2005 International Conference on Analysis of Algorithms DMTCS proc. **AD**, 2005, 223–230

Pattern distribution in various types of random trees

Gerard Kok [1]

[1]*Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien, Wiedner Hauptstraße 8-10/113, A-1040 Wien, Austria.*

Let $\mathcal{T}_n$ denote the set of unrooted unlabeled trees of size n and let $\mathcal{M}$ be a particular (finite) tree. Assuming that every tree of $\mathcal{T}_n$ is equally likely, it is shown that the number of occurrences X_n of $\mathcal{M}$ as an induced sub-tree satisfies $\mathbf{E}\,X_n \sim \mu n$ and $\mathbf{V}ar\,X_n \sim \sigma^2 n$ for some (computable) constants $\mu > 0$ and $\sigma \geq 0$. Furthermore, if $\sigma > 0$ then $(X_n - \mathbf{E}\,X_n)/\sqrt{\mathbf{V}ar\,X_n}$ converges to a limiting distribution with density $(A + Bt^2)e^{-Ct^2}$ for some constants A, B, C. However, in all cases in which we were able to calculate these constants, we obtained $B = 0$ and thus a normal distribution. Further, if we consider planted or rooted trees instead of $\mathcal{T}_n$ then the limiting distribution is always normal. Similar results can be proved for planar, labeled and simply generated trees.

Keywords: random trees, generating functions, limiting distributions

1 Introduction

By a *pattern* $\mathcal{M}$ we mean a given finite tree. Now we can consider the number of occurrences of $\mathcal{M}$ in other trees as induced subtree, cf. figure 1. Note that there might be overlaps of two or more copies of $\mathcal{M}$. More exactly, we'll consider the set $\mathcal{T}_n$ of unlabeled unrooted trees of size n , and compute the limiting distribution of the number of occurrences of $\mathcal{M}$ in trees in $\mathcal{T}_n$ as $n \to \infty$. This will also be done for planar and simply generated trees.

Chyzak, Drmota, Klausner and Kok already showed that this limiting distribution is normal for labeled trees (planted, rooted or unrooted), Chyzak et al. (manuscript). In this article we'll show that the same is true for patterns in planar or unlabeled non-planar trees which are planted or rooted and in simply generated trees. Furthermore, for unrooted trees we'll show that the limiting distribution has a density of the form $(A + Bt^2)e^{-Ct^2}$. However, for all examples we know (e.g. for *stars* and *chains*) $B = 0$, that is, the limiting distribution is normal. The case of *stars* (i.e. the number of nodes of degree k) was already explored by Drmota and Gittenberger (1999) for various types of trees. One gets that for any fixed k the number of nodes of degree k of labeled trees of size n satisfies a central limit theorem with mean $\sim \mu_k n$ and variance $\sim \sigma_k^2 n$ (for specific constants $\mu_k, \sigma_k > 0$).

As already mentioned the case of *stars* has been discussed by Drmota and Gittenberger (1999) for various types of trees and the case of labeled trees has been treated by Chyzak et al. (manuscript). Some previous work for unlabeled trees is due to Robinson and Schwenk (1975). Patterns in (rooted) trees have also been considered by Dershowitz and Zaks (1989). However, they only consider patterns starting at the root. There is also some work on patterns in random binary search trees by Flajolet, Gourdon and Martínez, Flajolet et al. (1997). They, too, obtain a central limit theorem. Flajolet and Steyaert also analyzed an algorithm for pattern matchings in trees Flajolet and Steyaert (1980); Steyaert and Flajolet (1983). Further Ruciński (1988) established conditions for when the number of occurrences of a given subgraph in random graphs follow a normal distribution.

The plan of the paper is as follows. In Section 2 we state our results. In Section 3 we show the combinatorial background of the problem, resulting in systems of equations for properly chosen generating functions. In Section 4 we discuss the analytic theorems that can be applied to these systems and we present the possible limiting distributions.

In this paper we only indicate the proof idea of most of the propositions. Detailed proofs can be found in the author's master thesis, Kok (2005), which can be found at http://www.dmg.tuwien.ac.at/kok/.

