MATHEMATICAL METHODS IN THE APPLIED SCIENCES
Math. Meth. Appl. Sci. 2009; **32**:1207–1245
Published online 7 November 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/mma.1085
MOS subject classification: 68 W 40; 05 C 05

On the shape of the fringe of various types of random trees

Michael Drmota[1], Bernhard Gittenberger[1], Alois Panholzer[1], Helmut Prodinger[2] and Mark Daniel Ward[3,*,†]

[1] *Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien,*
Wiedner Hauptstrasse. 8-10/104, A-1040 Wien, Austria
[2] *Department of Mathematics, Stellenbosch University, 7602 Stellenbosch, South Africa*
[3] *Department of Statistics, Purdue University, 150 North University Street, West Lafayette, IN 47907-2067, U.S.A.*

Communicated by W. Sprößig

SUMMARY

We analyze a fringe tree parameter w in a variety of settings, utilizing a variety of methods from the analysis of algorithms and data structures.

Given a tree t and one of its leaves a, the $\mathsf{w}(t,a)$ parameter denotes the number of internal nodes in the subtree rooted at a's father. The closely related $\overline{\mathsf{w}}(t,a)$ parameter denotes the number of leaves, excluding a, in the subtree rooted at a's father. We define the cumulative w parameter as $\mathsf{W}(t)=\sum_a \mathsf{w}(t,a)$, i.e. as the sum of $\mathsf{w}(t,a)$ over all leaves a of t. The w parameter not only plays an important rôle in the analysis of the Lempel–Ziv '77 data compression algorithm, but it is captivating from a combinatorial viewpoint too.

In this report, we determine the asymptotic behavior of the w and W parameters on a variety of types of trees. In particular, we analyze simply generated trees, recursive trees, binary search trees, digital search trees, tries and Patricia tries.

The final section of this report briefly summarizes and improves the previously known results about the $\overline{\mathsf{w}}$ parameter's behavior on tries and suffix trees, originally published in one author's thesis (see Analysis of the multiplicity matching parameter in suffix trees. *Ph.D. Thesis*, Purdue University, West Lafayette, IN, U.S.A., May 2005; *Discrete Math. Theoret. Comput. Sci.* 2005; **AD**:307–322; *IEEE Trans. Inform. Theory* 2007; **53**:1799–1813).

This survey of new results about the w parameter is very instructive since a variety of different combinatorial methods are used in tandem to carry out the analysis. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: analysis of algorithms; binary search trees; digital search trees; Patricia trees; recursive trees; simply generated trees; suffix trees; tries; pattern matching

[*] Correspondence to: Mark Daniel Ward, Department of Statistics, Purdue University, 150 North University Street, West Lafayette, IN 47907-2067, U.S.A.

[†] E-mail: mdw@purdue.edu

Contract/grant sponsor: FWF; contract/grant number: NFN S9600
Contract/grant sponsor: NRF; contract/grant number: 2053748
Contract/grant sponsor: NSF; contract/grant number: 0603821

Received 27 August 2007

1. INTRODUCTION

1.1. General remarks

Rooted planar trees are important combinatorial objects. On the one hand, they have a simple (global) structure and admit an easy recursive description, but nevertheless they usually have a nontrivial and combinatorially interesting local structure. Their simplicity makes the study of tree characteristics amenable to a wealth of methods from analytic combinatorics and probability theory. Trees play an important rôle in biology and computer science applications. For instance, trees are used in modelling the spread of epidemics and in the investigation of relationships of biological species. In computer science, they serve as data structures and are used in the analysis of the worst- and/or average-case behavior of algorithms.

The original idea of this paper was to present a survey of the asymptotic behavior of tree parameters for various tree classes. We are dealing with random trees, assuming that trees are selected uniformly from the set of all trees with n vertices. Our interest lies in the asymptotic behavior of tree parameters as n tends to infinity.

Owing to the rising importance of computer science during the last century, trees have been studied intensively for several decades. So there is a vast literature on tree parameters, which makes a comprehensive study impossible. Thus, we decided to focus on one fringe tree parameter as demonstration object. The parameter we chose, called the w parameter and described below, originates in the study of data compression algorithms. One advantage of this parameter is that it is easy to describe, yet sufficiently rich in complexity. A comprehensive study of its behavior in several tree classes requires the utilization of a variety of different methods, including: the symbolic method to set up generating functions, singularity analysis, asymptotic solutions of functional equations, the Mellin transform, probability theoretic arguments ranging from elementary to Poissonization.

1.2. The w parameter and its variants

For each leaf a in a tree t, the $\mathsf{w}(t,a)$ parameter denotes the number of internal nodes in the subtree rooted at a's father. Similarly, the $\overline{\mathsf{w}}$ parameter denotes the number of leaves, excluding a, in such a subtree. In a binary tree, for instance, where every branching node has exactly two children, we have obviously $\mathsf{w}(t,a)=\overline{\mathsf{w}}(t,a)$. We also study the cumulative $\mathsf{W}(t)$ parameter of a tree t, defined as $\mathsf{W}(t):=\sum_a \mathsf{w}(t,a)$, i.e. the sum of $\mathsf{w}(t,a)$ over all leaves a in the tree t.

The w and W parameters are well-defined on all rooted trees, i.e. on trees that have a specified root. We do not need a tree to be planar (i.e. embedded in the plane) in order to identify the $\mathsf{w}(t,a)$ parameter associated with leaf a.

The w parameter was originally applied to uncompressed suffix trees in a study about the Lempel–Ziv ’77 data compression algorithm. In particular, [1–3] discusses the behavior of the w parameter in both suffix trees and also in tries built over independent strings. On the other hand, the combinatorial aspects of the w parameter are very appealing in their own right.

The w parameter contains local information about the fringe of a tree. The w parameter does not increase as the total number of nodes in the tree increases. Instead, we expect the w parameter to exhibit a discrete limit law.

1.3. Example

We give an early example of a tree, depicted in Figure 1, to illustrate and clarify the definition of the $\mathsf{w}(t,a)$ and $\mathsf{W}(t)$ parameters. The tree is a planar tree; in other words, it has a natural embedding

DOI: 10.1002/mma

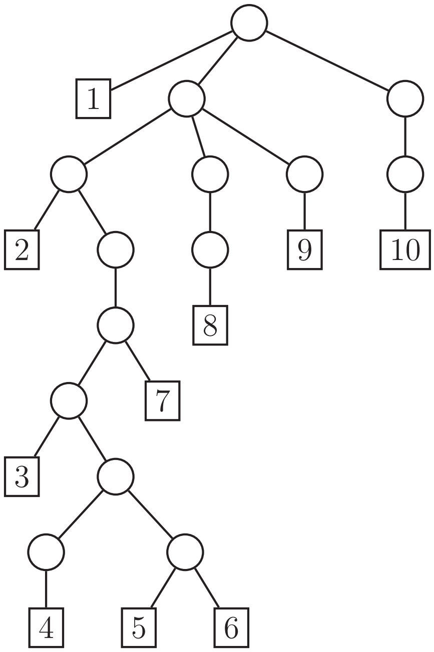


Figure 1. Example of a planar tree, with nodes numbered left to right.

in the plane, so its leaves can be counted from left to right. For this reason, we numbered the 14 leaf nodes from 1 to 14, reading left to right around the fringe of the tree. In this example tree t, the subtree rooted at node 1's father is the entire tree, which contains 14 nodes (in fact, node 1's father is the root), so $\mathsf{w}(t,1)=14$. Also, $\mathsf{w}(t,2)=7$, since the subtree rooted at node 2's father contains 7 nodes. Similarly, we compute

a	1	2	3	4	5	6	7	8	9	10
$\mathsf{w}(t,a)$	14	7	4	1	1	1	5	1	1	1

1.4. Plan of the paper

In Section 2, we give a brief description of each class of trees for which we study the w or W parameter. Section 3 presents the basic notations that will be used throughout the paper. Sections 4–9 contain the theoretical study of the w, $\overline{\mathsf{w}}$, and W parameters for the various tree classes. Finally, in Section 10 we present the application of the $\overline{\mathsf{w}}$ parameter in computer science, namely suffix trees and their relation to the Lempel–Ziv data compression algorithm.

2. THE TREE CLASSES

We will study the w (and W) parameters for the following tree classes.

DOI: 10.1002/mma

2.1. *Simply generated trees*

In the 19th century, Galton–Watson branching processes were introduced in order to study the evolution of family names. Such a process starts with a single particle which produces a random number of children (also particles), according to some *a priori* probability distribution D. Each child behaves in exactly the same way as the starting particle, and all children are independent. So if the starting particle has i children, then it gives rise to i independent Galton–Watson processes. Conditioning the family tree obtained by such a process on the total progeny to be n yields a simply generated tree of size n. Note that a simply generated tree can always be interpreted as a planted plane tree, that is, it is embedded into the plane and the root is planted. In particular, the successors (=children) of each node have a natural left to right order. Of course, this induces a left to right order for the leaves, too.

Combinatorially, these trees can be most easily described by their generating function. Let $\varphi(x)=\varphi_0+\varphi_1 x+\varphi_2 x^2+\cdots$ be the generating function describing the various ways that a node can have children. In the branching process setting $\varphi_j=\mathbb{P}\{D=j\}$ denotes the probability that a node has j children so that $\varphi(1)=1$. However, it is also possible to interpret φ_j as a nonnegative weight and not to assume that $\varphi_0+\varphi_1+\cdots$ sums up to 1. In all cases we introduce for each tree t the weight $\omega(t)=\prod_{j\geqslant 0}\varphi_j^{D_j(t)}$ where $D_j(t)$ denotes the number of nodes of t with outdegree (=number of children) j. Then the weighted number T_n of trees of size n is $T_n=\sum_{|t|=n}\omega(t)$ and the generating function of these weights is $T(z)=\sum_{n\geqslant 1}T_n z^n$. By definition of $\omega(t)$ and the recursive structure of trees it is clear that $T(z)$ satisfies the functional equation

$$T(z)=z\varphi(T(z))$$

(see, for instance, Meir and Moon [4]). The probability distribution on the trees of size n induced by the weights $\omega(t)$ is exactly that given by the conditioned Galton–Watson process if the offspring distribution D is given by $\mathbb{P}\{D=j\}=\varphi_j z_0^j/\varphi(z_0)$, where $z_0>0$ is arbitrary so that $\varphi(z_0)<\infty$.

Important special cases are binary trees ($\varphi(x)=(1+x)^2=1+2x+x^2$), planted plane trees ($\varphi(x)=1/(1-x)=1+x+x^2+\cdots$) and Cayley (i.e. labelled rooted) trees ($\varphi(x)=\mathrm{e}^x$).

2.2. *Recursive trees*

Recursive trees appear in various contexts. They are used to model the spread of epidemics (see [5]) or to investigate and construct family trees of preserved copies of ancient manuscripts (see [6]). Other applications are the study of the schemes of chain letters or pyramid games (see [7]).

They can be described as follows: start with a node carrying the label 1 as the root. Then attach a node with label 2. After having attached the nodes with labels $1,2,\ldots,k$, attach the node with label $k+1$ to one of the existing nodes, with each position being equally likely. This construction generates a nonplane labelled tree, with the nodes on each path starting at the root getting labelled monotonically. Moreover, each of the $(n-1)!$ possible trees of size n has the same probability.

2.3. *Binary search trees*

The origin of binary search trees dates to a fundamental problem in computer science, the dictionary problem. In this problem a set of records is given where each can be addressed by a key. The

DOI: 10.1002/mma

binary search tree is a data structure used for storing the records. Basic operations include *insert* and *search*.

Binary search trees are plane binary trees generated by a random permutation π of $\{1,2,\dots,n\}$. The elements of $\{1,2,\dots,n\}$ serve as keys. The data are stored in the internal nodes of the tree. Starting with a node labelled by $\pi(1)$, one first compares $\pi(1)$ with $\pi(2)$. If $\pi(2)<\pi(1)$, then $\pi(2)$ becomes root of the left subtree; otherwise, $\pi(2)$ becomes root of the right subtree. When having constructed a tree with nodes $\pi(1),\dots,\pi(k)$, the next node $\pi(k+1)$ is inserted by comparison with the existing nodes in the following way: start with the root as current node. If $\pi(k+1)$ is less than the current node, then descend into the left subtree, otherwise into the right subtree. Now continue with the root of the chosen subtree as current according to same rule. Finally, attach $n+1$ external nodes (= leaves) at the possible places and denote these leaves by $\{0,1,\dots,n\}$ from left to right.

2.4. *Digital search trees*

Digital search trees are intended for the same kinds of problems as binary search trees. The advantage of digital search trees is a much better worst-case performance. Furthermore it can be shown that the average-case performance is asymptotically optimal (see [8, 9]).

Digital search trees are constructed from a set of binary sequences, which serve as keys for data stored in the internal nodes of the tree. Consider a set of records, numbered from 1 to n, and generate a binary sequence for each item, e.g. by tossing a coin. This is a symmetric model, where zeros and ones are equally likely at each place in each sequence. We construct a binary tree from such a sequence as follows. Item 1 is the root of the tree. After having inserted the first k items we insert item $k+1$ as follows: choose the root as current node and look at the binary key of the current item. If the first digit is 1, descend into the right subtree, otherwise into the left one. If the root of the subtree is occupied, continue by looking at the next digit of the key.

2.5. *Tries and Patricia tries*

The idea is similar to that of digital search trees except that the records are stored in the leaves rather than in the internal nodes. Again a '1' indicates a descent into the right subtree, and '0' indicates a descent into the left subtree. Insertion causes some rearrangement of the tree, since a leaf becomes an internal node. In contrast to binary search trees and digital search trees, the shape of the trie is independent of the actual order of insertion. The position of each item is determined by the shortest unique prefix of its key.

An alternative description is as follows: Given a set $\mathscr{X}$ of strings, we partition $\mathscr{X}$ into two parts, $\mathscr{X}_L$ and $\mathscr{X}_R$, such that $X\in\mathscr{X}_L$ (respectively, $X\in\mathscr{X}_R$) if the first symbol of X is 0 (respectively, 1). The rest of the trie is defined recursively in the same manner, except that the splitting at the kth level depends on the kth symbol of each string. The first time that a branch contains exactly one string, a leaf is placed in the trie at that location (denoting the placement of the string into the trie), and no further branching takes place from such a portion of the trie.

Patricia tries are a slight modification of tries. Consider the case when several keys share the same prefix, but all other keys differ from this prefix already in their first position. Then the edges corresponding to this prefix may be contracted to one single edge. This method of construction leads to a more efficient structure.

DOI: 10.1002/mma

2.6. *Suffix trees*

Suffix trees are fundamental in data compression and pattern matching algorithms. Consider a string $X = X_1 X_2 X_3 \ldots$. Define $X^{(j)} = X_j X_{j+1} X_{j+2} \ldots$, i.e. $X^{(j)}$ denotes the jth suffix of X. Then a suffix tree is exactly a trie built from the strings $\mathscr{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$. Thus, a suffix tree is simply a trie built from the first n prefixes of a common string.

