time* (MFPT). This quantity has also been studied in the physics literature, see for example [3] for a recent paper on the mean first passage time of specific trees.

In this paper, we refine the aforementioned combinatorial analysis of Moon further by providing asymptotic formulas for all moments of the hitting time, which are sufficient to characterise the limiting distribution. In Section 3, we consider a uniformly random labelled tree with n vertices and two randomly chosen vertices x and y and study the distribution of the hitting time H_{xy}. We point out that H_{xy} is a deterministic quantity once the tree and the two vertices have been selected. As it turns out, the hitting time is typically of asymptotic order $n^{3/2}$, which is the same order of magnitude as that of the total distance of a random vertex. We will also be able to characterise the distribution obtained by normalising with a factor $n^{-3/2}$ by means of its moments. In its most basic form, our main result reads as follows:

THEOREM 1.1. *Let x and y be two randomly selected vertices of a uniformly random labelled tree with n vertices. The r-th moment of the hitting time between x and y satisfies the following asymptotic formula:*

$$\mathbb{E}(H_{xy}^r) \sim C_r n^{3r/2},$$

where

$$C_r = \frac{\sqrt{\pi}}{\Gamma(\frac{1}{3})} \cdot \Big(\frac{3}{\sqrt{2}}\Big)^r \cdot \frac{\Gamma(r+1)\Gamma(r+\frac{1}{3})}{\Gamma(\frac{3r+1}{2})}.$$

Consequently, the normalised random variable $n^{-3/2}H_{xy}$ converges weakly to a limit law that is characterised by the moment sequence C_r. This limit law has a continuous density on $[0,\infty)$.

Specifically, the mean of H_{xy} is asymptotically equal to $\sqrt{\frac{\pi}{2}}n^{3/2}$, while the variance is asymptotically equal to $\big(\frac{32}{15} - \frac{\pi}{2}\big)n^3$, which was already known to Moon [14].

In fact, this theorem holds in greater generality, and with the same limit law up to a constant factor, if we consider labelled trees with additional vertex weights that depend on the degrees (akin to simply generated trees, but without roots). See Section 4 for further details. We will also be able to characterise the joint moments with the distance $d(x,y)$ between the two random vertices x and y.

Finally, we consider a rather different model of random trees, namely *recursive trees*. These can be obtained by means of a growth process, where the n-th vertex is attached to one of the previous vertices chosen uniformly at random. In recursive trees, the root plays a special role, so instead of choosing two random vertices, we study the hitting time to and from the root.

Not only are the growth order and the distribution of the hitting time very different in recursive trees (as compared to random labelled trees), it also makes a major difference whether the hitting time from the root to a random vertex or the hitting time from a random vertex to the root are considered. As we will see in Section 5, the hitting time from the root is of average order $n \log n$ and satisfies a central limit theorem with a Gaussian limit law, while the hitting time to the root is only of order n and has a non-Gaussian limit law. Intuitively speaking, the reason for this remarkable phenomenon is the fact that the root of a recursive tree is generally very central and therefore "attracts" a random walk faster than a random vertex, which typically lies closer to the fringes. The limit behaviour is somewhat reminiscent of the *total path length*, the total distance (sum of all distances) from the root. In rooted labelled trees, its

mean order is $n^{3/2}$ [11, 17], while in recursive trees it is $n \log n$ [6, 13].

2 Combinatorial decomposition

We first describe a combinatorial decomposition of the hitting time between two vertices in a tree. To this end, consider first the hitting time along a single edge xy. Let S be the subtree consisting of all vertices that are closer to x than to y (including x itself), and let S consist of x and the subtrees $S_1, S_2, \dots, S_k$ (see Figure 1).

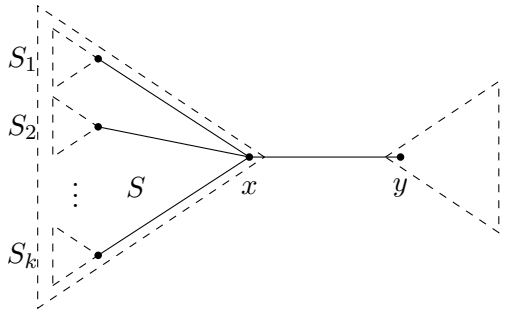

Figure 1: Hitting time along a single edge xy.

We first show that

$$H_{xy} = 2|S| - 1 \tag{1}$$

in this situation, where (here and in the following) $|S|$ denotes the number of vertices in S. This can be done by induction: the formula is trivial for $k = 0$ (and thus $|S| = 1$). For $k > 0$, the random walk moves to y with probability $\frac{1}{k+1}$ and otherwise moves a step back to one of the subtrees S_j. We then need to add 1 (for the first step), the expected time to return to x (which is $2|S_j| - 1$ by the induction hypothesis) and H_{xy} to obtain the expected time the random walk takes to reach y. Hence

$$H_{xy} = \frac{1}{k+1} + \frac{1}{k+1} \sum_{j=1}^{k} \Big(1 + 2|S_j| - 1 + H_{xy}\Big),$$

which yields

$$\begin{aligned} \frac{1}{k+1} H_{xy} &= \frac{1}{k+1} + \frac{1}{k+1} \sum_{j=1}^{k} 2|S_j| \\ &= \frac{1}{k+1} \Big(1 + 2(|S| - 1)\Big), \end{aligned}$$

and the desired formula follows.

