Combinatorics, Probability and Computing (2004) 13, 677–696. © 2004 Cambridge University Press
DOI: 10.1017/S0963548304006194 Printed in the United Kingdom

On Some Parameters in Heap Ordered Trees

KATE MORRIS[1], ALOIS PANHOLZER[2†]
and HELMUT PRODINGER[1]

[1]The John Knopfmacher Centre for Applicable Analysis and Number Theory, School of Mathematics,
University of the Witwatersrand, P. O. Wits, 2050 Johannesburg, South Africa
(e-mail: kate@maths.wits.ac.za, helmut@maths.wits.ac.za)

[2]Institut für Algebra und Computermathematik, TU Wien,
Wiedner Hauptstr. 8–10, 1040 Wien, Austria
(e-mail: Alois.Panholzer@tuwien.ac.at)

Received 7 November 2002; revised 31 December 2003

Heap ordered trees are planted plane trees, labelled in such a way that the labels always increase from the root to a leaf. We study two parameters, assuming that p of the n nodes are selected at random: the size of the ancestor tree of these nodes and the smallest subtree generated by these nodes. We compute expectation, variance, and also the Gaussian limit distribution, the latter as an application of Hwang's quasi-power theorem.

1. Introduction

A *heap ordered tree* with n nodes ('size n') can be described as a *planted plane tree* together with a bijection from the nodes to the set $\{1, \ldots, n\}$, which is *monotonically increasing* when going from the root to the leaves.

Some recent research papers [11, 12] deal with statistics of the height of the nodes in heap ordered trees. Now, the height of a given node is defined as the number of nodes lying on the unique path from the root to this node. In this paper we consider a simple generalization of the height: for p given nodes in a heap ordered tree T we consider the size of the ancestor tree of these selected nodes. To be more precise, the ancestor tree is the subtree of T which is spanned by the root and the p chosen nodes and hence it is defined as the tree containing all ascendants of the p given nodes.

Spanning tree size and the Wiener index for binary search trees have been computed in [7] and [10]. The *Wiener index* of a graph is the sum of all distances between pairs of

[†] Part of this work was done during the second author's visit to the John Knopfmacher Centre for Applicable Analysis and Number Theory at the University of the Witwatersrand, Johannesburg, South Africa.

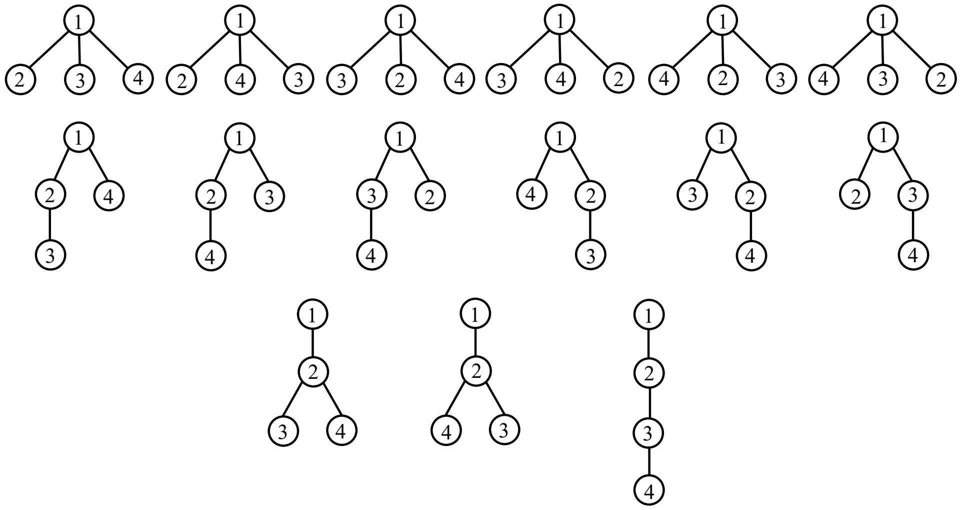

Figure 1. All 15 heap ordered trees with 4 nodes

nodes in the graph. It was introduced by the chemist H. Wiener in 1947 [13] in the study of organic compounds and their molecular graphs. The Wiener index of simply generated trees has been studied, for example, in the papers by Entringer, Meir, Moon and Szekely [4] and R. Neininger [8], and has numerous applications in chemistry and combinatorics.

A related parameter of interest is the *Steiner distance*. The Steiner distance of a graph is the expected distance of two random nodes in the graph. So, the Steiner distance is a scaled down version of the Wiener index; in a sense they behave roughly like path length versus (insertion) depth. For expectations, the concepts are equivalent, but not for higher moments and the limiting distribution. We consider a natural generalization: instead of selecting two random nodes and looking at the distance, we consider p randomly chosen nodes and look at the size of the subtree spanned by these nodes. A different generalization of the Steiner distance can, for example, be found in [3].

In this paper we aim to compute the expectation and variance for the size of the ancestor trees and the Steiner distance in heap ordered trees. Also, we will consider the limiting distributions involved. For the parameters we discuss the distributions turn out to be Gaussian and we will use Hwang's *quasi-power theorem* (see [6]) to determine them. For the convenience of the reader we include this important theorem here.

Theorem (H. K. Hwang). *Let $\{\Omega_n\}_{n\geqslant 1}$ be a sequence of integral random variables. Suppose that the moment-generating function satisfies the asymptotic expression*

$$M_n(s) = \mathbb{E}(e^{\Omega_n s}) = \sum_{m\geqslant 0} \mathbb{P}\{\Omega_n = m\} e^{ms} = e^{H_n(s)}\big(1 + O\big(\kappa_n^{-1}\big)\big),$$

the O-term being uniform for $|s| \leqslant \tau$, $s \in \mathbb{C}$, $\tau > 0$, where

(i) *$H_n = u(s)\phi(n) + v(s)$, with $u(s)$ and $v(s)$ analytic for $|s| \leqslant \tau$ and independent of n, $u''(0) \neq 0$,*

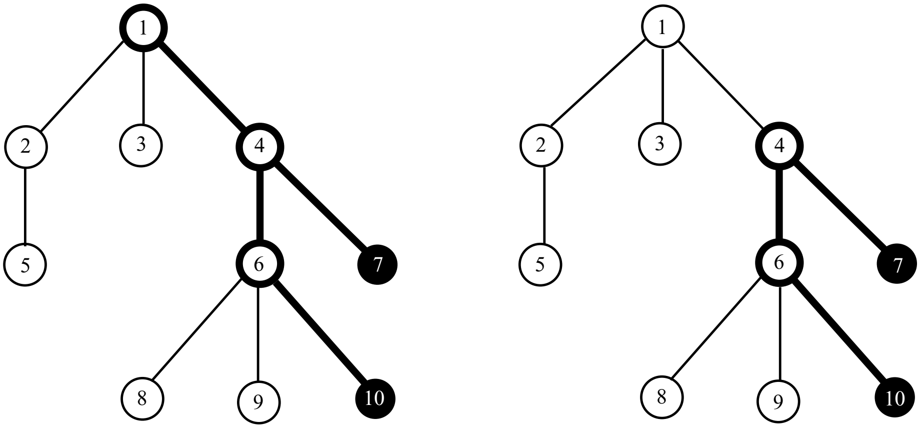

Figure 2. A heap ordered tree of size 10 with the two parameters under consideration; nodes 7 and 10 are labelled

(ii) $\phi(n) \to \infty$,

(iii) $\kappa_n \to \infty$.

Under these assumptions the distribution of Ω_n *is asymptotically Gaussian:*

$$\mathbb{P}\left\{ \frac{\Omega_n - u'(0)\phi(n)}{\sqrt{u''(0)\phi(n)}} < x \right\} = \Phi(x) + O\left(\frac{1}{\kappa_n} + \frac{1}{\sqrt{\phi(n)}} \right),$$

uniformly with respect to x, $x \in \mathbb{R}$. *Here* $\Phi(x)$ *denotes the distribution function of the standard normal distribution* $\mathcal{N}(0,1)$. *Moreover, the mean and variance of* Ω_n *satisfy*

$$\mathbb{E}(\Omega_n) = u'(0)\phi(n) + v'(0) + O\left(\kappa_n^{-1}\right), \qquad \mathbb{V}(\Omega_n) = u''(0)\phi(n) + v''(0) + O\left(\kappa_n^{-1}\right). \qquad \square$$

(We will also use the letters u and v in a different context in the paper, but there is no chance of confusion.)