3. NOTATIONS

We describe some of the notation utilized throughout this report. We utilize a quantity from partition enumeration:

$$Q_n := (1-1/2)(1-1/4)\cdots(1-1/2^n) \tag{1}$$

Note that $Q_\infty := \lim_{n\to\infty} Q_n = \prod_{i=1}^{\infty}(1-2^{-i}) \approx 0.2887880951$. We will utilize the following related functions as well:

$$Q(x) = \prod_{n\geqslant 1}\left(1-\frac{x}{2^n}\right) \quad \text{and} \quad Q_x = \frac{Q(1)}{Q(2^{-x})}$$

Note that for $x = n \in \mathbb{N}$ the notions Q_x and (1) coincide.

Since we frequently utilize analytic combinatorics, we will often present sequences of numbers and probabilities using generating functions. The (weighted) number of trees will be denoted by T_n; the corresponding generating function is

$$T(z) = \sum_{n\geqslant 0} T_n z^n$$

In Sections 4–9, the $\mathsf{w}(t,a)$ (respectively, $\overline{\mathsf{w}}(t,a)$) parameter will be evaluated for the leaf with number j, with the leaves numbered from left to right, in a *random* tree t of size n. (In other words, we will focus attention on 'a' denoting the jth leaf of the tree, when the leaves are numbered left-to-right in the tree's planar embedding.) Since the tree t is chosen at random from the class of all trees with size n, then we use $X_{n,j}$ (resp. $\overline{X}_{n,j}$) to denote the value of w (resp. $\overline{\mathsf{w}}$) on the jth leaf. If there is no jth leaf, then, of course, we set $X_{n,j} = \overline{X}_{n,j} = 0$.

In Section 10, we always consider the $\overline{\mathsf{w}}$ parameter on the last insertion in a trie or suffix tree with $n+1$ insertions, i.e. external nodes. So we simply refer to $\overline{\mathsf{w}}_n$ in Section 10.

The corresponding generating functions are

$$F(z,u,v) = \sum_t \omega(t) z^{|t|} \sum_a u^{\text{number of leaf } a} v^{\mathsf{w}(t,a)}$$

and

$$\overline{F}(z,u,v) = \sum_t \omega(t) z^{|t|} \sum_a u^{\text{number of leaf } a} v^{\overline{\mathsf{w}}(t,a)}$$

where the sums are over all trees t and over all leaves a. Using the random variable $X_{n,j}$ we can rewrite this generating function in probabilistic terms:

$$F(z,u,v) = \sum_{n,j} T_n \mathbb{E}[v^{X_{n,j}}] z^n u^j$$

DOI: 10.1002/mma

Sometimes it is more convenient to work with the generating function

$$P(z,u,v)=\sum_{n\geqslant 0}\sum_{j=0}^{n}\sum_{m\geqslant 0}\mathbb{P}\{X_{n,j}=m\}z^n u^j v^m$$

$$=\sum_{n\geqslant 0}\sum_{j=0}^{n}\mathbb{E}[v^{X_{n,j}}]z^n u^j$$

which differs from the previous one by the normalization w.r.t. the powers of z. In some instances it is easier to dispense with the tree of size 0, so we will then use

$$P^*(z,u,v)=P(z,u,v)-1$$

instead of $P(z,u,v)$.

The random variable giving the W parameter of a random tree will be denoted by X_n. Since this is a global tree parameter, we can define the numbers $y_{n\ell k}$ denoting the (weighted) number of trees t of size n with $\mathsf{W}(t)=k$ and ℓ internal nodes. Then we can write the corresponding generating function as

$$G(z,u,v)=\sum_{n\geqslant 0}\sum_{\ell\geqslant 0}\sum_{k\geqslant 0}y_{n\ell k}z^n u^\ell v^k=\sum_{t}\omega(t)z^{|t|}u^{\#\text{internal nodes in } t}v^{\mathsf{W}(t)}$$

Using the random variable X_n we can also write in probabilistic terms:

$$G(z,1,v)=\sum_{n\geqslant 0}T_n\mathbb{E}[v^{X_n}]z^n$$

For digital search trees and tries we will work with the exponential generating function of $\mathbb{E}[X_n]$, namely

$$L(z)=\sum_{n\geqslant 0}\mathbb{E}[X_n]\frac{z^n}{n!} \tag{2}$$

We also use the notation for harmonic numbers:

$$H_n=\sum_{k=1}^{n}\frac{1}{k} \quad \text{and} \quad H_n^{(2)}=\sum_{k=1}^{n}\frac{1}{k^2}$$

The jth falling power of Y is defined as

$$Y^{\underline{j}}:=Y(Y-1)(Y-2)\cdots(Y-j+1)$$

Analogously, the jth factorial moment of a random variable Y is defined as $\mathbb{E}[Y^{\underline{j}}]$.

We use Iverson notation (see [10]), namely $[\![A]\!]=1$ if statement A is true, and $[\![A]\!]=0$ otherwise.

4. SIMPLY GENERATED TREES

4.1. The number of trees

We recall that a simply generated family of trees is defined with help of a generating function $\varphi(x)=\varphi_0+\varphi_1 x+\varphi_2 x^2+\cdots$ describing the various ways that a node can have children. Each

DOI: 10.1002/mma

rooted tree t has then weight $\omega(t)=\prod_{j\geqslant 0}\varphi_j^{D_j(t)}$ where $D_j(t)$ denotes the number of nodes of t with outdegree (=number of children) j, and $T_n=\sum_{|t|=n}\omega(t)$ is the weighted number of trees of size n. The generating function $T(z)=\sum_{n\geqslant 1}T_nz^n$ of these weighted numbers satisfies the functional equation

$$T(z)=z\varphi(T(z)) \tag{3}$$

(see [4]).

We assume that $\varphi(x)$ in (3) has radius of convergence $R>0$ and that there exists $0<\tau<R$ with $\tau\varphi'(\tau)=\varphi(\tau)$. This is satisfied if the offspring distribution D of the underlying branching process has a finite exponential moment $\mathbb{E}[e^{\alpha D}]$ for some $\alpha>0$. Then it is known that $\rho=1/\varphi'(\tau)$ is the radius of convergence of $T(z)$ and that we have a local (singular) expansion of the form

$$T(z)=\tau-\sqrt{\frac{2\varphi(\tau)}{\varphi''(\tau)}}\sqrt{1-z\varphi'(\tau)}+O(|1-z\varphi'(\tau)|)$$

For simplicity we will assume that we are in the nonperiodic case, that is, $\gcd\{k\geqslant 0:\varphi_k>0\}=1$. This assumption assures that $z_0=\rho$ is the only singularity on the radius of convergence $|z|=\rho$ and that $T(z)$ can be analytically continued to (at least) to a region of the form $|z|\leqslant\rho+\eta$, $|\arg(z-\rho)|>0$. By using standard methods of singularity analysis (see [11]) this directly implies

$$T_n=\sqrt{\frac{\varphi(\tau)}{2\pi\varphi''(\tau)}}\frac{\rho^{-n}}{n^{3/2}}(1+O(n^{-1})) \tag{4}$$

Further note that

$$\max_{|z|\leqslant\rho}|T(z)|=T(\rho)=\tau<\infty$$

4.2. *The number of leaves*

For a rooted tree t let $\ell(t)$ denote the number of leaves of t. Then the generating function

$$L(z,u)=\sum_t\omega(t)z^{|t|}u^{\ell(t)}=\sum_{n\geqslant 1}T_n\mathbb{E}[u^{L_n}]z^n$$

satisfies the functional equation

$$L(z,u)=z\varphi(L(z,u))+\varphi_0z(u-1)$$

Note that $L(z,1)=T(z)$. If we now assume that u is a (positive real) parameter then it follows that the function $z\mapsto L(z,u)$ has a radius of convergence $z_0=r(u)=1/\varphi'(\tau(u))$ that is given by the system of equations

$$\tau(u)=r(u)\,\varphi(\tau(u))+\varphi_0z(u-1),\quad 1=r(u)\,\varphi'(\tau(u))$$

Theorem 1 (*cf. Kolchin* [*12*], *Drmota* [*13*] *and Hwang* [*14*])
Let L_n denote the random variable that counts the number of leaves in a random tree of size n (with respect to the probability distribution induced by the weights $\omega(t)$). Then

$$\mathbb{E}[L_n]=-\frac{\rho'(1)}{\rho}n+O(1)=\frac{\varphi_0}{\varphi(\tau)}n+O(1) \tag{5}$$

DOI: 10.1002/mma

and L_n satisfies a central limit theorem, that is, $(L_n - \mathbb{E}[L_n])/\mathbb{V}[L_n]$ converges weakly to the standard normal distribution $N(0,1)$.

We sketch the idea of the proof: We have the local expansions

$$L(z,u) = \tau(u) - \sqrt{\frac{2\varphi(\tau(u))}{\varphi''(\tau(u))}}\sqrt{1 - z\varphi'(\tau)} + O(|1 - z\varphi'(\tau(u))|)$$

and

$$r(u) = \rho - \rho\frac{\varphi_0}{\varphi(\tau)}(u-1) + O((u-1)^2)$$

Moreover, $r(u)$ is the only singularity on the circle of convergence $|z| = r(u)$ and we (again) have

$$\max_{|z| \leqslant r(u)} |L(z,u)| = L(r(u),u) = \tau(u) < \infty$$

Note also that $\tau(u)$ is continuous in u.

By extracting the nth coefficient of $L(z,u)$ we get (similarly to the above)

$$T_n \mathbb{E}[u^{L_n}] = \sqrt{\frac{\varphi(\tau(u))}{2\pi\varphi''(\tau(u))}}\frac{r(u)^{-n}}{n^{3/2}}(1 + O(n^{-1}))$$

Consequently,

$$\mathbb{E}[u^{L_n}] = \sqrt{\frac{\varphi''(\tau)\varphi(\tau(u))}{\varphi(\tau)\varphi''(\tau(u))}}\left(\frac{\rho}{r(u)}\right)^n (1 + O(n^{-1}))$$

this holds uniformly in a neighborhood of $u = 1$. Thus, by applying the local expansion of $r(u)$ (that we can extend to the complex plane), we directly get the central limit theorem and the asymptotic expansion (5).

4.3. Result on the w *parameter for simply generated trees*

Now we turn our attention to the w parameter. First we present an explicit formula for $F(z,u,v)$ (cf. Section 3).

Lemma 2

Let $T(z)$ and $L(z,u)$ be as above. Then

$$F(z,u,v) = \varphi_0 zu + \frac{\varphi_0 z^2 uv \dfrac{\varphi(L(zv, v^{-1})) - \varphi(L(zv, uv^{-1}))}{L(zv, v^{-1}) - L(zv, uv^{-1})}}{1 - z\dfrac{\varphi(T(z)) - \varphi(L(z,u))}{T(z) - L(z,u)}} \tag{6}$$

and

$$\overline{F}(z,u,v) = \varphi_0 zu + \frac{\varphi_0 z^2 u \dfrac{\varphi(L(z,v)) - \varphi(L(z,uv))}{L(z,v) - L(z,uv)}}{1 - z\dfrac{\varphi(T(z)) - \varphi(L(z,u))}{T(z) - L(z,u)}} \tag{7}$$

DOI: 10.1002/mma

Proof
Set $\tilde{F}(z,u,v)=F(z,u,v)-\varphi_0 zu$. Then the recursive description of simply generated trees leads to

$$\begin{aligned}\tilde{F}&=z\tilde{F}(\varphi_1+\varphi_2(T(z)+L(z,u))+\varphi_3(T(z)^2+T(z)L(z,u)+L(z,u)^2)+\cdots)\\&\quad+\varphi_0 z^2 uv(\varphi_1+\varphi_2(L(zv,v^{-1})+L(zv,uv^{-1}))\\&\quad+\varphi_3(L(zv,v^{-1})^2+L(zv,v^{-1})L(zv,uv^{-1})+L(zv,uv^{-1})^2)+\cdots)\\&=z\tilde{F}\frac{\varphi(T(z))-\varphi(L(z,u))}{T(z)-L(z,u)}+\varphi_0 z^2 uv\frac{\varphi(L(zv,v^{-1}))-\varphi(L(zv,uv^{-1}))}{L(zv,v^{-1})-L(zv,uv^{-1})}\end{aligned}$$

Of course, this proves (6). The proof of (7) is just a slight variation of that of (6). □

Theorem 3
Suppose that $\varepsilon>0$. Then

$$\mathbb{E}[v^{X_{n,j}}]=\rho v\varphi'(L(\rho v,v^{-1}))+O(n^{-1})$$

and

$$\mathbb{E}[v^{\overline{X}_{n,j}}]=\rho\varphi'(L(\rho,v))+O(n^{-1})$$

uniformly for

$$|v|\leqslant 1-\varepsilon \quad\text{and}\quad \varepsilon\leqslant\frac{j}{n}\leqslant\frac{\varphi_0}{\varphi(\tau)}(1-\varepsilon)$$

Consequently, $X_{n,j}$ and $\overline{X}_{n,j}$ have discrete limiting distributions X and $\overline{X}$ that are independent of the leaf number j as long as j is contained in the range given above.

Moreover, if p_m and $\overline{p}_m$ denote the probabilities $p_m=\mathbb{P}\{X=m\}$ and $\overline{p}_m=\mathbb{P}\{\overline{X}=m\}$ then

$$\mathbb{P}\{X_{n,j}=m\}=p_m+O((1-\varepsilon)^{-m}/n)$$

and

$$\mathbb{P}\{\overline{X}_{n,j}=m\}=\overline{p}_m+O((1-\varepsilon)^{-m}/n)$$

uniformly in the range $\varepsilon\leqslant j/n\leqslant(\varphi_0/\varphi(\tau))(1-\varepsilon)$.

Remark 4
Since the expected number of leaves is given by $\varphi_0/\varphi(\tau)n$, it is natural to assume this upper bound of the number of leaves. Note that $X_{n,j}$ is set zero if there is no jth leaf. This can of course occur but only with a very small probability that is hidden in the error term. Recall that it is expected that $X_{n,j}$ has a discrete limiting distribution, that is, it does not (really) depend on the tree size. There might be some *side effects* around the left most and right most leaves.

It should be further mentioned that the probability generating functions of the limiting distributions X and $\overline{X}$, namely $f(v)=\rho v^2\varphi'(T(\rho v))$ and $\overline{f}(v)=\rho v\varphi'(L(\rho,v))$, are singular at $v=1$. For example

$$f(v)=1-\sqrt{\frac{2\varphi(\tau)\varphi''(\tau)}{\varphi'(\tau)^2}\left(1-\frac{\varphi_0}{\varphi(\tau)}\right)}\sqrt{1-v}+O(|1-v|)$$

DOI: 10.1002/mma

which leads to

$$p_m \sim \sqrt{\frac{\varphi(\tau)\varphi''(\tau)}{2\pi\varphi'(\tau)^2}\left(1-\frac{\varphi_0}{\varphi(\tau)}\right)}m^{-3/2}$$

This asymptotic expansion also shows that the expected value of X resp. $\overline{X}$ is infinite.

In fact, we will also show that $\mathbb{E}[X_{n,j}]\to\infty$ and $\mathbb{E}[\overline{X}_{n,j}]\to\infty$.