Now we look at the more general situation where x and y are not necessarily neighbours. Consider a decomposition of a tree T along the path between the two vertices x and y. The vertices on this path are denoted by $x = w_0, w_1, w_2, \dots, w_d = y$. When the edges of this path are removed, we are left with $d + 1$ components, each of which can be interpreted as a rooted tree. We let T_j be the tree rooted at w_j, see Figure 2.

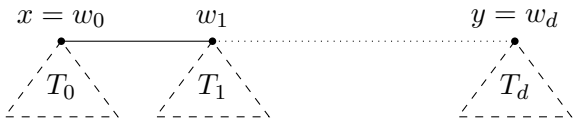

Figure 2: Decomposition along the path between x and y.

The hitting time can be broken up as follows:

$$H_{xy} = H_{w_0 w_1} + H_{w_1 w_2} + \cdots + H_{w_{d-1} w_d},$$

and since $H_{w_j w_{j+1}} = 2(|T_0| + |T_1| + \cdots + |T_j|) - 1$ by (1), we get

$$H_{xy} = \sum_{j=0}^{d-1} \Big(2 \sum_{h=0}^{j} |T_h| - 1\Big),$$

which reduces to

$$H_{xy} = 2 \sum_{j=0}^{d} (d - j)|T_j| - d. \tag{2}$$

Note in particular that the size of T_d is in fact irrelevant in this formula, since it is multiplied by 0. This is because a random walk starting at x never visits vertices of T_d before reaching y.

We remark that (2) can also be obtained in other ways, e.g. by means of a general formula due to Tetali (see [18]): for any two vertices x and y in a tree,

$$H_{xy} = \frac{1}{2} \sum_{w} d(w)\big(d(x, y) + d(w, y) - d(w, x)\big), \tag{3}$$

where the sum is over all vertices w of the graph and $d(w)$ denotes the degree of w. For more general graphs, the distance $d(\cdot, \cdot)$ in this formula needs to be replaced by the so-called *effective resistance*.

Our first task in the following section will be to translate the formula (2) for the hitting time to the world of generating functions. By means of singularity analysis (see [8], Chapter VI), we will then be able to derive Theorem 1.1. In Section 4, we briefly outline how to modify the argument in order to generalise the result to random labelled trees with vertex weights.

3 Random labelled trees

We first consider the most basic random tree model, where a labelled tree is chosen uniformly at random, and two vertices x and y are chosen uniformly and independently afterwards. The two vertices are allowed to coincide, in which case the hitting time is simply 0. Cayley's classical formula states that the number of labelled trees with n vertices is exactly n^{n-2}. Including the two vertices x and y, we have precisely n^n different choices, each of which is equally likely.

Let us now define a bivariate exponential generating function, where the second variable u marks the hitting time. Formally,

$$H(z,u) = \sum_{T,x,y} \frac{1}{|T|!} z^{|T|} u^{H_{xy}},$$

the sum being over all possible choices of a tree T and two vertices x and y. Recall also that $|T|$ denotes the number of vertices in T.

We can write $H(z,u) = \sum_{d\geq 0} H_d(z,u)$, where H_d is the generating function for the case that the distance between the two vertices is exactly d:

$$H_d(z,u) = \sum_{\substack{T,x,y \\ d(x,y)=d}} \frac{1}{|T|!} z^{|T|} u^{H_{xy}}.$$

For fixed d, we use the decomposition in Figure 2 and the resulting formula (2). If $|T| = n$, we can distribute the labels among the subtrees $T_0, \dots, T_d$ in $\binom{n}{|T_0|, |T_1|, \dots, |T_d|}$ ways. Each of the T_j's can be regarded as an independent rooted labelled tree (up to the canonical order-preserving relabelling). Hence we have

$$\begin{aligned} H_d(z,u) &= \sum_{T_0,T_1,\dots,T_d} \frac{u^{-d}}{n!} \binom{n}{|T_0|, |T_1|, \dots, |T_d|} \\ &\qquad \prod_{j=0}^{d} z^{|T_j|} u^{2(d-j)|T_j|} \\ &= u^{-d} \prod_{k=0}^{d} Y\left(zu^{2k}\right), \end{aligned}$$

where

$$Y(z) = \sum_{n=1}^{\infty} \frac{n^{n-1}}{n!} z^n$$

is the exponential generating function for rooted labelled trees. It is well known that $Y(z)$ (which is closely related to the Lambert W-function) has a dominant square-root singularity at e^{-1}, with an asymptotic expansion of the form

$$Y(z) = 1 - \sqrt{2(1-ez)} + O(|1-ez|).$$

We can derive the factorial moments of the hitting time from the probability generating function

$$\sum_k \mathbb{P}\left(H_{xy} = k\right) u^k = \frac{[z^n]H(z,u)}{[z^n]H(z,1)}$$

by taking partial derivatives with respect to u and setting $u = 1$. It follows from the product rule that

$$\begin{aligned} &\partial_u H(z,u)\Big|_{u=1} \\ &= \sum_{d=0}^{\infty} -dY(z)^{d+1} + \sum_{d=0}^{\infty} d(d+1) z Y'(z) Y(z)^d \\ &= -\frac{Y(z)^2}{(1-Y(z))^2} + \frac{2zY(z)Y'(z)}{(1-Y(z))^3} \\ &\sim \frac{1}{2(1-ez)^2}. \end{aligned}$$

Singularity analysis yields

$$\mathbb{E}\left(H_{xy}\right) = \frac{[z^n]\partial_u H(z,u)\Big|_{u=1}}{[z^n]H(z,1)} \sim \sqrt{\frac{\pi}{2}} n^{3/2}.$$

We require some auxiliary calculations to find the higher moments in general. In the following, it will be shown that $\mathbb{E}\left(H_{xy}^r\right)$ is of order $n^{3r/2}$ for all r.