For fixed p and $n \to \infty$, the expected value of both, the ancestor tree, and the Steiner distance, are asymptotic to $\frac{p}{2} \log n$, the difference being in the smaller order terms. To apply the quasi-power theorem, an inductive process (w.r.t. p) is used. Part of the difficulty is that a certain trivariate generating function is only implicitly given, and sufficient information must be 'pumped out' of this implicit equation.

2. Size of the ancestor tree

For a given tree family, let $X_{n,p}$ denote the random variable that counts the size of the ancestor tree of p randomly chosen nodes in a tree of size n and let T_n be the number of trees of size n.

A simple family of increasing trees (which includes heap ordered trees) is defined by labelled rooted trees in which labels along any branch from the root go in increasing order; see [2]. For this type of problem, it is natural to consider exponential generating functions. In this case, by introducing the generating functions

$$T(z) = \sum_{n \geqslant 0} \frac{T_n}{n!} z^n \quad \text{and} \quad G(z,u,v) = \sum_{n \geqslant 0, p \geqslant 0, m \geqslant 0} \mathbb{P}\{X_{n,p} = m\} T_n \binom{n}{p} \frac{z^n}{n!} u^p v^m,$$

we get the equations

$$T'(z) = \varphi(T(z)) \quad \text{and} \quad \frac{\partial}{\partial z} G(z,u,v) = v(1+u)\varphi(G(z,u,v)) + (1-v)\varphi(T(z)), \tag{2.1}$$

with initial values $T(0) = 0$ and $G(0,u,v) = 0$. The first term in (2.1) takes care of the instance where the root is labelled and the second term accounts for a non-labelled root. Here the *degree-generating function* $\varphi(t) = \sum_{n\geqslant 0} \varphi_n t^n$ satisfies $\varphi_i \geqslant 0$ for $i \geqslant 1$ and $\varphi_0 > 0$. This function is responsible for the recursive generation of these trees. Here, however, we are only concerned with the case where each degree can occur with weight one, *i.e.*, with heap ordered trees. We plan to treat the general case in a future publication.

Thus we have $\varphi(t) = \frac{1}{1-t}$, and we obtain the differential equation $T'(z) = \frac{1}{1-T(z)}$, $T(0) = 0$, which gives the well-known formula

$$T(z) = 1 - \sqrt{1-2z}$$

for the exponential generating function $T(z)$. By extracting coefficients we obtain the number of heap ordered trees,

$$T_n = \prod_{k=1}^{n-1} (2k-1) = \frac{(n-1)!}{2^{n-1}} \binom{2n-2}{n-1}.$$

The differential equation of interest for $G(z,u,v)$ in the case of heap ordered trees is thus

$$\frac{\partial}{\partial z} G(z,u,v) = \frac{v(1+u)}{1-G(z,u,v)} + \frac{1-v}{\sqrt{1-2z}},$$
$$G(0,u,v) = 0, \qquad G(z,u,1) = 1 - \sqrt{1-2z(1+u)}.$$

It turns out that it is advantageous to make the substitution

$$H(z,u,v) = \frac{1-G(z,u,v)}{\sqrt{1-2z}}.$$

Then the differential equation becomes

$$H(z,u,v) - \frac{v(1+u)}{H(z,u,v)} - 1 + v = (1-2z)\frac{\partial}{\partial z} H(z,u,v), \qquad H(0,u,v) = 1.$$

Using separation of variables we get the implicit solution

$$\frac{1}{2} \log \frac{1}{1-2z} = \int_{x=1}^{H(z,u,v)} \frac{x dx}{x^2 - (1-v)x - v(1+u)},$$

and by integration we obtain

$$\begin{aligned} \log \frac{1}{1-2z} = & \log\left(1 - \frac{(H(z,u,v)-1)(H(z,u,v)+v)}{vu}\right) \\ & - \frac{1-v}{\sqrt{4vu+(1+v)^2}} \log\left(1 + \frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+2-(1-v)}\right) \\ & + \frac{1-v}{\sqrt{4vu+(1+v)^2}} \log\left(1 - \frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+(1-v)-2}\right). \end{aligned} \tag{2.2}$$

Now we replace $H(z,u,v)$ with $\frac{1-G(z,u,v)}{\sqrt{1-2z}}$ in (2.2) and differentiate with respect to v. In the resulting equation we let $v=1$ and solve for $\frac{\partial}{\partial v}G(z,u,v)|_{v=1}$. We obtain

$$\begin{aligned}\frac{\partial}{\partial v}G(z,u,v)\Big|_{v=1} =&\frac{1}{2}\sqrt{1-2z}-\frac{1}{2}\sqrt{1-2z(1+u)}\\ &-\frac{1}{4}\frac{u\big(\log(2+u-4z(1+u)+2\sqrt{(1-2z(1+u))(1-2z)(1+u)}\big)}{\sqrt{(1+u)(1-2z(1+u))}}\\ &-\frac{1}{4}\frac{2u\log(1+\sqrt{1+u})}{\sqrt{(1+u)(1-2z(1+u))}}.\end{aligned}\tag{2.3}$$

From (2.3) we can also find $\frac{\partial^2}{\partial z\partial v}G(z,u,v)|_{v=1}$, which will be used in Section 3 to compute the expectation for the Steiner distance; see (3.3). We differentiate equation (2.2) to get

$$\begin{aligned}&\frac{\partial^2}{\partial z\partial v}G(z,u,v)\Big|_{v=1}\\ &=-\frac{1}{2\sqrt{1-2z}}-\frac{1+u}{2\sqrt{1-2z(1+u)}}\\ &+\left(4(1+u)-\frac{2(1+u)(1-2z)(1+u)+2(1-2z(1+u))(1+u)}{\sqrt{(1-2z(1+u))(1-2z)(1+u)}}\right)\\ &\times\frac{u}{4(2+u-4z(1+u))\sqrt{(1-2z(1+u))(1+u)}+8(1-2z(1+u))(1+u)\sqrt{1-2z}}\\ &-\bigg(\log\left(2+u-4z(1+u)+2\sqrt{(1-2z(1+u))(1-2z)(1+u)}\right)\\ &-\log(2+u+2\sqrt{1+u})\bigg)\frac{u\sqrt{1+u}}{4(1-2z(1+u))^{3/2}}.\end{aligned}$$

Next we consider the (formal) expansions

$$G(z,u,v)=\sum_{p\geqslant 0}G_p(z,v)u^p\quad\text{resp.}\quad H(z,u,v)=\sum_{p\geqslant 0}H_p(z,v)u^p,$$

where our aim is to describe the limiting behaviour of $[z^n]G_p(z,v)$ uniformly in a neighbourhood of $v=1$ and then apply a central limit theorem (Hwang's quasi-power theorem) to find the Gaussian limiting distribution of $X_{n,p}$ for fixed $p\geqslant 1$.

Obviously we have

$$\begin{aligned}G_p(z,v)&=\sum_{n\geqslant 0,m\geqslant 0}\mathbb{P}\{X_{n,p}=m\}T_n\binom{n}{p}\frac{z^n}{n!}v^m,\\ H_p(z,v)&=-\frac{G_p(z,v)}{\sqrt{1-2z}},\quad\text{for }\ p\geqslant 1,\end{aligned}$$

and

$$H_0(z,v)=\frac{1-G_0(z,v)}{\sqrt{1-2z}}.$$

Since $\mathbb{P}\{X_{n,0}=m\}=\delta_{m,n}$, we immediately get $G_0(z,v)=T(z)=1-\sqrt{1-2z}$ and $H_0(z,v)=1$.

The required expansion for $p\geqslant 1$ is stated as the following lemma.