Theorem 5
Set $\sigma := \tau\sqrt{\varphi''(\tau)/\varphi(\tau)}$. Then the expected value $\mathbb{E}[X_{n,j}]$ is given by

$$\mathbb{E}[X_{n,j}] = \frac{\varphi_0}{\varphi(\tau)}\left(1-\frac{\varphi_0}{\varphi(\tau)}\right)\frac{\sigma}{\sqrt{2\pi}}\frac{n^{3/2}}{\sqrt{j\left(\frac{\varphi_0}{\varphi(\tau)}n-j\right)}}(1+O(n^{-1/2})) \tag{8}$$

and

$$\mathbb{E}[\overline{X}_{n,j}] = \left(\frac{\varphi_0}{\varphi(\tau)}\right)^2\frac{\sigma}{\sqrt{2\pi}}\frac{n^{3/2}}{\sqrt{j\left(\frac{\varphi_0}{\varphi(\tau)}n-j\right)}}(1+O(n^{-1/2})) \tag{9}$$

uniformly for $\varepsilon \leqslant j/n \leqslant (\varphi_0/\varphi(\tau))(1-\varepsilon)$, where $\varepsilon>0$.

4.4. Proofs

The proof of these properties relies on an analysis of the generating functions $F(z,u,v)$ and $\overline{F}(z,u,v)$. For the sake of brevity we only present the first case here.

Proof of Theorem 3
We use a slightly different representation of $F(z,u,v)$. Since $T(z)=z\varphi(T(z))$ and $L(z,u)=z\varphi(L(z,u))+\varphi_0 z(u-1)$, we have

$$1-z\frac{\varphi(T(z))-\varphi(L(z,u))}{T(z)-L(z,u)} = \frac{z\varphi_0(1-u)}{T(z)-L(z,u)}$$

Consequently,

$$\begin{aligned} F(z,u,v) &= \varphi_0 zu + \frac{zuv(T(z)-L(z,u))}{1-u}\frac{\varphi(L(zv,v^{-1}))-\varphi(L(zv,uv^{-1}))}{L(zv,v^{-1})-L(zv,uv^{-1})} \\ &= \varphi_0 zu + \frac{zuvT(z)\varphi'(L(zv,v^{-1}))}{1-u} - \frac{zuvL(z,u)\varphi'(L(zv,v^{-1}))}{1-u} \\ &\quad - \frac{zuvT(z)}{1-u}\left(\varphi'(L(zv,v^{-1})) - \frac{\varphi(L(zv,v^{-1}))-\varphi(L(zv,uv^{-1}))}{L(zv,v^{-1})-L(zv,uv^{-1})}\right) \\ &\quad + \frac{zuv^2L(z,u)}{1-u}\left(\varphi'(L(zv,v^{-1})) - \frac{\varphi(L(zv,v^{-1}))-\varphi(L(zv,uv^{-1}))}{L(zv,v^{-1})-L(zv,uv^{-1})}\right) \end{aligned}$$

Our aim is to extract the coefficient of $u^j z^n$.

DOI: 10.1002/mma

We first discuss the term

$$\frac{zuvT(z)\varphi'(L(zv,v^{-1}))}{1-u}$$

The coefficient of u^j, $j\geqslant 1$, is trivially given by

$$zvT(z)\varphi'(L(zv,v^{-1}))$$

Note that we assume that $|v|\leqslant 1-\varepsilon$. This means that $L(zv,v^{-1})$ is regular if z is close to ρ. Hence the singular expansion of $zvT(z)\varphi'(L(zv,v^{-1}))$ is given by

$$\rho v\varphi'(L(\rho v,v^{-1}))\left(\tau-\sqrt{\frac{2\varphi(\tau)}{\varphi''(\tau)}}\sqrt{1-z\varphi'(\tau)}+c(1-z\varphi'(\tau))+O(|1-z\varphi'(\tau)|^{3/2})\right)$$

with some constant c. Of course this shows that the coefficient of z^n is asymptotically given by

$$T_n\rho v\varphi'(L(\rho v,v^{-1}))(1+O(n^{-1}))$$

It will turn out that this term is in fact the dominating term. Thus, it remains to show that the other terms are asymptotically small.

In order to extract the coefficient of u^jz^n in

$$\frac{zuvL(z,u)\varphi'(L(zv,v^{-1}))}{1-u}$$

we use Cauchy's formula and have to estimate the integral

$$\frac{1}{(2\pi\mathrm{i})^2}\int_{|z|=r(u')}\int_{|u|=u'}\frac{zuvL(z,u)\varphi'(L(zv,v^{-1}))}{1-u}\frac{\mathrm{d}u}{u^{j+1}}\frac{\mathrm{d}z}{z^{n+1}}$$

where $u'=1-\kappa/j$ with $\kappa=\log^2 n$. Since $j\geqslant\varepsilon n$ it is no loss of generality to assume that $j>\log^2 n$. Since we know that $|L(z,u)|\leqslant L(|z|,|u|)=L(r(u'),u')$ is bounded and that $L(zv,v^{-1})$ is regular (for sufficiently large $j\geqslant\varepsilon n$) we get a trivial upper bound of the form

$$C\frac{j}{\kappa}u'^{-j}r(u)^{-n}\ll\frac{j}{\kappa}\rho^{-n}\mathrm{e}^{-\kappa((\varphi_0/\varphi(\tau))(n/j)-1)}\ll n\rho^{-n}\frac{\mathrm{e}^{-\varepsilon\kappa}}{\kappa}$$

Recall that $\kappa=\log^2 n$. Hence, get an upper bound of the form T_n/n.

For the third term we use a similar procedure. Since $|v|\leqslant 1-\varepsilon$, it follows that there exists $\eta>0$ such that

$$\varphi'(L(zv,v^{-1}))-\frac{\varphi(L(zv,v^{-1}))-\varphi(L(zv,uv^{-1}))}{L(zv,v^{-1})-L(zv,uv^{-1})}=O(|u-1|)$$

uniformly for $|u|\leqslant 1+\eta$ and $|z|\leqslant\rho$. This implies that $u=1$ is not a singularity of the third term. Hence, we use Cauchy's formula for $|z|=\rho$ and $|u|=1+\eta$ and get an upper bound for the coefficient

DOI: 10.1002/mma

of $u^j z^n$ of the form

$$C\rho^{-n}(1+\eta)^{-j} \ll \rho^{-n}(1+\eta)^{-\varepsilon n} \ll \frac{T_n}{n}$$

Finally, for the fourth and last term we again use the contours $|u|=u'=1-\kappa/j$ with $\kappa=\log^2 n$ and $|z|=r(u')$ and get the same kind of upper bound of the form

$$C\rho^{-n}\frac{\mathrm{e}^{-\varepsilon\kappa}}{\kappa} \ll \frac{T_n}{n}$$

This completes the proof of Theorem 3 for $X_{n,j}$. □

Proof of Theorem 5
For the proof of (8) we work with

$$G(z,u) = \frac{\partial}{\partial v}F(z,u,v)\bigg|_{v=1} = \sum_{n,j} T_n \mathbb{E}[X_{n,j}]z^n u^j$$

By using the explicit representation for $F(z,u,v)$ (in terms of $T(z)$ and $L(z,u)$) and the functional equations for $T(z)$ and $L(z,u)$ we obtain an alternative representation:

$$G(z,u) = \varphi_0 uz + \frac{\varphi_0 uz}{T(z)-L(z,u)}\left(\frac{T(z)-\varphi_0 z}{1-z\varphi'(T(z))} - \frac{L(z,u)-\varphi_0 uz}{1-z\varphi'(L(z,u))}\right)$$

Note that

$$\frac{(T(z)-\varphi_0 z)(1-z\varphi'(L(z,u)))-(L(z,u)-\varphi_0 uz)(1-z\varphi'(T(z)))}{T(z)-L(z,u)}$$
$$=1-z\varphi'(T(z))+(T(z)-\varphi_0 z)z\varphi''(T(z))+O(|T(z)-L(z,u)|+|u-1|)$$

Hence, the asymptotic leading term of $G(z,u)$ (if z is close to ρ and u is close to 1) is given by

$$\frac{\varphi_0 uz^2(T(z)-\varphi_0 z)\varphi''(T(z))}{(1-z\varphi'(T(z)))(1-z\varphi'(L(z,u)))} \tag{10}$$

We concentrate on that term. It will then be an easy exercise to estimate the (minor) contribution of the other terms.

We are now in a situation where we can use the technique of [15], the *double Hankel contour method*. A Hankel contour[‡] $\mathscr{H}_M$, $0<M\leqslant\infty$ is of the form (compare also with Figure 2)

$$\mathscr{H}_M = \{t\in\mathbb{C}: |t|=1,\ \Re(t)<0\} \cup \{t\in\mathbb{C}: 0\leqslant\Re(t)\leqslant M,\ \Im(t)=\pm 1\}$$

For example, by using the substitution $v=w^2/2$ one easily shows that

$$\frac{1}{2\pi\mathrm{i}}\int_{\mathscr{H}_\infty}\frac{\mathrm{e}^{-v}}{\sqrt{-v}}\,\mathrm{d}v = \frac{1}{\sqrt{\pi}} \tag{11}$$

[‡] This notation is adopted from Hankel's representations of $\Gamma(z)$ and $1/\Gamma(z)$, where a 'Hankel contour' appears.

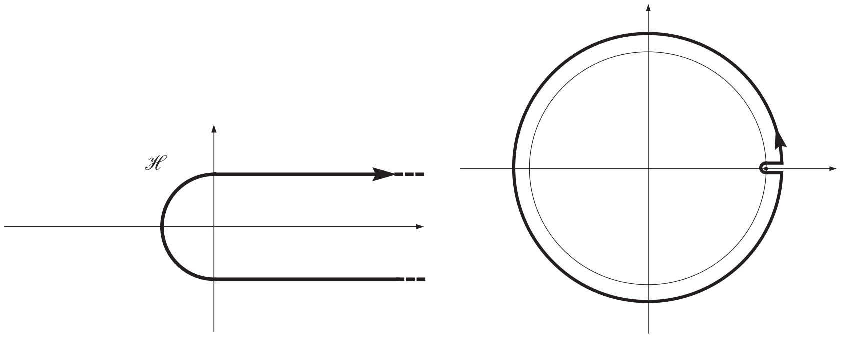


Figure 2. Hankel contour and path of integration.

We will use this formula in the sequel. By Cauchy's formula, the coefficient of $u^j z^n$ in (10) is given by

$$\frac{1}{(2\pi \mathrm{i})^2}\int_{|u|=u_0}\int_{|z|=z_0}\frac{\varphi_0 u z^2(T(z)-\varphi_0 z)\varphi''(T(z))}{(1-z\varphi'(T(z)))(1-z\varphi'(L(z,u)))}\frac{\mathrm{d}u}{u^{j+1}}\frac{\mathrm{d}z}{z^{n+1}} \tag{12}$$

where $u_0<1$ and $z_0<\rho$. We already know that $T(z)$ has a (square root) singularity at $z=\rho$ and $L(z,u)$ has (also a square root) singularity at $z=r(u)$, that is, if u is fixed then $z=r(u)$ is a singularity in z and, conversely, if z is fixed then $u=\rho^{-1}(z)$ is a singularity in u. It is a well-known fact that the main asymptotic behaviour of the integral (12) comes from that part of integration that is close to the singularity. More precisely, we will transform (first) the path of integration from $|z|=z_0$ into the form that is depicted in Figure 2; compare also with the *singularity analysis method* of Flajolet and Odlyzko [11]. In particular, for the *main part* that is close to the singularity one (usually) uses the substitution $z=\rho(1+t/n)$, where $t\in\tilde{\mathscr{H}}_1(n):=\mathscr{H}_{(\log n)^2}$. In (almost) the same way we modify the path of integration for u depending on $z=\rho(1+t/n)$ (for $t\in\mathscr{H}_{(\log n)^2}$). We set $u=1+s/j$, where

$$s\in\tilde{\mathscr{H}}_2(n,t):=\mathscr{H}_{K(\log n)^2}-\frac{m}{n}\frac{\varphi(\tau)}{\varphi_0}t$$

where K is a suitably chosen (large) constant depending on ε. This (small) part is completed by a *large circle* similar to the corresponding one for z, see also Figure 2. We will see in a moment that this choice ensures that the singularity of $L(z,u)$ is taken into account correctly. (If z is contained in the remaining part of the contour, that is, on the *large circle*, then we utilize a properly chosen circle for u, too; compare with [15]. However, this part of the integral will not contribute to the asymptotic leading term.)

For $z=\rho(1+t/n)$ and $u=1+s/j$ (with $t\in\tilde{\mathscr{H}}_1(n)$ and $s\in\tilde{\mathscr{H}}_2(n,t)$) we have

$$T(z)=\tau-\sqrt{\frac{2\varphi(\tau)}{\varphi''(\tau)}}\sqrt{-\frac{t}{n}}+O\left(\frac{|t|}{n}\right)$$

$$L(z,u)=\tau-\sqrt{\frac{2\varphi(\tau)}{\varphi''(\tau)}}\sqrt{-\frac{t}{n}-\frac{\varphi_0}{\varphi(\tau)}\frac{s}{j}}+O\left(\frac{|s|}{j}+\frac{|t|}{n}\right)$$

DOI: 10.1002/mma

and consequently

$$1-z\varphi'(T(z))=\sqrt{2\sigma^2}\sqrt{-\frac{t}{n}}+O\left(\frac{|t|}{n}\right)$$

$$1-z\varphi'(L(z,u))=\sqrt{2\sigma^2}\sqrt{-\frac{t}{n}-\frac{\varphi_0}{\varphi(\tau)}\frac{s}{j}}+O\left(\frac{|s|}{j}+\frac{|t|}{n}\right)$$

Hence, the asymptotic leading term of the integral in (12) is given by

$$\frac{\rho^{-n}}{jn}\frac{1}{(2\pi \mathrm{i})^2}\int_{\tilde{\mathscr{H}}_1(n)}\int_{\tilde{\mathscr{H}}_2(n,t)}\frac{\varphi_0\rho^2(\tau-\varphi_0\rho)\varphi''(\tau)}{2\sigma^2\sqrt{-\frac{t}{n}}\sqrt{-\frac{t}{n}-\frac{\varphi_0}{\varphi(\tau)}\frac{s}{j}}}\mathrm{e}^{-s-t}\left(1+O\left(\frac{|t|}{\sqrt{n}}+\frac{|s|}{\sqrt{j}}\right)\right)\mathrm{d}s\,\mathrm{d}t$$

$$=\frac{\rho^{-n}}{\sqrt{jn}}\frac{\sqrt{\varphi_0}(\tau-\varphi_0\rho)}{2\sqrt{\varphi(\tau)}}\frac{1}{(2\pi \mathrm{i})^2}\int_{\tilde{\mathscr{H}}_1(n)}\int_{\tilde{\mathscr{H}}_2(n,t)}\frac{\mathrm{e}^{-s-t}}{\sqrt{-t}\sqrt{-s-t\frac{\varphi(\tau)}{\varphi_0}\frac{j}{n}}}\left(1+O\left(\frac{|t|}{\sqrt{n}}+\frac{|s|}{\sqrt{j}}\right)\right)\mathrm{d}s\,\mathrm{d}t$$

Next we use the substitutions $v=s+t(\varphi(\tau)/\varphi_0)j/n$ and $w=t(1-(\varphi(\tau)/\varphi_0)j/n)$. We obtain

$$\frac{\rho^{-n}}{\sqrt{jn}}\frac{\sqrt{\varphi_0}(\tau-\varphi_0\rho)}{2\sqrt{\varphi(\tau)}}\frac{1}{(2\pi \mathrm{i})^2}\frac{1}{\sqrt{1-\frac{\varphi(\tau)}{\varphi_0}\frac{j}{n}}}\int_{\mathscr{H}}\int_{\mathscr{H}}\frac{\mathrm{e}^{-v-w}}{\sqrt{-v}\sqrt{-w}}\mathrm{d}v\,\mathrm{d}w=\frac{\varphi_0}{\varphi(\tau)}\frac{\rho^{-n}}{2\pi}\frac{\tau-\varphi_0\rho}{\sqrt{j\left(\frac{\varphi_0}{\varphi(\tau)}n-j\right)}}$$

if we replace the paths of integration by (infinite) Hankel contours, use formula (11), and neglect the (small) error term. Hence, if we divide by the asymptotics of T_n given in (4) we finally get

$$\mathbb{E}[X_{n,j}]\sim\frac{\varphi_0}{\varphi(\tau)}\left(1-\frac{\varphi_0}{\varphi(\tau)}\right)\frac{\sigma}{\sqrt{2\pi}}\frac{n^{3/2}}{\sqrt{j\left(\frac{\varphi_0}{\varphi(\tau)}n-j\right)}}$$

Of course, we must carefully handle the error terms; however, it is not difficult to show that they all are of order $1/\sqrt{n}$ smaller than the asymptotic leading term. We leave the details to the reader; compare also with [15].