It will be somewhat more convenient to work with the generating function for $H_{xy} + d(x, y)$ though, since it does not contain the additional factor u^{-d}:

$$\begin{aligned}\tilde{H}(z,u) &= \sum_{T,x,y} \frac{1}{|T|!} z^{|T|} u^{H_{xy}+d(x,y)} \\ &= \sum_{d=0}^{\infty} \prod_{k=0}^{d} Y\left(zu^{2k}\right).\end{aligned}$$

The r-th derivative with respect to u yields the r-th factorial moment. The asymptotic behaviour of this r-th derivative is determined in the following lemma:

LEMMA 3.1. *Around the dominant singularity* $\frac{1}{e}$, *we have*

$$\begin{aligned}&\partial_u^r \tilde{H}(z,u)\Big|_{u=1} \\ &\quad\sim 2^{-(r+1)/2}(1-ez)^{-\frac{3r+1}{2}} \\ &\quad\sum_{m=1}^{r} \sum_{h_1+\ldots+h_m=r} \binom{r}{h_1, h_2, \ldots, h_m} (r+m)! \\ &\quad\prod_{i=1}^{m} (2h_i-3)!! \prod_{i=1}^{m} \frac{1}{h_1+\ldots+h_i+i}. \qquad (4)\end{aligned}$$

Proof. The general Leibniz rule gives us

$$\begin{aligned}&\partial_u^r \prod_{k=0}^{d} Y\left(zu^{2k}\right)\Big|_{u=1} \\ &= \sum_{m=1}^{r} \sum_{\substack{h_1,h_2,\ldots,h_m\geq 1 \\ h_1+\ldots+h_m=r}} \binom{r}{h_1, h_2, \ldots, h_m} \\ &\quad\sum_{0\leq j_1<\ldots<j_m\leq d} Y(z)^{d+1-m} \prod_{i=1}^{m} \left[\partial_u^{h_i} Y\left(zu^{2j_i}\right)\right]\Big|_{u=1}.\end{aligned}$$

The asymptotic behaviour of the factors is given by

$$\begin{aligned}\partial_u^h Y\left(zu^{2j}\right)\Big|_{u=1} &\sim (2jz)^h Y^{(h)}(z) \\ &\sim \sqrt{2}(2h-3)!! j^h (1-ez)^{1/2-h}\end{aligned}$$

for all $h \geq 1$. The resulting sum over $j_1, j_2, \ldots, j_m$ is

$$\begin{aligned}\sum_{0\leq j_1<j_2<\ldots<j_m\leq d} \prod_{i=1}^{m} {j_i}^{h_i} &= \frac{d^{h_1+h_2+\ldots+h_m+m}}{\prod_{i=1}^{m}(h_1+\ldots+h_i+i)} \\ &\quad + O\left(d^{h_1+h_2+\ldots+h_m+m-1}\right),\end{aligned}$$

as can be seen by induction on m, and finally

$$\sum_{d=m-1}^{\infty} d^{r+m} Y(z)^{d+1-m} \sim \frac{(r+m)!}{(1-Y(z))^{r+m+1}}.$$

Putting all these together gives us the desired statement.

After some further manipulations, the double sum simplifies considerably:

LEMMA 3.2. *Around the dominant singularity* $\frac{1}{e}$, *we have*

$$\begin{aligned}&\partial_u^r \tilde{H}(z,u)\Big|_{u=1} \\ &\sim 2^{-(r+1)/2} r! (3r-2)!!! (1-ez)^{-(3r+1)/2}, \qquad (5)\end{aligned}$$

where the triple factorial $n!!!$ *is defined recursively as*

$$n!!! = \begin{cases} n & 0 < n \leq 3, \\ n \cdot (n-3)!!! & n > 3. \end{cases}$$

Proof. We can write

$$\begin{aligned}&\prod_{i=1}^{m} \frac{1}{h_1+\ldots+h_i+i} \\ &= \int_{0\leq x_1\leq\ldots\leq x_m\leq 1} \cdots \int x_1^{h_1} x_2^{h_2} \ldots x_m^{h_m} dx_1 dx_2 \ldots dx_m,\end{aligned}$$

which yields

$$\begin{aligned}&\sum_{m=1}^{r} \sum_{h_1+\ldots+h_m=r} r!(r+m)! \prod_{i=1}^{m} \frac{(2h_i-3)!!}{h_i!} \\ &\prod_{i=1}^{m} \frac{1}{h_1+\ldots+h_i+i} \\ &= \sum_{m=1}^{r} \frac{r!(r+m)!}{m!} [u^r] \left(\int_0^1 \sum_{h\geq 1} \frac{(2h-3)!! u^h x^h}{h!}\, dx\right)^m \\ &= \sum_{m=1}^{r} \frac{r!(r+m)!}{m!} [u^r] \left(\int_0^1 \left(1-\sqrt{1-2ux}\right) dx\right)^m \\ &= \sum_{m=1}^{r} \frac{r!(r+m)!}{m!} [u^r] \left(\frac{3u-1+(1-2u)^{3/2}}{3u}\right)^m \\ &= r!^2 [u^r] \left(\left(\frac{1-(1-2u)^{3/2}}{3u}\right)^{-r-1} - 1\right).\end{aligned}$$