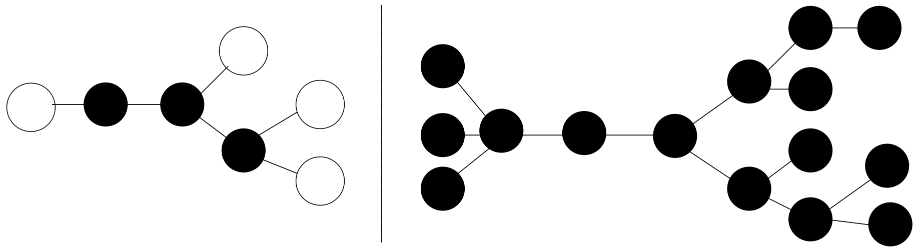

Fig. 1: The pattern on the left occurs twice in the tree on the right side. We call the black nodes of the pattern (non-leaves) the "internal nodes" of the pattern.

2 Results

We fix a finite tree $\mathcal{M}$ that we call *pattern* and say that $\mathcal{M}$ occurs in a tree T if $\mathcal{M}$ is a subtree of T such that the degrees of all *internal* nodes of $\mathcal{M}$ coincide with the corresponding node degrees of T (cf. fig. 1).

Now consider a class $\mathcal{T}_n$ of trees of size n (that might be rooted or unrooted) with a probability distribution (e.g. every tree in $\mathcal{T}_n$ is equally likely) then the number X_n of occurrences of $\mathcal{M}$ in $\mathcal{T}_n$ is a random variable.

2.1 Free trees

Theorem 1. *Let $\mathcal{R}_n$ denote the set of rooted unlabeled trees of size n and $\mathcal{T}_n$ the set of unrooted unlabeled trees of size n where we assume that every tree in $\mathcal{R}_n$ resp. $\mathcal{T}_n$ is equally likely. Then $\mathbf{E}\, X_n \sim \mu n$ and $\mathbf{V}ar\, X_n \sim \sigma^2 n$ for some constants $\mu > 0$ and $\sigma \geq 0$. Further, if $\sigma > 0$ then $(X_n - \mathbf{E}\, X_n)/\sqrt{\mathbf{V}ar\, X_n}$ converges to a limiting distribution with density $(A+Bt^2)e^{-Ct^2}$ for some (computable) constants $A, B, C \geq 0$.*

In particular, for rooted trees or if we consider stars or chains as patterns then $B = 0$, that is, we have a central limit theorem.

2.2 Planar trees

Theorem 2. *Let ${\mathcal{R}_n}^{(p)}$ denote the set of planar rooted unlabeled trees of size n and ${\mathcal{T}_n}^{(p)}$ the set of planar unrooted unlabeled trees of size n where we assume that every tree in ${\mathcal{R}_n}^{(p)}$ resp. ${\mathcal{T}_n}^{(p)}$ is equally likely. Then $\mathbf{E}\, X_n \sim \mu n$ and $\mathbf{V}ar\, X_n \sim \sigma^2 n$ for some (computable) constants $\mu > 0$ and $\sigma \geq 0$. Furthermore, if $\sigma > 0$ then $(X_n - \mathbf{E}\, X_n)/\sqrt{\mathbf{V}ar\, X_n}$ converges to a limiting distribution with density $(A + Bt^2)e^{-Ct^2}$ for some (computable) constants $A, B, C \geq 0$.*

In particular, for rooted trees or if we consider stars or chains as patterns then $B = 0$.

Remark. For both cases, for free trees and for planar trees it is conjectured that $B = 0$ (that is, one always has a central limit theorem) for all patterns. In Chyzak et al. (manuscript) this property is proved for rooted labeled trees.

2.3 Simply generated trees

Planted trees are rooted trees in which an edge without node is attached to the root. Simply generated trees have been introduced by Meir and Moon (1978) and are a generalization of several tree classes, including planted planar trees. One starts with a power series $\Psi(x) = \sum_{j\geq 0} \psi_j x^j$ of non-negative coefficients $\psi_j \geq 0$, where $\psi_0 > 0$ and $\psi_j > 0$ for some $j \geq 2$. We then define the weight $\omega(T)$ of a finite planted tree T by

$$\omega(T) = \prod_{j\geq 0} \psi_j^{D_j(T)},$$

where $D_j(T)$ denotes the number of nodes in T with j successors. It is well known that the generating function $y(x) = \sum_{n\geq 1} y_n x^n$ of the *weighted numbers*

$$y_n = \sum_{|T|=n} \omega(T)$$

(where T runs through all planted planar trees) satisfies the functional equation $y(x) = x\Psi(y(x))$. Furthermore it is natural to define $\omega(T)/y_n$ to be the *probability* of T (when T has n nodes).