The proof for (9) is very similar to that of (8). □

5. RECURSIVE TREES

There are $(n-1)!$ trees of size n; by definition there is no tree of size 0. Furthermore, we consider labelled objects, so the natural weight for a tree t is $\omega(t)=1/|t|!$. The combinatorial properties of this tree class translate into the functional-differential equation $T'(z)=\mathrm{e}^{T(z)}$ for its generating

DOI: 10.1002/mma

function. The generating function is therefore

$$T(z) = \log \frac{1}{1-z}$$

Since there is no order among the successors of a specific node, we have to consider the W parameter in this case. As in the previous section, let $L(z,u)$ denote the generating function counting size and number of leaves. We start with two ancillary results, one on $L(z,u)$ and one on $G(z,u,v)$.

Lemma 6
The generating function $L(z,u)$ is

$$L(z,u) = \log \frac{1}{\mathrm{e}^{-(u-1)z} - \dfrac{1}{u-1}(1-\mathrm{e}^{-(u-1)z})}$$

Proof
The function $L(z,u)$ is the solution of the differential equation $y' = \mathrm{e}^y + u - 1$ with initial condition $y(0) = 0$ where u can be viewed as a formal parameter. Solving this equation yields the assertion. □

Lemma 7
The generating function $G(z,u,v)$ satisfies the following functional-differential equation:

$$\frac{\partial G(z,u,v)}{\partial z} = \sum_{k \geqslant 0} \frac{1}{k!} \sum_{m=0}^{k} \binom{k}{m} z^m v^m u (G(z,uv^m,v) - z)^{k-m} \tag{13}$$

Proof
Consider a node x in some given recursive tree t. Assume that x has outdegree k and that exactly m subtrees of x are leaves. Then for each of these leaves the contribution to the W parameter equals the number of internal nodes of the subtree of t, which is rooted at x. Hence, x itself and all internal nodes of all nonleaves attached to x contribute m times. Now, the result can be easily seen by taking into account the additivity of the W parameter. □

It will eventually turn out that this equation is too complicated to obtain distributional results, but a pumping moment approach can be used to compute moments, though the expressions quickly become quite intricate, even for the second moment. So we will only compute the first two moments. In principle, higher moments can be extracted in the same manner.

We get

Theorem 8
The first two moments of the W parameter for recursive trees are

$$\mathbb{E}[X_n] = \begin{cases} 0, & n<2 \\ 1, & n=2 \\ \dfrac{n}{2}\left(H_n - \dfrac{5}{6}\right), & n \geqslant 3 \end{cases} \tag{14}$$

DOI: 10.1002/mma

and

$$\mathbb{E}[X_n^2] = \begin{cases} 0, & n<2 \\ 1, & n=2 \\ \dfrac{5}{2}, & n=3 \\ \dfrac{41}{6}, & n=4 \\ \dfrac{347}{24}, & n=5 \\ nL(n), & n\geqslant 6 \end{cases} \tag{15}$$

where

$$L(n) = \left(\frac{n}{4}-\frac{37}{12}\right)(H_n^2 - H_n^{(2)}) - \left(\frac{5n}{12}-\frac{1589}{360}\right)H_n + \frac{10}{3}H_n^{(2)} + \frac{46\,950n^3 - 160\,453n^2 - 14\,297n + 129\,600}{21\,600n(n-1)}$$

The asymptotic expressions for expectation and variance are given by

$$\mathbb{E}[X_n] = \frac{1}{2}n\log n + \left(\frac{\gamma}{2}-\frac{5}{6}\right)n + O(1)$$

and

$$\mathbb{V}[X_n] = \left(2-\frac{\pi^2}{24}\right)n^2 - \frac{37}{12}n\log^2 n + \frac{1589-2220\gamma}{360}n\log n + O(n)$$

Remark 9

Since the variance is smaller than the square of the mean, we observe that the distribution is concentrated around its mean. The complexity of the expressions and computations for the first two moments make it seem rather hopeless to compute all the other moments. Nevertheless the first two moments resemble those of the path length which is not normal, see [16].

Proof

Let

$$f_1(z) = \sum_{n\geqslant 1} \mathbb{E}[X_n]\frac{z^n}{n} = \frac{\partial G}{\partial v}(z,1,v)\bigg|_{v=1}$$

Then normalizing yields $\mathbb{E}[X_n] = n[z^n]f_1(z)$. Differentiating (13) w.r.t. v and standard simplifications lead to the first-order linear differential equation

$$f_1'(z) = \frac{1}{1-z}f_1(z) + \frac{z(G_u(z,1,1)+1)}{1-z} \tag{16}$$

DOI: 10.1002/mma

Note that

$$G(z,u,1)=L\left(zu,\frac{1}{u}\right) \tag{17}$$

and therefore $G_u(z,1,1)=zL_z(z,1)-L_u(z,1)$ where subscripts denote partial derivatives. Using $L(z,1)=T(z)$ and Lemma 6 we obtain

$$L_z(z,1)=\frac{1}{1-z},\quad L_u(z,1)=\frac{2z-z^2}{2(1-z)}$$

and using these formulae we get $G_u(z,1,1)=(2z-z^2)/(1-z)$ and thus Equation (16) becomes

$$f_1'(z)=\frac{1}{1-z}f_1(z)+\frac{2z-2z^2+z^3}{2(1-z)^2}$$

Solving this equation gives

$$f_1(z)=\frac{1}{1-z}\left(\frac{1}{2}\log\frac{1}{1-z}-\frac{z}{2}+\frac{z^2}{4}-\frac{z^3}{6}\right)$$

From this one easily extracts the coefficients and obtains (14).

For computing the second moment $\mathbb{E}[X_n^2]$ set

$$f_2(z)=\sum_{n\geqslant 1}\mathbb{E}[X_n^2]\frac{z^n}{n}=\left(v\frac{\partial}{\partial v}\right)^2 G(z,1,v)\bigg|_{v=1}$$

Hence,

$$\mathbb{E}[X_n^2]=n[z^n]f_2(z)=n[z^n]\left(v\frac{\partial}{\partial v}\right)^2 G(z,1,v)\bigg|_{v=1}$$

This formal differentiation can be done with the help of Maple. We get the differential equation

$$\begin{aligned}f_2'(z)=&\frac{1}{1-z}f_2(z)+\frac{(z+z^2)(1+G_{uu}(z,1,1)+3G_u(z,1,1)+G_u(z,1,1)^2)}{1-z}\\&+\frac{2z(f_1(z)+G_{uv}(z,1,1)+G_u(z,1,1)f_1(z)+f_1(z)^2}{1-z}\end{aligned} \tag{18}$$

Employing (17) again we obtain $G_{uu}(z,1,1)=z^2L_zz(z,1)-2zL_{zu}(z,1)+L_{uu}(z,1)+2L_u(z,1)$. Lemma 6 and the differential equation for $L(z,u)$ gives

$$L_zz(z,1)=\frac{1}{(1-z)^2},\quad L_{zu}(z,1)=\frac{2-2z+z^2}{2(1-z)^2},\quad L_uu(z,1)=\frac{(2z-z^2)^2}{4(1-z)^2}-\frac{3z^2-z^3}{3(1-z)}$$

and hence $G_{uu}(z,1,1)=z^3(4-z)/12(1-z)^2$. In order to compute $G_{uv}(z,1,1)$ we differentiate (13) w.r.t. u and v. Then $G_{uv}(z,1,1)$ is the solution of the differential equation

$$y'=\frac{y}{1-z}+\frac{(z+z^2)(G_{uu}(z,1,1)+G_u(z,1,1)^2)+z(1+3G_u(z,1,1))+f_1(z)(1+G_u(z,1,1))}{1-z}$$

DOI: 10.1002/mma

with initial condition $y(0)=0$. Plugging the solution of this equation into (18) and solving (18) afterwards gives the desired expression for $f_2(z)$:

$$\begin{aligned}f_2(z) &= \frac{1}{4(1-z)^2}\log^2\frac{1}{1-z}+\frac{1}{12(1-z)^2}\log\frac{1}{1-z}+\frac{253}{144(1-z)^2}-\frac{10}{3(1-z)}\log^2\frac{1}{1-z}\\&\quad+\frac{1559}{360(1-z)}\log\frac{1}{1-z}-\frac{151\,453}{21\,600(1-z)}-6\log\frac{1}{1-z}+\frac{z}{12}\log\frac{1}{1-z}\\&\quad+\frac{113\,503+109\,813z+33\,193z^2+8463z^3+2028z^4+240z^5}{21\,600}\end{aligned}$$

Extracting the coefficient by using the well-known formulas (see [17, p. 10])

$$\frac{1}{(1-z)^{m+1}}\log\frac{1}{1-z}=\sum_{n\geqslant 0}\binom{n+m}{m}(H_{n+m}-H_m)z^n$$

$$\frac{1}{(1-z)^{m+1}}\left(\log\frac{1}{1-z}\right)^2=\sum_{n\geqslant 0}\binom{n+m}{m}((H_{n+m}-H_m)^2-(H^{(2)}_{n+m}-H^{(2)}_m))z^n$$

and then normalizing, we obtain (15). The variance can now be computed from $\mathbb{V}[X_n]=\mathbb{E}[X_n^2]-\mathbb{E}[X_n]^2$. □

6. BINARY SEARCH TREES

Consider a binary search tree constructed from a random permutation π of $\{1,2,\ldots,n\}$. In that way an—at first—incomplete binary tree is constructed using the values $\pi(1),\pi(2),\ldots,\pi(n)$ and the rules described in Section 2. Then it is completed by attaching leaves to all nodes of outdegree less than two. These leaves do not store any data and correspond technically to null pointers.

To study the random variable $X_{n,j}$ we use the well-known combinatorial decomposition of binary search trees, i.e. viewing the tree as a single root, if it has size one, and as a root with two binary search trees attached to it, otherwise. This translates into the following differential equation for $P(z,u,v)$:

$$\frac{\partial P(z,u,v)}{\partial z}=\frac{P(z,u,v)-1}{1-z}+\frac{v}{1-zv}+\frac{u(P(z,u,v)-1)}{1-zu}+\frac{uv}{1-zuv} \tag{19}$$

with initial condition $P(0,u,v)=1$.

Via standard techniques or by using Maple one obtains the following solution of the differential equation (19):

$$\begin{aligned}P(z,u,v) &= \frac{-zu-z+zuv+zv+v}{v(1-zu)(1-z)}+\frac{(1-v)(u-v)}{v^2(1-zu)(1-z)}\log\left(\frac{1}{1-zv}\right)\\&\quad+\frac{(1-v)(1-uv)}{uv^2(1-zu)(1-z)}\log\left(\frac{1}{1-zuv}\right)\end{aligned} \tag{20}$$

Extracting the exact coefficients of (19) is an easy task, but one must take several cases into account. This leads to the following theorem.

Theorem 10
The exact probabilities $\mathbb{P}\{X_{n,j}=m\}$ of the w parameter are, for $n\geqslant 0$, $0\leqslant j\leqslant n$, and $m\geqslant 0$, given as follows (outside this range of n, j, m the probabilities are zero anyway).

For $m\geqslant 1$:

$$\mathbb{P}\{X_{n,j}=m\}=\begin{cases} \dfrac{4}{m(m+1)(m+2)} & \text{for } m\leqslant \min(j-1,n-j-1) \\ \dfrac{2}{m(m+1)(m+2)}+\dfrac{1}{m(m+1)} & \text{for } (j=m\leqslant n-j-1) \text{ or } (n-j=m\leqslant j-1) \\ \dfrac{2}{m(m+1)} & \text{for } j=m=n-j \\ \dfrac{2}{m(m+1)(m+2)} & \text{for } (1\leqslant n-j<m\leqslant j-1) \text{ or } (1\leqslant j<m\leqslant n-j-1) \\ \dfrac{1}{m(m+1)} & \text{for } (1\leqslant j<m=n-j) \text{ or } (1\leqslant n-j<m=j) \\ & \quad \text{or } (j=0 \text{ and } m\leqslant n-1) \text{ or } (j=n \text{ and } m\leqslant n-1) \\ \dfrac{1}{m} & \text{for } m=n \text{ and } (j=0 \text{ or } j=n) \\ 0 & \text{for } m>\max(j,n-j) \end{cases}$$

For $m=0$:

$$\mathbb{P}\{X_{n,j}=0\}=\begin{cases} 1 & \text{for } n=j=0 \\ 0 & \text{otherwise} \end{cases}$$

For all sequences $(j=j(n))_{n\in\mathbb{N}}$ such that $j\to\infty$ and $n-j\to\infty$, it follows from Theorem 10 that $X_{n,j}$ converges weakly as $n\to\infty$ to a discrete random variable X with distribution

$$\mathbb{P}\{X=m\}=\frac{4}{m(m+1)(m+2)} \quad \text{for } m\geqslant 1$$

Interestingly enough, the expectation of X is finite:

$$\mathbb{E}(X)=\sum_{m\geqslant 1}\frac{4}{(m+1)(m+2)}=2$$

but all higher moments $\mathbb{E}(X^s)$, $s\geqslant 2$—and thus, in particular, the variance $\mathbb{V}(X)$—do not exist.