Finally,

$$\left[u^r\right]\left(\frac{1-(1-2u)^{3/2}}{3u}\right)^{-r-1} = \frac{1}{2\pi i}\oint_{\mathcal{C}} u^{-r-1}\left(\frac{1-(1-2u)^{3/2}}{3u}\right)^{-r-1} du,$$

and the simple substitution $3t = 1-(1-2u)^{3/2}$ shows that this is equal to

$$\left[t^r\right](1-3t)^{-1/3} = \frac{(3r-2)!!!}{r!}.$$

Hence the formula of Lemma 3.1 simplifies to the statement of Lemma 3.2.

Given Lemma 3.2, a standard application of singularity analysis yields

LEMMA 3.3. *The r-th factorial moment of $H_{xy}+d(x,y)$ and thus also the r-th moment of $H_{xy}+d(x,y)$, are asymptotically given by*

$$\begin{aligned}\mathbb{E}\left((H_{xy}+d(x,y))^r\right) &\sim \mathbb{E}\left((H_{xy}+d(x,y))^{\underline{r}}\right)\\ &= \frac{[z^n]\partial_u^r \tilde{H}(z,u)\Big|_{u=1}}{[z^n]\tilde{H}(z,1)}\\ &\sim \frac{\sqrt{\pi}r!(3r-2)!!!}{2^{\frac{r}{2}}\Gamma\left(\frac{3r+1}{2}\right)} n^{\frac{3r}{2}}.\end{aligned}$$

Now Theorem 1.1 follows by a simple induction on r by noticing that

$$\begin{aligned}&\mathbb{E}\left((H_{xy}+d(x,y))^r\right)\\ &= \mathbb{E}\left(H_{xy}^r\right) + O\left(\sum_{k=0}^{r-1}\binom{r}{k}\mathbb{E}\left(H_{xy}^k\right) n^{r-k}\right).\end{aligned}$$

It only remains to show that the moment sequence C_r defines a limiting distribution with a continuous density on $[0,\infty)$. We shall make use of the fact that a random variable with absolutely integrable characteristic function has a continuous density on $\mathbb{R}$ (see e.g. [9], Chapter 13.7, Theorem 14). We consider the moment generating function associated with the moment sequence C_r, i.e.,

$$C(t) = \sum_{r=0}^{\infty}\frac{C_r}{r!}t^r.$$

We symmetrise it as follows: let Z be a random variable whose moment generating function is $C(t)$ and whose characteristic function is thus $C(it)$. An auxiliary random variable Y that is equal to Z with probability $\frac{1}{2}$ and $-Z$ otherwise has characteristic function

$$\frac{C(it)+C(-it)}{2} = {}_2F_2\left(1, \tfrac{2}{3}; \tfrac{5}{6}, \tfrac{1}{2}; -\tfrac{2t^2}{3}\right),$$

where ${}_2F_2$ denotes a generalised hypergeometric function. This is $O(|t|^{-4/3})$ as $t\to\pm\infty$ (see [5, §16.11(ii)]), so it represents an absolutely integrable function. Hence the random variable Y has a continuous density on $\mathbb{R}$, and Z has a continuous density on $[0,\infty)$.

Without going into detail, let us mention that it is possible to obtain joint moments of the hitting time H_{xy} with the distance $d(x,y)$ between the two randomly selected vertices. To this end, one considers the modified generating function

$$\begin{aligned}H^{(s)}(z,u) &= \sum_{T,x,y}\frac{1}{|T|!}z^{|T|}d(x,y)^s u^{H_{xy}}\\ &= \sum_{d\geq 0} d^s H_d(z,u).\end{aligned}$$

Now the r-th derivative with respect to u yields the joint moment $\mathbb{E}(d(x,y)^s H_{xy}^r)$. We have the following result:

THEOREM 3.1. *Let x and y be two randomly selected vertices of a random labelled tree with n vertices. The joint moments of the distance $d(x,y)$ and the hitting time H_{xy} between x and y satisfy the asymptotic formula*

$$\mathbb{E}(d(x,y)^s H_{xy}^r) \sim C_{r,s} n^{3r/2+s/2},$$

where

$$\begin{aligned}C_{r,s} = &\frac{\sqrt{\pi}(-3)^{r+s}r!(r+s)!}{2^{(r+3s)/2}\Gamma(\frac{3r+s+1}{2})}\\ &\cdot\sum_{k=0}^{s}(-1)^k\binom{s}{k}\binom{(2k-1)/3}{r+s}.\end{aligned}$$

Specifically, the covariance of distance and hitting time is asymptotically equal to $\left(2-\frac{\pi}{2}\right)n^2$. Details are left to the full version of this paper.

4 Weighted trees

In this section, we describe how Theorem 1.1 is generalised to labelled trees with vertex weights based on degrees. The resulting trees are similar to simply generated trees (which are in turn essentially equivalent to Galton-Watson trees), see [7] for a general reference. The main difference is that our trees do not have a root. We let w be a weight function on the set of positive integers and set

$$w(T) = \prod_{v \in T} w(d(v)).$$

Now we generate a random labelled tree in such a way that the probability is proportional to the weight. For example, if we set $w(j) = 1$ for $j \leq \Delta$ and $w(j) = 0$ otherwise, then we obtain uniformly random labelled trees whose maximum degree is at most Δ.