Lemma 2.1. *For $p \geqslant 1$, the coefficients $H_p(z,v)$ have, around their (only) dominant singularity $z = \frac{1}{2}$, the expansion*

$$H_p(z,v) = h_p(v) \frac{1}{(1-2z)^{\frac{p(v+1)}{2}}} + O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{(p-1)(v+1)}{2}}} \right),$$

uniformly for $|v-1| \leqslant \varepsilon$ and $\varepsilon > 0$. The coefficient-generating function $C(v,x) = \sum_{p \geqslant 1} h_p(v) x^p$ of the $h_p(v)$ is given implicitly by the equation

$$\frac{C(v,x)(1+v+C(v,x))}{vx} = -\left(\frac{1 + \frac{C(v,x)}{1+v}}{-\frac{1+v}{v} \frac{C(v,x)}{x}} \right)^{\frac{1-v}{1+v}}$$

and it holds for

$$h_p(1) = [x^p] C(1,x) = -\frac{2}{4^p p} \binom{2(p-1)}{p-1},$$

where $C(1,x) = -1 + \sqrt{1-x}$ and

$$C_v(1,x) = \frac{C(1,x)}{2} + \frac{x}{4} \frac{1}{1+C(1,x)} \log\left(\frac{1 + \frac{C(1,x)}{2}}{-2\frac{C(1,x)}{x}} \right).$$

Thus the expansion for the $G_p(z,v)$ for $p \geqslant 1$ is given by

$$G_p(z,v) = -h_p(v) \frac{1}{(1-2z)^{\frac{p(v+1)-1}{2}}} + O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{(p-1)(v+1)-1}{2}}} \right). \tag{2.4}$$

Proof. To obtain $H_1(z,v)$ and thus $G_1(z,v)$, we consider (2.2) and compare coefficients at u^0. We get

$$\begin{aligned}
&[u^0] \log\left(1 - \frac{(H(z,u,v)-1)(H(z,u,v)+v)}{vu} \right) \\
&\quad = [u^0] \log\left(1 - \frac{(H_1(z,v)u + O(u^2))(1+v+O(u))}{vu} \right) = \log\left(1 - \frac{1+v}{v} H_1(z,v) \right),
\end{aligned}$$

$$\begin{aligned}
&[u^0] \frac{1-v}{\sqrt{4vu+(1+v)^2}} \log\left(1 + \frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+2-(1-v)} \right) \\
&\quad = [u^0] \frac{1-v}{1+v} \frac{1}{\sqrt{1+\frac{4v}{(1+v)^2}u}} \log\left(1 + \frac{2(H_1(z,v)u + O(u^2))}{(1+v)\sqrt{1+\frac{4v}{(1+v)^2}u}+1+v} \right) \\
&\quad = [u^0] \frac{1-v}{1+v} (1+O(u)) \log(1+O(u)) = 0,
\end{aligned}$$

$$\begin{aligned}
&[u^0] \frac{1-v}{\sqrt{4vu+(1+v)^2}} \log\left(1 - \frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+(1-v)-2} \right) \\
&\quad = [u^0] \frac{1-v}{1+v} (1+O(u)) \log\left(1 - \frac{2(H_1(z,v)u + O(u^2))}{(1+v)\sqrt{1+\frac{4v}{(1+v)^2}u}-1-v} \right) \\
&\quad = [u^0] \frac{1-v}{1+v} (1+O(u)) \log\left(1 - \frac{1+v}{v} H_1(z,v) + O(u) \right) = \frac{1-v}{1+v} \log\left(1 - \frac{1+v}{v} H_1(z,v) \right),
\end{aligned}$$

and further

$$\log\left(\frac{1}{1-2z}\right) = \frac{2}{1+v}\log\left(1 - \frac{1+v}{v}H_1(z,v)\right),$$

which gives

$$H_1(z,v) = \frac{v}{1+v}\left(1 - \frac{1}{(1-2z)^{\frac{v+1}{2}}}\right) \quad \text{and} \quad G_1(z,v) = \frac{\sqrt{1-2z}\,v}{1+v}\left(\frac{1}{(1-2z)^{\frac{v+1}{2}}} - 1\right).$$

Therefore the asymptotic expansion given above holds for $p=1$ (although the bound for the remainder term is not tight here) with $h_1(v) = -\frac{v}{1+v}$ and thus the stated formula for $h_p(1)$ is also valid for $p=1$.

Now we assume that the lemma for $H_l(z,v)$ resp. $G_l(z,v)$ is true for all $1 \leqslant l \leqslant p$, and we will show that it then also holds for $p+1$. To prove the result for $H_{p+1}(z,v)$, we will consider the coefficients of u^p in the equation (2.2).

For the first term in (2.2), we use the expansion

$$\log\left(1 - \frac{(H(z,u,v)-1)(H(z,u,v)+v)}{vu}\right) = \log\left(1 - \frac{1+v}{v}H_1(z,v)\right) + \log\big(1 - \widetilde{H}(z,u,v)\big),$$

with

$$\begin{aligned}\widetilde{H}(z,u,v) &= \sum_{l\geqslant 1}\widetilde{H}_l(z,v)u^l \\ &= \frac{1}{1-\frac{1+v}{v}H_1(z,v)}\left(\frac{(H(z,u,v)-1)(H(z,u,v)+v)}{vu} - \frac{1+v}{v}H_1(z,v)\right).\end{aligned}$$

We then get

$$\begin{aligned}&[u^p]\log\left(1 - \frac{(H(z,u,v)-1)(H(z,u,v)+v)}{vu}\right) \\ &= -\sum_{j=1}^{p}\frac{1}{j}\sum_{p_1+\dots+p_j=pp_i\geqslant 1}\prod_{i=1}^{j}\widetilde{H}_{p_i}(z,v) \\ &= -\frac{\frac{1+v}{v}H_{p+1}(z,v)}{1-\frac{1+v}{v}H_1(z,v)} - \frac{\frac{1}{v}\sum_{k=1}^{p}H_k(z,v)H_{p+1-k}(z,v)}{1-\frac{1+v}{v}H_1(z,v)} - \sum_{j=2}^{p}\frac{1}{j}\sum_{\substack{p_1+\dots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\widetilde{H}_{p_i}(z,v),\end{aligned}$$

where

$$\widetilde{H}_l(z,v) = \frac{1}{1-\frac{1+v}{v}H_1(z,v)}\frac{1}{v}\left((1+v)H_{l+1}(z,v) + \sum_{k=1}^{l}H_k(z,v)H_{l+1-k}(z,v)\right).$$

Under the assumptions of the lemma we now obtain, for $1 \leqslant l \leqslant p-1$, around the dominant singularity $z = \frac{1}{2}$ in a neighbourhood of $v=1$, the uniform expansion

$$\begin{aligned}\widetilde{H}_l(z,v) &= (1-2z)^{\frac{v+1}{2}}\left(\frac{\frac{1+v}{v}h_{l+1}(v)}{(1-2z)^{\frac{(l+1)(v+1)}{2}}} + \frac{\frac{1}{v}\sum_{k=1}^{l}h_k(v)h_{l+1-k}(v)}{(1-2z)^{\frac{(l+1)(v+1)}{2}}} + O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{l(v+1)}{2}}}\right)\right) \\ &= \widetilde{h}_l(v)\frac{1}{(1-2z)^{\frac{l(v+1)}{2}}} + O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{(l-1)(v+1)}{2}}}\right),\end{aligned}$$

where

$$\widetilde{h}_l(v) = \frac{1}{v}\left((1+v)h_{l+1}(v) + \sum_{k=1}^{l} h_k(v)h_{l+1-k}(v)\right).$$

With the abbreviations

$$\widehat{H}(z,u,v) = \sum_{l\geqslant 1} \widehat{H}_l(z,v)u^l = \frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+2-(1-v)},$$

$$\widehat{a}_l(v) = [u^l]\frac{1}{\sqrt{1+\frac{4v}{(1+v)^2}u}}, \qquad \widehat{b}_l(v) = [u^l]\frac{2}{\sqrt{4vu+(1+v)^2}+2-(1-v)},$$

we get the expansion

$$\begin{aligned}
&[u^p]\frac{1-v}{\sqrt{4vu+(1+v)^2}}\log\left(1+\frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+2-(1-v)}\right)\\
&= \frac{1-v}{1+v}\sum_{k=1}^{p}\widehat{a}_{p-k}(v)\sum_{j=1}^{k}\frac{(-1)^{j+1}}{j}\sum_{\substack{k_1+\cdots+k_j=k\\ k_i\geqslant 1}}\prod_{i=1}^{j}\widehat{H}_{k_i}(z,v)\\
&= \frac{1-v}{1+v}\sum_{j=1}^{p}\frac{(-1)^{j+1}}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\widehat{H}_{p_i}(z,v)\\
&\quad + \frac{1-v}{1+v}\sum_{k=1}^{p-1}\widehat{a}_{p-k}(v)\sum_{j=1}^{k}\frac{(-1)^{j+1}}{j}\sum_{\substack{k_1+\cdots+k_j=k\\ k_i\geqslant 1}}\prod_{i=1}^{j}\widehat{H}_{k_i}(z,v),
\end{aligned}$$