We remark that results concerning X could be obtained also by arguments that can be found in [18], where local counters, as, e.g. the number of leaves or the number of nodes with m descendants, are studied for binary search trees.

We will also give exact formulae for the expectation $\mathbb{E}(X_{n,j})$ and the second moment $\mathbb{E}(X_{n,j}^2)$. To do this, we differentiate $P(z,u,v)$ as given by (20) once and twice w.r.t. v and evaluate at

DOI: 10.1002/mma

$v=1$. We obtain

$$G_1(z,u) := \left.\frac{\partial P(z,u,v)}{\partial v}\right|_{v=1} = \frac{1-u}{(1-zu)(1-z)}\log\left(\frac{1}{1-z}\right)+\frac{u-1}{u(1-zu)(1-z)}\log\left(\frac{1}{1-zu}\right) + \frac{z(u+1)}{(1-zu)(1-z)}$$

$$G_2(z,u) := \left.\frac{\partial^2 P(z,u,v)}{\partial v^2}\right|_{v=1} = \frac{4u-2}{(1-zu)(1-z)}\log\left(\frac{1}{1-z}\right)+\frac{4-2u}{u(1-zu)(1-z)}\log\left(\frac{1}{1-zu}\right) + \frac{2(2-z)}{(1-z)^2}+\frac{2}{(1-zu)^2}-\frac{2(z+3)}{(1-zu)(1-z)}$$

Extracting coefficients then leads to the following result:

$$\mathbb{E}(X_{n,j}) = [z^n u^j]G_1(z,u) = \begin{cases} 2-\dfrac{1}{j+1}-\dfrac{1}{n-j+1} & \text{for } 1\leqslant j\leqslant n-1 \\ H_n & \text{for } (j=0\vee j=n)\wedge n\geqslant 1 \\ 0 & \text{for } j=0\wedge n=0 \end{cases}$$

$$\mathbb{E}(X_{n,j}^2) = [z^n u^j]G_2(z,u)+\mathbb{E}(X_{n,j}) = \begin{cases} 2H_j+2H_{n-j}+\dfrac{3}{j+1}+\dfrac{3}{n-j+1}-6 & \text{for } 1\leqslant j\leqslant n-1 \\ 2n-H_n & \text{for } (j=0\vee j=n)\wedge n\geqslant 1 \\ 0 & \text{for } j=0\wedge n=0 \end{cases}$$

7. THE W PARAMETER FOR DIGITAL SEARCH TREES

Recall from Section 2 that digital search trees are built in a similar way as binary search trees, but the keys are 0–1-strings rather than integers. That means that the keys are used to construct an incomplete binary tree, which is completed by attached leaves to all nodes of outdegree less than two. Internal nodes where both successors are leaves are called *internal endnodes*.

When studying the w parameter in digital search trees, we confine ourselves to the W parameter. Here we compute the expected value of the W parameter. Studying this parameter is similar—but slightly more involved—than studying the number of internal endnodes, which was performed by Flajolet and Sedgewick, who solved a problem of Knuth [9].

We consider the symmetric model of digital search trees only, but the modification to the asymmetric model would be straightforward, for the expected value. The reader should note that in [19] the variance of the number of internal endnodes was computed, and this project was of daunting complexity.

DOI: 10.1002/mma

Theorem 11

The expected value of the W parameter in the symmetric model of digital search trees is

$$\mathbb{E}[X_n] \sim n(A+\delta(\log_2 n))$$

with

$$A := -\sum_{l\geqslant 0}\frac{2^{-l}(l+1)(l-2)}{Q_l}+\frac{1}{L}\sum_{l\geqslant 0}\frac{2^{-l}(2l-1)}{Q_l}+\sum_{k\geqslant 1, l\geqslant 0}\frac{2^{-l}}{2^{l+k}-1}\frac{(l+1)(l-2)}{Q_l}=1.1030266984\ldots$$

The periodic function $\delta(x)$ of period 1 has mean zero and computable Fourier coefficients.[§]

Proof

For brevity, set $l_n := \mathbb{E}[X_n]$. Then the following recursion for $n\geqslant 3$ is straightforward:

$$l_{n+1}=2^{-n}\sum_{k=2}^{n-2}\binom{n}{k}(l_k+l_{n-k})+2^{1-n}l_n+2^{1-n}n(n+1+l_{n-1})$$

and $l_0=l_1=0$, $l_2=2$, $l_3=4$.

We can rewrite it:

$$\begin{aligned} l_{n+1} &= 2^{1-n}\sum_{k=0}^{n-2}\binom{n}{k}l_k+2^{1-n}l_n+2^{1-n}n(n+1+l_{n-1}) \\ &= 2^{1-n}\sum_{k=0}^{n}\binom{n}{k}l_k+2^{1-n}n(n+1) \end{aligned}$$

Recalling $L(z)=\sum_{n\geqslant 0} l_n z^n/n!$ this translates into

$$L'(z)=2\mathrm{e}^{z/2}L\left(\frac{z}{2}\right)+\mathrm{e}^{z/2}\left(2z+\frac{z^2}{2}\right)$$

With the *Poisson transformed* function $M(z)=\mathrm{e}^{-z}L(z)$, this is

$$M(z)+M'(z)=2M\left(\frac{z}{2}\right)+\mathrm{e}^{-z/2}\left(2z+\frac{z^2}{2}\right)$$

Reading off the coefficients $m_n := n![z^n]M(z)$ we get:

$$m_n+m_{n+1}=2^{1-n}m_n+n![z^n]\mathrm{e}^{-z/2}\left(\frac{z^2}{2}+2z\right)$$

[§] Note also that we use this notation in other parts of this paper in a *generic sense*, i.e. $\delta(x)$ always denotes a periodic function of period 1 with mean zero, but in different contexts it could mean a different function.

DOI: 10.1002/mma

or

$$\begin{aligned}m_{n+1}&=-(1-2^{1-n})m_n+n!\left(\frac{1}{2}[z^{n-2}]\mathrm{e}^{-z/2}+2[z^{n-1}]\mathrm{e}^{-z/2}\right)\\&=-(1-2^{1-n})m_n+\left(\frac{n(n-1)}{2}2^{2-n}(-1)^n+n2^{2-n}(-1)^{n-1}\right)\\&=-(1-2^{1-n})m_n+2^{1-n}(-1)^n n(n-3)\end{aligned}$$

This holds for $n\geqslant 2$; $m_0=m_1=0$. It follows that

$$\begin{aligned}\frac{m_{n+1}}{Q_{n-1}(-1)^{n+1}}&=\frac{m_n}{Q_{n-2}(-1)^n}-\frac{2^{1-n}n(n-3)}{Q_{n-1}}\\&=2-\sum_{k=2}^{n}\frac{2^{1-k}k(k-3)}{Q_{k-1}}\\&=-\sum_{k=0}^{n-1}\frac{2^{-k}(k+1)(k-2)}{Q_k}\end{aligned}$$

Hence,

$$m_n=Q_{n-2}(-1)^{n-1}\sum_{k=0}^{n-2}\frac{2^{-k}(k+1)(k-2)}{Q_k}$$

and

$$l_n=\sum_{k=2}^{n}\binom{n}{k}(-1)^k f(k-2)$$

with

$$\begin{aligned}f(k)&=-Q_k\sum_{l=0}^{k}\frac{2^{-l}(l+1)(l-2)}{Q_l}\\&=-Q_k\left[\sum_{l\geqslant 0}\frac{2^{-l}(l+1)(l-2)}{Q_l}-\sum_{l-k>0}\frac{2^{-l}(l+1)(l-2)}{Q_l}\right]\\&=-Q_k\left[\rho-\sum_{l\geqslant 1}\frac{2^{-l-k}(l+k+1)(l+k-2)}{Q_{l+k}}\right]\end{aligned}$$

This can be continued to a complex function:

$$f(z)=-Q_z\left[\rho-\sum_{l\geqslant 1}\frac{2^{-l-z}(l+z+1)(l+z-2)}{Q_{l+z}}\right]$$

with

$$\rho := \sum_{l\geqslant 0} \frac{2^{-l}(l+1)(l-2)}{Q_l}$$

and the usual $Q_z = Q_\infty / Q(2^{-z})$.

Following Rice's method [20], we can write

$$l_n = \frac{1}{2\pi \mathrm{i}} \int_{\mathscr{C}} \frac{n!(-1)^n}{z(z-1)\cdots(z-n)} f(z-2)\,\mathrm{d}z$$

where the curve $\mathscr{C}$ encircles the poles $2, \ldots, n$ and no others. Shifting the line of integration, we find that the main contribution of the asymptotic expansion of l_n comes from the pole at $z=1$.

We need $f(z)$ around $z=-1$. In the following, we use the shorthand notation $L = \log 2$; there is no danger to confuse this with the generating function $L(z)$!

$$f(z) \sim \frac{1}{L} \frac{\mathrm{d}}{\mathrm{d}z} \sum_{l\geqslant 1} \frac{2^{-l-z}(l+z+1)(l+z-2)}{Q_{l+z}} \bigg|_{z=-1}$$

With some effort, this differentiation can be made explicit:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}z} \sum_{l\geqslant 1} \frac{2^{-l-z}(l+z+1)(l+z-2)}{Q_{l+z}} &= \frac{1}{Q_\infty} \frac{\mathrm{d}}{\mathrm{d}z} \sum_{l\geqslant 1} 2^{-l-z}(l+z+1)(l+z-2)Q(2^{-l-z}) \\ &= -\frac{L}{Q_\infty} \sum_{l\geqslant 1} 2^{-l-z}(l+z+1)(l+z-2)Q(2^{-l-z}) \\ &\quad + \frac{1}{Q_\infty} \sum_{l\geqslant 1} 2^{-l-z}(2l+2z-1)Q(2^{-l-z}) \\ &\quad - \frac{L}{Q_\infty} \sum_{l\geqslant 1} 2^{-l-z}(l+z+1)(l+z-2)Q'(2^{-l-z})\end{aligned}$$

Note that

$$Q'(x) = -Q(x) \sum_{k\geqslant 1} \frac{1}{2^k - x}$$

Consequently we have

$$\lim_{z\to -1} f(z) = -\sum_{l\geqslant 1} \frac{2^{1-l} l(l-3)}{Q_{l-1}} + \frac{1}{L} \sum_{l\geqslant 1} \frac{2^{1-l}(2l-3)}{Q_{l-1}} + \sum_{l,k\geqslant 1} \frac{2^{1-l}}{2^{l+k-1}-1} \frac{l(l-3)}{Q_{l-1}}$$

Putting everything together, the theorem follows. □

From Flajolet–Sedgewick's work [9, pp. 759–763], we know that (apart from small fluctuations), there are about $\beta \cdot n$ internal endnodes ($\beta \approx 0.372046812$). Dividing A by β, we get the average size of the W parameter, averaged over all trees and over all internal endnodes. The numerical constant is ≈ 2.9647525.

8. THE W PARAMETER FOR TRIES

Recall (Section 2) that tries are constructed in basically the same way as digital search trees, but data are stored in the leaves rather than the internal nodes. We consider a trie built of n random data (more precisely n random 0–1-strings). In general, a trie constructed from n data has more than n such that not all the leaves contain data. We will study the W parameter, where the sum is over those leaves that *contain* data only!

Let a_n denote the (well known) average number of internal nodes, in a random trie built of n random data. Note that

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

for $n \geqslant 2$; $a_0 = a_1 = 0$.

Theorem 12
The average $\mathbb{E}[X_n]$ of the W parameter, over random tries of n nodes, is asymptotic to

$$n\left[1+\frac{1}{L}\sum_{k\geqslant 2}\frac{a_k}{k2^k}\right]+n\delta(\log_2 n)$$

The numerical value of the constant in the bracket is $1.782784867\ldots$; $\delta(x)$ is (again) a periodic function of period 1 with mean zero.

Proof
Let l_n be as in the previous section.

The following recursion is straightforward:

$$l_n = 2^{-n}\sum_{k=0}^{n}\binom{n}{k}(l_k+l_{n-k}) + 2\cdot 2^{-n} n(1+a_{n-1})$$

for $n \geqslant 2$; $l_0 = l_1 = 0$.

Now let

$$L(z) := \sum_{n\geqslant 0} l_n \frac{z^n}{n!}, \quad M(z) = \mathrm{e}^{-z} L(z), \quad A(z) := \sum_{n\geqslant 0} a_n \frac{z^n}{n!}, \quad B(z) = \mathrm{e}^{-z} A(z)$$

The recursion translates into

$$\begin{aligned} L(z) &= 2\mathrm{e}^{z/2} L\left(\frac{z}{2}\right) + 2\sum_{n\geqslant 2}\frac{(z/2)^n}{n!} n(1+a_{n-1}) \\ &= 2\mathrm{e}^{z/2} L\left(\frac{z}{2}\right) + z\sum_{n\geqslant 1}\frac{(z/2)^n}{n!} + z\sum_{n\geqslant 1}\frac{(z/2)^n}{n!} a_n \\ &= 2\mathrm{e}^{z/2} L\left(\frac{z}{2}\right) + z(\mathrm{e}^{z/2}-1) + zA\left(\frac{z}{2}\right) \end{aligned}$$

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/mma

Therefore,

$$\begin{aligned} M(z) &= 2M\left(\frac{z}{2}\right)+z(\mathrm{e}^{-z/2}-\mathrm{e}^{-z})+z\,\mathrm{e}^{-z}A\left(\frac{z}{2}\right) \\ &= 2M\left(\frac{z}{2}\right)+R(z) \end{aligned}$$

The technique of depoissonization now produces $l_n \sim M(n)$; we can determine $M(n)$ using the Mellin transform. The Mellin transform (see [21, 22]) of $f(x)$ is defined as $f^*(s)=\mathscr{M}(f(x);s)=\int_0^\infty f(x)x^{s-1}\,\mathrm{d}x$;

$$M^*(s)=2^{1+s}M^*(s)+R^*(s)=\frac{R^*(s)}{1-2^{1+s}}$$

The inversion formula tells us that

$$M(z)\sim-\sum_{\text{poles } s_k}\mathrm{Res}_{s=s_k}\frac{R^*(s)}{1-2^{1+s}}z^{-s}$$

We consider only the pole at $s=-1$ (there are also some at $-1+\chi_k$, leading to the fluctuations that are commonly found in such asymptotic estimates).

So $M(z)\sim(1/L)R^*(-1)z$, for z large (apart from fluctuations).

But

$$\begin{aligned} R^*(-1) &= \int_0^\infty R(z)z^{-2}\,\mathrm{d}z \\ &= \int_0^\infty \left[\mathrm{e}^{-z/2}-\mathrm{e}^{-z}+\mathrm{e}^{-z}A\left(\frac{z}{2}\right)\right]\frac{\mathrm{d}z}{z} \\ &= L+\int_0^\infty \mathrm{e}^{-z}\sum_{n\geqslant 2}a_n\frac{z^{n-1}}{n!2^n}\,\mathrm{d}z \\ &= L+\sum_{n\geqslant 2}\frac{a_n}{n!2^n}\Gamma(n) \\ &= L+\sum_{n\geqslant 2}\frac{a_n}{n2^n} \end{aligned}$$

□

Alternative model: Now we sum over all leaves, even if they do *not* contain data. The treatment is very similar.

$$l_n=2^{-n}\sum_{k=0}^{n}\binom{n}{k}(l_k+l_{n-k})+2^{1-n}n(1+a_{n-1})+2^{1-n}(1+a_n)$$

for $n\geqslant 2$; $l_0=l_1=0$.