Theorem 3. *Let ${\mathcal{R}_n}^{\Psi}$ denote the set of simply generated trees of size n with probability distribution defined by Ψ. Assume that the radius of convergence R of $\Psi(x)$ is positive and that there exists $0 < \tau < R$ with $\Psi(\tau) = \tau\Psi'(\tau)$. Then $\mathbf{E}\, X_n \sim \mu n$ and $\mathbf{V}ar\, X_n \sim \sigma^2 n$ for some (computable) constants $\mu > 0$ and $\sigma \geq 0$. Furthermore, if $\sigma > 0$ then $(X_n - \mathbf{E}\, X_n)/\sqrt{\mathbf{V}ar\, X_n}$ is asymptotically normal.*

3 Combinatorial background

In this section we'll treat the combinatorial background of the problem. We'll proceed similarly in the three cases. Note that the labeled case is already treated by Chyzak et al. (manuscript). For reasons of shortness we'll only bring our new results for unlabeled trees, planar trees and simply generated trees.

We'll make use of the concept of bivariate generating functions (BGF). We say that $p(x, u)$ is a generating function where x counts size and u counts the number of occurrences of the pattern if

$$p(x, u) = \sum_T x^{|T|} u^{X(T)} = \sum_{n,k} p_{n,k} x^n u^k,$$

where $X(T)$ denotes the number of occurrences of the pattern $\mathcal{M}$ in T and $p_{n,k}$ denotes the number of trees of size n with k occurrences of $\mathcal{M}$. In this article $p(x, u)$ will always denote a BGF of planted trees, $r(x, u)$ of rooted trees and $t(x, u)$ of unrooted trees. Furthermore, $Z(S_l; .)$ will denote the cycle index of the symmetric group S_l and with $Z(S_l; a(x, u))$ we'll mean $Z(S_l; a(x, u), a(x^2, u^2), \ldots, a(x^l, u^l))$.

3.1 Free trees

Proposition 1. *Let $\mathcal{P}$ be the class of planted unlabeled non-planar trees and let $\mathcal{M}$ be a pattern. Let $p(x, u)$ be the bivariate generating function of $\mathcal{P}$ where x counts size and u counts the number of occurrences of the pattern. Then there exists a certain number $L + 1$ of auxiliary generating functions $a_i(x, u)$ $(0 \leq i \leq L)$ with*

$$p(x, u) = \sum_{i=0}^{L} a_i(x, u),$$

a power series with infinitely many variables and non-negative coefficients

$$P_0(\ (y_{i,j})_{0\leq i\leq L,\ 1\leq j<\infty}\ )$$

and a number H and polynomials

$$P_j(y_{0,1}, \ldots, y_{L,1}, \ldots, y_{0,H}, \ldots, y_{L,H}, u) \quad (1 \leq j \leq L)$$

with non-negative coefficients such that

$$\begin{aligned}
a_0(x, u) &= xP_0(\ (y_{i,j})\ )|_{y_{i,j}=a_i(x^j,u^j)} \\
a_1(x, u) &= xP_1(a_0(x, u), \ldots, a_L(x, u), \ldots, a_0(x^H, u^H), \ldots, a_L(x^H, u^H), u) \\
&\vdots \\
a_L(x, u) &= xP_L(a_0(x, u), \ldots, a_L(x, u), \ldots, a_0(x^H, u^H), \ldots, a_L(x^H, u^H), u)
\end{aligned} \tag{1}$$

Furthermore,

$$P_0(\ (y_{i,j})\ ) + \sum_{i=1}^{L} P_i(y_{0,1}, \ldots y_{L,H}, 1) = \exp\Big(\sum_{k\geq 1} \frac{1}{k}(y_{0,k} + y_{1,k} + \cdots + y_{L,k})\Big)$$