Let us only describe the main differences to the previous section. The decomposition of Figure 2 can still be used, but there is now one small difference: the trees T_0 and T_d give rise to factors that are slightly different from the others. The reason is that the degree of w_0 in T is equal to the degree of w_0 in T_0 plus 1 (and likewise for w_d), while for $j \notin \{0, d\}$, the degree of w_j in T is equal to the degree of w_j in T_j plus 2. The generating function $A(z)$ associated with the subtrees T_0 and T_d satisfies

$$\begin{aligned} A(z) &= z \sum_{d=0}^{\infty} \frac{w(d+1)}{d!} A(z)^d \\ &= z\phi(A(z)), \end{aligned}$$

where we set $\phi(t) = \sum_{d=0}^{\infty} \frac{w(d+1)}{d!} t^d$. The generating function $\tilde{A}(z)$ associated with all other subtrees $T_1, T_2, \ldots, T_{d-1}$, on the other hand, can be written in terms of $A(z)$:

$$\begin{aligned} \tilde{A}(z) &= z \sum_{d=0}^{\infty} \frac{w(d+2)}{d!} A(z)^d \\ &= z\phi'(A(z)) \\ &= 1 - \frac{A(z)}{zA'(z)}. \end{aligned}$$

Translating the decomposition of Figure 2 to generating functions in the setup of weighted trees now gives us

$$\begin{aligned} HW_d(z,u) &= \sum_{\substack{T,x,y \\ d(x,y)=d}} \frac{w(T)}{|T|!} z^{|T|} u^{H_{xy}} \\ &= u^{-d} A(z) A\left(zu^{2d}\right) \prod_{k=1}^{d-1} \tilde{A}\left(zu^{2k}\right) \end{aligned}$$

for $d \geq 1$, and we set

$$HW(z,u) = \sum_{d=0}^{\infty} HW_d(z,u).$$

In order to proceed in the same way as in the previous section, we need information on the singular behaviour of $A(z)$ and $\tilde{A}(z)$. It is well known that the functional equation $A(z) = z\phi(A(z))$, which characterises simply generated families of trees, gives rise to a square-root singularity under some technical conditions (see [7], Chapter 3.1.4, or [8], Chapter VII.4). Suppose that the power series that defines ϕ has positive radius of convergence, and that there exists a solution τ of the equation $\phi(t) = t\phi'(t)$. Moreover, we assume for simplicity that ϕ is aperiodic, which means here that $\gcd\{d : w(d+1) \neq 0\} = 1$. Then $A(z)$ has a square-root singularity at $\rho = \tau/\phi(\tau) = 1/\phi'(\tau)$, and the expansion is of the form

$$A(z) = \tau - \lambda\Big(1 - \frac{z}{\rho}\Big)^{1/2} + O\Big(\Big|1 - \frac{z}{\rho}\Big|\Big),$$

where $\lambda = \sqrt{\frac{2\phi(\tau)}{\phi''(\tau)}}$. From this, we derive that

$$\tilde{A}(z) = 1 - \frac{2\tau}{\lambda}\Big(1 - \frac{z}{\rho}\Big)^{1/2} + O\Big(\Big|1 - \frac{z}{\rho}\Big|\Big).$$

Now we can continue in a similar way as in the previous section to obtain

$$\begin{aligned} \mathbb{E}\left(H_{xy}\right) &= \frac{[z^n]\partial_u HW(z,u)\Big|_{u=1}}{[z^n]HW(z,1)} \\ &\sim \frac{\frac{\phi(\tau)}{2\phi''(\tau)}\phi'(\tau)^n n}{\tau\sqrt{\frac{\phi(\tau)}{2\phi''(\tau)\pi}}\phi'(\tau)^n n^{-1/2}} \\ &\sim \frac{1}{\tau}\sqrt{\frac{\pi\phi(\tau)}{2\phi''(\tau)}}n^{3/2} \end{aligned}$$

and in general

$$\mathbb{E}(H_{xy}^r) \sim C_r \left(\frac{1}{\tau} \sqrt{\frac{\phi(\tau)}{\phi''(\tau)}} \right)^r n^{3r/2},$$

with C_r as in Theorem 1.1. Moreover, an analogue of Theorem 3.1 holds as well. We leave the details of the procedure to the full version.

5 Random recursive trees

A *recursive tree* is a rooted labelled tree with the property that the sequence of labels along any path starting at the root is increasing. Random recursive trees can be generated by a simple growth process: starting with the root (labelled 1), the n-th vertex is attached to one of the previous $n-1$ vertices, chosen uniformly at random. It is easy to see that there are precisely $(n-1)!$ recursive trees with n vertices.

The structure of random recursive trees differs substantially from that of random labelled trees; for instance, it is known that the height of recursive trees with n vertices is typically of order $\log n$, while the height of rooted labelled trees is typically of order $\sqrt{n}$ (see [7], Chapters 4.2.7 and 6.4, respectively). We observe a similar phenomenon for hitting times. As it has been mentioned in the introduction, the root plays a special role in recursive trees, so we will consider only a single random vertex and study its hitting time to and from the root.