Only the R-function changes; now it is

$$R_{\text{alt}}(z)=R(z)+2\,\mathrm{e}^{-z/2}-(2+z)\,\mathrm{e}^{-z}+2\,\mathrm{e}^{-z}\sum_{n\geqslant 2}a_n\frac{z^n}{2^n n!}$$

Furthermore

$$R^*_{\mathrm{alt}}(-1) = R^*(-1) + 1 - L + 2\sum_{n\geqslant 2} \frac{a_n}{2^n n(n-1)} = 1 + \sum_{n\geqslant 2} \frac{a_n(n+1)}{2^n n(n-1)}$$

Theorem 13

The average l_n of the W parameter (alternative model), over random tries of n nodes, is asymptotic to

$$\frac{n}{L}\left[1 + \sum_{k\geqslant 2} \frac{a_k(k+1)}{k(k-1)2^k}\right] + n\delta(\log_2 n)$$

The numerical value of the constant of n (the quantity in the bracket) is $3.266982603\ldots$

9. THE W PARAMETER FOR PATRICIA TRIES

A very similar treatment applies for Patricia tries, but the computations are simpler! We list the key steps:

$$l_n = 2^{1-n} \sum_{k=0}^{n} \binom{n}{k} l_k + 2^{1-n} n(n-1)$$

for $n\geqslant 2$; $l_0 = l_1 = 0$. Thus,

$$L(z) = 2\mathrm{e}^{z/2} L\left(\frac{z}{2}\right) + \frac{z^2}{2}\mathrm{e}^{z/2}$$

and

$$M(z) = 2M\left(\frac{z}{2}\right) + \frac{z^2}{2}\mathrm{e}^{-z/2}$$

whence

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

and

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

The asymptotic evaluation of this is conveniently done by Rice's method, with the result:

Theorem 14

The average l_n of the W parameter over random Patricia tries of n nodes, is asymptotic to

$$\frac{2n}{L} + n\delta(\log_2 n)$$

Note once again that we use $\delta(x)$ in a generic sense; in each appearance it will denote a (possibly) different periodic function (with mean zero).

10. THE w PARAMETER IN TRIES AND SUFFIX TREES

The material presented in this section was originally presented in one of these author's thesis (see [1, 2]) and was utilized in an analysis of the Lempel–Ziv '77 algorithm (see [3]). Here, we make several improvements. For instance, we avoid a recurrence relation in the proof of Theorem 15, by instead using a direct combinatorial argument at the outset. Similarly, in the proof of Theorem 16, we now utilize a very general combinatorial pattern matching technique (concerning correlations of words with borders).

10.1. Setup

Throughout our discussion, we work with the binary alphabet $\mathscr{A}=\{0,1\}$.

In this section, we analyze the $\overline{\mathsf{w}}$ parameter in two cases.

In the first case, we analyze the $\overline{\mathsf{w}}$ parameter for a particular insertion into a trie built over $n+1$ independent strings. In other words, we consider the scenario where $n+1$ independent strings $X^{(1)},\dots,X^{(n+1)}$ are used to build a trie. We study the $\overline{\mathsf{w}}$ parameter associated with the external node containing $X^{(i)}$ for some fixed i. Since the inserted strings are i.i.d., then without loss of generality, we let $i=n+1$; in other words, we study the $\overline{\mathsf{w}}$ parameter of the last insertion into the trie. [This is in contrast to the previous sections of this report, where we study the ath leaf, listed in order from left to right.] We let $\overline{\mathsf{w}}_n$ denote the $\overline{\mathsf{w}}$ parameter of the $(n+1)$th insertion in the trie. In other words, $\overline{\mathsf{w}}_n$ enumerates the number of leaves in the subtree rooted at the $(n+1)$th leaf inserted, excluding the $(n+1)$th leaf itself.

In the second case, we analyze the $\overline{\mathsf{w}}$ parameter for the last (i.e. $(n+1)$th) insertion into a suffix tree built from the first $n+1$ suffixes of a common string. We write this as $\overline{\mathsf{w}}_n$.

10.2. Main results

We obtain similar results in both of the cases just mentioned. In fact, we will utilize the results of Theorem 15 to establish the results of Theorem 16.

Theorem 15

Consider a binary memoryless source with probabilities p and $q:=1-p$ of generating '0' and '1', respectively, and let $h:=-p\log p-q\log q$ denote the entropy rate. Insert $n+1$ strings $X^{(1)},\dots,X^{(n+1)}$ into a trie. Then there exist $\delta>0$ and $\varepsilon>0$ such that

$$\begin{aligned}\mathbb{E}[u^{\overline{\mathsf{w}}_n}]&=-\frac{q\log(1-pu)+p\log(1-qu)}{h}+\gamma(\log_{1/p}n,u)+O(n^{-\delta})\\ \mathbb{E}[(\overline{\mathsf{w}}_n)^{\underline{j}}]&=\frac{(j-1)!(q(p/q)^j+p(q/p)^j)}{h}+\gamma_j(\log_{1/p}n)+O(n^{-\varepsilon})\end{aligned}\tag{21}$$

where $\gamma(\cdot,u)$ and γ_j are periodic functions with mean 0 and small modulus if $\log p/\log q$ is rational, or asymptotically zero otherwise.

Thus, $\overline{\mathsf{w}}_n$ asymptotically follows the logarithmic series distribution, i.e. the leading asymptotic term is

$$\mathbb{P}\{\overline{\mathsf{w}}_n=j\}\approx\frac{p^jq+q^jp}{jh}\tag{22}$$

plus some additional small, fluctuating terms if $\log p/\log q$ is rational.

Theorem 16
Consider a binary memoryless source with probabilities p and $q := 1-p$ of generating ‘0’ and ‘1’, respectively, and let $h := -p\log p - q\log q$ denote the entropy rate. Insert the first $n+1$ suffixes of a common string into a suffix tree. In other words, if $X = X_1X_2X_3\ldots$, then define $X^{(j)} = X_jX_{j+1}X_{j+2}\ldots$ for each j, and construct a suffix tree by building a trie from the strings $X^{(1)}, X^{(2)}, \ldots, X^{(n+1)}$. Then, as in Theorem 15, there exist $\delta>0$ and $\varepsilon>0$ such that (21) and (22) hold.

10.3. Proof of Theorem 15

The proof technique of Theorem 15 is basically as follows: We first determine $\mathbb{P}\{\overline{\mathsf{w}}_n = k\}$ exactly, using a combinatorial observation. Then we define two generating functions associated with the distribution and the jth moment of $\overline{\mathsf{w}}_n$, respectively. We utilize a Poisson transform to change from a model with fixed n (recall that the number of strings inserted into the trie is $n+1$) to a model where there are $N+1$ strings, where N is a random variable with Poisson distribution and mean n. Afterwards, we utilize the Mellin transform and its inverse to find the distribution and jth moment in the Poissonized model. Finally, we use depoissonization techniques to find the analogous asymptotic behavior of $\mathbb{E}[u^{\overline{\mathsf{w}}_n}]$ and of $\mathbb{E}[(\overline{\mathsf{w}}_n)^{\underline{j}}]$ in the original model.

Lemma 17
Let w denote the *longest* prefix of both $X^{(n+1)}$ and at least one other $X^{(i)}$. Write $\beta := X^{(n+1)}_{|w|+1}$ (i.e. β denotes the $(|w|+1)$st symbol of $X^{(n+1)}$) and $\alpha = 1-\beta$. Then $\overline{\mathsf{w}}_n = k$ if and only if k of the strings $X^{(1)}, \ldots, X^{(n)}$ have $w\alpha$ as a prefix and the other $n-k$ of these strings do not have w as a prefix.

Proof
By definition, $X^{(n+1)}$ has common prefix w with some other $X^{(i)}$, but $X^{(n+1)}$ has no longer prefix in common with any other $X^{(i)}$. So a string is placed in the subtree rooted at the father of $X^{(n+1)}$’s leaf if and only if the string begins with w. Only $X^{(n+1)}$ begins with $w\beta$, so $\overline{\mathsf{w}}_n$ enumerates exactly the words that begin with prefix $w\alpha$. □

It follows immediately from Lemma 17 that

$$\mathbb{P}\{\overline{\mathsf{w}}_n = k\} = \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \mathbb{P}(w\beta)\binom{n}{k}\mathbb{P}(w\alpha)^k(1-\mathbb{P}(w))^{n-k} \tag{23}$$

The following two generating functions will aid us in computing the asymptotics of (respectively) the distribution of $\overline{\mathsf{w}}_n$ and the jth factorial moment of $\overline{\mathsf{w}}_n$. We define

$$\begin{aligned} G(z,u) &:= \sum_{n\geqslant 0}\mathbb{E}[u^{\overline{\mathsf{w}}_n}]\frac{z^n}{n!} \\ F_j(z) &:= \sum_{n\geqslant 0}\mathbb{E}[(\overline{\mathsf{w}}_n)^{\underline{j}}]\frac{z^n}{n!} \end{aligned} \tag{24}$$

with $\overline{\mathsf{w}}_0 := 0$ by convention. Applying (23) to (24) for $n\geqslant 1$ yields

$$G(z,u) = 1 + \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \mathbb{P}(w\beta)(\mathrm{e}^{z(1-\mathbb{P}(w)+u\mathbb{P}(w\alpha))} - \mathrm{e}^{z(1-\mathbb{P}(w))})$$

$$F_j(z) = \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \mathbb{P}(w\beta)\mathrm{e}^{z(1-\mathbb{P}(w\beta))}(\mathbb{P}(w\alpha)z)^j$$

Now we consider a model where, instead of inserting $n+1$ strings into the trie, we insert $N+1$ strings into the trie, where N is a Poisson random variable with mean n. We emphasize that the results we obtain in this poissonized model must be translated back to the original model at the end of the proof.

If we insert $N+1$ strings into the trie, where N is a Poisson random variable with mean n, then the generating functions analogous to $G(z,u)$ and $F_j(z)$ are

$$\begin{aligned}\widetilde{G}(z,u) &:= \sum_{n\geqslant 0,\ k\geqslant 0} \mathbb{P}\{\overline{\mathrm{w}}_n = k\} u^k \frac{z^n}{n!} \mathrm{e}^{-z} \\ \widetilde{F}_j(z) &:= \sum_{n\geqslant 0} \mathbb{E}[(\overline{\mathrm{w}}_n)^{\underline{j}}] \frac{z^n}{n!} \mathrm{e}^{-z}\end{aligned} \tag{25}$$

We observe that

$$\widetilde{G}(z,u) = \mathrm{e}^{-z} + \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \mathbb{P}(w\beta)(\mathrm{e}^{-z\mathbb{P}(w)(1-u\mathbb{P}(\alpha))} - \mathrm{e}^{-z\mathbb{P}(w)})$$

$$\widetilde{F}_j(z) = \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \mathbb{P}(w\beta)\mathrm{e}^{-z\mathbb{P}(w\beta)}(\mathbb{P}(w\alpha)z)^j$$

by applying (23) to (25).

In order to provide for the fundamental strip of $\widetilde{G}(x,u)$, we hope to have $\widetilde{G}(x,u) = O(x)$ as $x \to 0$, so we replace $\widetilde{G}(x,u)$ by defining $\widehat{G}(x,u) := \widetilde{G}(x,u) - 1$. So for $|u| \leqslant \min\{p^{-1}, q^{-1}\}$ and $\Re(s) \in \langle -1, 0\rangle$, we have

$$\widehat{G}^*(s,u) = \Gamma(s) \frac{q(1-pu)^{-s} + p(1-qu)^{-s} - p^{-s+1} - q^{-s+1}}{1 - p^{-s+1} - q^{-s+1}}$$

Similarly, if $j \in \mathbb{N}$ and $\Re(s) \in \langle -j, 0\rangle$, then

$$\widetilde{F}_j^*(s) = \Gamma(s+j) \frac{p^j q^{-s-j+1} + q^j p^{-s-j+1}}{1 - p^{-s+1} - q^{-s+1}}$$

The corresponding inverse Mellin transforms are

$$\widetilde{F}_j(x) = \frac{1}{2\pi\mathrm{i}} \int_{-1/2-\mathrm{i}\infty}^{-1/2+\mathrm{i}\infty} \widetilde{F}_j^*(s) x^{-s}\, \mathrm{d}s$$

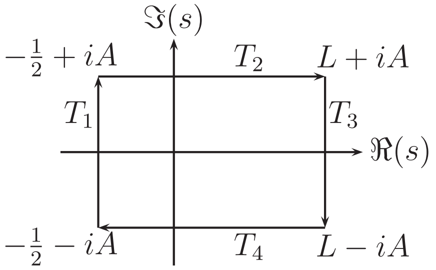
$$\widehat{G}(x,u) = \frac{1}{2\pi\mathrm{i}} \int_{-1/2-\mathrm{i}\infty}^{-1/2+\mathrm{i}\infty} \widehat{G}^*(s,u) x^{-s}\, \mathrm{d}s$$

since $c = -\frac{1}{2}$ is in the fundamental strip of $\widehat{G}(x,u)$ and $\widetilde{F}_j(x)$.

We restrict attention to the case where $\log p/\log q$ is rational, say $\log p/\log q = r/t$. Then, a theorem of Jacquet and Schachinger (see [22, Lemma 8.22]) states that the poles of $1/(1-p^{-s+1}-q^{-s+1})$ are $\{a_k = 2kr\pi\mathrm{i}/\log p \,|\, k \in \mathbb{Z}\}$. So $\widehat{G}^*(s,u)x^{-s}$ and $\widetilde{F}_j^*(s)x^{-s}$ each have poles only at the a_k's, each of which is a simple pole.

DOI: 10.1002/mma

We define T_1, T_2, T_3, T_4 as follows:



By Cauchy's theorem and the smallness property of the Mellin transform (see [22]),

$$\widetilde{G}(x,u) = -\frac{q\log(1-pu)+p\log(1-qu)}{h} + \gamma(\log_{1/p} x, u) + O(x^{-L})$$

$$\widetilde{F}_j(x) = \frac{(j-1)!(q(p/q)^j + p(q/p)^j)}{h} + \gamma_j(\log_{1/p} x) + O(x^{-L})$$

where $h=-p\log p - q\log q$ denotes the entropy of the underlying probability source. Also, if $\log p/\log q = r/t$ is rational, we have

$$\gamma(t,u) = \sum_{k\in\mathbb{Z}\setminus\{0\}} -\frac{\mathrm{e}^{2kr\pi\mathrm{i}t}\Gamma(a_k)(q(1-pu)^{-a_k}+p(1-qu)^{-a_k}-p^{-a_k+1}-q^{-a_k+1})}{p^{-a_k+1}\log p + q^{-a_k+1}\log q}$$

$$\gamma_j(t) = \sum_{k\in\mathbb{Z}\setminus\{0\}} -\frac{\mathrm{e}^{2kr\pi\mathrm{i}t}\Gamma(a_k+j)(p^j q^{-a_k-j+1}+q^j p^{-a_k-j+1})}{p^{-a_k+1}\log p + q^{-a_k+1}\log q}$$

If $\log p/\log q$ is irrational, then $\gamma_j(x)\to 0$ as $x\to\infty$ and $\gamma(x,u)\to 0$ uniformly for $|u|\leqslant \min\{p^{-1}, q^{-1}\}$ as $x\to\infty$. So γ and $\gamma_j(\gamma,u)$ do not exhibit fluctuation when $\log p/\log q$ is irrational.