for the coefficients of the second term in (2.2), where

$$\widehat{H}_l(z,v) = \sum_{k=1}^{l} H_k(z,v)\widehat{b}_{l-k}(v).$$

Under the assumptions of the lemma we obtain, for $1 \leqslant l \leqslant p$, the uniform expansion

$$\begin{aligned}
\widehat{H}_l(z,v) &= \sum_{k=1}^{l}\left(\frac{h_k(v)}{(1-2z)^{\frac{k(v+1)}{2}}} + O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{(k-1)(v+1)}{2}}}\right)\right)\widehat{b}_{l-k}(v)\\
&= \widehat{h}_l(v)\frac{1}{(1-2z)^{\frac{l(v+1)}{2}}} + O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{(l-1)(v+1)}{2}}}\right),
\end{aligned}$$

where

$$\widehat{h}_l(v) = \frac{1}{1+v}h_l(v).$$

Finally, for the third term in (2.2) we use the expansion

$$\begin{aligned}
&\frac{1-v}{\sqrt{4vu+(1+v)^2}}\log\left(1-\frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+(1-v)-2}\right)\\
&= \frac{1-v}{\sqrt{4vu+(1+v)^2}}\log\left(1-\frac{1+v}{v}H_1(z,v)\right) + \frac{1-v}{\sqrt{4vu+(1+v)^2}}\log\big(1-\overline{H}(z,u,v)\big),
\end{aligned}$$

with

$$\begin{aligned}\overline{H}(z,u,v) &= \sum_{l\geqslant 1} \overline{H}_l(z,v)u^l\\ &= \frac{1}{1-\frac{1+v}{v}H_1(z,v)}\left(\frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+(1-v)-2} - \frac{1+v}{v}H_1(z,v)\right).\end{aligned}$$

Further, we use the abbreviations

$$\overline{a}_l(v) = [u^l]\frac{1}{\sqrt{1+\frac{4v}{(1+v)^2}u}}, \qquad \overline{b}_l(v) = [u^l]\frac{2u}{\sqrt{4vu+(1+v)^2}+(1-v)-2}.$$

We get the expansion

$$\begin{aligned}&[u^p]\frac{1-v}{\sqrt{4vu+(1+v)^2}}\log\left(1-\frac{2(H(z,u,v)-1)}{\sqrt{4vu+(1+v)^2}+(1-v)-2}\right)\\ &=\frac{1-v}{1+v}\overline{a}_p(v)\log\left(1-\frac{1+v}{v}H_1(z,v)\right)\\ &\quad-\frac{1-v}{1+v}\sum_{k=1}^{p}\overline{a}_{p-k}(v)\sum_{j=1}^{k}\frac{(-1)^{j+1}}{j}\sum_{\substack{k_1+\cdots+k_j=k\\ k_i\geqslant 1}}\prod_{i=1}^{j}\overline{H}_{k_i}(z,v)\\ &=\frac{1-v}{1+v}\overline{a}_p(v)\log\left(1-\frac{1+v}{v}H_1(z,v)\right)-\frac{1-v}{1+v}\frac{\frac{1+v}{v}H_{p+1}(z,v)}{1-\frac{1+v}{v}H_1(z,v)}\\ &\quad-\frac{1-v}{1+v}\frac{\sum_{k=0}^{p-1}H_{k+1}(z,v)\overline{b}_{p-k}(v)}{1-\frac{1+v}{v}H_1(z,v)}-\frac{1-v}{1+v}\sum_{j=2}^{p}\frac{1}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\overline{H}_{p_i}(z,v)\\ &\quad-\frac{1-v}{1+v}\sum_{k=1}^{p-1}\overline{a}_{p-k}(v)\sum_{j=1}^{k}\frac{1}{j}\sum_{\substack{k_1+\cdots+k_j=k\\ k_i\geqslant 1}}\prod_{i=1}^{j}\overline{H}_{k_i}(z,v),\end{aligned}$$

where

$$\overline{H}_l(z,v) = \frac{1}{1-\frac{1+v}{v}H_1(z,v)}\sum_{k=0}^{l}H_{k+1}(z,v)\overline{b}_{l-k}(v).$$

Now, under the assumptions of the lemma, we obtain, for $1\leqslant l\leqslant p-1$, the uniform expansion

$$\begin{aligned}\overline{H}_l(z,v) &= (1-2z)^{\frac{v+1}{2}}\sum_{k=0}^{l}\left(\frac{h_{k+1}(v)}{(1-2z)^{\frac{(k+1)(v+1)}{2}}}+O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{k(v+1)}{2}}}\right)\right)\overline{b}_{l-k}(v)\\ &= \overline{h}_l(v)\frac{1}{(1-2z)^{\frac{l(v+1)}{2}}}+O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{(l-1)(v+1)}{2}}}\right),\end{aligned}$$

where

$$\overline{h}_l(v) = \frac{1+v}{v}h_{l+1}(v).$$

Comparing coefficients leads to the following equation for $H_{p+1}(z,v)$:

$$\begin{aligned}
&\frac{2}{v}\frac{1}{1-\frac{1+v}{v}H_1(z,v)}H_{p+1}(z,v)\\
&\quad=-\frac{\frac{1}{v}\sum_{k=1}^{p}H_k(z,v)H_{p+1-k}(z,v)}{1-\frac{1+v}{v}H_1(z,v)}-\sum_{j=2}^{p}\frac{1}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\widetilde{H}_{p_i}(z,v)\\
&\quad-\frac{1-v}{1+v}\sum_{j=1}^{p}\frac{(-1)^{j+1}}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\widehat{H}_{p_i}(z,v)\\
&\quad-\frac{1-v}{1+v}\sum_{k=1}^{p-1}\widehat{a}_{p-k}(v)\sum_{j=1}^{k}\frac{(-1)^{j+1}}{j}\sum_{\substack{k_1+\cdots+k_j=k\\ k_i\geqslant 1}}\prod_{i=1}^{j}\widehat{H}_{k_i}(z,v)\\
&\quad+\frac{1-v}{1+v}\overline{a}_p(v)\log\left(1-\frac{1+v}{v}H_1(z,v)\right)-\frac{1-v}{1+v}\frac{\sum_{k=0}^{p-1}H_{k+1}(z,v)\overline{b}_{p-k}(v)}{1-\frac{1+v}{v}H_1(z,v)}\\
&\quad-\frac{1-v}{1+v}\sum_{j=2}^{p}\frac{1}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\overline{H}_{p_i}(z,v)\\
&\quad-\frac{1-v}{1+v}\sum_{k=1}^{p-1}\overline{a}_{p-k}(v)\sum_{j=1}^{k}\frac{1}{j}\sum_{\substack{k_1+\cdots+k_j=k\\ k_i\geqslant 1}}\prod_{i=1}^{j}\overline{H}_{k_i}(z,v).
\end{aligned}$$

The asymptotic expansion

$$H_{p+1}(z,v)=h_{p+1}(v)\frac{1}{(1-2z)^{\frac{(p+1)(v+1)}{2}}}+O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{p(v+1)}{2}}}\right) \tag{2.5}$$

follows by inspection, where

$$\begin{aligned}
h_{p+1}(v)=\frac{v}{2}&\left[-\frac{1}{v}\sum_{k=1}^{p}h_k(v)h_{p+1-k}(v)-\sum_{j=2}^{p}\frac{1}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\widetilde{h}_{p_i}(v)\right.\\
&\left.-\frac{1-v}{1+v}\sum_{j=1}^{p}\frac{(-1)^{j+1}}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\widehat{h}_{p_i}(v)-\frac{1-v}{1+v}\sum_{j=2}^{p}\frac{1}{j}\sum_{\substack{p_1+\cdots+p_j=p\\ p_i\geqslant 1}}\prod_{i=1}^{j}\overline{h}_{p_i}(v)\right],
\end{aligned} \tag{2.6}$$

and this part of the lemma is proved. The expansion of $G_p(z,v)$ given in (2.4) follows immediately. It should be remarked that this detailed inductive description of $H_{p+1}(z,v)$ also proves that the assumptions necessary for the application of singularity analysis are satisfied. The logarithmic remainder term appears for $p=2$ owing to $\log\big(1-\frac{1+v}{v}H_1(z,v)\big)=-\frac{v+1}{2}\log(1-2z)$, and thus also for $p\geqslant 2$.