It is well known that the depth (distance from the root) of a random vertex in a random recursive tree with n vertices is equidistributed with the number of cycles (minus 1) of a random permutation of n elements, see [7, Lemma 6.16]. Therefore, the probability generating function of the depth factorises into a simple product of $n-1$ linear factors, and its coefficients are the Stirling numbers of the first kind. Surprisingly, there is still such a factorisation if the hitting time is included in the probability generating function. Let y be a vertex of a recursive tree; the root always carries label 1, so H_{1y} is the hitting time from the root, while H_{y1} is the hitting time to the root. The key result in this section is the following theorem:

THEOREM 5.1. *The joint probability generating function of the depth $d(1,y)$ of a random vertex y in a random recursive tree and the hitting time H_{1y} from the root to y is given by*

$$\begin{aligned} p_n(u,v) &= \frac{1}{n!} \sum_{T,y} u^{H_{1y}} v^{d(1,y)} \\ &= \frac{1}{n!} \prod_{j=1}^{n-1} \big(n - j + u^{2j-1} v\big). \end{aligned} \tag{6}$$

Likewise, the joint probability generating function of the depth $d(y,1)$ of a random vertex y in a random recursive tree and the hitting time H_{y1} from y to the root is given by

$$\begin{aligned} q_n(u,v) &= \frac{1}{n!} \sum_{T,y} u^{H_{y1}} v^{d(y,1)} \\ &= \frac{1}{n!} \prod_{j=1}^{n-1} \big(j + u^{2j-1} v\big). \end{aligned} \tag{7}$$

This theorem can be proven by setting up a recursion for p_n and verifying the formula by induction. An alternative approach uses a combinatorial argument: we count the total number of pairs of a recursive tree T and a vertex y such that the subtrees in the decomposition of Figure 2, where x is the root, have prescribed sizes $a_0, a_1, \dots, a_d$. The root $x = w_0$ must be labelled 1, and there are $\binom{n-1}{a_0-1}$ choices for the remaining labels of T_0, as well as $(a_0-1)!$ possibilities for the shape of T_0. The label of the next vertex w_1 must be the minimum of the remaining labels. There are $\binom{n-a_0-1}{a_1-1}$ possibilities for the other labels of T_1, and $(a_1-1)!$ for the shape of T_1. Repeating the argument, we find that there are

$$\begin{aligned} &\prod_{j=0}^{d} \binom{n - a_0 - \cdots - a_{j-1} - 1}{a_j - 1} (a_j - 1)! \\ &\quad = \frac{(n-1)!}{\prod_{j=1}^{d} (n - a_0 - a_1 - \cdots - a_{j-1})} \\ &\quad = \frac{(n-1)!}{\prod_{j=1}^{d} (a_j + a_{j+1} + \cdots + a_d)} \end{aligned}$$

possible tree-vertex pairs, given $a_0, a_1, \dots, a_d$. This is also exactly the coefficient of

$x_{a_d} x_{a_{d-1}+a_d} \cdots x_{a_1+a_2+\cdots+a_d}$ in the expansion of

$$\prod_{j=1}^{n-1}(j+x_j).$$

Hence the probability that the sizes of the subtrees $T_0, T_1, \ldots, T_d$ from the root to a random vertex of a random recursive tree are precisely $a_0, a_1, \ldots, a_d$ is the coefficient of $x_{a_d} x_{a_{d-1}+a_d} \cdots x_{a_1+a_2+\cdots+a_d}$ in the expansion of

$$\frac{1}{n!}\prod_{j=1}^{n-1}(j+x_j). \tag{8}$$

Now recall that the hitting time from the root to y is

$$H_{xy} = 2\sum_{j=0}^{d}(d-j)a_j - d$$

by (2), which can be rewritten as

$$H_{xy} = 2\sum_{j=0}^{d-1}\sum_{h=0}^{j} a_h - d = \sum_{j=0}^{d-1}\Big(2\Big(n - \sum_{h=j+1}^{d} a_h\Big) - 1\Big).$$

Thus we get the desired bivariate probability generating function for the hitting time from the root to y (marked by u) and the depth of y (marked by v) by replacing each x_j in (8) by $u^{2(n-j)-1}v$. Formula (6) follows immediately, and (7) is obtained analogously.

We remark that the hitting times H_{xy} and H_{yx} are connected by the identity

$$H_{xy} + H_{yx} = 2(n-1)d(x,y)$$

in trees (the sum $H_{xy} + H_{yx}$ is also known as *commute time*). This follows directly from (2), or also from Tetali's formula (3). Thus the two formulas in Theorem 5.1 are in fact equivalent. We see that both H_{1y} and H_{y1} can be written as a sum of $n-1$ independent, but not identically distributed, random variables.

Our next goal is to infer information on the distribution of H_{1y} and H_{y1} from these product representations. Let us first have a look at mean and variance: here, Theorem 5.1 immediately yields

COROLLARY 5.1. *Let H_{1y} be the hitting time from the root of a random recursive tree with n vertices to a random vertex, and H_{y1} the hitting time from a random vertex to the root. The mean and variance of H_{1y} are given by*

$$\mathbb{E}(H_{1y}) = (2n+1)H_n - 4n + 1 \sim 2n\log n$$

and

$$\begin{aligned}\mathbb{V}(H_{1y}) &= (2n+1)(2n+5)H_n - (2n+1)^2 H_n^{(2)} \\ &\quad - 6n(n+1) \\ &\sim 4n^2 \log n\end{aligned}$$

respectively. The mean and variance of H_{y1} are given by

$$\mathbb{E}(H_{y1}) = (2n+1) - 3H_n \sim 2n$$

and

$$\mathbb{V}(H_{y1}) = 2n(n-7) + 21H_n - 9H_n^{(2)} \sim 2n^2$$

respectively. Here, $H_n = \sum_{j=1}^{n}\frac{1}{j}$ and $H_n^{(2)} = \sum_{j=1}^{n}\frac{1}{j^2}$ are harmonic numbers of first and second order, respectively.