Finally, we must translate these results from the current model (the Poisson model) back to the original model, using depoissonization. We refer to the depoissonization lemmas of [22, 23]. (See [1] for details.) This allows us to translate the results about $\widetilde{G}(z,u)$ and $\widetilde{F}_j(z)$ into analogous results about $\mathbb{E}[u^{\overline{\mathrm{w}}_n}]$ and $\mathbb{E}[(\overline{\mathrm{w}}_n)^{\underline{j}}]$, up to lower-order terms, which completes the proof of Theorem 15.

10.4. Proof of Theorem 16

The proof technique of Theorem 16 is ultimately made by comparing the generating functions for $\overline{\mathrm{w}}_n$ in the trie and suffix tree cases. We first define $M^T(z,u)$ and $M^S(z,u)$ to denote the bivariate generating functions associated with $\overline{\mathrm{w}}_n$ in the independent trie and suffix tree cases, respectively. We first prove some results about the autocorrelation polynomial $S_w(z)$, which is used to precisely describe the extent to which a word $w\in\mathscr{A}^*$ has overlaps with itself. We also must prove that $M^S(z,u)$ can be analytically extended from the unit disk to a slightly larger disk. Then we determine the poles of $M^{\mathrm{T}}(z,u)$ and $M^{\mathrm{S}}(z,u)$. Finally, we use residue analysis to prove that $M^{\mathrm{T}}(z,u)$ and $M^{\mathrm{S}}(z,u)$ are closely related in a very narrow sense. We conclude that the $\overline{\mathrm{w}}$ parameter has asymptotically the same behavior, up to first order, in both tries built over independent strings and in suffix trees.

DOI: 10.1002/mma

We use $M^{\mathrm{T}}(z,u)$ to denote the bivariate generating function for $\overline{\mathsf{w}}_n$ in the trie setup above:

$$M^{\mathrm{T}}(z,u):=\sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\mathbb{P}\{\overline{\mathsf{w}}_n^{(\mathrm{trie})}=k\}u^k z^n$$

Using the observation from (23), it follows very easily that

$$M^{\mathrm{T}}(z,u)=\sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}}\frac{u\mathbb{P}(\beta)\mathbb{P}(w)}{1-z(1-\mathbb{P}(w))}\frac{z\mathbb{P}(w)\mathbb{P}(\alpha)}{1-z(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))}\tag{26}$$

We define $M^{\mathrm{S}}(z,u)$ in an analogous way to $M^{\mathrm{T}}(z,u)$:

$$M^{\mathrm{S}}(z,u):=\sum_{n=1}^{\infty}\sum_{k=1}^{\infty}\mathbb{P}\{\overline{\mathsf{w}}_n^{(\text{suffix tree})}=k\}u^k z^n$$

In order to understand $M^{\mathrm{S}}(z,u)$, we need to define the autocorrelation polynomial of a word $w\in\mathscr{A}^*$. We write $m=|w|$, and we let $\mathscr{P}(w)$ denote the set of positions k of w satisfying $w_1\dots w_k=w_{m-k+1}\dots w_m$; in other words, w's prefix of length k exactly matches w's suffix of length k. For ease of notation, we also write $w_{k+1}^m:=w_{k+1}\dots w_m$. Now we define the autocorrelation polynomial of w as

$$S_w(z):=\sum_{k\in\mathscr{P}(w)}\mathbb{P}(w_{k+1}^m)z^{m-k}$$

The autocorrelation polynomial $S_w(z)$ records the extent to which a word w has overlaps with itself. Informally, with high probability, we expect $S_w(z)$ to be close to 1, because most words w have only a trivial overlap (where w entirely overlaps with itself) and perhaps some very small overlaps of a short prefix of w with a short suffix of w; long overlaps of w with itself are very rare. The autocorrelation polynomial has been discussed extensively; see, for instance, [24–27].

We are able to use the autocorrelation polynomial to compute the following useful result about $M^{\mathrm{S}}(z,u)$.

Theorem 18
Let $D_w(z)=(1-z)S_w(z)+z^m\mathbb{P}(w)$, where $S_w(z)$ denotes the autocorrelation polynomial of w. Then

$$M^{\mathrm{S}}(z,u)=\sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}}\frac{u\mathbb{P}(\beta)\mathbb{P}(w)}{D_w(z)}\frac{D_{w\alpha}(z)-(1-z)}{D_w(z)-u(D_{w\alpha}(z)-(1-z))}$$

for $|u|<1$ and $|z|<1$.

Proof
We let $w\in\mathscr{A}^*$ denote the longest word that occurs both as a prefix of $X^{(n+1)}$ and as a prefix of at least one other $X^{(i)}$. We also let $\beta\in\mathscr{A}$ denote the $(|w|+1)$th symbol of $X^{(n+1)}$. Then $\overline{\mathsf{w}}_n=k$ if and only if exactly k strings $X^{(i)}$ $(1\leqslant i\leqslant n)$ have $w\alpha$ as a prefix and the other $n-k$ such strings do not have w as a prefix. So we use combinatorics on words to enumerate strings X that have exactly k occurrences of $w\alpha$, followed by an occurrence of $w\beta$, with no other occurrences of w. In other words, we enumerate the language $\mathscr{R}_w\alpha(\mathscr{T}_w^{(\alpha)})^{k-1}\mathscr{T}_w^{(\alpha)}\beta$, where

$$\mathscr{R}_w=\{v\,|\,v \text{ contains exactly one occurrence of } w,\ \text{located at the right end}\}$$

$$\mathscr{T}_w^{(\alpha)}=\{v\,|\,w\alpha v \text{ contains exactly two occurrences of } w,\ \text{located at the left and right ends}\}$$

These languages have generating functions

$$R_w(z) := \sum_{v\in\mathscr{R}_w} \mathbb{P}(v)z^{|v|} \quad \text{and} \quad T_w^{(\alpha)}(z) := \sum_{v\in\mathscr{T}_w^{(\alpha)}} \mathbb{P}(v)z^{|v|}$$

It follows immediately that

$$M^{\mathrm{S}}(z,u) = \sum_{k=1}^{\infty} \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \sum_{s\in\mathscr{R}_w} \mathbb{P}(s\alpha)z^{|s|+1}u \left(\sum_{t\in\mathscr{T}_w^{(\alpha)}} \mathbb{P}(t\alpha)z^{|t|+1}u \right)^{k-1} \sum_{v\in\mathscr{T}_w^{(\alpha)}} \mathbb{P}(v\beta)z^{|v|+1-|w|-1}$$

From [26], we know $R_w(z)/z^{|w|} = \mathbb{P}(w)/D_w(z)$, so we simplify to obtain

$$M^{\mathrm{S}}(z,u) = \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \frac{u\mathbb{P}(\beta)\mathbb{P}(w)}{D_w(z)} \frac{\mathbb{P}(\alpha)zT_w^{(\alpha)}(z)}{1-\mathbb{P}(\alpha)zuT_w^{(\alpha)}(z)} \tag{27}$$

To obtain an explicit form of $T_w^{(\alpha)}(z)$, we define

$$\mathscr{M}_w := \{v \mid wv \text{ contains exactly two occurrences of } w, \text{ located at the left and right ends}\}$$

$$\mathscr{H}_w^{(\alpha)} := \mathscr{M}_w \cap (\alpha\mathscr{A}^*)$$

We observe that $\alpha\mathscr{T}_w^{(\alpha)} = \mathscr{H}_w^{(\alpha)}$. Thus, (27) simplifies to

$$M^{\mathrm{S}}(z,u) = \sum_{w\in\mathscr{A}^*,\ \alpha\in\mathscr{A}} \frac{u\mathbb{P}(\beta)\mathbb{P}(w)}{D_w(z)} \frac{H_w^{(\alpha)}(z)}{1-uH_w^{(\alpha)}(z)} \tag{28}$$

So to complete the proof of Theorem 18, it suffices to prove

$$H_w^{(\alpha)}(z) = \frac{D_{w\alpha}(z)-(1-z)}{D_w(z)} \tag{29}$$

To see this, we use correlation of words with borders, as discussed in [26].

We define $\mathscr{H} = \{w\alpha, w\beta\}$; also let $H_1 = w\alpha$ and $H_2 = w\beta$. We write

$$\mathbb{H} = \begin{bmatrix} \mathbb{P}(H_1) & \mathbb{P}(H_1) \\ \mathbb{P}(H_2) & \mathbb{P}(H_2) \end{bmatrix}$$

We define $\mathscr{A}_{H,F} = \{F_{k+1}^m \mid H_{m-k+1}^m = F_1^k\}$ as a generalization of the autocorrelation polynomial, describing the overlap of H with F. This yields

$$A_{w\alpha,w\alpha}(z) = S_{w\alpha}(z), \quad A_{w\alpha,w\beta}(z) = (S_{w\alpha}(z)-1)\mathbb{P}(\beta)/\mathbb{P}(\alpha)$$

$$A_{w\beta,w\alpha}(z) = (S_{w\beta}(z)-1)\mathbb{P}(\alpha)/\mathbb{P}(\beta), \quad A_{w\beta,w\beta}(z) = S_{w\beta}(z)$$

Next we define $\mathbb{D}(z) = (1-z)\mathbb{A}(z) + z^{m+1}\mathbb{H}^T$, where $\mathbb{H}^T$ denotes the transpose of $\mathbb{H}$, and where

$$\mathbb{A}(z) := \begin{bmatrix} A_{w\alpha,w\alpha}(z) & A_{w\alpha,w\beta}(z) \\ A_{w\beta,w\alpha}(z) & A_{w\beta,w\beta}(z) \end{bmatrix}$$

DOI: 10.1002/mma

We also define $\mathbb{M}(z)=(\mathbb{D}(z)+(z-1)\mathbb{I})\mathbb{D}(z)^{-1}$, where $\mathbb{I}$ denotes the 2×2 identity matrix. Then

$$\mathbb{M}_{1,2}(z)=\frac{(1-z)(S_{w\alpha}(z)-1)\mathbb{P}(\beta)/\mathbb{P}(\alpha)+z^{m+1}\mathbb{P}(w\beta)}{(1-z)S_w(z)+z^m\mathbb{P}(w)} \tag{30}$$

Because $\mathbb{M}_{1,2}=\mathscr{T}_w^{(\alpha)}\cdot\beta$ and $\mathscr{H}_w^{(\alpha)}=\alpha\mathscr{T}_w^{(\alpha)}$, then (30) implies (29), which completes the proof of Theorem 18. □

Now we assume, without loss of generality, that $p\geqslant q$. Note $p\leqslant\sqrt{p}<1$ so there exists $\rho>1$ such that $\rho\sqrt{p}<1$, and thus $\rho p<1$ also. Finally, for ease of notation, we define $\delta:=\sqrt{p}$.

We now make precise the notation that the autocorrelation polynomial

$$S_w(z)=\sum_{k\in\mathscr{P}(w)}\mathbb{P}(w_{k+1}^m)z^{m-k}$$

is close to 1 with high probability. (Recall that $\mathscr{P}(w)$ denotes the set of positions k of w such that $w_1\ldots w_k=w_{m-k+1}\ldots w_m$.)

Lemma 19
If $\theta=(1-p\rho)^{-1}>1$, then

$$\sum_{w\in\mathscr{A}^k}[\![|S_w(\rho)-1|\leqslant(\rho\delta)^k\theta]\!]\mathbb{P}(w)\geqslant 1-\delta^k\theta$$

Proof
See [28] or [29]; the proof is rather straightforward. □

The proof can very easily be strengthened to show that both $S_w(z)$ and $S_{w\alpha}(z)$ are simultaneously close to 1 with high probability:

Lemma 20
If $\theta=(1-p\rho)^{-1}+1$ and $\alpha\in\mathscr{A}$, then

$$\sum_{w\in\mathscr{A}^k}[\![\max\{|S_w(\rho)-1|,|S_{w\alpha}(z)-1|\}\leqslant(\rho\delta)^k\theta]\!]\mathbb{P}(w)\geqslant 1-\delta^k\theta$$

where $[\![A]\!]=1$ if A is true, and $[\![A]\!]=0$ otherwise.

Proof
This is an easy enhancement of the lemma above. □

Since $S_w(z)$ is very close to 1 with high probability, then we expect that $|S_w(z)|$ can be bounded away from 0. To formalize this notation, we claim that if $0<r<1$, then there exists $C>0$ (depending on r) such that

$$|D_w(z)-u(D_{w\alpha}(z)-(1-z))|\geqslant C \tag{31}$$

for $|z|\leqslant r$ and $|u|\leqslant\delta^{-1}$.

It follows that $M^S(z,u)$ can be analytically continued for all z and u with $|u|\leqslant\delta^{-1}$ and $|z|<1$.

Now we find the zeroes of $D_w(z)-u(D_{w\alpha}(z)-(1-z))$ for $|u|\leqslant\delta^{-1}$ and (in particular) the zeroes of $D_w(z)$ (by setting $u=0$). First we determine that (for $|w|$ sufficiently large), there is a unique such zero.

DOI: 10.1002/mma

Lemma 21
There exists an integer $K_2 \geqslant 1$ such that, for u fixed (with $|u| \leqslant \delta^{-1}$) and $|w| \geqslant K_2$, there is exactly one root of $D_w(z) - u(D_{w\alpha}(z) - (1-z))$ in the closed disk $\{z \mid |z| \leqslant \rho\}$.

Proof
Apply Rouché's Theorem; see [1] for the details. □

When $u=0$, this lemma implies (for $|w| \geqslant K_2$) that $D_w(z)$ has exactly one root in the disk $\{z \mid |z| \leqslant \rho\}$.

We let A_w and $C_w(u)$ denote the roots of $D_w(z)$ and $D_w(z) - u(D_{w\alpha}(z) - (1-z))$, respectively. Also, we define

$$B_w = D'_w(A_w)$$

$$E_w(u) = D'_w(C_w) - u(D'_{w\alpha}(C_w) + 1)$$

Now we have identified the relevant singularities of $M^{\mathrm{S}}(z,u)$. We are ready to compare $M^{\mathrm{S}}(z,u)$ to $M^{\mathrm{T}}(z,u)$.

10.4.1. Comparing suffix trees to tries. Our ultimate goal is to show that $M^{\mathrm{S}}(z,u)$ and $M^{\mathrm{T}}(z,u)$ have asymptotically similar behaviors.