Proof. (Sketch) We can proceed similarly to Chyzak et al. (manuscript), where a corresponding property for labeled trees is presented. It is well known (see Otter) that planted unlabeled non-planar trees can be recursively described: a planted tree is a planted root to which are attached $0, 1, 2, \ldots$ planted subtrees. In our case we are also considering pattern occurrences. Therefore, we have to split up $\mathcal{P}$ in several (but finitely many) subclasses a_i. These subclasses are such, that the number of occurrences of the pattern at the root of a tree is the same for every tree of a given subclass. We get a set of subclasses with recursive descriptions. Now we can translate these descriptions to a system of functional equations for the generating functions $a_i(x, u)$ of these classes, as is stated in the theorem. Algorithms for calculating the tree classes a_i and for calculating the system of equations can be found in Kok (2005). $\square$

Proposition 2. *Let $\mathcal{R}$ be the class of rooted unlabeled non-planar trees and let $\mathcal{M}$ be a pattern. Let $r(x,u)$ be the BGF of $\mathcal{R}$ where x counts size and u counts the number of occurrences of $\mathcal{M}$. Let $a_i(x,u), 0 \leq i \leq L$ be the solutions of the system of equations* (1). *Then $r(x,u)$ is given by:*

$$r(x,u) \quad = \quad x\exp(\sum_{k=1}^{\infty}\frac{1}{k}p(x^k,u^k)) + \ x\sum_{d\in D}\sum_{\substack{l_0,\dots,l_L\geq 0\\ l_0+\cdots+l_L=d}} Z(S_{l_0};a_0(x,u))\cdots Z(S_{l_L};a_n(x,u))(u^{k_r(l_0,\dots,l_L)}-1)$$

with $D \subseteq \mathbb{N}$ finite and where $k_r(l_0,\dots,l_L)$ is a computable function.

Proof. The generating function of all unordered l_j-tuples of trees of class a_j is $Z(S_{l_j};a_j(x,u))$ (*multiset* construction for unlabeled objects). For a partially ordered d-tuple with l_j trees of class a_j we get the product of the different cycle indices (*sequence* construction). D is the set of internal node degrees of the pattern. $k_r(l_0,\dots,l_L)$ equals the number of occurrences of the pattern at the root of any tree of a class with recursive description $a_{j_1}\otimes\cdots\otimes a_{j_d}$ in which the factor a_i occurs l_i times ($a_{j_1}\otimes\cdots\otimes a_{j_d}$ denotes the class of trees which consist of a root to which are attached d subtrees, which are of the class $a_{j_1},\dots,a_{j_d}$ respectively). □

Proposition 3. *Let $\mathcal{T}$ be the class of unrooted unlabeled non-planar trees and let $\mathcal{M}$ be a pattern. Let $t(x,u)$ be the BGF of $\mathcal{T}$ where x counts size and u counts the number of occurrences of $\mathcal{M}$. Let $a_i(x,u), 0 \leq i \leq L$ be the solutions of the system of equations* (1). *Then $t(x,u)$ is given by:*

$$t(x,u) = r(x,u) - \frac{1}{2}\sum_{0\leq i,j\leq L} a_i(x,u)a_j(x,u)u^{k_t(i,j)} + \frac{1}{2}\sum_{i=0}^{L} a_i(x^2,u^2)u^{k_t(i,i)}$$

where $k_t(i,j)$ is a computable function.

Proof. This follows from a bijection discovered by Otter (1948). Let $\mathcal{R}$ denote the set of rooted trees, $\mathcal{T}$ the set of unrooted trees and $\mathcal{P}^{(2)}$ the set consisting of pairs of two different planted trees. Then there exists a bijection between $\mathcal{R}$ and $\mathcal{T}\cup\mathcal{P}^{(2)}$. To get an equality for the generating functions we have to consider the number of *additional occurrences* $k_t(i,j)$ of the pattern when joining two planted trees $T_1 \in a_i,\ T_2 \in a_j$ to form an unrooted trees T. □

3.1.1 Chains

If $\mathcal{M}$ is a chain (that is, a tree consisting of a finite number of consecutive nodes) then the above system of equations can be reduced to a single equation for $p(x,u)$, for details see Kok (2005).