To get an equation for the coefficient generating function $C(v,x)=\sum_{p\geqslant 1}h_p(v)x^p$ one could of course use equation (2.6), but it follows much more easily direct from (2.2), when considering which terms give contributions to the main term of $H_p(z,v)$. Then we get

$$\log\left(1-\frac{\frac{C(v,x)}{x}(1+v+C(v,x))-(v+1)h_1(v)}{v}\right)$$
$$-\frac{1-v}{1+v}\log\left(1+\frac{C(v,x)}{1+v}\right)+\frac{1-v}{1+v}\log\left(1-\frac{1+v}{v}\left(\frac{C(v,x)}{x}-h_1(v)\right)\right)=0,$$

or

$$\frac{C(v,x)(1+v+C(v,x))}{vx}=-\left(\frac{1+\frac{C(v,x)}{1+v}}{-\frac{1+v}{v}\frac{C(v,x)}{x}}\right)^{\frac{1-v}{1+v}}. \tag{2.7}$$

We easily obtain from (2.7) the equation $\frac{C(1,x)(2+C(1,x))}{x}=-1$, which gives

$$C(1,x)=-1+\sqrt{1-x} \quad \text{and}$$
$$h_p(1)=[x^p]C(1,x)=-\frac{2}{4^p p}\binom{2(p-1)}{p-1}, \quad \text{for } p\geqslant 1. \tag{2.8}$$

This completes the proof of the lemma. □

Using singularity analysis, we immediately get from the above lemma the following expansion, which is uniform for $|v-1|\leqslant\varepsilon$ and $\varepsilon>0$,

$$\begin{aligned}\sum_{m\geqslant 0}\mathbb{P}\{X_{n,p}=m\}v^m &= \frac{n!}{\binom{n}{p}T_n}[z^n]G_p(z,v)\\ &= -\frac{p!h_p(v)2\sqrt{\pi}}{\Gamma\left(\frac{p(v+1)-1}{2}\right)}n^{\frac{p(v-1)}{2}}\left(1+O\left(\frac{1}{n^{1-\varepsilon}}\right)\right)\\ &= \exp\left[\frac{p(v-1)}{2}\log n+\log\left(\frac{-2\sqrt{\pi}p!h_p(v)}{\Gamma\left(\frac{p(v+1)-1}{2}\right)}\right)\right]\left(1+O\left(\frac{1}{n^{1-\varepsilon}}\right)\right),\end{aligned}$$

where we have used the asymptotic expansion for the number

$$T_n=\frac{n!2^{n-1}n^{-\frac{3}{2}}}{\sqrt{\pi}}\left(1+O\left(\frac{1}{n}\right)\right)$$

of heap ordered trees.

With the notations of the quasi-power theorem, we get

$$u(s)=\frac{p(e^s-1)}{2} \quad \text{and} \quad v(s)=\log\left(\frac{-2\sqrt{\pi}p!h_p(e^s)}{\Gamma\left(\frac{p(e^s+1)-1}{2}\right)}\right).$$

To apply the quasi-power theorem, we need $v(s)$ to be analytic around $s=0$, which is true, since $h_p(1)=-\frac{2}{4^p p}\binom{2(p-1)}{p-1}\neq 0$.

Further, we have

$$u'(s)=\frac{p}{2}e^s, \quad u''(s)=\frac{p}{2}e^s, \quad \text{thus} \quad u'(0)=\frac{p}{2}, \quad u''(0)=\frac{p}{2}.$$

Therefore we get the following theorem.

Theorem 2.2. *The distribution of the random variable $X_{n,p}$, which counts the size of the ancestor tree of p randomly chosen nodes in a random heap ordered tree of size n, is for $p \geqslant 1$ asymptotically Gaussian, where the convergence rate is of order $O(\frac{1}{\sqrt{\log n}})$, that is,*

$$\mathbb{P}\left\{ \frac{X_{n,p} - \frac{p}{2}\log n}{\sqrt{\frac{p}{2}\log n}} < x \right\} = \Phi(x) + O\left(\frac{1}{\sqrt{\log n}}\right),$$

and the expectation $E_{n,p} = \mathbb{E}(X_{n,p})$ and the variance $V_{n,p} = \mathbb{V}(X_{n,p})$ satisfy

$$E_{n,p} = \frac{p}{2}\log n + v'(0) + O\left(\frac{1}{n^{1-\varepsilon}}\right),$$
$$V_{n,p} = \frac{p}{2}\log n + v''(0) + O\left(\frac{1}{n^{1-\varepsilon}}\right). \qquad \square$$

Remark. By inspection we can get the following expansions:

$$[u^p]\frac{\partial}{\partial v}G(z,u,v)\Big|_{v=1} = \sum_{i=1}^{p}(-1)^{p+i}(p-1)^{\underline{i-1}}\,\frac{(2i-2)!}{(i-1)!\,4^i}\,\frac{1}{(1-2z)^{i-1/2}}\,\log\frac{1}{1-2z} + \sum_{i=0}^{p-1} b_i(p)\frac{1}{(1-2z)^{p-i-1/2}}.$$

The computation of the $b_i(p)$s is cumbersome as they become increasingly involved. However, we were able to obtain $b_1(p)$ and $b_2(p)$ explicitly:

$$b_1(p) = 2^{-2p-1}\binom{2p}{p}\left(H_{2p} - H_p\right),$$
$$b_2(p) = -H_{2p-1}\left(2^{2p-3} + \frac{1}{2}\binom{2p-2}{p} + \binom{2p-2}{p-1}\right) + \sum_{k=0}^{p}(p+1-k)\binom{2p-2}{k}H_{2p-1-k}.$$

The constant $v'(0)$ in the expectation can also be computed. We get

$$v'(s) = \frac{h_p'(e^s)e^s}{h_p(e^s)} - \frac{p}{2}e^s\Psi\left(\frac{p(e^s+1)-1}{2}\right), \quad \text{thus} \quad v'(0) = \frac{h_p'(1)}{h_p(1)} - \frac{p}{2}\Psi\left(\frac{2p-1}{2}\right).$$

Here $\Psi(x)$ denotes the digamma function $\Psi(x) = (\log\Gamma(x))'$. For properties of this function we refer the reader to [1]. There remains the calculation of $h_p'(1) = [x^p]C_v(1,x)$. We get the equation

$$C_v(1,x) = \frac{C(1,x)}{2} + \frac{x}{4(1+C(1,x))}\log\left(\frac{1+\frac{C(1,x)}{2}}{-2\frac{C(1,x)}{x}}\right)$$
$$= \frac{\sqrt{1-x}-1}{2} + \frac{x}{4\sqrt{1-x}}\log\left(\frac{1+\frac{\sqrt{1-x}-1}{2}}{-2\frac{\sqrt{1-x}-1}{x}}\right). \qquad (2.9)$$

To extract coefficients, we consider

$$[x^p]\log\left(\frac{1+\frac{\sqrt{1-x}-1}{2}}{-2\frac{\sqrt{1-x}-1}{x}}\right)=\frac{1}{p}[x^{p-1}]\left[\log\left(\frac{1+\frac{\sqrt{1-x}-1}{2}}{-2\frac{\sqrt{1-x}-1}{x}}\right)\right]'$$
$$=\frac{1}{p}[x^{p-1}]\left(-\frac{1}{x\sqrt{1-x}}+\frac{1}{x}\right)=-\frac{1}{4^pp}\binom{2p}{p},$$

and we find with Lemma 2.3 (below)

$$h_p'(1)=-\frac{1}{4^pp}\binom{2(p-1)}{p-1}-\frac{1}{4^p}\sum_{j=1}^{p-1}\frac{1}{j}\binom{2j}{j}\binom{2(p-1-j)}{p-1-j}$$
$$=-\frac{1}{4^pp}\binom{2(p-1)}{p-1}-\frac{2}{4^p}\binom{2(p-1)}{p-1}(H_{2p-2}-H_{p-1})$$
$$=-\frac{1}{4^p}\binom{2(p-1)}{p-1}\left(\frac{1}{p}+2(H_{2p-2}-H_{p-1})\right). \tag{2.10}$$

However, the way (2.9) is expressed is ungainly and the substitution $x=\frac{4t}{(1+t)^2}$ is useful for the following computations:

$$C_v(1,x)=\frac{2t}{1-t^2}\log\left(\frac{1}{1+t}\right)-\frac{t}{1+t},$$
$$C_{vv}(1,x)=-\frac{2t(t^2+1)}{(1-t)^3(1+t)}\log^2\left(\frac{1}{1+t}\right)+\frac{2t}{(1-t)^2}\log\left(\frac{1}{1+t}\right)+\frac{t}{1-t^2}.$$

Lemma 2.3.