We define

$$Q(z,u) = M^{\mathrm{S}}(z,u) - M^{\mathrm{T}}(z,u)$$

For ease of notation, we write

$$M^{\mathrm{T}}_{w,\alpha}(z,u) = \frac{u\mathbb{P}(\beta)\mathbb{P}(w)}{1-z(1-\mathbb{P}(w))} \frac{z\mathbb{P}(w)\mathbb{P}(\alpha)}{1-z(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))}$$

$$M^{\mathrm{S}}_{w,\alpha}(z,u) = \frac{u\mathbb{P}(\beta)\mathbb{P}(w)}{D_w(z)} \frac{D_{w\alpha}(z)-(1-z)}{D_w(z)-u(D_{w\alpha}(z)-(1-z))}$$

We recall that, by (26) and Theorem 18,

$$Q(z,u) = \sum_{w \in \mathscr{A}^*,\ \alpha \in \mathscr{A}} (M^{\mathrm{S}}_{w,\alpha}(z,u) - M^{\mathrm{T}}_{w,\alpha}(z,u))$$

We also define $Q_n(u) = [z^n]Q(z,u)$. We denote the contribution to $Q_n(u)$ from a specific pair $w \in \mathscr{A}^*$ and $\alpha \in \mathscr{A}$ as

$$Q_n^{(w,\alpha)}(u) := [z^n](M^{\mathrm{S}}_{w,\alpha}(z,u) - M^{\mathrm{T}}_{w,\alpha}(z,u))$$

Then we observe that

$$Q_n^{(w,\alpha)}(u) = \frac{1}{2\pi \mathrm{i}} \oint (M^{\mathrm{S}}_{w,\alpha}(z,u) - M^{\mathrm{T}}_{w,\alpha}(z,u)) \frac{\mathrm{d}z}{z^{n+1}}$$

where the path of integration is a circle about the origin with counterclockwise orientation.

We define

$$I_n^{(w,\alpha)}(\rho,u) = \frac{1}{2\pi \mathrm{i}} \int_{|z|=\rho} (M^{\mathrm{S}}_{w,\alpha}(z,u) - M^{\mathrm{T}}_{w,\alpha}(z,u)) \frac{\mathrm{d}z}{z^{n+1}} \tag{32}$$

DOI: 10.1002/mma

By Cauchy's theorem, it follows that

$$Q_n^{(w,\alpha)}(u) = I_n^{(w,\alpha)}(\rho,u) - \mathrm{Res}_{z=A_w}\frac{M_{w,\alpha}^{\mathrm{S}}(z,u)}{z^{n+1}} - \mathrm{Res}_{z=C_w(u)}\frac{M_{w,\alpha}^{\mathrm{S}}(z,u)}{z^{n+1}} + \mathrm{Res}_{z=1/(1-\mathbb{P}(w))}\frac{M_{w,\alpha}^{\mathrm{T}}(z,u)}{z^{n+1}} + \mathrm{Res}_{z=1/(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))}\frac{M_{w,\alpha}^{\mathrm{T}}(z,u)}{z^{n+1}} \tag{33}$$

We compute

$$\begin{aligned}
\mathrm{Res}_{z=A_w}\frac{M_{w,\alpha}^{\mathrm{S}}(z,u)}{z^{n+1}} &= -\frac{\mathbb{P}(\beta)\mathbb{P}(w)}{B_w}\frac{1}{A_w^{n+1}}\\
\mathrm{Res}_{z=C_w(u)}\frac{M_{w,\alpha}^{\mathrm{S}}(z,u)}{z^{n+1}} &= \frac{\mathbb{P}(\beta)\mathbb{P}(w)}{E_w(u)}\frac{1}{C_w(u)^{n+1}}\\
\mathrm{Res}_{z=1/(1-\mathbb{P}(w))}\frac{M_{w,\alpha}^{\mathrm{T}}(z,u)}{z^{n+1}} &= \mathbb{P}(\beta)\mathbb{P}(w)(1-\mathbb{P}(w))^n\\
\mathrm{Res}_{z=1/(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))}\frac{M_{w,\alpha}^{\mathrm{T}}(z,u)}{z^{n+1}} &= -\mathbb{P}(\beta)\mathbb{P}(w)(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))^n
\end{aligned} \tag{34}$$

and then it follows from (33) that

$$Q_n^{(w,\alpha)}(u) = I_n^{(w,\alpha)}(\rho,u) + \frac{\mathbb{P}(\beta)\mathbb{P}(w)}{B_w}\frac{1}{A_w^{n+1}} - \frac{\mathbb{P}(\beta)\mathbb{P}(w)}{E_w(u)}\frac{1}{C_w(u)^{n+1}} + \mathbb{P}(\beta)\mathbb{P}(w)(1-\mathbb{P}(w))^n - \mathbb{P}(\beta)\mathbb{P}(w)(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))^n \tag{35}$$

We next determine the contribution of the $z=A_w$ terms of $M^{\mathrm{S}}(z,u)$ and the $z=1/(1-\mathbb{P}(w))$ terms of $M^{\mathrm{T}}(z,u)$ to the difference $Q_n(u)=[z^n](M(z,u)-M^I(z,u))$. The proof involves a straightforward application of the Mellin transform. All details can be found in [1].

Lemma 22

The 'A_w terms' and the '$1/(1-\mathbb{P}(w))$ terms' (for $|w|\geqslant K_2$) altogether have only $O(n^{-\varepsilon})$ contribution to $Q_n(u)$, i.e.

$$\sum_{|w|\geqslant K_2,\ \alpha\in\mathscr{A}}\left(-\mathrm{Res}_{z=A_w}\frac{M_{w,\alpha}^{\mathrm{S}}(z,u)}{z^{n+1}} + \mathrm{Res}_{z=1/(1-\mathbb{P}(w))}\frac{M_{w,\alpha}^{\mathrm{T}}(z,u)}{z^{n+1}}\right) = O(n^{-\varepsilon})$$

for some $\varepsilon>0$.

Now we bound the contribution to $Q_n(u)$ from the $C_w(u)$ terms of $M^{\mathrm{S}}(z,u)$ and the $z=1/(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))$ terms of $M^T(z,u)$.

Lemma 23

The '$C_w(u)$ terms' and the '$1/(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))$ terms' (for $|w|\geqslant K_2$) altogether have only $O(n^{-\varepsilon})$ contribution to $Q_n(u)$, for some $\varepsilon>0$. More precisely,

$$\sum_{|w|\geqslant K_2,\ \alpha\in\mathscr{A}}\left(-\mathrm{Res}_{z=C_w(u)}\frac{M_{w,\alpha}^{\mathrm{S}}(z,u)}{z^{n+1}} + \mathrm{Res}_{z=1/(1+u\mathbb{P}(w)\mathbb{P}(\alpha)-\mathbb{P}(w))}\frac{M_{w,\alpha}^{\mathrm{T}}(z,u)}{z^{n+1}}\right) = O(n^{-\varepsilon})$$

DOI: 10.1002/mma

Next, we note that the $I_n^{(w,\alpha)}(\rho,u)$ terms in (35) have $O(n^{-\varepsilon})$ contribution to $Q_n(u)$.

Lemma 24
The '$I_n^{(w,\alpha)}(\rho,u)$ terms' (for $|w|\geqslant K_2$) altogether have only $O(n^{-\varepsilon})$ contribution to $Q_n(u)$, for some $\varepsilon>0$. More precisely,

$$\sum_{|w|\geqslant K_2,\ \alpha\in\mathscr{A}} I_n^{(w,\alpha)}(\rho,u)=O(n^{-\varepsilon})$$

Proof
The proof relies on Lemmas 20 and 21. A complete proof is given in [1]. □

Words w with short length also have altogether asymptotically small contribution to $Q_n(u)$. To see this, we note that $|w|$ has a normal distribution with mean $\frac{1}{h}\log n$ and variance $\theta\log n$, where $h=-p\log p-q\log q$ denotes the entropy of the source, and θ is a constant. So the probability of having $|w|\leqslant K_2$ is extremely small, and as a result, the contribution to $Q_n(u)$ from words w with $|w|\leqslant K_2$ is very small.

Lemma 25
The terms $\sum_{|w|<K_2,\ \alpha\in\mathscr{A}}(M_{w,\alpha}^{\mathrm{S}}(z,u)-M_{w,\alpha}^{\mathrm{T}}(z,u))$ altogether have only $O(n^{-\varepsilon})$ contribution to $Q_n(u)$.

All contributions to (35) have now been analyzed. We are finally prepared to summarize our results.

10.4.2. Summary and conclusion. Combining the last four lemmas, we see that $Q_n(u)=O(n^{-\varepsilon})$ uniformly for $|u|\leqslant\delta^{-1}$, where $\delta^{-1}=p^{-1/2}>1$ (recall we are assuming, without loss of generality, that $p>q$). Finally, we apply Cauchy's theorem again. We compute

$$\mathbb{P}\{M_n=k\}-\mathbb{P}\{M_n^I=k\}=[u^kz^n]Q(z,u)=[u^k]Q_n(u)=\frac{1}{2\pi\mathrm{i}}\int_{|u|=p^{-1/2}}\frac{Q_n(u)}{u^{k+1}}\,\mathrm{d}u$$

Since $Q_n(u)=O(n^{-\varepsilon})$, it follows that

$$|\mathbb{P}\{M_n=k\}-\mathbb{P}\{M_n^I=k\}|\leqslant\frac{1}{|2\pi\mathrm{i}|}(2\pi p^{-1/2})\frac{O(n^{-\varepsilon})}{(p^{-1/2})^{k+1}}=O(n^{-\varepsilon}p^{k/2})$$

So M_n^{S} and M_n^{T} have asymptotically the same distribution, and therefore M_n^{S} and M_n^{T} asymptotically have the same factorial moments. Thus, Theorem 16 follows from Theorem 15.

11. CONCLUSION

We have studied the w parameter and its variants, in order to describe what happens on the fringe of a random tree. This analysis has been carried out for a variety of tree classes. It is captivating from an combinatorial view point since many different methods are necessary to study the w parameter for all the classes we considered. We refrained from studying more tree classes (for instance, plane-oriented recursive trees), because this report is already quite lengthy. The emphasis of our study is on illustrating how different methods of analytic combinatorics are combined, rather

DOI: 10.1002/mma

than on a thorough analysis. Therefore, for several tree classes, we dispensed with computing the limiting distributions and confined ourselves with expectation and/or variance only. We hope that this report will inspire further research on this intriguing parameter[¶] and its variants, as well as other tree parameters.

ACKNOWLEDGEMENTS

We are grateful to the referees for suggesting detailed comments and for finding several misprints and inconsistencies.

REFERENCES

1. Ward MD. Analysis of the multiplicity matching parameter in suffix trees. *Ph.D. Thesis*, Purdue University, West Lafayette, IN, U.S.A., May 2005.
2. Ward MD, Szpankowski W. Analysis of the multiplicity matching parameter in suffix trees. *Discrete Mathematics and Theoretical Computer Science* 2005; **AD**:307–322.
3. Lonardi S, Szpankowski W, Ward MD. Error resilient LZ'77 data compression: algorithms, analysis, and experiments. *IEEE Transactions on Information Theory* 2007; **53**:1799–1813.
4. Meir A, Moon JW. On the altitude of nodes in random trees. *Canadian Journal of Mathematics* 1978; **30**:997–1015.
5. Moon JW. The distance between nodes in recursive trees. *Combinatorics*, *Proceedings of the British Combinatorial Conference*, University of Wales College, Aberystwyth, 1973, London Mathematical Society, Lecture Note Series, vol. 13. Cambridge University Press: London, 1974; 125–132.
6. Najock D, Heyde CC. On the number of terminal vertices in certain random trees with an application to stemma construction in philology. *Journal of Applied Probability* 1982; **19**(3):675–680.
7. Gastwirth JL, Bhattacharya PK. Two probability models of pyramid or chain letter schemes demonstrating that their promotional claims are unreliable. *Operations Research* 1984; **32**(3):527–536.
8. Coffman Jr EG, Eve J. File structures using hashing functions. *Communications of the ACM* 1974; **13**(7):427–436.
9. Flajolet P, Sedgewick R. Digital search trees revisited. *SIAM Journal on Computing* 1986; **15**(3):748–767.
10. Knuth DE. Two notes on notation. *American Mathematical Monthly* 1992; **99**(5):403–422.
11. Flajolet P, Odlyzko AM. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics* 1990; **3**:216–240.
12. Kolchin VF. *Random Mappings*. Springer: New York, 1986.
13. Drmota M. Asymptotic distributions and a multivariate Darboux method in enumeration problems. *Journal of Combinatorial Theory*, *Series A* 1994; **67**:169–184.
14. Hwang H-K. Théoremes limites pour les structures combinatoires et les fonctions arithmetiques. *Ph.D. Thesis*, École Polytechnique, December 1994.
15. Drmota M. The height distribution of leaves in rooted trees. *Discrete Mathematics and Applications* 1994; **4**:45–58.
16. Dobrow RP, Fill JA. Total path length for random recursive trees. *Combinatorics*, *Probability and Computing* 1999; **8**:317–333.
17. Greene DH, Knuth DE. *Mathematics for the Analysis of Algorithms* (3rd edn). Birkhäuser: Boston, 1990.
18. Devroye L. Limit laws for local counters in random binary search tree. *Random Structures and Algorithms* 1991; **2**:303–316.
19. Kirschenhofer P, Prodinger H. Eine Anwendung der Theorie der Modulfunktionen in der Informatik. *Sitzungsberichte der Österreichischen Akademie der Wissenschaften* 1988; **197**:339–366.
20. Flajolet P, Sedgewick R. Mellin transforms and asymptotics: finite differences and Rice's integrals. *Theoretical Computer Science* 1995; **144**:101–124.
21. Flajolet P, Gourdon X, Dumas P. Mellin transforms and asymptotics: harmonic sums. *Theoretical Computer Science* 1995; **144**:3–58.

[¶] Originally, H. P. suggested to call it Ward-parameter, but this idea has been dropped for a variety of reasons.

DOI: 10.1002/mma

22. Szpankowski W. *Average Case Analysis of Algorithms on Sequences*. Addison-Wesley, Wiley: Reading, MA, New York, 2001.
23. Jacquet P, Szpankowski W. Analytical depoissonization and its applications. *Theoretical Computer Science* 1998; **201**:1–62.
24. Guibas L, Odlyzko AM. Periods in strings. *Journal of Combinatorial Theory* 1981; **30**:19–43.
25. Guibas L, Odlyzko AM. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory* 1981; **30**:183–208.
26. Régnier M, Szpankowski W. On pattern frequency occurrences in a Markovian sequence. *Algorithmica* 1998; **22**:631–649.
27. Jacquet P, Szpankowski W. Analytic approach to pattern matching. In *Applied Combinatorics on Words*, Chapter 7, Lothaire M (ed.). Cambridge, 2005.
28. Fayolle J, Ward MD. Analysis of the average depth in a suffix tree under a Markov model. *Discrete Mathematics and Theoretical Computer Science* 2005; **AD**:95–104.
29. Jacquet P, Szpankowski W. Autocorrelation on words and its applications. Analysis of suffix trees by string-ruler approach. *Journal of Combinatorial Theory* 1994; **A66**:237–269.

DOI: 10.1002/mma