Proposition 4 (Chains). *Let $\mathcal{P}$ resp. $\mathcal{T}$ be the class of planted resp. unrooted non-planar unlabeled trees. Let $\mathcal{M}$ be a linear pattern (chain) with m internal nodes. Then the bivariate generating functions $p(x,u)$, resp. $t(x,u)$ counting nodes (x) and pattern occurrences (u) in trees of $\mathcal{P}$ resp. $\mathcal{T}$ fulfill*

$$p(x,u) = x\exp\left(\sum_{k=1}^{\infty} p(x^k,u^k)\right) - \frac{x^m(x-1)(u-1)}{1-xu+x^m(u-1)}p(x,u) \tag{2}$$

$$t(x,u) = x\exp\left(\sum_{k=1}^{\infty} p(x^k,u^k)\right) - \frac{1}{2}p(x,u)^2 + \left(\frac{(1-x)(1-xu)}{1-xu+x^m(u-1)}\right)^2\left(\frac{1}{2}x^m u\frac{1-x^m u^m}{1-xu} - x^m\frac{x^m-1}{x-1}\right. \tag{3}$$

$$\left. + x^m(u-1)\frac{x^m u}{1-xu} + \frac{1}{2}(xu^m - x - u^{m-1}+1)(\frac{x^m u}{1-xu})^2\right)p(x,u)^2 + \sum_{i=0}^{L}\sum_{j\geq 2} c_{ij}a_i(x^j,u^j)$$

where c_{ij} are some computable real numbers.

3.2 Planar and simply generated trees

In this section we consider planar and simply generated trees. Note that planted planar trees can be seen as simply generated trees with weights $\psi_j = 1$ for all $j \geq 0$. Further, the notion of a *weighted* generating function $p(x,u)$ will be used as

$$p(x,u) = \sum_{T}\omega(T)x^{|T|}u^{X(T)}.$$

Proposition 5. *Let $\mathcal{P}$ be the class of simply generated trees with power series $\Psi(x) = \sum_{j\geq 0}\psi_j x^j$ ($\psi_j \geq 0\ \forall j, \psi_0 > 0,\ \exists j \geq 2 : \psi_j > 0$) and let $\mathcal{M}$ be a pattern. Let $p(x,u)$ be the (weighted) bivariate generating function where x counts size and u counts the number of occurrences $\mathcal{M}$. Then there exists a certain number $L+1$ of auxiliary generating functions $a_i(x,u)$ ($0 \leq i \leq L$) with*

$$p(x,u) = \sum_{i=0}^{L} a_i(x,u),$$

a power series $P_0(y_0,\dots,y_L)$ and polynomials $P_i(y_0,\dots,y_L,u)$ ($1 \leq i \leq L$) all with nonnegative coefficients such that

$$\begin{aligned} a_0(x,u) &= xP_0(a_0(x,u),\dots,a_L(x,u)) \\ a_1(x,u) &= xP_1(a_0(x,u),\dots,a_L(x,u),u) \\ &\vdots \\ a_L(x,u) &= xP_L(a_0(x,u),\dots,a_L(x,u),u) \end{aligned} \tag{4}$$

and

$$P_0(y_0,\dots,y_L) + \sum_{j=1}^{L} P_j(y_0,\dots,y_L,1) = \Psi(y_0+y_1+\cdots+y_L) \tag{5}$$

Proof. (Sketch) We can proceed similarly as in the non-planar case. However, the construction is a bit simpler, because in the planar case we don't have to take care of *overlapping* patterns. □

Simply generated trees are (of course) *planted* planar trees, that is, there is a natural left to right order. Usually there is no rooted planar and definitely no unrooted planar version of simply generated trees in general. Nevertheless, for planar trees (where $\psi_j = 1$) rooted and unrooted version make sense.

Proposition 6. *Let $\mathcal{R}$ be the class of rooted planar trees and let $\mathcal{M}$ be a pattern. Let $r(x,u)$ be the (weighted) bivariate generating function where x counts size and u counts the number of occurrences $\mathcal{M}$. Let $a_i(x,u), 0 \leq i \leq L$ be the solutions of the system of equations* (4) *with $\psi_j = 1\ \forall j$. Then $r(x,u)$ is given by:*

$$r(x,u) = x + x\sum_{k=1}^{\infty}\frac{\varphi(k)}{k}\log\frac{1}{1-p(x^k,u^k)} + x\sum_{d\in D}\sum_{\substack{s=(i_1,\dots,i_d)\\ 0\leq i_1,\dots,i_d\leq L}}\frac{1}{p(s)}Z(C_{d/p(s)};a_{i_1}(x,u)\cdots a_{i_d}(x,u))(u^{k_r(s)}-1)$$

where $\varphi(k)$ is Euler's totient function, $D \subseteq \mathbb{N}$ is finite, $k_r(s)$ is a computable function (number of additional occurrences) and $p(s)$ is the smallest period of the sequence $(i_1,\dots,i_d)$. (E.g. $(4,2,3,4,2,3)$ has period 3.)