(i) $\sum_{j\geqslant 1}\frac{1}{j}\binom{2j}{j}z^j=2\log\left(\frac{1-\sqrt{1-4z}}{2z}\right),$

(ii) $\sum_{j=1}^{p-1}\frac{1}{j}\binom{2j}{j}\binom{2(p-1-j)}{p-1-j}=\binom{2(p-1)}{p-1}(H_{2p-2}-H_{p-1}).$

Proof. **(i)** It is easier to prove the equivalent result

$$\sum_{j\geqslant 1}\binom{2j}{j}z^{j-1}=2\frac{d}{dz}\left[\log\left(\frac{1-\sqrt{1-4z}}{2z}\right)\right]$$
$$=\frac{4}{\sqrt{1-4z}(1-\sqrt{1-4z})}-\frac{2}{z}=\frac{1}{z}\left[\frac{1}{\sqrt{1-4z}}-1\right].$$

Now, it is well known that

$$\sum_{j\geqslant 0}\binom{2j}{j}z^j=\frac{1}{\sqrt{1-4z}},$$

and thus

$$\sum_{j\geqslant 1}\binom{2j}{j}z^{j-1}=\frac{1}{z}\left[\frac{1}{\sqrt{1-4z}}-1\right],$$

which proves the first part of the lemma.

(ii) We use the substitution

$$z = \frac{u}{(1+u)^2}, \quad dz = \frac{1-u}{(1+u)^3}du, \quad \sqrt{1-4z} = \frac{1-u}{1+u},$$

to simplify the given summation as follows:

$$\begin{aligned}\sum_{j=1}^{p-1} \frac{1}{j}\binom{2j}{j}\binom{2(p-1-j)}{p-1-j} &= [z^{p-1}]\frac{1}{\sqrt{1-4z}}\log\left(\frac{1-\sqrt{1-4z}}{2z}\right)^2 \\ &= \frac{1}{2\pi i}\oint \frac{(1+u)^{2p-2}}{u^p} 2\log(1+u)du \\ &= [u^{p-1}]2(1+u)^{2p-2}\log(1+u) \\ &= (-1)^p[u^{p-1}]2(1-u)^{2p-2}\log\left(\frac{1}{1-u}\right) \\ &= 2(-1)^p\binom{-p}{p-1}(H_{-p} - H_{-2p+1}) \\ &= 2\binom{2p-2}{p-1}(H_{2p-2} - H_{p-1}). \qquad \square\end{aligned}$$

We can determine the constant term $v'(0)$ in the asymptotic expansion of the expectation $E_{n,p}$ given above:

$$\begin{aligned}v'(0) &= \frac{1}{2} + p(H_{2p-2} - H_{p-1}) - \frac{p}{2}\Psi\left(\frac{2p-1}{2}\right) \\ &= \frac{1}{2} + p(H_{2p-2} - H_{p-1}) - \frac{p}{2}\left(2H_{2p-2} - H_{p-1} + \Psi\left(\frac{1}{2}\right)\right) = -\frac{p}{2}H_p + \frac{p}{2}\gamma + p\log 2.\end{aligned}$$

Next we compute $v''(0)$ in the variance. We obtain

$$\begin{aligned}v''(s) &= \frac{h_p''(e^s)e^{2s}}{h_p(e^s)} + \frac{h_p'(e^s)e^s}{h_p(e^s)} - \frac{(h_p'(e^s))^2e^{2s}}{h_p^2(e^s)} \\ &\quad - \frac{p}{2}e^s\Psi\left(\frac{p(e^s+1)-1}{2}\right) - \frac{p^2}{4}e^{2s}\Psi'\left(\frac{p(e^s+1)-1}{2}\right), \\ v''(0) &= \frac{h_p''(1)}{h_p(1)} + \frac{h_p'(1)}{h_p(1)} - \frac{(h_p'(1))^2}{h_p^2(1)} - \frac{p}{2}\Psi\left(\frac{2p-1}{2}\right) - \frac{p^2}{4}\Psi'\left(\frac{2p-1}{2}\right).\end{aligned}$$

Firstly, we are required to calculate $h_p''(1) = [x^p]C_{vv}(1,x)$, namely

$$[x^p]\left(-\frac{2t(t^2+1)}{(1-t)^3(1+t)}\log^2\left(\frac{1}{1+t}\right) + \frac{2t}{(1-t)^2}\log\left(\frac{1}{1+t}\right) + \frac{t}{1-t^2}\right). \qquad (2.11)$$

We confine ourselves to considering the first few terms only. From the series expansion of (2.11) we can produce the local expansion around the dominant singularity $x = 1$ and

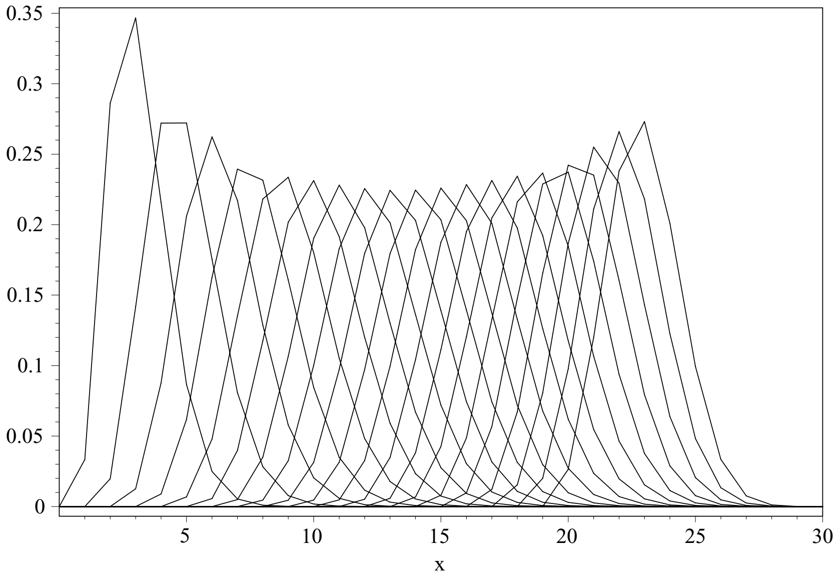

Figure 3. The probability distributions of the ancestor tree for $n = 30$, $p = 1, \ldots, 20$

use singularity analysis [5]:

$$\begin{aligned} h_p''(1) &= [x^p]\Big(-\frac{\log^2 2}{4}(1-x)^{-3/2} + \Big(\frac{1}{2} - \frac{\log 2}{4}\Big)(1-x)^{-1/2} + O(1)\Big) \\ &= -\frac{\log^2 2}{4}\binom{-3/2}{p} + \Big(\frac{1}{2} - \frac{\log 2}{4}\Big)\binom{-1/2}{p} + O(1). \end{aligned} \tag{2.12}$$

From this it follows that

$$v''(0) = -\frac{1}{2}p\log p + p\Big(\log 2 - \frac{5}{4}\Big) + \frac{1}{8}\log 2 + \frac{15}{16} - \frac{1}{4}\log^2 2 + O(p^{-1}). \tag{2.13}$$

3. The Steiner distance

An analogous approach works for the Steiner distance. Here $Y_{n,p}$ will denote the random variable that counts the Steiner distance of p randomly chosen nodes in a tree of size n of a given tree family.

For increasing trees we get, by introducing the generating function

$$F(z,u,v) = \sum_{n\geqslant 0, p\geqslant 0, m\geqslant 0} \mathbb{P}\{Y_{n,p} = m\} T_n \binom{n}{p}\frac{z^n}{n!}u^p v^m,$$

the equation

$$\frac{\partial}{\partial z}F(z,u,v) = \varphi'(T(z))F(z,u,v) + \frac{\partial}{\partial z}G(z,u,v) - v\varphi'(T(z))G(z,u,v) - (1-v)\varphi'(T(z))T(z), \tag{3.1}$$

with initial value $F(0,u,v)=0$. The generating functions $T(z)$ and $G(z,u,v)$ are as defined in Section 2. The first two terms in (3.1) arise when the root is labelled and the last two terms represent the corrections arising when the root is not labelled.