Proof. In view of (6), H_{1y} can be regarded as a sum of independent random variables. The j-th of these variables is equal to $2j-1$ with probability $\frac{1}{n-j+1}$ and 0 otherwise. Consequently, the expected value is

$$\mathbb{E}(H_{1y}) = \sum_{j=1}^{n-1}\frac{2j-1}{n-j+1} = (2n+1)H_n - 4n + 1.$$

The other formulas follow analogously.

We observe that the asymptotic orders of mean and variance differ between H_{1y} and H_{y1}: getting to the root from a vertex y of a recursive tree is typically faster than getting from the root to y. It is known in general that the vertices of any graph can be arranged in a preorder $\prec$ in such a way that $x \prec y$ implies $H_{xy} \le H_{yx}$. In a tree, there is a simple characterisation in terms of distances (see [10]): $H_{xy} \le H_{yx}$ if and only if $D(x) \ge D(y)$ (recall that $D(x) = \sum_w d(x,w)$). Since the centroid (i.e., the vertex that minimises the total distance D)

is typically close to the root of a recursive tree (as shown by Moon in [15]), one can certainly expect that $\mathbb{E}(H_{1y}) \geq \mathbb{E}(H_{y1})$. Nevertheless, the difference in the order of magnitude is somewhat surprising.

Moreover, we find that H_{1y} is concentrated, while H_{y1} is not. Thus we also have different limiting distributions, which are described in the following theorem:

THEOREM 5.2. *The hitting time from the root is asymptotically normally distributed:*

$$\frac{H_{1y} - \mathbb{E}(H_{1y})}{\sqrt{\mathbb{V}(H_{1y})}} \xrightarrow{d} N(0,1).$$

On the other hand, the hitting time to the root converges upon normalisation by a factor n^{-1} to a random variable Z:

$$n^{-1} H_{y1} \xrightarrow{d} Z.$$

This random variable has a continuous density on $[0,\infty)$.

Proof. In both cases, we can determine the limit of the moment generating function of the normalised random variable by a direct calculation. First, we consider the hitting time from the root. Let

$$P_n(t) := \mathbb{E}\left(e^{tH_n^*}\right)$$

be the moment generating function of the normalised variable

$$H_n^* := \frac{H_{1y} - 2n\log n}{2n\sqrt{\log n}}.$$

The product representation (6) for the probability generating function $p_n(u,v)$ gives us (after applying a change of variable $j \mapsto n+1-j$)

$$P_n(t) = e^{-t\sqrt{\log n}} \prod_{j=2}^{n} \left(\frac{j-1}{j} + \frac{e^{\frac{t(2n-2j+1)}{2n\sqrt{\log n}}}}{j} \right).$$

The expression in the exponent inside the product is uniformly bounded, hence we can approximate the exponential function by its Taylor expansion:

$$e^{\frac{t(2n-2j+1)}{2n\sqrt{\log n}}} = 1 + \frac{t(2n-2j+1)}{2n\sqrt{\log n}} + \frac{t^2(2n-2j+1)^2}{8n^2\log n} + O\big((\log n)^{-3/2}\big).$$

This yields

$$\begin{aligned} P_n(t) &= e^{-t\sqrt{\log n}} \prod_{j=2}^{n} \left(1 + \frac{t(2n-2j+1)}{2jn\sqrt{\log n}} \right. \\ &\qquad \left. + \frac{t^2(2n-2j+1)^2}{8jn^2\log n} + O\big(j^{-1}(\log n)^{-3/2}\big)\right) \\ &= \exp\left(-t\sqrt{\log n} + \sum_{j=2}^{n} \frac{t(2n-2j+1)}{2jn\sqrt{\log n}} \right. \\ &\qquad \left. + \sum_{j=2}^{n} \frac{t^2(2n-2j+1)^2}{8jn^2\log n} + O\big((\log n)^{-1/2}\big)\right). \end{aligned}$$

Evaluating the sums, we arrive at

$$P_n(t) = \exp\left(\frac{t^2}{2} + O\big((\log n)^{-1/2}\big)\right).$$

Hence the limit as $n \to \infty$ is exactly the moment generating function of a standard normal distribution, which proves the first part of the theorem.

For the second part, we follow essentially the same lines. Again, we consider the moment generating function of a suitably normalised random variable, namely

$$Q_n(t) = \mathbb{E}\left(e^{tH_{y1}/n}\right).$$

Now the product representation (7) yields

$$Q_n(t) = \exp\left(\sum_{j=2}^{n} \log\left(1 + \frac{e^{\frac{(2j-3)t}{n}} - 1}{j}\right)\right).$$

We note that

$$\frac{e^{\frac{(2j-3)t}{n}} - 1}{j} = \frac{e^{\frac{2jt}{n}} - 1}{j} + O\Big(\frac{1}{jn}\Big) = O\Big(\frac{1}{n}\Big)$$

holds uniformly in j, so

$$Q_n(t) = \exp\left(\sum_{j=1}^{n} \frac{e^{\frac{2jt}{n}} - 1}{j} + O\Big(\frac{\log n}{n}\Big)\right).$$