Proposition 7. *Let $\mathcal{T}$ be the class of unrooted planar trees and let $\mathcal{M}$ be a pattern. Let $t(x,u)$ be the (weighted) bivariate generating function where x counts size and u counts the number of occurrences $\mathcal{M}$. Let $a_i(x,u), 0 \leq i \leq L$ be the solutions of the system of equations* (4) *with $\psi_j = 1\ \forall j$. Then the bivariate generating function $t(x,u)$ of unrooted planar trees is given by:*

$$t(x,u) = r(x,u) - \frac{1}{2}\sum_{0\leq i,j\leq L} a_i(x,u)a_j(x,u)u^{k_t(i,j)} + \frac{1}{2}\sum_{i=0}^{L} a_i(x^2,u^2)u^{k_t(i,i)}$$

where $k_t(i,j)$ is a computable function (number of additional occurrences).

The proofs are very similar to the *free* case.

3.2.1 Chains

Proposition 8 (Chains). *Let $\mathcal{P}$ resp. $\mathcal{T}$ be the class of planted resp. unrooted planar trees. Let $\mathcal{M}$ be a linear pattern with m internal nodes. Then the bivariate generating functions $p(x,u)$, resp. $t(x,u)$*

counting nodes (x) and pattern occurrences (u) in trees of $\mathcal{P}$ resp. $\mathcal{T}$ fulfill

$$p(x,u) = x(1-p(x,u))^{-1} - \frac{x^m(x-1)(u-1)}{1-xu+x^m(u-1)}p(x,u) \tag{6}$$

$$t(x,u) = x(1-p(x,u))^{-1} - \frac{1}{2}p(x,u)^2 + \left(\frac{(1-x)(1-xu)}{1-xu+x^m(u-1)}\right)^2 \left(\frac{1}{2}x^m u\frac{1-x^m u^m}{1-xu} - x^m\frac{x^m-1}{x-1}\right. \tag{7}$$

$$\left. +x^m(u-1)\frac{x^m u}{1-xu} + \frac{1}{2}(xu^m - x - u^{m-1}+1)(\frac{x^m u}{1-xu})^2\right) p(x,u)^2 + \sum_{i=0}^{L}\sum_{j\ge 2} c_{ij}a_i(x^j,u^j)$$

where c_{ij} are some computable real numbers.

4 Analytic background

4.1 Singularity Analysis

It is a well known fact (see Drmota and Gittenberger (1999)) that the generating functions $p(x) = p(x,1)$ resp. $r(x) = r(x,1)$ that count the numbers p_n resp. r_n of planted resp. rooted trees of size n have a square root singularity of the following kind:

$$p(x) = g_1(x) - h_1(x)\sqrt{1-\frac{x}{x_0}} \tag{8}$$

$$r(x) = g_2(x) - h_2(x)\sqrt{1-\frac{x}{x_0}}, \tag{9}$$

where x_0 is the radius of convergence of $p(x)$ and $r(x)$ and where $g_i(x)$ and $h_i(x)$, $i = 1,2$ are analytic functions (locally around $x = x_0$). Further, $x = x_0$ is the only singularity on the circle of convergence. This is true for labeled, unlabeled, planar and simply generated trees (that are non-periodic). Thus, the numbers p_n and r_n are asymptotically always of the form $y_n \sim h_i(x_0)/(2\sqrt{\pi})x_0^{-n}n^{-3/2}$. For planted planar trees we get for example the (explicit) expression $p(x) = 1/2 - 1/2\sqrt{1-4x}$ and $p_n = \frac{1}{n}\binom{2n-2}{n-1} \sim (1/\sqrt{\pi})4^{n-1}n^{-3/2}$.