In this paper we only look at $\varphi(t)=\frac{1}{1-t}$, which is the special case of heap ordered trees. Therefore the Steiner distance requires the study of the differential equation

$$\frac{\partial}{\partial z}F(z,u,v)=\frac{\partial}{\partial z}G(z,u,v)+F(z,u,v)\frac{1}{1-2z}-G(z,u,v)\frac{v}{1-2z}-\frac{1-v}{1-2z}(1-\sqrt{1-2z}).$$

This is a first-order differential equation. We solve for $F(z,u,v)$ and get

$$F(z,u,v)=\frac{1}{\sqrt{1-2z}}\int\limits_0^z\sqrt{1-2t}\left[\frac{\partial}{\partial t}G(t,u,v)-G(t,u,v)\frac{v}{1-2t}-\frac{1-v}{1-\sqrt{1-2t}}\right]dt. \quad (3.2)$$

For the expectation we differentiate $F(z,u,v)$ with respect to v and let $v=1$, to obtain

$$\begin{aligned}\frac{\partial}{\partial v}F(z,u,v)\Big|_{v=1}=\frac{1}{\sqrt{1-2z}}\int\limits_0^z\sqrt{1-2t}\bigg[&\frac{\partial^2}{\partial v\partial t}G(t,u,v)\Big|_{v=1}-\frac{\partial}{\partial v}G(t,u,v)\Big|_{v=1}\frac{1}{1-2t}\\&-\frac{1-\sqrt{1-2t(1+u)}}{1-2t}+\frac{1}{1-\sqrt{1-2t}}\bigg]dt,\end{aligned} \quad (3.3)$$

since $G(z,u,v)|_{v=1}=1-\sqrt{1-2z(1+u)}$. This integration is cumbersome, so instead of performing it we find the coefficients u^p in (3.3) and then we consider the dominant term

$$\begin{aligned}[u^p]\frac{\partial}{\partial v}F(z,u,v)\Big|_{v=1}&=[u^p]\frac{1}{\sqrt{1-2z}}\int\limits_0^z\sqrt{1-2t}\bigg[\frac{\partial^2}{\partial v\partial t}G_p(t,v)\Big|_{v=1}\\&\qquad-\frac{\partial}{\partial v}G_p(t,v)\Big|_{v=1}\frac{1}{1-2t}-\frac{1-\sqrt{1-2t(1+u)}}{1-2t}+\frac{1}{1-\sqrt{1-2t}}\bigg]dt\\&=\left(\frac{ph_p(1)\log(1-2z)}{(1-2z)^{p-1/2}}-\frac{h_p'(1)}{(1-2z)^{p-1/2}}+\frac{h_p'(1)}{(1-2z)^{1/2}}\right)\\&\qquad+O\left(\frac{\log(1-2z)}{(1-2z)^{p-3/2}}\right)-[u^p]\frac{1}{\sqrt{1-2z}}\int\limits_0^z\frac{1-\sqrt{1-2t(1+u)}}{\sqrt{1-2t}}dt,\end{aligned} \quad (3.4)$$

where $h_p(1)$ and $h_p'(1)$ were computed in (2.8) and (2.10) respectively. It is not difficult to see that

$$[u^p]\frac{1}{\sqrt{1-2z}}\int\limits_0^z\frac{1-\sqrt{1-2t(1+u)}}{\sqrt{1-2t}}dt=O\left(\frac{1}{(1-2z)^{p-1/2}}\right),$$

therefore the main contribution comes from $ph_p(1)\log(1-2z)/(1-2z)^{p-1/2}$.

We find the expected value of the Steiner distance, $\mathbb{E}(Y_{n,p})$, by dividing (3.4) by our normalizing constant $\binom{n}{p}^{-1}\frac{n!}{1\cdot3\cdots(2n-3)}$ and then reading off the coefficient of z^n in the resulting equation. Firstly, looking at the dominant term in (3.4), we see that

$$\begin{aligned} E_{n,p} &= [z^n]\frac{\binom{n}{p}1\cdot 3\cdots(2n-3)}{n!}\,\frac{ph_p(1)\log(1-2z)}{(1-2t)^{p-1/2}} \\ &= [z^n]\frac{\binom{n}{p}1\cdot 3\cdots(2n-3)}{n!}\,\frac{-p\frac{2}{p4^p}\binom{2(p-1)}{p-1}\log(1-2z)}{(1-2z)^{p-1/2}} \sim \frac{p}{2}\log n, \end{aligned} \tag{3.5}$$

since we have

$$\begin{aligned} [z^n]\frac{1}{(1-2z)^{p-1/2}}\log(1-2z) &= -2^n[z^n]\frac{1}{(1-z)^{p-1/2}}\log\frac{1}{1-z} \\ &= -2^n\binom{n+p-3/2}{n}(H_{n+p-3/2}-H_{p-3/2}) \\ &\sim -2^n\frac{n^{p-3/2}}{\Gamma(p-\frac{1}{2})}\log n \quad (n\to\infty,\ p \text{ fixed}), \end{aligned}$$

as well as $\frac{n!}{1\cdot 3\cdots(2n-3)} \sim 2^{1-n}n^{3/2}\sqrt{\pi}$ and $\binom{n}{p} \sim \frac{n^p}{p!}$.

To obtain limiting theorems for the distribution of $Y_{n,p}$, we want to apply the quasi-power theorem again and will therefore require for $|v-1| \leqslant \varepsilon$ a uniform expansion of $F_p(z,v) = [u^p]F(z,u,v)$ around the dominant singularity $z = \frac{1}{2}$. From equation (3.2) we immediately obtain

$$F_p(z,v) = \frac{1}{\sqrt{1-2z}}\int_{t=0}^{z}\sqrt{1-2t}\left(\frac{\partial}{\partial t}G_p(t,v) - \frac{v}{1-2t}G_p(t,v)\right)dt. \tag{3.6}$$

We will now use the following more detailed expansion of $G_p(z,v)$, which follows from the proof of Lemma 2.1:

$$\begin{aligned} G_p(z,v) = &-h_p(v)\frac{1}{(1-2z)^{\frac{p(v+1)-1}{2}}} \\ &+ \sum_{\substack{1\leqslant k\leqslant p-1,\\ 0\leqslant j\leqslant p-k}} \alpha_{p,k,j}(v)\frac{\log^j(1-2z)}{(1-2z)^{\frac{k(v+1)-1}{2}}} + \alpha_{p,0,0}(v)\sqrt{1-2z}. \end{aligned}$$

This is also used to obtain the bound for the remainder term given below.

The integrand in (3.6) is then given by

$$\begin{aligned} &\sqrt{1-2t}\left(\frac{\partial}{\partial t}G_p(t,v) - \frac{v}{1-2t}G_p(t,v)\right) \\ &\quad = \sqrt{1-2t}\left(\frac{-h_p(v)(p(v+1)-1)}{(1-2t)^{\frac{p(v+1)+1}{2}}} + \frac{vh_p(v)}{(1-2t)^{\frac{p(v+1)+1}{2}}} + O\left(\frac{\log(1-2t)}{(1-2t)^{\frac{(p-1)(v+1)+1}{2}}}\right)\right) \\ &\quad = -\frac{h_p(v)(p-1)(v+1)}{(1-2t)^{\frac{p(v+1)}{2}}} + O\left(\frac{\log(1-2t)}{(1-2t)^{\frac{(p-1)(v+1)}{2}}}\right), \end{aligned}$$

and thus, for $p \geqslant 2$, we get the expansion

$$F_p(z,v) = -\frac{h_p(v)(p-1)(v+1)}{p(v+1)-2}\,\frac{1}{(1-2z)^{\frac{p(v+1)-1}{2}}} + O\left(\frac{\log(1-2z)}{(1-2z)^{\frac{(p-1)(v+1)-1}{2}}}\right).$$

Using singularity analysis to extract coefficients leads to

$$[z^n]F_p(z,v) = -\frac{h_p(v)(p-1)(v+1)}{p(v+1)-2}\,\frac{2^n n^{\frac{p(v+1)-1}{2}-1}}{\Gamma\left(\frac{p(v+1)-1}{2}\right)}\left(1+O\left(\frac{1}{n^{1-\varepsilon}}\right)\right),$$

and furthermore

$$\begin{aligned}\sum_{m\geqslant 0}\mathbb{P}(Y_{n,p}=m)v^m &= \frac{n!}{\binom{n}{p}T_n}[z^n]F_p(z,v)\\ &= -\frac{2\sqrt{\pi}p!(p-1)(v+1)h_p(v)}{\Gamma(\frac{p(v+1)-1}{2})(p(v+1)-2)}n^{\frac{p(v-1)}{2}}\left(1+O\left(\frac{1}{n^{1-\varepsilon}}\right)\right).\end{aligned}$$

With the notation used in the quasi-power theorem, we have

$$u(s) = \frac{p(e^s-1)}{2}, \qquad v(s) = \log\left(\frac{-2\sqrt{\pi}p!(p-1)(e^s-1)h_p(e^s)}{\Gamma\left(\frac{p(e^s+1)-1}{2}\right)(p(e^s+1)-2)}\right),$$

which gives

$$u'(0) = \frac{p}{2}, \qquad u''(0) = \frac{p}{2}.$$

For $p \geqslant 2$, $v(1) \neq 0$ since $h_p(1) < 0$ and thus the quasi-power theorem is applicable. On the other hand, for $p = 1$ we know *a priori*, from the combinatorial description, that $\mathbb{P}\{Y_{n,1} = 1\} = 1$ for $n \geqslant 1$.