The remaining sum

$$\sum_{j=1}^{n} \frac{e^{\frac{2jt}{n}} - 1}{j} = \frac{2}{n} \sum_{j=1}^{n} \frac{e^{\frac{2jt}{n}} - 1}{2j/n}$$

can be interpreted as a Riemann sum for the integral

$$\int_0^{2t} \frac{e^x - 1}{x} dx,$$

so it follows that

$$\lim_{n\to\infty} Q_n(t) = \exp\Big(\int_0^{2t} \frac{e^x - 1}{x} dx \Big).$$

Again, the sequence of moment generating functions has a limit (we denote it by $Q(t)$), which implies weak convergence. In order to show that the limiting distribution has a continuous density on $[0, \infty)$, we follow the same approach as in the proof of Theorem 1.1. Let Z be a random variable whose characteristic function is

$$\begin{aligned} Q(it) &= \exp\left(\int_0^{2it} \frac{e^x - 1}{x} dx \right) \\ &= \frac{e^{-\gamma}}{2t} \exp\big(\operatorname{Ci}(2t) + i \operatorname{Si}(2t) \big), \end{aligned}$$

where γ is the Euler-Mascheroni constant and Ci and Si denote cosine integral and sine integral, respectively. The auxiliary random variable Y that is equal to Z with probability $\frac{1}{2}$ and $-Z$ otherwise has characteristic function

(9)
$$\frac{Q(it) + Q(-it)}{2} = \frac{e^{-\gamma}}{2t} \exp\big(\operatorname{Ci}(2t) \big) \cos\big(\operatorname{Si}(2t) \big).$$

Since $\operatorname{Ci}(2t) = O(t^{-1})$ and $\operatorname{Si}(2t) = \frac{\pi}{2} + O(t^{-1})$ as $t \to \infty$ (see [5, §6.12(ii)]), it follows that the expression in (9) is $O(t^{-2})$ as $t \to \pm\infty$, so it represents an absolutely integrable function. Hence the random variable Y has a continuous density on $\mathbb{R}$, and Z has a continuous density on $[0, \infty)$.

Let us finally remark that it is possible to study the joint distribution of the hitting times H_{1y} and H_{y1} and the depth $d(1, y)$ as well. For instance, the covariance of H_{1y} and $d(1, y)$ is

$$(2n + 3)H_n - (2n + 1)H_n^{(2)} - 2n \sim 2n \log n,$$

while the covariance of H_{y1} and $d(1, y)$ is

$$2n - 5H_n + 3H_n^{(2)} \sim 2n.$$

References

[1] David Aldous, *The continuum random tree. II. An overview*, Stochastic analysis (Durham, 1990), London Math. Soc. Lecture Note Ser., vol. 167, Cambridge Univ. Press, Cambridge, 1991, pp. 23–70.

[2] ———, *Random walk covering of some special trees*, J. Math. Anal. Appl. **157** (1991), no. 1, 271–283.

[3] Francesc Comellas and Alicia Miralles, *Mean first-passage time for random walks on generalized deterministic recursive trees*, Physical Review E **81** (2010), 061103.

[4] David Croydon, *Convergence of simple random walks on random discrete trees to Brownian motion on the continuum random tree*, Ann. Inst. Henri Poincaré Probab. Stat. **44** (2008), no. 6, 987–1019.

[5] *NIST Digital library of mathematical functions*, `http://dlmf.nist.gov/`, Release 1.0.10 of 2015-08-07, 2015, Online companion to [16].

[6] Robert P. Dobrow and James Allen Fill, *Total path length for random recursive trees*, Combin. Probab. Comput. **8** (1999), no. 4, 317–333.

[7] Michael Drmota, *Random trees*, SpringerWienNewYork, Vienna, 2009.

[8] Philippe Flajolet and Robert Sedgewick, *Analytic combinatorics*, Cambridge University Press, Cambridge, 2009.

[9] Bert Fristedt and Lawrence Gray, *A modern approach to probability theory*, Probability and its Applications, Birkhäuser Boston, Inc., Boston, MA, 1997.

[10] Agelos Georgakopoulos and Stephan Wagner, *Hitting times, cover cost, and the Wiener index of a tree*, Journal of Graph Theory, to appear, 2016.

[11] Svante Janson, *The Wiener index of simply generated random trees*, Random Structures Algorithms **22** (2003), no. 4, 337–358.

[12] Matthias Löwe and Felipe Torres, *A note on hitting times for simple random walk on rooted, subcritical Galton-Watson trees*, Preprint, 2014.

[13] Hosam M. Mahmoud, *Limiting distributions for path lengths in recursive trees*, Probab. Engrg. Inform. Sci. **5** (1991), no. 1, 53–59.

[14] J. W. Moon, *Random walks on random trees*, J. Austral. Math. Soc. **15** (1973), 42–53.

[15] ———, *On the centroid of recursive trees*, Australas. J. Combin. **25** (2002), 211–219.

[16] Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark (eds.), *NIST*

Handbook of mathematical functions, Cambridge University Press, New York, 2010.

[17] Lajos Takács, *The asymptotic distribution of the total heights of random rooted trees*, Acta Sci. Math. (Szeged) **57** (1993), no. 1-4, 613–625.

[18] Prasad Tetali, *Random walks and the effective resistance of networks*, J. Theoretical Prob. **4** (1991), 101–109.