The situation is a little bit different for unrooted trees. Here one has a representation of the form

$$t(x) = g_3(x) + h_3(x)\left(1-\frac{x}{x_0}\right)^{3/2} \tag{10}$$

and consequently $t_n \sim h_3(x_0)/((4/3)\sqrt{\pi})x_0^{-n}n^{-5/2}$. In fact, (10) follows from (8) and (9) since $t(x) = r(x) - \frac{1}{2}p(x)^2 + \frac{1}{2}p(x^2)$ (or $t(x) = r(x) - \frac{1}{2}p(x)^2$).

Interestingly, $p(x,u)$, $r(x,u)$ and $t(x,u)$ behave almost the same.

Proposition 9. *There exists functions $g_1(x,u), h_1(x,u)$ and $f(u)$ that are analytic around $x = x_0$ and $u = 1$ such that*

$$p(x,u) = g_1(x,u) - h_1(x,u)\sqrt{1-\frac{x}{f(u)}}.$$

Furthermore, $x = f(u)$ is the only singularity on the circle $|x| \le |f(u)|$ if u is sufficiently close to 1.

Proof. We can apply the concept of Drmota (1997) (compare also with the appendix of Chyzak et al. (manuscript)) for systems of functional equations with strongly connected dependency graphs. □

Of course, Proposition 9 immediately implies a corresponding property for $r(x,u)$ and $t(x,u)$.

Proposition 10. *There exists functions $g_2(x,u), g_3(x,u), h_2(x,u), h_3(x,u)$ that are analytic around $x = x_0$ and $u = 1$ such that*

$$r(x,u) = g_2(x,u) - h_2(x,u)\sqrt{1-\frac{x}{f(u)}}$$

$$t(x,u) = g_3(x,u) - h_3(x,u)\sqrt{1-\frac{x}{f(u)}},$$

where $h_3(x_0,1) = 0$ and $f(u)$ is the same function as in Proposition 9. Furthermore, $x = f(u)$ is the only singularity on the circle $|x| \le |f(u)|$ if u is sufficiently close to 1.

Note that $h_i(x,1) = h_i(x)$, $i = 1,2,3$ and thus $h_1(x_0,1) \neq 0$, $h_2(x_0,1) \neq 0$, $h_3(x_0,1) = 0$, $\frac{\partial}{\partial x} h_3(x_0,1) \neq 0$.

4.2 Limiting Distributions

By definition it is clear the $t(x,u)$ can be interpreted as

$$t(x,u) = \sum_{n \geq 0} t_n \mathbf{E}\, u^{X_n}\, x^n.$$

Thus, if we set $u = e^{it}$ then the n-th coefficient of $t(x,e^{it})$ is (despite of the asymptotically known factor t_n) precisely the characteristic function of X_n.

Our main (analytic) theorem is the following one:

Theorem 4. *Suppose that X_n is a sequence of random variables and t_n a sequence of positive numbers such that $t(x,u) = \sum_{n\geq 0} t_n \mathbf{E}\, u^{X_n}\, x^n$ has the form*

$$t(x,u) = g(x,u) - h(x,u)\sqrt{1 - \frac{x}{f(u)}}, \tag{11}$$

where $g(x,u), h(x,u)$ and $f(u)$ are analytic functions around $x = f(1)$ and $u = 1$ that satisfy $h(f(1),1) = 0$, $h_x(f(1),1) \neq 0$, $f(1) > 0$, and $f'(1) < 0$. Furthermore, $x = f(u)$ is the only singularity on the circle $|x| \leq |f(u)|$ if u is sufficiently close to 1.

Then $\mathbf{E}\, X_n \sim \mu n$ and $\mathbf{V}ar\, X_n \sim \sigma^2 n$, where $\mu = -f'(1)/f(1)$ and $\sigma \geq 0$. Furthermore, if $\sigma > 0$ then $(X_n - \mathbf{E}\, X_n)/\sqrt{\mathbf{V}ar\, X_n}$ converges to a limiting distribution with density $(A + Bt^2)e^{-Ct^2}$ for some (computable) constants $A, B, C \geq 0$. We have $B = 0$ if and only if $\frac{d^2}{du^2} h(f(u),u)|_{u=1} = 0$.