For the constant $v'(0)$ in the expectation $E_{n,p} = \mathbb{E}(Y_{n,p})$ we compute

$$\begin{aligned}v'(s) &= \left[\log(e^s+1)+\log(h_p(e^s))-\log(p(e^s+1)-2)-\log\Gamma\left(\frac{p(e^s+1)-1}{2}\right)\right]'\\ &= \frac{e^s}{e^s+1}+\frac{h_p'(e^s)e^s}{h_p(e^s)}-\frac{pe^s}{p(e^s+1)-2}-\frac{pe^s}{2}\Psi\left(\frac{p(e^s+1)-1}{2}\right),\end{aligned}$$

and further

$$v'(0) = \frac{h_p'(1)}{h_p(1)}-\frac{p}{2}\Psi\left(\frac{2p-1}{2}\right)-\frac{1}{2(p-1)} = -\frac{p}{2}H_p+\frac{p}{2}\gamma+p\log 2-\frac{1}{2(p-1)}.$$

We note that this gives us the expected value with higher accuracy than (3.5) and it leads to the following theorem.

Theorem 3.1. *The distribution of the random variable $Y_{n,p}$, which counts the Steiner distance of p randomly chosen nodes in a random heap ordered tree of size n, is for $p \geqslant 2$ asymptotically Gaussian, where the convergence rate is of order $O\left(\frac{1}{\sqrt{\log n}}\right)$, that is,*

$$\mathbb{P}\left\{\frac{Y_{n,p}-\frac{p}{2}\log n}{\sqrt{\frac{p}{2}\log n}} < x\right\} = \Phi(x)+O\left(\frac{1}{\sqrt{\log n}}\right),$$

and the expectation $E_{n,p} = \mathbb{E}(Y_{n,p})$ and variance $V_{n,p} = \mathbb{V}(X_{n,p})$ satisfy

$$\begin{aligned}E_{n,p} &= \frac{p}{2}\log n-\frac{p}{2}H_p+\frac{p}{2}\gamma+p\log 2-\frac{1}{2(p-1)}+O\left(\frac{1}{n^{1-\varepsilon}}\right),\\ V_{n,p} &= \frac{p}{2}\log n+v''(0)+O\left(\frac{1}{n^{1-\varepsilon}}\right).\end{aligned}$$

□

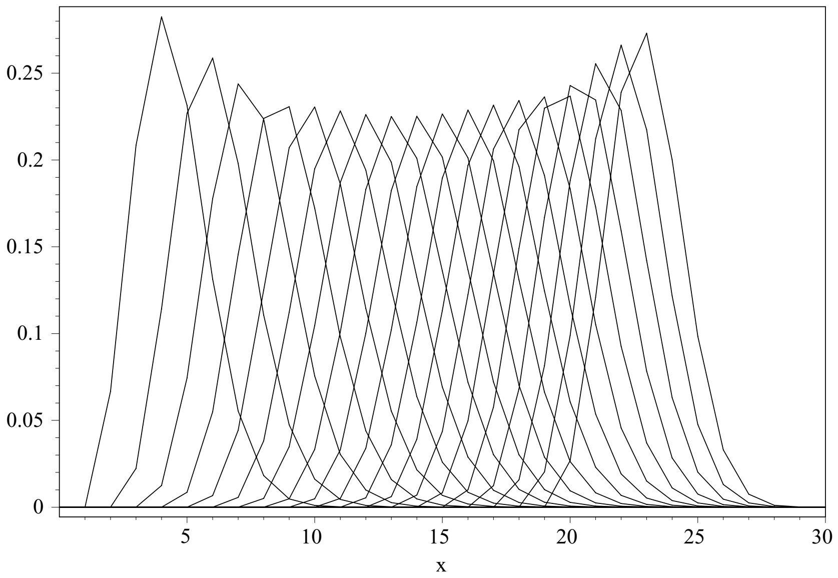

Figure 4. The probability distributions of the Steiner distance for $n = 30$, $p = 2, \ldots, 20$

For the proof, it remains to discuss the variance. Since we have obtained the variance of the size of the ancestor tree in (2.13), we can easily get the variance of the Steiner distance. It follows that

$$\begin{aligned} v''(s) = &\frac{e^s}{e^s+1} - \frac{e^{2s}}{e^s+1} + \frac{h_p''(e^s)e^s}{h_p(e^s)} + \frac{h_p'(e^s)e^s}{h_p(e^s)} - \frac{(h_p'(e^s))^2e^{2s}}{h_p^2(e^s)} \\ &- \frac{pe^s}{p(e^s+1)-2} + \frac{p^2e^{2s}}{(p(e^s+1)-2)^2} \\ &- \frac{pe^s}{2}\Psi\left(\frac{p(e^s+1)-1}{2}\right) - \frac{p^2e^{2s}}{4}\Psi'\left(\frac{p(e^s+1)-1}{2}\right), \end{aligned}$$

where $h_p''(1)$ is given by (2.11), and furthermore

$$\begin{aligned} v''(0) &= \frac{3}{4} + \frac{h_p''(1)}{h_p(1)} + \frac{h_p'(1)}{h_p(1)} - \frac{(h_p'(1))^2}{h_p^2(1)} - \frac{p}{2(p-1)} + \frac{p^2}{4(p-1)^2} \\ &\quad - \frac{p}{2}\Psi\left(\frac{2p-1}{2}\right) - \frac{p^2}{4}\Psi'\left(\frac{2p-1}{2}\right) \\ &= -\frac{p}{2}\log p + p\left(\log 2 - \frac{5}{4}\right) + \frac{1}{8}\log 2 + \frac{23}{16} - \frac{1}{4}\log^2 2 + O(p^{-1}). \end{aligned}$$

References

[1] Abramowitz, M. and Stegun, I. A. (1972) *Handbook of Mathematical Functions*, Dover.

[2] Bergeron, F., Flajolet, P. and Salvy, B. (1992) *Varieties of Increasing Trees*, Vol. 581 of *Lecture Notes in Computer Science*, Springer, pp. 24–48.

[3] Dankelmann, P., Oellermann, O. R. and Swart, H. C. (1996) The average Steiner distance of a graph. *J. Graph Theory* **22** 15–22.

[4] Entringer, R., Meir, A., Moon, J. and Szekely, L. (1994) The Wiener index of trees from certain families. *Australasian J. Combin.* **10** 211–224.

[5] Flajolet, P. and Odlyzko, A. (1990) Singularity analysis of generating functions. *SIAM J. Discrete Math.* **3** 216–240.

[6] Hwang, H.-K. (1998) On convergence rates in the central limit theorems for combinatorial structures. *Europ. J. Combin.* **19** 329–343.

[7] Mahmoud, H. and Neininger, R. (2003) Distribution of distances in random binary search trees. *Ann. Appl. Probab.* **13** 253–276.

[8] Neininger, R. (2002) Wiener index of random trees. *Combin. Probab. Comput.* **11** 587–597.

[9] Panholzer, A. (2002) The distribution of the size of the ancestor tree and of the induced spanning subtree for random trees. To appear in *Random Struct. Alg.*

[10] Panholzer, A. and Prodinger, H. (2003) Spanning tree size in random binary search trees. To appear in *Ann. Appl. Probab.*

[11] Prodinger, H. (1996) Depth and path length of heap ordered trees. *Internat. J. Foundations Comput. Sci.* **7** 293–299.

[12] Prodinger, H. (1996) The level of nodes in heap ordered trees. Available from:
`http://www.wits.ac.za/helmut/abstract/abs_126.htm`.

[13] Wiener, H. (1947) Structural determination of paraffin boiling points. *J. Amer. Chem. Soc.* **69** 17–20.