Proof. Set $C_0(u) = h(f(u),u) = D_1(u-1) + D_2(u-1)^2 + O((u-1)^3)$, $C_1(u) = f(u)\frac{\partial}{\partial x} h(f(u),u)$, $\mu = -f'(1)/f(1)$ and $\mu_2 = \mu^2 + \mu - f''(1)/f(1)$. Then the assumption (11) on $t(x,u)$ and Flajolet and Odlyzko (1990) directly imply that $t_n = C_1(1)/(4/3\sqrt{\pi}) f(1)^{-n} n^{-5/2}(1 + \mathcal{O}(1/n))$ and

$$\mathbf{E}\, e^{itX_n} = \left(\frac{2C_0(e^{it})}{3C_1(1)} n + \frac{C_0(e^{it})}{4C_1(1)} + \frac{C_1(e^{it})}{C_1(1)} + \mathcal{O}\left(\frac{1}{n}\right)\right) e^{(i\mu t - \frac{1}{2}\mu_2 t^2 + \mathcal{O}(t^3))n}.$$

In particular we get $\mathbf{E}\, e^{itX_n/n} \to (1 + it(2D_1)/(3C_1(1)))e^{i\mu t}$. Because the absolute value of the left side is at most 1, it follows that $D_1 = 0$ and that X_n/n is concentrated at μ. More precisely we get, as $n \to \infty$,

$$\mathbf{E}\, e^{it(X_n - \mu n)/\sqrt{n}} \to \left(1 - \frac{2D_2}{3C_1(1)} t^2\right) e^{-\frac{1}{2}\mu_2 t^2}.$$

Of course, this (limiting) characteristic function corresponds to a distribution with density of the form $(A + Bt^2)e^{-Ct^2}$ and $B = 0$ if and only if $D_2 = 0$. Finally expected value and variance can be easily computed. □

Remark. If $h(f(1),1) \neq 0$ then it is even easier to show that X_n satisfies a central limit theorem with $\mu = -f'(1)/f(1)$ and $\sigma^2 = \mu^2 + \mu - f''(1)/f(1)$.

By combining Propositions 9 and 10 and Theorem 4 our main results follow (Theorems 1–3).

In the case of *chains* we can be more precise since the system of equations can be reduced to a single one. A precise analysis similar to the *star case* (see Drmota and Gittenberger (1999)) yields $h_3(f(u),u) = 0$ (see Kok (2005)). Thus, the limiting distribution is surely normal.

Acknowledgements

The author would like to thank Michael Drmota for his advice as supervisor of the author's master thesis. This thesis formed the basis for the article.

References

F. Chyzak, M. Drmota, T. Klausner, and G. Kok. The distribution of patterns in random trees. manuscript.

N. Dershowitz and S. Zaks. Patterns in trees. *Discrete Appl. Math.*, 25(3):241–255, 1989.

M. Drmota. Systems of functional equations. *Random Structures Algorithms*, 10(1-2):103–124, 1997. Average-case analysis of algorithms.

M. Drmota and B. Gittenberger. The distribution of nodes of given degree in random trees. *J. Graph Theory*, 31(3):227–253, 1999.

P. Flajolet, X. Gourdon, and C. Martínez. Patterns in random binary search trees. *Random Structures Algorithms*, 11(3):223–244, 1997. ISSN 1042-9832.

P. Flajolet and A. Odlyzko. Singularity analysis of generating functions. *SIAM J. Discrete Math.*, 3(2): 216–240, 1990.

P. Flajolet and J.-M. Steyaert. On the analysis of tree-matching algorithms. In *Trees in algebra and programming (Proc. 5th Lille Colloq., Lille, 1980)*, pages 22–40. Univ. Lille I, Lille, 1980.

G. Kok. The distribution of patterns in random trees. *master thesis, TU Wien, http://www.dmg.tuwien.ac.at/kok/*, 2005.

A. Meir and J. W. Moon. On the altitude of nodes in random trees. *Canad. J. Math.*, 30(5):997–1015, 1978.

R. Otter. The number of trees. *Ann. of Math. (2)*, 49:583–599, 1948.

R. W. Robinson and A. J. Schwenk. The distribution of degrees in a large random tree. *Discr. Math.*, 12: 359–372, 1975.

A. Ruciński. When are small subgraphs of a random graph normally distributed? *Probab. Theory Related Fields*, 78(1):1–10, 1988.

J.-M. Steyaert and P. Flajolet. Patterns and pattern-matching in trees: an analysis. *Inform. and Control*, 58(1-3):19–58, 1983.